A Probability Metrics Approach to Financial Risk Measures

A Probability Metrics Approach to Financial Risk Measures

Svetlozar T. Rachev
Stoyan V. Stoyanov
Frank J. Fabozzi

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2011

Blackwell Publishing was acquired by John Wiley & Sons in February 2007. Blackwell's publishing program has been merged with Wiley's global Scientific, Technical, and Medical business to form Wiley-Blackwell.

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA

9600 Garsington Road, Oxford, OX4 2DQ, UK

The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data

Rachev, S. T. (Svetlozar Todorov)
A probability metrics approach to financial risk measures / Svetlozar T. Rachev, Stoyan V. Stoyanov, Frank J. Fabozzi, CFA.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-4051-8369-7 (hardback)
1. Financial risk management. 2. Probabilities. I. Stoyanov, Stoyan V. II. Fabozzi, Frank J. III. Title.
HD61.R33 2010
332.01'5192–dc22

2010040519

A catalogue record for this book is available from the British Library.

Set in 10.5/13.5pt Palatino by Thomson Digital, Noida, India
Printed in Malaysia

01 2011

STR

To my grandchildren Iliana, Zoya, and Zari

SVS

To my parents Veselin and Evgeniya Kolevi and
my brother Pavel Stoyanov

FJF

To my wife Donna and
my children Francesco, Patricia, and Karly

Contents

Preface **xiii**

About the Authors **xv**

1 Introduction **1**
1.1 Probability Metrics 1
1.2 Applications in Finance 2

2 Probability Distances and Metrics **7**
2.1 Introduction 9
2.2 Some Examples of Probability Metrics 9
2.2.1 Engineer's metric 10
2.2.2 Uniform (or Kolmogorov) metric 10
2.2.3 Lévy metric 11
2.2.4 Kantorovich metric 14
2.2.5 L_p-metrics between distribution functions 15
2.2.6 Ky Fan metrics 16
2.2.7 L_p-metric 17
2.3 Distance and Semidistance Spaces 19
2.4 Definitions of Probability Distances and Metrics 24
2.5 Summary 28
2.6 Technical Appendix 28
2.6.1 Universally measurable separable metric spaces 29
2.6.2 The equivalence of the notions of p. (semi-)distance on $\mathcal{P}_2$ and on $\mathfrak{X}$ 35

3 Choice under Uncertainty **40**
3.1 Introduction 41
3.2 Expected Utility Theory 44
3.2.1 St Petersburg Paradox 44
3.2.2 The von Neumann–Morgenstern expected utility theory 46
3.2.3 Types of utility functions 48
3.3 Stochastic Dominance 51
3.3.1 First-order stochastic dominance 52
3.3.2 Second-order stochastic dominance 53
3.3.3 Rothschild–Stiglitz stochastic dominance 55
3.3.4 Third-order stochastic dominance 56
3.3.5 Efficient sets and the portfolio choice problem 58
3.3.6 Return versus payoff 59
3.4 Probability Metrics and Stochastic Dominance 63
3.5 Cumulative Prospect Theory 66
3.6 Summary 70
3.7 Technical Appendix 70
3.7.1 The axioms of choice 71
3.7.2 Stochastic dominance relations of order n 72
3.7.3 Return versus payoff and stochastic dominance 74
3.7.4 Other stochastic dominance relations 76

4 A Classification of Probability Distances **83**
4.1 Introduction 86
4.2 Primary Distances and Primary Metrics 86
4.3 Simple Distances and Metrics 90
4.4 Compound Distances and Moment Functions 99
4.5 Ideal Probability Metrics 105
4.5.1 Interpretation and examples of ideal probability metrics 107
4.5.2 Conditions for boundedness of ideal probability metrics 112
4.6 Summary 114
4.7 Technical Appendix 114
4.7.1 Examples of primary distances 114
4.7.2 Examples of simple distances 118
4.7.3 Examples of compound distances 131
4.7.4 Examples of moment functions 135

5 Risk and Uncertainty **146**
5.1 Introduction 147
5.2 Measures of Dispersion 150
5.2.1 Standard deviation 151
5.2.2 Mean absolute deviation 153
5.2.3 Semi-standard deviation 154
5.2.4 Axiomatic description 155
5.2.5 Deviation measures 156
5.3 Probability Metrics and Dispersion Measures 158
5.4 Measures of Risk 159
5.4.1 Value-at-risk 160
5.4.2 Computing portfolio VaR in practice 165
5.4.3 Back-testing of VaR 172
5.4.4 Coherent risk measures 175
5.5 Risk Measures and Dispersion Measures 179
5.6 Risk Measures and Stochastic Orders 181
5.7 Summary 182
5.8 Technical Appendix 183
5.8.1 Convex risk measures 183
5.8.2 Probability metrics and deviation measures 184
5.8.3 Deviation measures and probability quasi-metrics 187

6 Average Value-at-Risk **191**
6.1 Introduction 192
6.2 Average Value-at-Risk 193
6.2.1 AVaR for stable distributions 200
6.3 AVaR Estimation from a Sample 204
6.4 Computing Portfolio AVaR in Practice 207
6.4.1 The multivariate normal assumption 207
6.4.2 The historical method 208
6.4.3 The hybrid method 208
6.4.4 The Monte Carlo method 209
6.4.5 Kernel methods 211
6.5 Back-testing of AVaR 218
6.6 Spectral Risk Measures 220
6.7 Risk Measures and Probability Metrics 223
6.8 Risk Measures Based on Distortion Functionals 226
6.9 Summary 227

6.10 Technical Appendix 228
6.10.1 Characteristics of conditional loss distributions 228
6.10.2 Higher-order AVaR 232
6.10.3 The minimization formula for AVaR 234
6.10.4 ETL vs AVaR 237
6.10.5 Kernel-based estimation of AVaR 242
6.10.6 Remarks on spectral risk measures 245

7 Computing AVaR through Monte Carlo 252
7.1 Introduction 253
7.2 An Illustration of Monte Carlo Variability 256
7.3 Asymptotic Distribution, Classical Conditions 259
7.4 Rate of Convergence to the Normal Distribution 262
7.4.1 The effect of tail thickness 263
7.4.2 The effect of tail truncation 268
7.4.3 Infinite variance distributions 271
7.5 Asymptotic Distribution, Heavy-tailed Returns 277
7.6 Rate of Convergence, Heavy-tailed Returns 283
7.6.1 Stable Paretian distributions 283
7.6.2 Student's *t* distribution 286
7.7 On the Choice of a Distributional Model 290
7.7.1 Tail behavior and return frequency 290
7.7.2 Practical implications 295
7.8 Summary 297
7.9 Technical Appendix 298
7.9.1 Proof of the stable limit result 298

8 Stochastic Dominance Revisited 304
8.1 Introduction 306
8.2 Metrization of Preference Relations 308
8.3 The Hausdorff Metric Structure 310
8.4 Examples 314
8.4.1 The Lévy quasi-semidistance and first-order stochastic dominance 315
8.4.2 Higher-order stochastic dominance 317
8.4.3 The H-quasi-semidistance 320
8.4.4 AVaR generated stochastic orders 322
8.4.5 Compound quasi-semidistances 324
8.5 Utility-type Representations 325

8.6 Almost Stochastic Orders and Degree of Violation 328
8.7 Summary 330
8.8 Technical Appendix 332
8.8.1 Preference relations and topology 332
8.8.2 Quasi-semidistances and preference relations 334
8.8.3 Construction of quasi-semidistances on classes of investors 335
8.8.4 Investors with balanced views 338
8.8.5 Structural classification of probability distances 339

Index **357**

Preface

The theory of probability metrics is a branch of probability theory. It finds application in different theoretical and applied fields such as probability theory, queuing theory, insurance risk theory, and finance. The theory of probability metrics looks for answers to the following basic question: How can one measure the difference between random quantities? In finance, for example, we assume a stochastic model for asset return distributions and, in order to estimate the risk of a portfolio of assets, we sample from the fitted distribution. Then, we use the generated simulations to calculate portfolio risk. In this context, there are two issues arising on two different levels. First, the assumed stochastic model should be "close" to the empirical data. In this sense, we say that we need a realistic model in the first place. Second, since the risk estimate is essentially computed from random scenarios, we have to be aware of the variability of the estimator and how it depends on the assumed asset return distributions.

Although based on universal principles and ideas, the field of probability metrics is very specialized. Most of the literature is highly technical and is accessible mostly to specialists in probability theory. As far as applications are concerned, apart from our book *Advanced Stochastic Models, Risk Assessment, and Portfolio Optimization: Ideal Risk, Uncertainty, and Performance Measures* (John Wiley & Sons, 2008), we are unaware of other literature describing applications in finance.

This book has two goals. The first goal is to describe applications in finance and extend them where possible. The second goal is to present the theory of probability metrics in a more accessible form which would be appropriate for non-specialists in the field. Topics requiring more mathematical rigor and detail are included in technical appendices to chapters.

The book is organized in the following way. Chapter 1 provides a conceptual description of the method of probability metrics and reviews direct and indirect applications in the field of finance. Chapter 2 provides an introduction to the theory of probability metrics. The classical theory describing investor choice under uncertainty is provided in Chapter 3. Chapter 4 discusses the classification of probability distances to primary, simple, and compound types. The information in Chapter 2 is a prerequisite. Chapters 5, 6, and 7 are devoted to risk and uncertainty measures and discuss in detail AVaR and the Monte Carlo method for AVaR estimation. Chapter 6 is a prerequisite to Chapter 7. Finally, Chapter 8 considers the problem of quantifying stochastic dominance relations and takes advantage of the terms introduced in Chapter 3.

Svetlozar T. Rachev
Stoyan V. Stoyanov
Frank J. Fabozzi

About the Authors

Svetlozar (Zari) T. Rachev completed his Ph.D. degree in 1979 from Moscow State (Lomonosov) University, and his Doctor of Science Degree in 1986 from Steklov Mathematical Institute in Moscow. Currently he is Chair-Professor in Statistics, Econometrics and Mathematical Finance at the University of Karlsruhe in the School of Economics and Business Engineering. He is also Professor Emeritus at the University of California, Santa Barbara in the Department of Statistics and Applied Probability. He has published seven monographs, eight handbooks and special-edited volumes, and over 300 research articles. His recently coauthored books published by Wiley in mathematical finance and financial econometrics include *Fat-Tailed and Skewed Asset Return Distributions: Implications for Risk Management, Portfolio selection, and Option Pricing* (2005), *Operational Risk: A Guide to Basel II Capital Requirements, Models, and Analysis* (2007), *Financial Econometrics: From Basics to Advanced Modeling Techniques* (2007), and *Bayesian Methods in Finance* (2008). Professor Rachev is cofounder of Bravo Risk Management Group, specializing in financial risk-management software. Bravo Group was recently acquired by FinAnalytica, for which he currently serves as Chief-Scientist.

Stoyan V. Stoyanov is a Professor of Finance at EDHEC Business School and Scientific Director for EDHEC-Risk Institute in Asia. Prior to joining EDHEC, he was the Head of Quantitative Research

at FinAnalytica, specializing in financial risk management software. He completed his Ph.D. degree with honors in 2005 from the School of Economics and Business Engineering (Chair of Statistics, Econometrics and Mathematical Finance) at the University of Karlsruhe and is author and co-author of numerous papers. His research interests include probability theory, heavy-tailed modeling in the field of finance, and optimal portfolio theory. His articles have recently appeared in *Economics Letters*, *Journal of Banking and Finance*, *Applied Mathematical Finance*, *Applied Financial Economics*, and *International Journal of Theoretical and Applied Finance*. He is a co-author of the mathematical finance book *Advanced Stochastic Models, Risk Assessment and Portfolio Optimization: The Ideal Risk, Uncertainty and Performance Measures* (2008) published by Wiley.

Frank J. Fabozzi is Professor in the Practice of Finance in the School of Management at Yale University. Prior to joining the Yale faculty, he was a Visiting Professor of Finance in the Sloan School of Management at MIT. Professor Fabozzi is a Fellow of the International Center for Finance at Yale University and on the Advisory Council for the Department of Operations Research and Financial Engineering at Princeton University. He is the editor of the *Journal of Portfolio Management*. His recently co-authored books published by Wiley in mathematical finance and financial econometrics include *The Mathematics of Financial Modeling and Investment Management* (2004), *Financial Modeling of the Equity Market: From CAPM to Cointegration* (2006), *Robust Portfolio Optimization and Management* (2007), *Financial Econometrics: From Basics to Advanced Modeling Techniques* (2007), and *Bayesian Methods in Finance* (2008). He earned a doctorate in economics from the City University of New York in 1972. In 2002 Professor Fabozzi was inducted into the Fixed Income Analysts Society's Hall of Fame and he is the 2007 recipient of the C. Stewart Sheppard Award given by the CFA Institute. He earned the designation of Chartered Financial Analyst and Certified Public Accountant.

Chapter 1

Introduction

In this chapter, we provide a conceptual description of the method of probability metrics and discuss direct and indirect applications in the field of finance, which are described in more detail throughout the book.

1.1 Probability Metrics

The development of the *theory of probability metrics* started with the investigation of problems related to limit theorems in probability theory. Limit theorems occupy a very important place in probability theory, statistics, and all their applications. A well-known example is the celebrated central limit theorem (CLT) but there are many other limit theorems, such as the generalized CLT, the max-stable CLT, functional limit theorems, etc. In general, the applicability of the limit theorems stems from the fact that the limit law can be regarded as an approximation to the stochastic model under consideration and, therefore, can be accepted as an approximate substitute. The central question arising is how large an error we make by adopting the approximate model and this question can be investigated by

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

studying the distance between the limit law and the stochastic model. It turns out that this distance is not influenced by the particular problem. Rather, it can be studied by a theory based on some universal principles.

Generally, the theory of probability metrics studies the problem of measuring distances between random quantities. On one hand, it provides the fundamental principles for building probability metrics – the means of measuring such distances. On the other, it studies the relationships between various classes of probability metrics. Another realm of study concerns problems which require a particular metric while the basic results can be obtained in terms of other metrics. In such cases, the metrics relationship is of primary importance.

Certainly, the problem of measuring distances is not limited to random quantities only. In its basic form, it originated in different fields of mathematics. Nevertheless, the theory of probability metrics was developed due to the need of metrics with specific properties. Their choice is very often dictated by the stochastic model under consideration and to a large extent determines the success of the investigation. Rachev (1991) provides more details on the methods of the theory of probability metrics and its numerous applications in both theoretical and more practical problems.

1.2 Applications in Finance

There are no limitations in the theory of probability metrics concerning the nature of the random quantities. This makes its methods fundamental and appealing. Actually, in the general case, it is more appropriate to refer to the random quantities as random *elements*. They can be random variables, random vectors, random functions or random elements in general spaces. For instance, in the context of financial applications, we can study the distance between two random stocks prices, or between vectors of financial variables that are used to construct portfolios, or between yield curves which are much more complicated objects. The methods of the theory remain

the same, irrespective of the nature of the random elements. This represents the most direct application of the theory of probability metrics in finance: that is, it provides a method for measuring how different two random elements are. We explain the axiomatic construction of probability metrics and provide financial interpretations in Chapter 2.

Financial economics, like any other science relying on statistical methods, considers statistical information about the objects it studies on several levels. In some theories in the area of finance, conclusions are drawn only on the basis of certain characteristics of the corresponding distributions. For example, an investor would oftentimes use a risk-reward ratio to rank investment opportunities. Essentially, this reduces to computing the measure of reward (e.g., the expected return) and the measure of risk (e.g., value-at-risk, conditional value-at-risk, standard deviation). Both the measure of reward and the measure of risk represent two characteristics of the corresponding distributions. In effect, the final decision is made on the basis of these two characteristics which, from the investor's perspective, aggregate the information available in the distribution functions.

The theory describing investor choice under uncertainty, the fundamentals of which we discuss in Chapter 3, uses a different approach. Various criteria were developed for first-, second-, and higher-order stochastic dominance based on the distributions themselves. As a consequence, investment opportunities are compared directly through their distribution functions, which is a superior approach from the standpoint of the utilized information.

As another example, consider the problem of building a diversified portfolio. The investor would be interested not only in the marginal distribution characteristics (i.e., the characteristics of the assets on a stand-alone basis), but also in how the assets depend on each another. This requires an additional piece of information which cannot be recovered from the distribution functions of the asset returns. The notion of stochastic dependence can be described by considering the joint behavior of assets returns.

The theory of probability metrics offers a systematic approach towards such a hierarchy of ways to utilize statistical information.

It distinguishes between *primary*, *simple*, and *compound* types of distances which are defined on the space of characteristics, the space of distribution functions, and the space of joint distributions, respectively. Therefore, depending on the particular problem, one can choose the appropriate distance type and this represents another direct application of the theory of probability metrics in the field of finance. This classification of probability distances is explained in Chapter 4.

Besides direct applications, there are also a number of indirect ones. For instance, one of the most important problems in risk estimation is formulating a realistic hypothesis for the asset return distributions. This is largely an empirical question because no arguments exist that can be used to derive a model from some general principles. Therefore, we have to hypothesize a model that best describes a number of empirically confirmed phenomena about asset returns: (1) volatility clustering, (2) autoregressive behavior, (3) short- and long-range dependence, and (4) fat-tailed behavior of the building blocks of the time-series model which varies depending on the frequency (e.g., intra-day, daily, monthly). The theory of probability metrics can be used to suggest a solution to (4). The fact that the degree of heavy-tailedness varies with the frequency may be related to the process of aggregation of higher-frequency returns to obtain lower frequency returns. Generally, the residuals from higher-frequency return models tend to have heavier tails and this observation together with a result known as a *pre-limit theorem* can be used to derive a suggestion for the overall shape of the return distribution. Furthermore, the probability distance used in the pre-limit theorem indicates that the derived shape is most relevant for the body of the distribution. As a result, through the theory of probability metrics we can obtain an approach to construct reasonable models for asset return distributions. We discuss in more detail limit and pre-limit theorems in Chapter 7.

Another central topic in finance is quantification of risk and uncertainty. The two notions are related but are not synonymous. Functionals quantifying risk are called *risk measures* and functionals quantifying uncertainty are called *deviation measures* or *dispersion*

measures. Axiomatic constructions are suggested in the literature for all of them. It turns out that the axioms defining measures of uncertainty can be linked to the axioms defining probability distances, however, with one important modification. The axiom of symmetry, which every distance function should satisfy, appears unnecessarily restrictive. Therefore, we can derive the class of deviation measures from the axiomatic construction of asymmetric probability distances which are also called *probability quasi-distances*. The topic is discussed in detail in Chapter 5.

As far as risk measures are concerned, we consider in detail advantages and disadvantages of value-at-risk, average value-at-risk (AVaR), and spectral risk measures in Chapter 5 and Chapter 6. Since Monte Carlo-based techniques are quite common among practitioners, we discuss in Chapter 7 Monte Carlo-based estimation of AVaR and the problem of stochastic stability in particular. The discussion is practical, based on simulation studies, and is inspired by the classical application of the theory of probability metrics in estimating the stochastic stability of probabilistic models. We apply the CLT and the Generalized CLT to derive the asymptotic distribution of the AVaR estimator under different distributional hypotheses and we discuss approaches to improve its stochastic stability.

We mentioned that adopting stochastic dominance rules for prospect selection rather than rules based on certain characteristics leads to a more efficient use of the information contained in the corresponding distribution functions. Stochastic dominance rules, however, are of the type "X dominates Y" or "X does not dominate Y": that is, the conclusion is qualitative. As a consequence, computational problems are hard to solve in this setting. A way to overcome this difficulty is to transform the nature of the relationship from qualitative to quantitative. We describe how this can be achieved in Chapter 8, which is the last chapter in the book. Our approach is fundamental and is based on asymmetric probability semidistances, which are also called *probability quasi-semidistances*.

The link with probability metrics theory allows a classification of stochastic dominance relations in general. They can be primary, simple, or compound but also, depending on the underlying structure,

they may or may not be generated by classes of investors, which is a typical characterization in the classical theory of choice under uncertainty. This is also a topic discussed in Chapter 8.

References

Rachev, S. T. (1991), *Probability Metrics and the Stability of Stochastic Models*, Wiley, New York.

Chapter 2

Probability Distances and Metrics

The goals of this chapter are the following:

- To provide examples of metrics in probability theory and interpretations from a financial economics perspective.
- To introduce formally the notions of a probability metric and a probability distance.
- To consider the general setting of random variables defined on a given probability space $(\Omega, \mathcal{A}, \Pr)$ taking values in a separable metric space U, allowing a unified treatment of problems involving one-dimensional random variables, random vectors or stochastic processes, for example.
- To consider the alternative setting of probability distances on the space of probability measures $\mathcal{P}_2$ defined on the σ-algebras of Borel subsets of $U^2 = U \times U$ where U is a separable metric space.
- To examine the equivalence of the notion of a probability distance on the space of probability measures $\mathcal{P}_2$ and on the space of joint distributions $\mathcal{LX}_2$ generated by pairs of random variables (X, Y) taking values in a separable metric space U.

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

Notation introduced in this chapter:

Notation	*Description*
EN	The engineer's metric
$\mathfrak{X}^p$	The space of real-valued r.v. with $E\|X\|^p < \infty$
$\boldsymbol{\rho}$	The uniform (Kolmogorov) metric
$\mathfrak{X} = \mathfrak{X}(\mathbb{R})$	The space of real-valued r.v.s
L	The Lévy metric
$\boldsymbol{\kappa}$	The Kantorovich metric
$\boldsymbol{\theta}_p$	The L_p-metric between distribution functions
K, **K***	The Ky Fan metrics
$\mathcal{L}_p$	The L_p-metric between r.v.s
$\mathbf{MOM}_p$	The metric between p-th moments
(S, ρ)	Metric space with a metric ρ
$\mathbb{R}^n$	The n-dimensional vector space
$r(C_1, C_2)$	The Hausdorff metric (semimetric between sets)
$s(F, G)$	The Skorokhod metric
$\mathbb{K} = \mathbb{K}_\rho$	Parameter of a distance space
$\mathcal{H}$	The class of Orlicz's functions
ρ_H	The Birnbaum–Orlicz distance
Kr	The Kruglov distance
(U, d)	Separable metric space with metric d
s.m.s.	Separable metric space
U^k	The k-fold Cartesian product of U
$\mathcal{B}_k = \mathcal{B}_k(U)$	The Borel σ-algebra on U^k
$\mathcal{P}_k = \mathcal{P}_k(U)$	The space of probability laws on $\mathcal{B}_k$
$T_{\alpha,\beta,\ldots,\gamma}P$	The marginal of $P \in \mathcal{P}_k$ on the coordinates α, $\beta, \ldots, \gamma$
Pr_X	The distribution of X
μ	A probability semidistance
$\mathfrak{X} := \mathfrak{X}(U)$	The set of U-valued random variables
$\mathcal{L}\mathfrak{X}_2 := \mathcal{L}\mathfrak{X}_2(U)$	The space of $\mathrm{Pr}_{X,Y}$, $X, Y \in \mathfrak{X}(U)$
u.m.	Universally measurable
u.m.s.m.s.	Universally measurable separable metric space

Important terms introduced in this chapter:

Term	*Concise explanation*
(semi)metric function	A special function satisfying properties making it uniquely positioned for computing distances
(semi)metric space	A space equipped with a (semi)metric function for measuring distances between space elements
probability (semi)metric	A (semi)metric function designed to measure distances between random elements

2.1 Introduction

Generally speaking, a functional which measures the distance between random quantities is called a *probability metric*. These random quantities can be of a very general nature. For instance, they can be random variables, such as the daily returns of equities, the daily change of an exchange rate, etc., or stochastic processes, such as a price evolution in a given period, or much more complex objects, such as the daily movement of the shape of the yield curve.

In this chapter, we provide examples of probability metrics and interpretations from the perspective of financial economics, limiting the discussion to one-dimensional random variables. Then we proceed with the axiomatic definition of probability metrics. In the appendix, we provide a more technical discussion of the axiomatic construction in a much more general context.

2.2 Some Examples of Probability Metrics

Below is a list of various metrics commonly found in probability and statistics. In this section, we limit the discussion to one-dimensional variables only.

2.2.1 Engineer's metric

The engineer's metric is

$$\mathbf{EN}(X, Y) := |E(X) - E(Y)| \quad X, Y \in \mathfrak{X}^1 \tag{2.2.1}$$

where $\mathfrak{X}^p$ is the space of all real-valued random variables (r.v.s) with $E|X|^p < \infty$. In the case of the engineer's metric, we measure the distance between the random variables X and Y only in terms of the deviation of their means. For example, if X and Y describe the return on two common stocks, then the engineer's metric computes the distance between their expected returns.

2.2.2 Uniform (or Kolmogorov) metric

The uniform (or Kolmogorov) metric is

$$\boldsymbol{\rho}(X, Y) := \sup\{|F_X(x) - F_Y(x)| : x \in \mathbb{R}\} \quad X, Y \in \mathfrak{X} = \mathfrak{X}(\mathbb{R}) \tag{2.2.2}$$

where F_X is the distribution function (d.f.) of X, $\mathbb{R} = (-\infty, +\infty)$, and $\mathfrak{X}$ is the space of all real-valued r.v.s.

Figure 2.1 illustrates the Kolmogorov metric. The c.d.f.s of two random variables are plotted on the top plot and the bottom plot shows the absolute difference between them, $|F_X(x) - F_Y(x)|$, as a function of x. The Kolmogorov metric is equal to the largest absolute difference between the two c.d.f.s. A arrow shows where it is attained.

If the random variables X and Y describe the return distribution of the common stocks of two corporations, then the Kolmogorov metric has the following interpretation. The distribution function $F_X(x)$ is by definition the probability that X loses more than a level x, $F_X(x) = P(X \leq x)$. Similarly, $F_Y(x)$ is the probability that Y loses more than x. Therefore, the Kolmogorov distance $\boldsymbol{\rho}(X, Y)$ is the maximum deviation between the two probabilities that can be attained

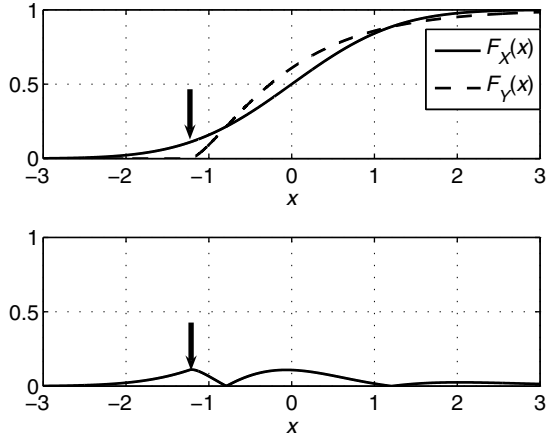

Figure 2.1: Illustration of the Kolmogorov metric. The bottom plot shows the absolute difference between the two c.d.f.s plotted on the top plot. The arrow indicates where the largest absolute difference is attained.

by varying the loss level x. If $\boldsymbol{\rho}(X, Y) = 0$, then the probabilities that X and Y lose more than a loss level x coincide for all loss levels.

Usually, the loss level x, for which the maximum deviation is attained, is close to the mean of the return distribution, i.e. the mean return. Thus, the Kolmogorov metric is completely insensitive to the tails of the distribution which describe the probabilities of extreme events – extreme returns or extreme losses.

2.2.3 Lévy metric

The Lévy metric is

$$\begin{aligned}\mathbf{L}(X, Y) := \inf \{\varepsilon > 0 : F_X(x - \varepsilon) - \varepsilon \leq F_Y(x) \leq F_X(x + \varepsilon) \\ + \varepsilon \quad \forall x \in \mathbb{R}\}. \end{aligned} \tag{2.2.3}$$

The Lévy metric is difficult to calculate in practice. Figure 2.2 contains an illustration. The Lévy metric has important theoretic

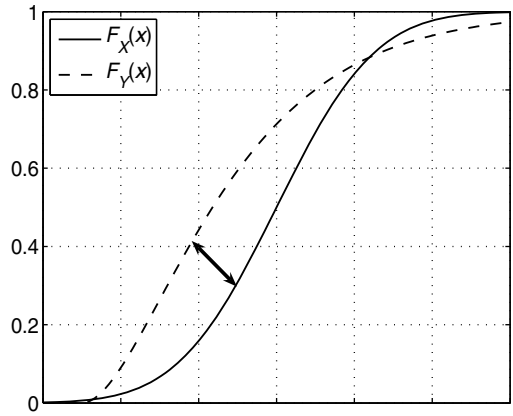

Figure 2.2: Illustration of the Lévy metric. $\mathbf{L}(X, Y)\sqrt{2}$ is the maximum distance between the graphs of F_X and F_Y along a 45 degrees direction. The arrow indicates where the maximum is attained.

application in probability theory as it metrizes the weak convergence. It can be viewed as measuring the closeness between the graphs of the distribution functions while the Kolmogorov metric is a uniform metric between the distribution functions. The general relationship between the two is

$$\mathbf{L}(X, Y) \leq \boldsymbol{\rho}(X, Y) \tag{2.2.4}$$

For example, suppose that X is a random variable describing the return distribution of a portfolio of stocks and Y is a deterministic benchmark with a return of 2.5% ($Y = 2.5\%$). (The deterministic benchmark in this case could be either the cost of funding over a specified time period or a target return requirement to satisfy a liability such as a guaranteed investment contract.) Assume also that the portfolio return has a normal distribution with mean equal to 2.5% and a volatility σ, $X \in N(2.5\%, \sigma^2)$. Since the expected portfolio return is exactly equal to the deterministic benchmark, the Kolmogorov distance between them is always equal to $1/2$ irrespective of how small

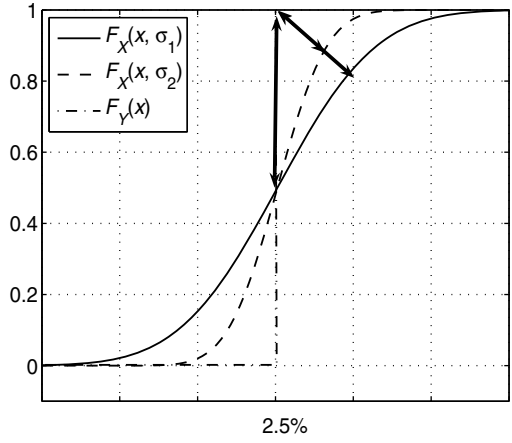

Figure 2.3: Illustration of the relationship between the Lévy and Kolmogorov metrics. The length of the vertical arrow equals $\boldsymbol{\rho}(X, Y)$, while the length of the tilted indicate equals $\mathbf{L}(X, Y)\sqrt{2}$ where $X \in N(2.5\%, \sigma^2)$, $\sigma_1 > \sigma_2$ and $Y = 2.5\%$.

the volatility is,

$$\boldsymbol{\rho}(X, 2.5\%) = 1/2, \quad \forall\, \sigma > 0.$$

Thus, if we rebalance the portfolio and reduce its volatility, the Kolmogorov metric will not register any change in the distance between the portfolio return and the deterministic benchmark. In contrast to the Kolmogorov metric, the Lévy metric will indicate that the rebalanced portfolio is closer to the benchmark. This is illustrated in Figure 2.3.

Remark 2.2.1. We see that $\boldsymbol{\rho}$ and $\mathbf{L}$ may actually be considered as metrics on the space of all distribution functions. However, this cannot be done for $\mathbf{EN}$ simply because $\mathbf{EN}(X, Y) = 0$ does not imply the coincidence of F_X and F_Y, while $\boldsymbol{\rho}(X, Y) = 0 \iff \mathbf{L}(X, Y) = 0 \iff F_X = F_Y$. The Lévy metric metrizes weak convergence (convergence in distribution) in the space $\mathcal{F}$, whereas $\boldsymbol{\rho}$ is often applied in the CLT, cf. Hennequin and Tortrat (1965).

2.2.4 Kantorovich metric

The Kantorovich metric is

$$\boldsymbol{\kappa}(X, Y) = \int_{\mathbb{R}} |F_X(x) - F_Y(x)| \mathrm{d}x \qquad X, Y \in \mathfrak{X}^1. \tag{2.2.5}$$

The Kantorovich metric can be interpreted along the lines of the Kolmogorov metric. Suppose that X and Y are random variables describing the return distribution of two common stocks. Then, as we explained, $F_X(x)$ and $F_Y(x)$ are the probabilities that X and Y, respectively, lose more than the level x. The Kantorovich metric sums the absolute deviation between the two probabilities for all possible values of the loss level x. Thus, the Kantorovich metric provides aggregate information about the deviations between the two probabilities. This is illustrated in Figure 2.4.

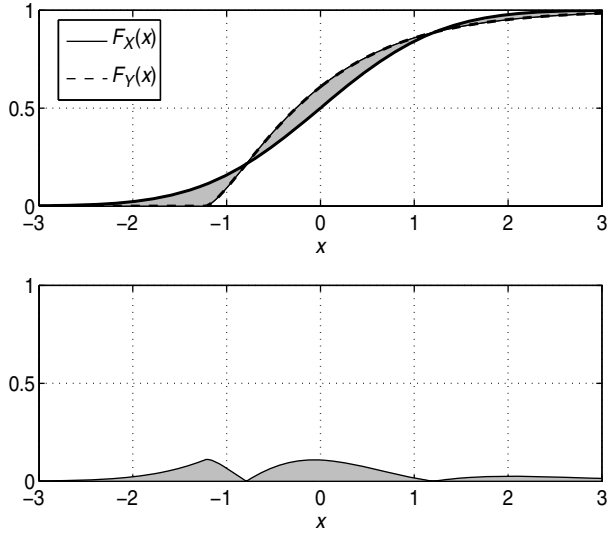

Figure 2.4: Illustration of the Kantorovich metric. The bottom plot shows the absolute difference between the two c.d.f.s plotted on the top plot. The Kantorovich metric equals the shaded area.

In contrast to the Kolmogorov metric, the Kantorovich metric is sensitive to the differences in the probabilities corresponding to extreme profits and losses but to a small degree. This is because the difference $|F_X(x) - F_Y(x)|$ converges to zero as the loss level (x) increases or decreases and, therefore, the contribution of the terms corresponding to extreme events to the total sum is small. As a result, the differences in the tail behavior of X and Y will be reflected in $\boldsymbol{\kappa}(X, Y)$ but only to a small extent.

2.2.5 L_p-metrics between distribution functions

The L_p-metrics between distribution functions is

$$\boldsymbol{\theta}_p(X, Y) := \left(\int_{-\infty}^{\infty} |F_X(t) - F_Y(t)|^p \mathrm{d}t \right)^{1/p} \quad p \geq 1 \quad X, Y \in \mathfrak{X}^1. \tag{2.2.6}$$

The financial interpretation of $\boldsymbol{\theta}_p(X, Y)$ is similar to the interpretation of the Kantorovich metric, which appears as a special case, $\boldsymbol{\kappa}(X, Y) = \boldsymbol{\theta}_1(X, Y)$. The metric $\boldsymbol{\theta}_p(X, Y)$ is an aggregate metric of the difference between the probabilities that X and Y lose more than the level x. The power p exercises a very special effect. It makes the smaller contributors to the total sum of the Kantorovich metric become even smaller contributors to the total sum in (2.2.6). Thus, as p increases, only the largest absolute differences $|F_X(x) - F_Y(x)|$ start to matter. At the limit, as p approaches infinity, only the largest difference $|F_X(x) - F_Y(x)|$ becomes significant and the metric $\boldsymbol{\theta}_\infty(X, Y)$ turns into the Kolmogorov metric. Therefore, if we would like to accentuate on the differences between the two return distributions in the body of the distribution, we can choose a large value of p.

Remark 2.2.2. Clearly the Kantorovich metric arises as a special case, $\boldsymbol{\kappa} = \boldsymbol{\theta}_1$. Moreover, we can extend the definition of $\boldsymbol{\theta}_p$ when $p = \infty$ by setting $\boldsymbol{\theta}_\infty = \boldsymbol{\rho}$. One reason for this extension is the following dual

representation for $1 \leq p \leq \infty$:

$$\theta_p(X, Y) = \sup_{f \in \mathcal{F}_p} |Ef(X) - Ef(Y)|, \quad X, Y \in \mathfrak{X}^1$$

where $\mathcal{F}_p$ is the class of all measurable functions f with $\|f\|_q < 1$. Here, $\|f\|_q (1/p + 1/q = 1)$ is defined, as usual, by

$$\|f\|_q := \begin{cases} \left(\int |f|^q \right)^{1/q} & 1 \leq q < \infty \\ \\ \operatorname*{ess\,sup}_{\mathbb{R}} |f| & q = \infty. \end{cases}$$

(The proof of the above representation is given by Dudley (1989), p. 333.)

2.2.6 Ky Fan metrics

The Ky Fan metrics are

$$\mathbf{K}(X, Y) := \inf\{\varepsilon > 0 : \Pr(|X - Y| > \varepsilon) < \varepsilon\} \qquad X, Y \in \mathfrak{X}. \quad (2.2.7)$$

and

$$\mathbf{K}^*(X, Y) := E \frac{|X - Y|}{1 + |X - Y|}. \quad (2.2.8)$$

Both metrics metrize convergence in probability on $\mathfrak{X} = \mathfrak{X}(\mathbb{R})$, the space of real random variables (Lukacs (1968), Chapter 3, and Dudley (1976), Theorem 3.5).

Assume that X is a random variable describing the return distribution of a portfolio of stocks and Y describes the return distribution of a benchmark portfolio. The probability

$$P(|X - Y| > \varepsilon) = P\left(\{X < Y - \epsilon\} \bigcup \{X > Y + \epsilon\} \right)$$

concerns the event that either the portfolio will outperform the benchmark by ϵ or it will underperform the benchmark by ϵ. Therefore, the quantity 2ϵ can be interpreted as the width of a performance

band. The probability $1 - P(|X - Y| > \varepsilon)$ is actually the probability that the portfolio stays within the performance band, i.e. it does not deviate from the benchmark more than ϵ in an upward or downward direction.

As the width of the performance band decreases, the probability $P(|X - Y| > \varepsilon)$ increases because the portfolio returns will be more often outside a smaller band. The metric $\mathbf{K}(X, Y)$ calculates the width of a performance band such that the probability of the event that the portfolio return is outside the performance band is smaller than half of it.

2.2.7 L_p-metric

The L_p-metric is

$$\mathcal{L}_p(X, Y) := \{E\,|X - Y|^p\}^{1/p} \quad p \geq 1 \quad X, Y \in \mathfrak{X}^p. \tag{2.2.9}$$

From a financial economics viewpoint, we can recognize two widely used measures of deviation which belong to the family of the p-average compound metrics. If p is equal to 1, we obtain the mean absolute deviation between X and Y,

$$\mathcal{L}_1(X, Y) = E|X - Y|.$$

Suppose that X describes the returns of a stock portfolio and Y describes the returns of a benchmark portfolio. Then the mean absolute deviation is a way to measure how closely the stock portfolio tracks the benchmark. If p is equal to 2, we obtain

$$\mathcal{L}_2(X, Y) = \sqrt{E(X - Y)^2}$$

which is a quantity very similar to the tracking error between the two portfolios.

Remark 2.2.3. Certain relations can be obtained between the Ky Fan metric, the L_p-metric, and a metric which is similar in nature to

the engineer's metric. Define

$$m^p(X) := \{E\,|X|^p\}^{1/p} \quad p > 1 \quad X \in \mathfrak{X}^p \tag{2.2.10}$$

and

$$\mathbf{MOM}_p(X, Y) := |m^p(X) - m^p(Y)| \quad p \geq 1 \quad X, Y \in \mathfrak{X}^p. \tag{2.2.11}$$

The metric $\mathbf{MOM}_p(X, Y)$ measures the distance between the corresponding absolute moments of the two random variables and, thus, it is called the *absolute moments metric*. For example, if $p = 2$ then $\mathbf{MOM}_2(X, Y)$ calculates the distance between the standard deviations of X and Y. From a financial economics perspective, if we adopt standard deviation as a proxy for risk, $\mathbf{MOM}_2(X, Y)$ can be interpreted as measuring the deviation between the risk profiles of X and Y.

The relationship between $\mathcal{L}_p(X, Y)$, $\mathbf{K}(X, Y)$, and $\mathbf{MOM}_p(X, Y)$ can be summarized in the following way. Choose a sequence of random variables, $X_0, X_1, \ldots \in \mathfrak{X}^p$. Then,

$$\mathcal{L}_p(X_n, X_0) \to 0 \iff \begin{cases} \mathbf{K}(X_n, X_0) \to 0 \\ \mathbf{MOM}_p(X_n, X_0) \to 0. \end{cases} \tag{2.2.12}$$

See, for example, Lukacs (1968), Chapter 3.

All of the (semi-)metrics on subsets of $\mathfrak{X}$ mentioned above may be divided into three main groups:

- primary;
- simple;
- compound.

A metric μ is *primary* if $\mu(X, Y) = 0$ implies that certain moment characteristics of X and Y agree. As examples, we have **EN** (2.2.1) and $\mathbf{MOM}_p$ (2.2.11). For these metrics

$$\begin{aligned} \mathbf{EN}(X, Y) = 0 &\iff E\,X = E\,Y \\ \mathbf{MOM}_p(X, Y) = 0 &\iff m^p(X) = m^p(Y). \end{aligned} \tag{2.2.13}$$

A metric μ is *simple* if $\mu(X, Y) = 0$ implies complete coincidence between the distribution functions F_X and F_Y,

$$\mu(X, Y) = 0 \iff F_X = F_Y. \tag{2.2.14}$$

Examples are $\boldsymbol{\rho}$ (2.2.2), $\mathbf{L}$ (2.2.3), and $\boldsymbol{\theta}_p$ (2.2.6). These metrics essentially measure the distance between the distribution functions F_X and F_Y. Simple metrics imply a stronger form of identity than primary metrics since $F_X = F_Y$ implies coincidence of all moment characteristics.

The third group, the *compound* (semi-)metrics have the property

$$\mu(X, Y) = 0 \iff \Pr(X = Y) = 1. \tag{2.2.15}$$

Some examples are $\mathbf{K}$ (2.2.7), $\mathbf{K}^*$ (2.2.8), and $\mathcal{L}_p$ (2.2.9). Compound metrics imply a stronger form of identity than primary metrics. If X and Y are two random variables which coincide in all states of the world, possibly except for some states of the world with total probability equal to zero, their distributions functions agree completely.

Later on, precise definitions of these classes are given, and we study the relationships between them. Now we begin with a common definition of probability metric which will include the types mentioned above.

2.3 Distance and Semidistance Spaces

In section 2.2, we considered examples of probability metrics and provided interpretations from a financial economics perspective. All examples concerned one-dimensional random variables and, therefore, the probability metric was regarded as an object related to the space of one-dimensional random variables. In a certain sense, the random variables were considered points in an abstract space and the probability metric appears as a function measuring the distance between these abstract points.

In fact, we considered one-dimensional random variables but the general idea of using a special function which can measure distances

between abstract points belonging to a certain space does not depend on this assumption. We can consider random elements which could be of very general nature, such as multivariate variables and stochastic processes, without changing much the general framework. From a practical viewpoint, by extracting the general principles, we are able to treat equally easy one-dimensional random variables describing, for example, stochastic returns on investments, multivariate random variables describing, for instance, the multi-dimensional behavior of positions participating in two different portfolios, or much more complex objects such as yield curves.

To this end, we begin with a slightly more technical discussion concerning metric spaces, metric and semimetric functions without involving the notion of random elements. The discussion is extended further in section 2.4.

We begin with the notions of metric and semimetric space. Generalizations of these notions will be needed in the Theory of Probability Metrics (TPM).

Definition 2.3.1. A set $S := (S, \rho)$ is said to be a *metric* space with the metric ρ if ρ is a mapping from the product $S \times S$ to $[0, \infty)$ having the following properties for each $x, y, z \in S$:

(1) *Identity property*: $\rho(x, y) = 0 \iff x = y$;
(2) *Symmetry*: $\rho(x, y) = \rho(y, x)$;
(3) *Triangle inequality*: $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$.

Some well-known examples of metric spaces are the following.

(a) *The n-dimensional vector space* $\mathbb{R}^n$ endowed with the metric $\rho(x, y) := \|x - y\|_p$, where

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{\min(1,1/p)} \qquad x = (x_1, \ldots, x_n) \in \mathbb{R}^n \quad 0 < \rho < \infty$$

$$\|x\|_\infty := \sup_{1 \leq i \leq n} |x_i|.$$

(b) *The Hausdorff metric between closed sets*

$$r(C_1, C_2) = \max\left\{\sup_{x_1\in C_1}\inf_{x_2\in C_2}\rho(x_1, x_2), \sup_{x_2\in C_2}\inf_{x_1\in C_1}\rho(x_1, x_2)\right\}$$

where C_is are closed sets in a bounded metric space (S, ρ) (Hausdorff 1949).

(c) *The H-metric.* Let $D(\mathbb{R})$ be the space of all bounded functions $f : \mathbb{R} \to \mathbb{R}$, continuous from the right and having limits from the left, $f(x-) = \lim_{t\uparrow x} f(t)$. For any $f \in D(\mathbb{R})$ define the graph Γ_f as the union of the sets $\{(x, y) : x \in \mathbb{R}, y = f(x)\}$ and $\{(x, y) : x \in \mathbb{R}, y = f(x-)\}$. *The H-metric* $H(f, g)$ in $D(\mathbb{R})$ is defined by the Hausdorff distance between the corresponding graphs, $H(f, g) := r(\Gamma_f, \Gamma_g)$. Note that in the space $\mathcal{F}(\mathbb{R})$ of distribution functions, H metrizes the same convergence as the *Skorokhod metric*:

$$s(F, G) = \inf\Big\{\varepsilon > 0 : \text{there exists a strictly increasing continuous function } \lambda : \mathbb{R} \to \mathbb{R}, \text{ such that } \lambda(\mathbb{R}) = \mathbb{R}, \sup_{t\in\mathbb{R}}|\lambda(t) - t| < \varepsilon \text{ and } \sup_{t\in\mathbb{R}}|F(\lambda(t)) - G(t)| < \varepsilon\Big\}.$$

Moreover, H-convergence in $\mathcal{F}$ implies convergence in distributions (the weak convergence). Clearly, $\boldsymbol{\rho}$-convergence (see (2.2.2)) implies H-convergence.

If the identity property in Definition 2.3.1 is weakened by changing (1) to

$$x = y \Rightarrow \rho(x, y) = 0, \tag{1*}$$

then S is said to be a *semimetric space* (or *pseudometric space*) and ρ a *semimetric* (or *pseudometric*) in S. For example, the Hausdorff metric r is only a semimetric in the space of all Borel subsets of a bounded metric space (S, ρ).

Obviously, in the space of real numbers **EN** (see (2.2.1)) is the usual uniform metric on the real line $\mathbb{R}$, i.e. $\mathbf{EN}(a, b) := |a - b|$, $a, b \in$

$\mathbb{R}$. For $p \geq 0$, define $\mathcal{F}^p$ as the space of all distribution functions F with $\int_{-\infty}^{0} F(x)^p dx + \int_{0}^{\infty}(1 - F(x))^p dx < \infty$. The distribution function space $\mathcal{F} = \mathcal{F}^0$ can be considered as a metric space with metrics $\boldsymbol{\rho}$ and $\mathbf{L}$, while $\boldsymbol{\theta}_p (1 \leq p < \infty)$ is a metric in $\mathcal{F}^p$. The Ky-Fan metrics (see (2.2.7), (2.2.8)) [resp. $\mathcal{L}_p$-metric (see (2.2.9))] may be viewed as semimetrics in $\mathfrak{X}$ (resp. $\mathfrak{X}^1$) as well as metrics in the space of all Pr-equivalence classes:

$$\widetilde{X} := \{Y \in \mathfrak{X} : \Pr(Y = X) = 1\} \quad \forall X \in \mathfrak{X}[\text{resp. } \mathfrak{X}^p]. \tag{2.3.1}$$

$\mathbf{EN}$, $\mathbf{MOM}_p$, $\boldsymbol{\theta}_p$, $\mathcal{L}_p$ can take infinite values in $\mathfrak{X}$ so we shall assume, in the next generalization of the notion of metric, that ρ may take infinite values; at the same time we shall extend also the notion of triangle inequality.

Definition 2.3.2. The set $\mathbf{S}$ is called a *distance space* with distance ρ and parameter $\mathbb{K} = \mathbb{K}_\rho$ if ρ is a function from $S \times S$ to $[0, \infty]$, $\mathbb{K} \geq 1$ and for each $x, y, z \in S$ the identity property (1) and the symmetry property (2) hold as well as the following version of the triangle inequality: (3*) (*Triangle inequality with parameter* $\mathbb{K}$)

$$\rho(x, y) \leq \mathbb{K}[\rho(x, z) + \rho(z, y)]. \tag{2.3.2}$$

If, in addition, the identity property (1) is changed to (1*) then S is called a semidistance space and ρ is called a semidistance (with parameter $\mathbb{K}_\rho$).

Here and in the following we shall distinguish the notions 'metric' and 'distance', using 'metric' only in the case of 'distance with parameter $\mathbb{K} = 1$, taking finite or infinite values'.

Remark 2.3.1. It is not difficult to check that each distance ρ generates a topology in S with a basis of open sets $B(a, r) := \{x \in S; \rho(x, a) < r\}, \in S, r > 0$. We know, of course, that every metric space is normal and that every separable metric space has a countable basis. In much the same way, it is easily shown that the same is true for distance space. Hence, by Urysohn's Metrization

Theorem (Dunford and Schwartz (1988), 1.6.19), every separable distance space is metrizable.

Actually, distance spaces have been used in functional analysis for a long time, as is seen by the following examples.

Example 2.3.1. Let $\mathcal{H}$ be the class of all nondecreasing continuous functions H from $[0, \infty)$ onto $[0, \infty)$ which vanish at the origin and satisfy Orlicz's condition

$$K_H := \sup_{t>0} \frac{H(2t)}{H(t)} < \infty. \tag{2.3.3}$$

Then $\widetilde{\rho} := H(\rho)$ is a distance in S for each metric ρ in S and $\mathbb{K}_{\widetilde{\rho}} = K_H$.

Example 2.3.2. (*Birnbaum–Orlicz distance space*, Birnbaum and Orliz (1931), and Dunford and Schwartz (1988), p. 400.)

The *Birnbaum–Orlicz space* $L^H (H \in \mathcal{H})$ consists of all integrable functions on $[0, 1]$ endowed with Birnbaum–Orlicz distance:

$$\rho_H(f_1, f_2) := \int_0^1 H(|f_1(x) - f_2(x)|)\mathrm{d}x. \tag{2.3.4}$$

Obviously, $\mathbb{K}_{\rho_H} = K_H$.

Example 2.3.3. Similarly to (2.3.4), Kruglov (1973) introduced the following distance in the space of distribution functions:

$$\mathbf{Kr}(F, G) = \int \phi(F(x) - G(x))\mathrm{d}x \tag{2.3.5}$$

where the function ϕ satisfies the following conditions:

(a) ϕ is even and strictly increasing on $[0, \infty)$, $\phi(0) = 0$;
(b) for any x and y and some fixed $A \geq 1$

$$\phi(x + y) \leq A(\phi(x) + \phi(y)). \tag{2.3.6}$$

Obviously, $\mathbb{K}_{\mathbf{Kr}} = A$.

2.4 Definitions of Probability Distances and Metrics

While we gave examples with random variables in section 2.2, probability metrics are naturally defined on spaces of probability measures. Thus, a probability metric calculates the distance between two specified probability measures. The probability measures themselves can be defined on a variety of spaces, such as the space of real numbers, which was the context we considered in section 2.2. The nature of the space the probability measures are defined on determines the interpretation.

There is a subtle question of consistency which arises. If we are naturally thinking of random elements in terms of their real-world application (i.e., random variables, random vectors, stochastic processes, and the definition of probability metrics involves the more abstract concept of a probability measure), is there a loss of generality? It turns out that the answer to this question is essentially negative under some regularity conditions and it will be considered in the appendix to this chapter.

The definition of probability metrics involves two probability measures and it turns out that it may be important how they depend on each other; that is, how they are *coupled* or *joined* in a joint probability measure. From a practical viewpoint, consider two random variables X and Y, which are dependent. For example, they may represent the stochastic returns of two common stocks in one and the same industry. Calculating the distance between X and Y in terms of $\mathcal{L}_p(X, Y)$, for example, is influenced by the way X and Y depend on each other. For this reason, probability metrics are defined on pairs of random variables, in this case the pair (X, Y), in which X and Y are the corresponding one-dimensional *projections* of the pair.

In this section, we provide a formal definition of probability metrics in the space of probability measures. We assume a very general nature of the space U on which the probability measures are defined, which includes as special cases all examples we have considered. We introduce the following notation:

U	a *separable metric space* (s.m.s.) with metric d
$U^k = U \times \cdots \times U$ (k times)	the k-fold Cartesian product of U
$\mathcal{B}_k = \mathcal{B}_k(U)$	the σ algebra of Borel subsets of U^k.
$\mathcal{P}_k = \mathcal{P}_k(U)$	the space of all probability measures defined on the σ algebra $\mathcal{B}_k = \mathcal{B}_k(U)$

We shall use the terms "probability measure" and "law" interchangeably. For any set $\{\alpha, \beta, \ldots, \gamma\} \subseteq \{1, 2, \ldots, k\}$ and for any $P \in \mathcal{P}_k$ let us define the marginal of P on the coordinates $\alpha, \beta, \ldots, \gamma$ by $T_{\alpha,\beta,\ldots,\gamma}P$. For example, for any Borel subsets A and B of U, $T_1P(A) = P(A \times U \times \cdots \times U)$, $T_{1,3}P(A \times B) = P(A \times U \times B \times \cdots \times U)$. Let $\mathbb{B}$ be the operator in U^2 defined by $\mathbb{B}(x, y) := (y, x)$ $(x, y \in U)$. All metrics $\mu(X, Y)$ cited in section 2.2 (see (2.2.1)–(2.2.11)) are completely determined by the joint distributions $\Pr_{X,Y}$ ($\Pr_{X,Y} \in \mathcal{P}_2(\mathbb{R})$) of the random variables $X, Y \in \mathfrak{X}(\mathbb{R})$. In the next definition we shall introduce the notion of probability distance and thus we shall describe the primary, simple, and compound metrics in a uniform way. Moreover, the space where the r.v.s X and Y take values will be extended to U, an arbitrary s.m.s.

Definition 2.4.1. A mapping μ defined on $\mathcal{P}_2$ and taking values in the extended interval $[0, \infty]$ is said to be a *probability semidistance with parameter* $\mathbb{K} := \mathbb{K}_\mu \geq 1$ (or briefly, *p. semidistance*) in $\mathcal{P}_2$, if it possesses the three properties listed below:

(1) *Identity Property* (**ID**). If $P \in \mathcal{P}_2$ and $P(\cup_{x \in U}\{(x, x)\}) = 1$ then $\mu(P) = 0$

(2) *Symmetry* (**SYM**). If $P \in \mathcal{P}_2$ then $\mu(P \circ \mathbb{B}^{-1}) = \mu(P)$

(3) *Triangle Inequality* (**TI**). If $P_{13}, P_{12}, P_{23} \in \mathcal{P}_2$ and there exists a law $Q \in \mathcal{P}_3$ such that the following 'consistency' condition holds:

$$T_{13}Q = P_{13} \qquad T_{12}Q = P_{12} \qquad T_{23}Q = P_{23}, \tag{2.4.1}$$

then

$$\mu(P_{13}) \leq \mathbb{K}[\mu(P_{12}) + \mu(P_{23})].$$

If $\mathbb{K} = 1$ then μ is said to be a *probability semimetric*. If we strengthen the condition ID to $\widetilde{\mathbf{ID}}$: If $P \in P_2$, then

$$P(\cup\{(x, x) : x \in U\}) = 1 \iff \mu(P) = 0,$$

then we say that μ is a *probability distance with parameter* $\mathbb{K} = \mathbb{K}_\mu \geq 1$ (or briefly, *p. distance*).

Definition 2.4.1 acquires a visual form in terms of random variables, namely, let $\mathfrak{X} := \mathfrak{X}(U)$ be the set of all r.v.s on a given probability space $(\Omega, \mathcal{A}, \Pr)$ taking values in $(U, \mathcal{B}_1)$. By $\mathcal{L}\mathfrak{X}_2 := \mathcal{L}\mathfrak{X}_2(U) := \mathcal{L}\mathfrak{X}_2(U; \Omega, \mathcal{A}, \Pr)$ we denote the space of all joint distributions $\Pr_{X,Y}$ generated by the pairs $X, Y \in \mathfrak{X}$. Since $\mathcal{L}\mathfrak{X}_2 \subseteq \mathcal{P}_2$, then the notion of p. (semi-)distance is naturally defined on $\mathcal{L}\mathfrak{X}_2$. Considering μ on the subset $\mathcal{L}\mathfrak{X}_2$, we shall put

$$\mu(X, Y) := \mu(\Pr_{X,Y})$$

and call μ a *p. semidistance on* $\mathfrak{X}$. If μ is a p. distance, then we use the phrase *p. distance* on $\mathfrak{X}$. Each p. semidistance [resp. distance] μ on $\mathfrak{X}$ is a semidistance [resp. distance] on $\mathfrak{X}$ in the sense of Definition 2.3.2. Then the relationships **ID**, $\widetilde{\mathbf{ID}}$, **SYM**, and **TI** have simple 'metrical' interpretations:

$\mathbf{ID}^{(*)}$	$\Pr(X = Y) = 1 \Rightarrow \mu(X, Y) = 0$
$\widetilde{\mathbf{ID}}^{(*)}$	$\Pr(X = Y) = 1 \iff \mu(X, Y) = 0$
$\mathbf{SYM}^{(*)}$	$\mu(X, Y) = \mu(Y, X)$
$\mathbf{TI}^{(*)}$	$\mu(X, Z) \leq \mathbb{K}[\mu(X, Z) + \mu(Z, Y)]$.

Definition 2.4.2. A mapping $\mu : \mathcal{L}\mathfrak{X}_2 \to [0, \infty]$ is said to be a *probability semidistance* in $\mathfrak{X}$ [resp. *distance*] *with parameter* $\mathbb{K} := \mathbb{K}_\mu \geq 1$, if $\mu(X, Y) = \mu(\Pr_{X,Y})$ satisfies the properties $\mathbf{ID}^{(*)}$ [resp. $\widetilde{\mathbf{ID}}^{(*)}$], $\mathbf{SYM}^{(*)}$ and $\mathbf{TI}^{(*)}$ for all r.v.s $X, Y, Z \in \mathfrak{X}(U)$.

Example 2.4.1. Let $H \in \mathcal{H}$ (see Example 2.3.1) and (U, d) be a s.m.s. Then $\mathcal{L}_H(X, Y) = E\,H(d(Z, V))$ is a p. distance in $\mathfrak{X}(U)$. Clearly, $\mathcal{L}_H$ is finite in the subspace of all X with finite moment $E\,H(d(X, a))$ for some $a \in U$. The Kruglov's distance $\mathbf{Kr}(X, Y) := \mathbf{Kr}(F_X, F_Y)$ is a p. semidistance in $\mathfrak{X}(\mathbb{R})$.

Examples of p. metrics in $\mathfrak{X}(U)$ are the Ky Fan metric

$$\mathbf{K}(X, Y) := \inf\{\varepsilon > 0 : \Pr(d(X, Y) > \varepsilon) < \varepsilon\} \quad (X, Y \in \mathfrak{X}(U)) \tag{2.4.2}$$

and the $\mathcal{L}_p$-*metrics* ($0 \le p \le \infty$)

$$\mathcal{L}_p(X, Y) := \{E\, d^p(X, Y)\}^{\min(1,1/p)} \quad 0 < p < \infty, \tag{2.4.3}$$

$$\mathcal{L}_\infty(X, Y) := \operatorname{ess\,sup} d(X, Y) := \inf\{\varepsilon > 0 : \Pr(d(X, Y) > \varepsilon) = 0\} \tag{2.4.4}$$

$$\mathcal{L}_0(X, Y) := E\, I\{X, Y\} := \Pr(X, Y). \tag{2.4.5}$$

The engineer's metric **EN**, Kolmogorov metric $\boldsymbol{\rho}$, Kantorovitch metric $\boldsymbol{\kappa}$, and the Lévy metric **L** (see section 2.2) are p. semimetrics in $\mathfrak{X}(\mathbb{R})$.

Remark 2.4.1. Unlike Definition 2.4.2, Definition 2.4.1 is free of the choice of the initial probability space, and depends only on the structure of the metric space U. The main reason for considering not arbitrary but separable metric spaces (U, d) is that we need the measurability of the metric d in order to connect the metric structure of U with that of $\mathfrak{X}(U)$. In particular, the measurability of d enables us to handle, in a well-defined way, p. metrics such as the Ky Fan metric **K** and $\mathcal{L}_p$-metrics. Note that $\mathcal{L}_0$ does not depend on the metric d, so one can define $\mathcal{L}_0$ on $\mathfrak{X}(U)$, where U is an arbitrary measurable space, while in (2.4.2)–(2.4.4) we need $d(X, Y)$ to be a random variable. Thus the natural class of spaces appropriate to our investigation is the class of s.m.s.

One of the axioms defining probability semidistances is the symmetry axiom $\mathbf{SYM}^{(*)}$. In applications in financial economics, the symmetry axiom is not important and we can omit it. Thus, we extend the treatment of the defining axioms of probability semidistances in the same way as it is done in the field of functional analysis. In case the symmetry axiom, $\mathbf{SYM}^{(*)}$, is omitted, then *quasi-* is added to the name.

Definition 2.4.3. A mapping $\mu : \mathcal{LX}_2 \to [0, \infty]$ is said to be

- a *probability quasi-distance* in $\mathfrak{X}$ with parameter $\mathbb{K} := \mathbb{K}_\mu \geq 1$ if $\widetilde{\mathbf{ID}}^{(*)}$ and $\mathbf{TI}^{(*)}$ hold;
- a *probability quasi-semidistance* in $\mathfrak{X}$ with parameter $\mathbb{K} := \mathbb{K}_\mu \geq 1$ if $\mathbf{ID}^{(*)}$ and $\mathbf{TI}^{(*)}$ hold.

Note that by removing the symmetry axiom we obtain a larger class in which semidistances appear as symmetric quasi-semidistances. A probability quasi-semidistance with parameter $\mathbb{K} = 1$ is called a probability quasi-semimetric.

2.5 Summary

We considered a number of examples of probability metrics and distances. We provided interpretations from a financial economics perspective. Probability (semi-)distances were formally introduced on the space of probability measures $\mathcal{P}_2$ defined on the σ-algebras of Borel subsets of U^2.

Probability (semi-)distances were formally introduced on the space of joint distributions $\mathcal{LX}_2$ generated by pairs of U-valued random variables defined on a probability space $(\Omega, \mathcal{A}, \Pr)$ where U is a separable metric space.

Since the space of joint distributions $\mathcal{LX}_2$ forms a subspace of $\mathcal{P}_2$ by construction, we considered the question of when the notions of a probability distance on $\mathcal{P}_2$ and on $\mathcal{LX}_2$ are the same. For every s.m.s. U, we can find a probability space $(\Omega, \mathcal{A}, \Pr)$ such that this equivalence holds. Moreover, if U is a s.m.s., then the equivalence holds for every non-atomic probability space only if U is universally measurable.

2.6 Technical Appendix

In section 2.4, we discussed that there may be a loss of generality when considering probability metrics defined on the space of pairs

of random elements and space of joint probability measures. While the latter notion is more general, we are accustomed to thinking in terms of the corresponding real-world application which is directly linked to the interpretation of the random elements. Therefore, it is interesting to verify if thinking about probability metrics defined on the space of pairs of random elements is not restrictive in some way. In this appendix, we find out that the answer to this question is negative under some regularity assumptions.

2.6.1 Universally measurable separable metric spaces

What follows is an exposition of some basic results regarding universally measurable separable metric spaces (u.m.s.m.s.). As we shall see, the notion of u.m.s.m.s. plays an important role in TPM.

Definition 2.6.1. Let P be a Borel probability measure on a metric space (U, d). We say that P is *tight* if for each $\varepsilon > 0$, there is a compact $K \subseteq U$ with $P(K) \geq 1 - \varepsilon$. See Dudley (1989), section 11.5.

Definition 2.6.2. A s.m.s. (U, d) is *universally measurable* (u.m.) if every Borel probability measure on U is tight.

Definition 2.6.3. A s.m.s. (U, d) is *Polish* if it is topologically complete (i.e. there is a topologically equivalent metric e such that (U, e) is complete). Here the topological equivalence of d and e simply means that for any $x, x_1, x_2, \ldots$ in U

$$d(x_n, x) \to 0 \iff e(x_n, x) \to 0.$$

Theorem 2.6.1. Every Borel subset of a Polish space is u.m.

Proof. See Billingsley (1968), Theorem 1.4, Cohn (1980), Proposition 8.1.10, and Dudley (1989), p. 391.

Remark 2.6.1. Theorem 2.6.1 provides us with many examples of u.m. spaces, but does not exhaust this class. The topological

characterization of u.m. s.m.s. is a well-known open problem (see Billingsley (1968), Appendix III, p. 234).

In his famous paper on measure theory, Lebesgue (1905) claimed that the projection of any Borel subset of $\mathbb{R}^2$ onto $\mathbb{R}$ is a Borel set. As noted by Souslin and his teacher Lusin (1930), this is in fact not true. As a result of the investigations surrounding this discovery, a theory of such projections (the so-called "analytic" or "Souslin" sets) was developed. Although not a Borel set, such a projection was shown to be Lebesgue-measurable, in fact u.m. This train of thought leads to the following definition.

Definition 2.6.4. Let S be a Polish space and suppose that f is a measurable function mapping S onto a separable metric space U. In this case, we say that U is *analytic*.

Theorem 2.6.2. Every analytic s.m.s. is u.m.

Proof. See Cohn (1980), Theorem 8.6.13, p. 294, and Dudley (1989), Theorem 13.2.6.

Example 2.6.1. Let $\mathbb{Q}$ be the set of rational numbers with the usual topology. Since $\mathbb{Q}$ is a Borel subset of the Polish space R, then $\mathbb{Q}$ is u.m., however, $\mathbb{Q}$ is not itself a Polish space.

Example 2.6.2. In any uncountable Polish space, there are analytic (hence u.m.) non-Borel sets. See Cohn (1980), Corollary 8.2.17 and Dudley (1989), Proposition 13.2.5.

Example 2.6.3. Let $C[0,1]$ be the space of continuous functions $f : [0,1] \to \mathbb{R}$ under the uniform norm. Let $E \subseteq C[0,1]$ be the set of f which fail to be differentiable at some $t \in [0,1]$. Then a theorem of Mazukiewicz (1936) says that E is an analytic, non-Borel subset of $C[0,1]$. In particular, E is u.m.

Recall again the notion of *Hausdorjf metric* $r := r_\rho$ in the space of all subsets of a given metric space (S, ρ)

$$\begin{aligned} r(A, B) &= \max\left\{\sup_{x\in A}\inf_{y\in B}\rho(x, y), \sup_{y\in B}\inf_{x\in A}\rho(x, y)\right\} \\ &= \inf\{\varepsilon > 0 : A^\varepsilon \supseteq B, B^\varepsilon \supseteq A\} \end{aligned} \tag{2.6.1}$$

where A^ε is the open ε-neighborhood of A, $A^\varepsilon = \{x : d(x.A) < \varepsilon\}$.

As we noticed in the space 2^S of all subsets $A \neq \emptyset$ of S, the Hausdorff distance r is actually only a semidistance. However, in the space $\mathcal{C} = \mathcal{C}(S)$ of all closed non-empty subsets, r is a metric (see Definition 2.3.1) and takes on both finite and infinite values, and if S is a bounded set then r is a finite metric on $\mathcal{C}$.

Theorem 2.6.3. Let (S, ρ) be a metric space, and let $(\mathcal{C}(S), r)$ be the space described above. If (S, ρ) is separable [resp. complete; resp. totally bounded], then $(\mathcal{C}(S), r)$ is separable [resp. complete; resp. totally bounded].

Proof. See Hausdorff (1949), section 29, and Kuratowski (1969), sections 21 and 23. □

Example 2.6.4. Let $S = [0, 1]$ and let ρ be the usual metric on S. Let $\mathcal{R}$ be the set of all finite complex-valued Borel measures m on S such that the Fourier transform

$$\widehat{m}(t) = \int_0^1 \exp(\mathrm{i}ut)m(\mathrm{d}u)$$

vanishes at $t = \pm\infty$. Let $\mathcal{M}$ be the class of sets $E \in \mathcal{C}(S)$ such that there is some $m \in \mathcal{R}$ concentrated on E. Then $\mathcal{M}$ is an analytic, non-Borel subset of $(\mathcal{C}(S), r_\rho)$, see Kaufman (1984).

We seek a characterization of u.m. s.m.s. in terms of their Borel structure.

Definition 2.6.5. A measurable space M with σ-algebra $\mathcal{M}$ is *standard* if there is a topology $\mathcal{T}$ on M such that $(M, \mathcal{T})$ is a compact metric space and the Borel σ-algebra generated by $\mathcal{T}$ coincides with $\mathcal{M}$.

A s.m.s. is standard if it is a Borel subset of its completion (see Dudley (1989), p. 347). Obviously, every Borel subset of a Polish space is standard.

Definition 2.6.6. Say that two s.m.s. U and V are called Borel-isomorphic if there is a one-to-one correspondence f of U onto V such that $B \in \mathcal{B}(U)$ if and only if $f(B) \in \mathcal{B}(V)$.

Theorem 2.6.4. Two standard s.m.s. are Borel-isomorphic if and only if they have the same cardinality.

Proof. See Cohn (1980), Theorem 8.3.6 and Dudley (1989), Theorem 13.1.1. □

Theorem 2.6.5. Let U be a separable metric space. The following are equivalent:

(1) U is u.m.

(2) For each Borel probability m on U, there is a standard set $S \in \mathcal{B}(U)$ such that $m(S) = 1$.

Proof. $1 \Rightarrow 2$: Let m be a law on U. Choose compact $K_n \subseteq U$ with $m(K_n) \geq 1 - 1/n$. Put $S = \cup_{n \geq 1} K_n$. Then S is σ-compact, and hence standard. So $m(S) = 1$, as desired.

$2 \Leftarrow 1$: Let m be a law on U. Choose a standard set $S \in \mathcal{B}(U)$ with $m(S) = 1$. Let $\overline{U}$ be the completion of U. Then S is Borel in its completion $\overline{S}$, which is closed in $\overline{U}$. Thus, S is Borel in $\overline{U}$. It follows from Theorem 2.6.1 that

$$1 = m(S) = \sup\{m(K) : K \text{ compact}\}.$$

Thus, every law m on U is tight, so that U is u.m. □

Corollary 2.6.1. Let (U, d) and (V, e) be Borel-isomorphic separable metric spaces. If (U, d) is u.m., then so is (V, e).

Proof. Suppose that m is a law on V. Define a law n on U by $n(A) = m(f(A))$ where $f : U \to V$ is a Borel-isomorphism. Since U is u.m.

there is a standard set $\subseteq U$ with $n(S) = 1$. Then $f(S)$ is a standard subset of V with $m(f(S)) = 1$. Thus, by Theorem 2.6.5, V is u.m. □

The following result, which is in essence due to Blackwell (1956), will be used in an important way later on (cf. the basic theorem of section 4.3, Theorem 4.3.1).

Theorem 2.6.6. Let U be a u.m. separable metric space and suppose that Pr is a probability measure on U. If $\mathcal{A}$ is a countably generated sub-σ-algebra of $\mathcal{B}(U)$, then there is a real-valued function $P(B|x)$, $B \in \mathcal{B}(U)$, $x \in U$ such that

(1) for each fixed $B \in \mathcal{B}(U)$, the mapping $x \to P(B|x)$ is an $\mathcal{A}$-measurable function on U;
(2) for each fixed $x \in U$, the set function $B \to P(B|x)$ is a law on U;
(3) for each $A \in \mathcal{A}$ and $B \in \mathcal{B}(U)$, we have $\int_A P(B|x) \Pr(\mathrm{d}x) = \Pr(A \cap B)$;
(4) there is a set $N \in \mathcal{A}$ with $\Pr(N) = 0$ such that $P(B|x) = 1$ whenever $x \in U - N$.

Proof. Choose a sequence $F_1, F_2, \ldots$ of sets in $\mathcal{B}(U)$ which generates $\mathcal{B}(U)$ and is such that a subsequence generates $\mathcal{A}$. We shall prove that there exists a metric e on U such that (U, d) and (U, e) are Borel-isomorphic and for which the sets $F_1, F_2, \ldots$ are *clopen*, i.e., open and closed. □

Claim 1. If (U, d) is a s.m.s. and $A_1, A_2, \ldots$ is a sequence of Borel subsets of U, then there is some metric e on U such that

(i) (U, e) is a separable metric space isometric with a closed subset of $\mathbb{R}$;
(ii) $A_1, A_2, \ldots$ are clopen subsets of (U, e);
(iii) (U, d) and (U, e) are Borel-isomorphic (see Definition 2.6.6).

Proof of claim. Let $B_1, B_2, \ldots$ be a countable base for the topology of (U, d). Define sets $C_1, C_2, \ldots$ by $C_{2n-1} = A_n$ and $C_{2n} = B_n$ $(n = 1, 2, \ldots)$ and $f : U \to \mathbb{R}$ by $f(x) = \sum_{n=1}^{\infty} 2I_{C_n}(x)/3^n$. Then f is a

Borel-isomorphism of (U, d) onto $f(U) \subseteq K$, where K is the Cantor set,

$$K := \left\{ \sum_{n=1}^{\infty} \alpha_n / 3^n : \alpha_n \text{take value 0 or 2} \right\}.$$

Define the metric e by $e(x, y) = |f(x) - f(y)|$, so that (U, e) is isometric with $f(U) \subseteq K$. Then $A_n = f^{-1}\{x \in K; x(n) = 2\}$, where $x(n)$ is the nth digit in the ternary expansion of $x \in K$. Thus, A_n is clopen in (U, e), as required.

Now (U, e) is (Corollary 2.6.1) u.m., so there are compact sets $K_1 \subseteq K_2 \subseteq \cdots$ with $\Pr(K_n) \to 1$. Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be the (countable) algebras generated by the sequences $F_1, F_2, \ldots$ and $F_1, F_2, \ldots, K_1, K_2, \ldots$, respectively. Then define $P_1(B|x)$ so that (1) and (3) are satisfied for $B \in \mathcal{G}_2$. Since $\mathcal{G}_2$ is countable, there is some set $N \in \mathcal{A}$ with $\Pr(N) = 0$ and such that for $x \in N$,

(a) $P_1(\cdot|x)$ is a finitely additive probability on $\mathcal{G}_2$;
(b) $P_1(A|x) = 1$ for $A \in \mathcal{A} \cap \mathcal{G}_2$ and $x \in A$;
(c) $P_1(K_n|x) \to 1$ as $n \to \infty$.

Claim 2. For $x \in N$, the set function $B \to P_1(B|x)$ is countably additive on $\mathcal{G}_1$.

Proof of claim. Suppose that $H_1, H_2, \ldots$ are disjoint sets in $\mathcal{G}_1$ whose union is U. Since the H_n are clopen and the K_n are compact in (U, e), there is, for each n, some $M = M(n)$ such that $K_n \subseteq H_1 \cup H_2 \cup \cdots \cup H_M$. Finite additivity of $P_1(x, \cdot)$ on $\mathcal{G}_2$ yields, for $x \notin N$, $P_1(K_n|x) \leq \sum_{i=1}^{M} P_1(H_i|x) \leq \sum_{i=1}^{\infty} P_1(H_i|x)$. Let $n \to \infty$ and apply (c) to obtain $\sum_{i=1}^{\infty}(P_1(H_i|x) = 1$, as required.

In view of the claim, for each $x \in N$, we define $B \to P(B|x)$ as the unique countably additive extension of P_1 from $\mathcal{G}_1$ to $\mathcal{B}(U)$. For $x \in N$, put $P(B|x) = \Pr(B)$. Clearly, (2) holds. Now the class of sets in $\mathcal{B}(U)$ for which (1) and (3) hold is a monotone class containing $\mathcal{G}_1$, and so coincides with $\mathcal{B}(U)$.

Claim 3. Condition (4) holds.

Proof of claim. Suppose that $A \in \mathcal{A}$ and $x \in A - N$. Let A_0 be the $\mathcal{A}$-atom containing x. Then $A_0 \subseteq A$ and there is a sequence $A_1, A_2, \ldots$ in $\mathcal{G}_1$ such that $A_0 = A_1 \cap A_2 \cap \cdots$. From (b), $P(A_n|x) = 1$ for $n \geq 1$, so that $P(A_0|x) = 1$, as desired. □

Corollary 2.6.2. Let U and V be u.m. s.m.s. and let Pr be a law on $U \times V$. Then there is a function $P : \mathcal{B}(V) \times U \to \mathbb{R}$ such that

(1) for each fixed $B \in \mathcal{B}(V)$ the mapping $x \to P(B|x)$ is measurable on U;
(2) for each fixed $x \in U$, the set function $B \to P(B|x)$ is a law on V;
(3) for each $A \in \mathcal{B}(U)$ and $B \in \mathcal{B}(V)$, we have

$$\int_{\mathcal{A}} P(B|x) P_1(\mathrm{d}x) = \Pr(A \cap B)$$

where P_1 is the marginal of Pr on U.

Proof. Apply the preceding theorem with $\mathcal{A}$ the σ-algebra of rectangles $A \times U$ for $A \in \mathcal{B}(U)$. □

2.6.2 The equivalence of the notions of p. (semi-)distance on $\mathcal{P}_2$ and on $\mathfrak{X}$

As we have seen in section 2.4, every p. (semi-)distance on $\mathcal{P}_2$ induces (by restriction) a p. (semi-)distance on $\mathfrak{X}$. It remains to be seen whether every p. (semi-)distance on $\mathfrak{X}$ arises in this way. This will certainly be the case whenever

$$\mathcal{L}\mathfrak{X}_2(U, (\Omega, \mathcal{A}, \Pr)) = \mathcal{P}_2(U). \tag{2.6.2}$$

Note that the left member depends not only on the structure of (U, d) but also on the underlying probability space.

In this section we will prove the following facts.

(i) There is some probability space $(\Omega, \mathcal{A}, \Pr)$ such that (2.6.2) holds for every separable metric space U.

(ii) If U is a separable metric space, then (2.6.2) holds for every non-atomic probability space $(\Omega, \mathcal{A}, \Pr)$ if and only if U is universally measurable.

We need a few preliminaries.

Definition 2.6.7. (see Loeve (1963), p. 99, and Dudley (1989), p. 82). If $(\Omega, \mathcal{A}, \Pr)$ is a probability space, we say that $A \in \mathcal{A}$ is an *atom* if $\Pr(A) > 0$ and $\Pr(B) = 0$ or $\Pr(A)$ for each measurable $B \subseteq A$. A probability space is *non-atomic* if it has no atoms.

Lemma 2.6.1. (Berkes and Phillip (1979)). Let v be a law on a complete s.m.s. (U, d) and suppose that $(\Omega, \mathcal{A}, \Pr)$ is a non-atomic probability space. Then there is a U-valued random variable X with distribution $\mathcal{L}(X) = v$.

Proof. Denote by d^* the following metric on $U^2 : d^*(x, y) := d(x_1, x_2) + d(y_1, y_2)$ for $x = (x_1, y_1)$ and $y = (x_2, y_2)$. For each k, there is a partition of U^2 comprising non-empty Borel sets $\{A_{ik} : i = 1, 2, \dots\}$ with $\operatorname{diam}(A_{ik}) < 1/k$ and such that A_{ik} is a subset of some $A_{j,k-1}$.

Since $(\Omega, \mathcal{A}, \Pr)$ is non-atomic, we see that for each $\mathcal{C} \in \mathcal{A}$ and for each sequence p_i of non-negative numbers such that $p_1 + p_2 + \dots = \Pr(\mathcal{C})$, there exists a partitioning $\mathcal{C}_1, \mathcal{C}_2, \dots$ of $\mathcal{C}$ such that $\Pr(\mathcal{C}_i) = p_i$, $i = 1, 2, \dots$ (see e.g. Loeve (1963), p. 99).

Therefore, there exist partitions $\{B_{ik} : i = 1, 2, \dots\} \subseteq \mathcal{A}$, $k = 1, 2, \dots$ such that $B_{ik} \subseteq B_{jk-1}$ for some $j = j(i)$ and $\Pr(B_{ik}) = v(A_{ik})$ for all i, k. For each pair (i, j), let us pick a point $x_{ik} \in A_{ik}$ and define U^2-valued $X_k(\omega) = x_{ik}$ for $\omega \in B_{ik}$. Then $d^*(X_{k+m}(\omega), X_k(\omega)) < 1/k$, $m = 1, 2, \dots$ and since (U^2, d^*) is a complete space, then there exists the limit $X(\omega) = \lim_{k\to\infty} X_k(\omega)$. Thus

$$d^*(X(\omega), X_k(\omega)) \leq \lim_{m\to\infty} [d^*(X_{k+m}(\omega), X(\omega)) + d^*(X_{k+m}(\omega), X_k(\omega))] \leq \frac{1}{k}.$$

Let $P_k := \Pr_{X_k}$ and $P^* := \Pr_X$. Further, our aim is to show that $P^* = v$. For each closed subset $A \subseteq U$

$$P_k(A) = \Pr(X_k \in A) \leq \Pr(X \in A^{1/k}) = P^*(A^{1/k}) \leq P_k(A^{2/k}) \quad (2.6.3)$$

where $A^{1/k}$ is the open $1/k$-neighborhood of A. On the other hand,

$$\begin{aligned} P_k(A) &= \sum\{P_k(x_{ik}) : x_{ik} \in A\} = \sum\{\Pr(B_{ik}) : x_{ik} \in A\} \\ &= \sum\{v(A_{ik}) : x_{ik} \in A\} \leq \sum\{v(A_{ik} \cap A^{1/k}) : x_{ik} \in A\} \\ &\leq v(A^{1/k}) \leq \sum\{v(A_{ik}) : x_{ik} \in A^{2/k}\} \leq P_k(A^{2/k}). \end{aligned} \tag{2.6.4}$$

Further, we can estimate the value $P_k(A^{2/k})$ in the same way as in (2.6.3) and (2.6.4) and thus, we get the inequalities

$$P^*(A^{1/k}) \leq P_k(A^{2/k}) \leq P^*(A^{2/k}) \tag{2.6.5}$$

$$v(A^{1/k}) \leq P_k(A^{2/k}) \leq v(A^{3/k}). \tag{2.6.6}$$

Since $v(A^{1/k})$ tends to $v(A)$ with $k \to \infty$ for each closed set A and analogously $P^*(A^{1/k}) \to P^*(A)$ as $k \to \infty$, then by (2.6.5) and (2.6.6) we obtain the equality

$$P^*(A) = \lim_{k\to\infty} P_k(A^{2/k}) = v(A)$$

for each closed A and hence, $P^* = v$. □

Theorem 2.6.7. There is a probability space $(\Omega, \mathcal{A}, \Pr)$ such that for every separable metric space U and every Borel probability μ on U, there is a random variable $X : \Omega \to U$ with $\mathcal{L}(X) = \mu$.

Proof. Define $(\Omega, \mathcal{A}, \Pr)$ as the measure-theoretic (von Neumann) product (see Hewitt and Stromberg (1965), Theorems 22.7 and 22.8, pp. 432–3) of the probability spaces $(C, \mathcal{B}(C), v)$, where C is some non-empty subset of $\mathbb{R}$ with Borel σ-algebra $\mathcal{B}(C)$ and v is some Borel probability on $(C, \mathcal{B}(C))$.

Now, given a separable metric space U, there is some set $C \subseteq \mathbb{R}$ Borel-isomorphic with U (cf. Claim 1 in Theorem 2.6.6). Let $f : C \to U$ supply the isomorphism. If μ is a Borel probability on U, let v be a probability on C such that $f(v) := vf^{-1} = \mu$. Define $X : \Omega \to U$ as $X = f \circ \pi$, where $\pi : \Omega \to C$ is a projection onto the factor $(C, \mathcal{B}(C), v)$. Then $\mathcal{L}(X) = \mu$, as desired. □

Remark 2.6.2. The result above establishes the claim (i) made at the beginning of the section. It provides one way of ensuring (2.6.2): simply insist that all r.v.s be defined on a 'super-probability space' as in Theorem 2.6.7. We make this assumption throughout the sequel.

The next theorem extends the Berkes and Phillips's Lemma 2.6.1 to the case of u.m. s.m.s. U.

Theorem 2.6.8. Let U be a separable metric space. The following are equivalent.

(1) U is u.m.
(2) If $(\Omega, \mathcal{A}, \Pr)$ is a non-atomic probability space, then for every Borel probability P on U, there is a random variable $X : \Omega \to U$ with law $\mathcal{L}(X) = P$.

Proof. $1 \Rightarrow 2$: Since U is u.m. there is some standard set $S \in \mathcal{B}(U)$ with $P(S) = 1$ (Theorem 2.6.5). Now there is a Borel-isomorphism f mapping S onto a Borel subset B of $\mathbb{R}$ (Theorem 2.6.4). Then $f(P) := P \circ f^{-1}$ is a Borel probability on $\mathbb{R}$. Thus, there is a random variable $g : \Omega \to \mathbb{R}$ with $\mathcal{L}(g) = f(P)$ and $g(\Omega) \subseteq B$ (Lemma 2.6.1 with $(U, d) = (\mathbb{R}, |\cdot|)$). We may assume that $g(\Omega) \subseteq B$ since $\Pr(g^{-1}(B)) = 1$. Define $x : \Omega \to U$ by $x(\omega) = f^{-1}(g(\omega))$. Then $\mathcal{L}(X) = v$, as claimed.

$2 \Rightarrow 1$: Now suppose that v is a Borel probability on U. Consider a random variable $X : \Omega \to U$ on the (non-atomic) probability space $((0,1), \mathcal{B}(0,1), \lambda)$ with $\mathcal{L}(X) = v$. Then range(X) is an analytic subset of U with $v^*(\text{range}(X)) = 1$. Since range(X) is u.m. (Theorem 2.6.2), there is some standard set $S \subseteq \text{range}(X)$ with $P(S) = 1$. This follows from Theorem 2.6.5. The same theorem shows that U is u.m. □

Remark 2.6.3. If U is u.m. s.m.s., we operate under the assumption that all U-valued r.v.s are defined on a non-atomic probability space. Then (2.6.2) will be valid.

References

Berkes, I. and W. Phillip (1979), 'Approximation theorems for independent and weakly independent random vectors', *Ann. Prob.* **7**, 29–54.

Billingsley, P. (1968), *Convergence of Probability Measures*, John Wiley, New York.

Birnbaum, Z. W. and W. Orliz (1931), 'Über die verallgemeinerung des begriffes der zueinander Konjugierten Potenzen', *Studia Math.* **3**, 1–67.

Blackwell, D. (1956), *On a class of probability spaces. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 2, University of California Press, Berkeley, pp. 1–6.

Cohn, D. L. (1980), *Measure Theory*, Birkhauser, Boston.

Dudley, R. M. (1976), *Probabilities and Metrics: Convergence of Laws on Metric Spaces, With a View to Statistical Testing*, Aarhus University Mathematics Institute Lecture Notes Series no. 45, Aarhus.

Dudley, R. M. (1989), *Real Analysis and Probability*, Wadsworth & Brooks-Cole, Pacific Grove, California.

Dunford, N. and J. Schwartz (1988), *Linear Operators. Vol. 1*, Wiley, New York.

Hausdorff, F. (1949), *Set Theory*, Dover, New York.

Hennequin, P. L. and A. Tortrat (1965), *Théorie des probabilités et quelques applications*, Masson, Paris.

Hewitt, E. and K. Stromberg (1965), *Real and Abstract Analysis*, Springer, New York.

Kaufman, R. (1984), 'Fourier transforms and descriptive set theory', *Mathematika* **31**, 336–339.

Kruglov, V. M. (1973), 'Convergence of numerical characteristics of independent random variables with values in a Hilbert space', *Theory Prob. Appl.* **18**, 694–712.

Kuratowski, K. (1969), *Topology,* Vol. II, Academic, New York.

Lebesgue, H. (1905), 'Sur les fonctions representables analytiquement', *J. Math. Pures Appl.* **V**, 139–216.

Loeve, M. (1963), *Probability Theory,* 3rd edn, Van Nostrand, Princeton.

Lukacs, E. (1968), *Stochastic Convergence*, D. C. Heath. Lexington, Mass.

Lusin, N. (1930), *Lecons Sur les Ensembles Analytiaues*, Gauthier-Villars, Paris.

Mazukiewicz, S. (1936), 'Uberdie Menge der differenzierbaren Funktionen', *Fund Math.* **27**, 247–248.

Chapter 3

Choice under Uncertainty

The goals of this chapter are the following:

- To describe the expected utility theory which prescribes the rational behavior of economic agents.
- To relate the notion of stochastic dominance to probability metrics.
- To consider the cumulative prospect theory which is believed to be most successful in explaining individuals' behavior.

Notation introduced in this chapter:

Notation	*Description*
$P_X \succeq P_Y$	Lottery Y is not preferred to lottery X
$u(x)$	The utility function of an economic agent
$Eu(X)$	The expected utility of a lottery X
$\succeq_{FSD}$	First-order stochastic dominance order
$\succeq_{SSD}$	Second-order stochastic dominance order
$\succeq_{TSD}$	Third-order stochastic dominance order
$\succeq_n$	n-th order stochastic dominance order
$v(x)$	The value function of an individual
$w^-(p)$	A function weighting the cumulative probabilities $P(X < x)$

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

Notation	*Description*
$w^+(p)$	A function weighting the tail $P(X > x)$
$V(X)$	A value assigned to the prospect X by an individual depending on the corresponding value function and the weighting functions

Important terms introduced in this chapter:

Term	*Concise explanation*
utility function	A function defining the utility gained by an agent from a given elementary outcome of a random variable
stochastic dominance order	A partial order on the space of random variables introduced according to the preferences of a class of investors
status quo	The current state which represents a reference point relative to which individuals compare future possible outcomes
value function	A function defining the subjective value gained by an individual from a change in profits relative to the status quo

3.1 Introduction

Agents in financial markets operate in a world in which they make choices under risk and uncertainty. Portfolio managers, for example, make investment decisions in which they take risks and expect rewards. They choose to invest in a given portfolio because they believe it is "better" than any other they can buy. Thus, the chosen portfolio is the most preferred one among all portfolios which are admissible for investment. Not all portfolio managers invest in the same portfolio because their expectations and preferences vary.

The theory of how choices under risk and uncertainty are made was introduced by John von Neumann and Oskar Morgenstern in 1944 in their book *Theory of Games and Economic Behavior*. They gave an explicit representation of investors' preferences in terms of an

investor's *utility function*. If no uncertainty is present, the utility function can be interpreted as a mapping between the available alternatives and real numbers indicating the "relative happiness" the investor gains from a particular alternative. If an individual prefers good "A" to good "B", then the utility of "A" is higher than the utility of "B". Thus, the utility function characterizes an individual's preferences. Von Neumann and Morgenstern showed that if there is uncertainty, then it is the *expected utility* which characterizes the preferences. The expected utility of an uncertain prospect, often called a *lottery*, is defined as the probability weighted average of the utilities of the simple outcomes. In fact, the expected utility model was first proposed by Daniel Bernoulli in 1738 as a solution to the famous St Petersburg Paradox but von Neumann and Morgenstern proved that only the expected utility can characterize preferences over lotteries.

The expected utility theory in von Neumann and Morgenstern (1944) defines the lotteries by means of the elementary outcomes and their probability distribution. In this sense, the lotteries can also be interpreted as random variables which can be discrete, continuous, or mixed, and the preference relation is defined on the probability distributions of the random variables. The probability distributions are regarded as *objective*: that is, the theory is consistent with the classical view that, in some sense, the randomness is inherent in Nature and all individuals observe the same probability distribution of a given random variable.

In 1954, a decade after the pioneering von Neumann–Morgenstern theory was published, a new theory of decision making under uncertainty appeared. It was based on the concept that probabilities are not objective, rather they are *subjective* and are a numerical expression of the decision maker's beliefs that a given outcome will occur. This theory was developed by Leonard Savage in his book *The Foundations of Statistics*. Savage (1954) showed that individuals' preferences in the presence of uncertainty can be characterized by an expected utility calculated as a weighted average of the utilities of the simple outcomes, and the weights are the *subjective* probabilities of the outcomes. The subjective probabilities and the utility function arise as a pair from the individual's preferences. Thus, it is possible to modify

the utility function and to obtain another subjective probability measure so that the resulting expected utility also characterizes the individual's preferences. In some aspects, Savage's approach is considered to be more general than the von Neumann–Morgenstern theory.

Another mainstream utility theory describing choices under uncertainty is the state-preference approach of Kenneth Arrow and Gérard Debreu. The basic principle is that the choice under uncertainty is reduced to a choice problem without uncertainty by considering state-contingent bundles of commodities. The agent's preferences are defined over bundles in all states of the world and the notion of randomness is almost ignored. This construction is quite different from the theories of von Neumann–Morgenstern and Savage because preferences are not defined over lotteries. The Arrow–Debreu approach is applied in general equilibrium theories where the payoffs are not measured in monetary amounts but are actual bundles of goods.[1]

In 1992, a new version of the expected utility theory was advanced by Amos Tversky and Daniel Kahneman – the *cumulative prospect theory*. Instead of a utility function, they introduce a *value function* which measures the payoff relative to a reference point. Tversky and Kahneman (1992) also introduce a weighting function which changes the cumulative probabilities of the prospect. The cumulative prospect theory is believed to be a superior alternative to the von Neumann–Morgenstern expected utility theory as it resolves some of the puzzles related to it. Nevertheless, the cumulative prospect theory is a *positive theory*, explaining individuals' behavior, in contrast to the expected utility theory which is a *normative theory* prescribing the rational behavior of agents.

The appeal of utility theories stems from the generality in which the choice under uncertainty is considered. On the basis of such general thinking, it is possible to characterize classes of investors by the shape of their utility function, such as non-satiable investors, risk-averse investors, and so on. Moreover, we are able to identify general rules that a class of investors would follow in choosing between two risky ventures. If all investors of a given class prefer one prospect from another, we say that this prospect *dominates* the

other. In this fashion, the first-, second-, and third-order stochastic dominance relations arise.[2] The stochastic dominance rules characterize the efficient set of a given class of investors; the efficient set consists of all risky ventures which are not dominated by other risky ventures according to the corresponding stochastic dominance relation. Finally, the consequences of stochastic dominance relations are so powerful that any newly formed theory of choice under risk and uncertainty is tested as to whether it is consistent with them.

In this chapter, we briefly describe expected utility theory and the stochastic dominance relations that result. We apply the stochastic dominance relations to the portfolio choice problem and check how the theory of probability metrics can be combined with the stochastic dominance relations.

3.2 Expected Utility Theory

We start with the well-known St Petersburg Paradox, which is historically the first application of the concept of the expected utility function. As a next step, we describe the essential result of von Neumann–Morgenstern characterization of the preferences of individuals.

3.2.1 St Petersburg Paradox

The St Petersburg Paradox is a lottery game presented to Daniel Bernoulli by his cousin Nicolas Bernoulli in 1713. Daniel Bernoulli published the solution in 1734 but another Swiss mathematician, Gabriel Cramer, had already discovered parts of the solution in 1728.

The lottery goes as follows. A fair coin is tossed until a head appears. If the head appears on the first toss, the payoff is \$1.[3] If it appears on the second toss, then the payoff is \$2. After that, the payoff increases sharply. If the head appears on the third toss, the payoff is \$4, on the fourth toss it is \$8, etc. Generally, if the head appears on the n-th toss, the payoff is 2^{n-1} dollars.

At that time, it was commonly accepted that the fair value of a lottery should be computed as the expected value of the payoff. Since

a fair coin is tossed, the probability of having a head on the n-th toss equals $1/2^n$,

$$P(\text{“First head on trial } n\text{”}) = P(\text{“Tail on trial 1”}) \cdot P(\text{“Tail on trial 2”}) \cdot \ldots \cdot P(\text{“Tail on trial } n\text{-1”}) \cdot P(\text{“Head on trial } n\text{”}) = \frac{1}{2^n}$$

Therefore, the expected payoff is calculated as

$$\begin{aligned} \text{Expected payoff} &= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + \ldots + 2^{n-1} \cdot \frac{1}{2^n} + \ldots \\ &= \frac{1}{2} + \frac{1}{2} + \ldots + \frac{1}{2} + \ldots \\ &= \infty. \end{aligned}$$

This result means that people should be willing to participate in the game no matter how large the price of the ticket. Any price makes the game worthwhile because the expected payoff is infinite. Nevertheless, in reality very few people would be ready to pay as much as \$100 for a ticket.

In order to explain the paradox, Daniel Bernoulli suggested that instead of the actual payoff, the utility of the payoff should be considered. Thus, the fair value is calculated by

$$\begin{aligned} \text{Fair value} &= u(1) \cdot \frac{1}{2} + u(2) \cdot \frac{1}{4} + \ldots + u(2^{n-1}) \cdot \frac{1}{2^n} + \ldots \\ &= \sum_{k=1}^{\infty} \frac{u(2^{k-1})}{2^n} \end{aligned}$$

where the function $u(x)$ is the utility function. The general idea is that the value is determined by the utility an individual gains and not directly by the monetary payoff.

Daniel Bernoulli considered utility functions with diminishing marginal utility: that is, the utility gained from one extra dollar diminishes with the sum of money one has. In the solution of the paradox, Bernoulli considered the logarithmic utility function, $u(x) = \log x$, and showed that the fair value of the lottery equals approximately \$2.

The solutions of Bernoulli and Cramer are not completely satisfactory because the lottery can be changed in such a way that the fair value becomes infinite even with their choice of utility functions. Nevertheless, their attempt to solve the problem uses concepts which were later developed into theories of decision making under uncertainty.

3.2.2 The von Neumann–Morgenstern expected utility theory

The St Petersburg Paradox shows that the naive approach to calculate the fair value of a lottery can lead to counter-intuitive results. A deeper analysis shows that it is the utility gained by an individual which should be considered and not the monetary value of the outcomes. The theory of von Neumann–Morgenstern gives a numerical representation of individuals' preferences over lotteries. The numerical representation is obtained through the expected utility and it turns out that this is the only possible way of obtaining a numerical representation.

We used the term "lottery" in the discussion of the game behind the St Petersburg Paradox without providing a definition. Technically, a lottery is a probability distribution defined on the set of payoffs. In fact, the lottery in the St Petersburg Paradox is given in Table 3.1. Generally, lotteries can be discrete, continuous and mixed. Table 3.1 provides an example of a discrete lottery. In the continuous case, the lottery is described by the cumulative distribution function (c.d.f.) of the random payoff. Any portfolio of common stocks, for example, can be regarded as a continuous lottery defined by the c.d.f. of the portfolio payoff. We use the notation P_X to denote the lottery (or the probability distribution), the payoff of which is the random variable

Table 3.1 The lottery in the St Petersburg Paradox.

Probability	$1/2$	$1/4$	$1/8$	$\ldots$	$1/2^n$	$\ldots$
Payoff	1	2	4	$\ldots$	2^{n-1}	$\ldots$

X. The particular values of the random payoff (the outcomes) we denote by lower-case letters, x, and the probability that the payoff is below x is denoted by $P(X \leq x) = F_X(x)$, which is in fact the c.d.f.

Denote by $\mathcal{X}$ the set of all lotteries. Any element of $\mathcal{X}$ is considered a possible choice of an economic agent. If $P_X \in \mathcal{X}$ and $P_Y \in \mathcal{X}$, then there are the following possible cases:

- The economic agent may prefer P_X to P_Y or be indifferent between them, denoted by $P_X \succeq P_Y$.
- The economic agent may prefer P_Y to P_X or be indifferent between them, denoted by $P_Y \succeq P_X$.
- If both relations hold, $P_Y \succeq P_X$ and $P_X \succeq P_Y$, then we say that the economic agent is indifferent between the two choices, $P_X \sim P_Y$.

Sometimes, for notational convenience, we will use $X \succeq Y$ instead of $P_X \succeq P_Y$ without changing the assumption that we are comparing the probability distributions.

A *preference relation* or a *preference order* of an economic agent on the set of all lotteries $\mathcal{X}$ is a relation concerning the ordering of the elements of $\mathcal{X}$, which satisfies certain axioms called the *axioms of choice*. A more detailed description of the axioms of choice is provided in the appendix to this chapter. A numerical representation of a preference order is a real-valued function U defined on the set of lotteries, $U : \mathcal{X} \to \mathbb{R}$, such that $P_X \succeq P_Y$ if and only if $U(P_X) \geq U(P_Y)$,

$$P_X \succeq P_Y \quad \Longleftrightarrow \quad U(P_X) \geq U(P_Y).$$

Thus, the numerical representation characterizes the preference order. In fact, we can take advantage of the numerical representation as comparing real numbers is easier than dealing with the preference order directly.

The von Neumann–Morgenstern theory states that if the preference order satisfies certain technical continuity conditions, then the numerical representation U has the form

$$U(P_X) = \int_{\mathbb{R}} u(x) dF_X(x) \tag{3.2.1}$$

where $u(x)$ is the utility function of the economic agent defined over the elementary outcomes of the random variable X, the probability distribution function of which is $F_X(x)$. Equation (3.2.1) is actually the mathematical expectation of the random variable $u(X)$,

$$U(P_X) = Eu(X),$$

and for this reason the numerical representation of the preference order is, in fact, the expected utility.

Note that the preference order is defined by the economic agent; various agents may have different preference orders. In the equivalent numerical representation, it is the utility function $u(x)$ which characterizes U and, therefore, determines the preference order. In effect, the utility function can be regarded as the fundamental building block which describes the agent's preferences.

As we explained, lotteries may be discrete, continuous, or mixed. If the lottery is discrete, then the payoff is a discrete random variable and equation (3.2.1) becomes

$$U(P_X) = \sum_{j=1}^{n} u(x_j)p_j \tag{3.2.2}$$

where x_j are the outcomes and p_j is the probability that the j-th outcome occurs, $p_j = P(X = x_j)$. The formula for the fair value in the St Petersburg Paradox given by Daniel Bernoulli has the form of equation (3.2.2). Thus, the St Petersburg Paradox is resolved by calculating the fair value through the expected utility of the lottery. If the lottery is such that it has only one possible outcome (i.e., the profit is equal to x with certainty), then the expected utility coincides with the utility of the corresponding payoff, $u(x)$.

3.2.3 Types of utility functions

Some properties of the utility function are derived from common arguments valid for investors belonging to a certain category. For example, concerning certain prospects, all investors who prefer more to less are called *non-satiable*. If there are two prospects, one with a

certain payoff of \$100, and another with a certain payoff of \$200, a non-satiable investor would never prefer the first opportunity. Therefore, the utility function of any such investor should indicate that the utility corresponding to the first prospect should not be less than the utility of the second one, $u(200) \geq u(100)$. We can generalize that the utility functions of non-satiable investors should be non-decreasing,

Non-decreasing property $\quad u(x) \leq u(y)$, if $x \leq y$ for any $x, y \in \mathbb{R}$.

The outcomes x and y can be interpreted as the payoffs of two opportunities without an element of uncertainty, i.e. both x and y occur with probability 1. If the utility function is differentiable, then the non-decreasing property translates as a non-negative first derivative, $u'(x) \geq 0, \ x \in \mathbb{R}$.

Other characteristics of investors' preferences can also be described by the shape of the utility function. Suppose that the investor gains a lower utility from a venture with some expected payoff and a prospect with a certain payoff, equal to the expected payoff of the venture: that is, the investor is *risk averse*. Assume that the venture has two possible outcomes: x_1 with probability p, and x_2 with probability $1 - p$, $p \in [0, 1]$. Thus, the expected payoff of the venture equals $px_1 + (1 - p)x_2$. In terms of the utility function, the risk-aversion property is expressed as

$$u(px_1 + (1 - p)x_2) \geq pu(x_1) + (1 - p)u(x_2), \ \forall x_1, x_2 \text{ and } p \in [0, 1] \tag{3.2.3}$$

where the left-hand side corresponds to the utility of the certain prospect and the right-hand side is the expected utility of the venture. By definition, if a utility function satisfies (3.2.3), then it is called *concave* and, therefore, the utility functions of risk-averse investors should be concave:

Concavity $\quad u(x)$ with support on a set S is said to be a concave function if S is a convex set and if $u(x)$ satisfies (3.2.3) for all $x_1, x_2 \in S$ and $p \in [0, 1]$.

If the utility function is twice differentiable, the concavity property translates as a negative second derivative, $u''(x) \leq 0, \ \forall x \in S$.

A formal measure of absolute risk aversion is the *coefficient of absolute risk aversion*[4] defined by

$$r_A(x) = -\frac{u''(x)}{u'(x)}, \tag{3.2.4}$$

which indicates that the more curved the utility function is, the higher the risk-aversion level of the investor (the more pronounced the inequality in (3.2.3) becomes).

Some common examples of utility functions are listed below.

(a) *Linear utility function*

$$u(x) = a + bx$$

The linear utility function always satisfies (3.2.3) with equality and, therefore, represents a risk-neutral investor. If $b > 0$, then it represents a non-satiable investor.

(b) *Quadratic utility function*

$$u(x) = a + bx + cx^2$$

If $c < 0$, then the quadratic utility function is concave and represents a risk-averse investor.

(c) *Logarithmic utility function*

$$u(x) = \log x, \qquad x > 0$$

The logarithmic utility represents a non-satiable, risk-averse investor. It exhibits a decreasing absolute risk aversion since $r_A(x) = 1/x$ and the coefficient of absolute risk aversion decreases with x.

(d) *Exponential utility function*

$$u(x) = -e^{-ax}, \qquad a > 0$$

The exponential utility represents a non-satiable, risk-averse investor. It exhibits a constant absolute risk aversion since $r_A(x) = a$ and the coefficient of absolute risk aversion does not depend on x.

(e) *Power utility function*

$$u(x) = \frac{-x^{-a}}{a}, \qquad x > 0,\ a > 0$$

The power utility represents a non-satiable, risk-averse investor. It exhibits a decreasing absolute risk aversion since $r_A(x) = a/x$ and the coefficient of absolute risk aversion decreases with x.

3.3 Stochastic Dominance

In section 3.2.3, we noted that key characteristics of investors' preferences determine the shape of the utility function. For example, all non-satiable investors have non-decreasing utility functions and all risk-averse investors have concave utility functions. Thus, different classes of investors can be defined through the general unifying properties of their utility functions.

Suppose that there are two portfolios X and Y, such that all investors from a given class do not prefer Y to X. This means that the probability distributions of the two portfolios differ in a special way that, no matter the particular expression of the utility function, if an investor belongs to the given class, then Y is not preferred by that investor. In this case, we say that portfolio X dominates portfolio Y with respect to the class of investors. Such a relation is often called a *stochastic dominance relation* or a *stochastic ordering*.

Since it is only a relationship between the probability distributions of X and Y which determines whether X dominates Y for a given class of investors, it appears possible to obtain a criterion characterizing the stochastic dominance, involving only the cumulative distribution functions (c.d.f.s) of X and Y. Thus, we are able to identify by only looking at distribution functions of X and Y if either of the two portfolios is preferred by an investor from the class. This section discusses such criteria for three important classes of investors.

3.3.1 First-order stochastic dominance

Suppose that X is an investment opportunity with two possible outcomes – the investor receives \$100 with probability $1/2$ and \$200 with probability $1/2$. Similarly, Y is a venture with two payoffs – \$150 with probability $1/2$ and \$200 with probability $1/2$. A non-satiable investor would never prefer the first opportunity because of the following relationship between the corresponding expected utilities:

$$U(P_X) = u(100)/2 + u(200)/2 \leq u(150)/2 + u(200)/2 = U(P_Y).$$

The inequality arises because $u(100) \leq u(150)$ as a non-satiable investor by definition prefers more to less.

Denote by $\mathcal{U}_1$ the set of all utility functions representing non-satiable investors: that is, the set contains all non-decreasing utility functions. We say that the venture X dominates the venture Y in the sense of the *first-order stochastic dominance* (FSD), $X \succeq_{FSD} Y$, if a non-satiable investor would not prefer Y to X. In terms of the expected utility,

$$X \succeq_{FSD} Y \quad \text{if} \quad Eu(X) \geq Eu(Y), \text{ for any } u \in \mathcal{U}_1.$$

The condition in terms of the c.d.f.s of X and Y characterizing the FSD order is

$$X \succeq_{FSD} Y \quad \text{if and only if} \quad F_X(x) \leq F_Y(x), \ \forall\, x \in \mathbb{R}. \tag{3.3.1}$$

where $F_X(x)$ and $F_Y(x)$ are the c.d.f.s of the two ventures.

Figure 3.1 provides an illustration of the relationship between the two c.d.f.s. If X and Y describe the payoff of two portfolios with distribution functions such as the ones plotted in Figure 3.1, then we can conclude that a non-satiable investor would never invest in Y.

A necessary condition for FSD is that the expected payoff of the preferred venture should exceed the expected payoff of the alternative, $EX \geq EY$ if $X \succeq_{FSD} Y$. This is true because the utility function $u(x) = x$ represents a non-satiable investor as it is non-decreasing and, therefore, it belongs to the set $\mathcal{U}_1$. Consequently, if X is preferred by all non-satiable investors, then it is preferred by the investor with

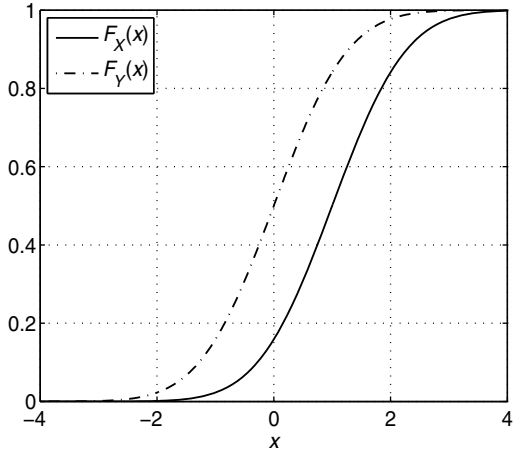

Figure 3.1: An illustration of the first-order stochastic dominance condition in terms of the distribution functions, $X \succeq_{FSD} Y$.

utility function $u(x) = x$ which means that the expected utility of X is not less than the expected utility of Y, $EX \geq EY$.

In general, the converse statement does not hold. If the expected payoff of a portfolio exceeds the expected payoff of another portfolio, it does not follow that any non-satiable investor would necessarily choose the portfolio with the larger expected payoff. This is because the inequality between the c.d.f.s of the two portfolios given in (3.3.1) may not hold. In effect, there will be non-satiable investors who would choose the portfolio with the larger expected payoff and other non-satiable investors who would choose the portfolio with the smaller expected payoff. It depends on the particular expression of the utility function: for example, whether it is a logarithmic or a power utility function.

3.3.2 Second-order stochastic dominance

For decision making under risk, the concept of first-order stochastic dominance is not very useful because the condition in (3.3.1) is rather restrictive. According to the analysis in the previous section,

if the distribution functions of two portfolios satisfy (3.3.1), then a non-satiable investor would never prefer portfolio Y. This conclusion also holds for the subcategory of the non-satiable investors who are also risk-averse. Therefore, the condition in (3.3.1) is only a sufficient condition for this subcategory of investors but is unable to characterize completely their preferences. This is demonstrated in the following example.

Consider a venture Y with two possible payoffs: \$100 with probability $1/2$ and \$200 with probability $1/2$, and a prospect X yielding \$180 with probability 1. A non-satiable, risk-averse investor would never prefer Y to X because the expected utility of Y is not larger than the expected utility of X,

$$Eu(X) = u(180) \geq u(150) \geq u(100)/2 + u(200)/2 = Eu(Y)$$

where $u(x)$ satisfies property (3.2.3) and is assumed to be non-decreasing. The distribution functions of X and Y do not satisfy (3.3.1). Nevertheless, a non-satiable, risk-averse investor would never prefer Y.

Denote by $\mathcal{U}_2$ the set of all utility functions which are non-decreasing and concave. Thus, the set $\mathcal{U}_2$ represents the non-satiable, risk-averse investors and is a subset of $\mathcal{U}_1$, $\mathcal{U}_2 \subset \mathcal{U}_1$. We say that a venture X dominates venture Y in the sense of *second-order stochastic dominance* (SSD), $X \succeq_{SSD} Y$, if a non-satiable, risk-averse investor does not prefer Y to X. In terms of the expected utility,

$$X \succeq_{SSD} Y \quad \text{if} \quad Eu(X) \geq Eu(Y), \text{ for any } u \in \mathcal{U}_2.$$

The condition in terms of the c.d.f.s of X and Y characterizing the SSD order is

$$X \succeq_{SSD} Y \quad \Longleftrightarrow \quad \int_{-\infty}^{x} F_X(t)dt \leq \int_{-\infty}^{x} F_Y(t)dt, \ \forall\, x \in \mathbb{R} \qquad (3.3.2)$$

where $F_X(t)$ and $F_Y(t)$ are the c.d.f.s of the two ventures.

Similarly to FSD, inequality between the expected payoffs is a necessary condition for SSD, $EX \geq EY$ if $X \succeq_{SSD} Y$, because the utility function $u(x) = x$ belongs to the set $\mathcal{U}_2$. In contrast to the FSD, the condition in (3.3.2) allows the distribution functions to intersect. It

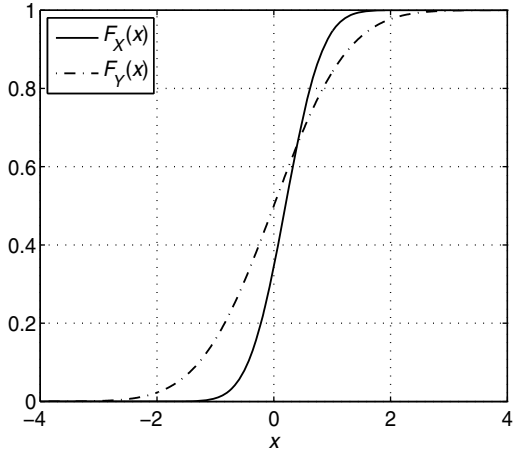

Figure 3.2: An illustration of the second-order stochastic dominance condition in terms of the distribution functions, $X \succeq_{SSD} Y$.

turns out that if the distribution functions cross only once, then X dominates Y with respect to SSD if $F_X(x)$ is below $F_Y(x)$ to the left of the crossing point. Such an illustration is provided in Figure 3.2.

3.3.3 Rothschild–Stiglitz stochastic dominance

In the SSD order, we considered the class of all non-satiable and risk-averse investors. Rothschild and Stiglitz (1970) introduce a slightly different order by dropping the requirement that the investors are non-satiable. A venture X is said to dominate a venture Y in the sense of *Rothschild–Stiglitz stochastic dominance* (RSD),[5] $X \succeq_{RSD} Y$, if no risk-averse investor prefers Y to X. In terms of the expected utility,

$$X \succeq_{RSD} Y \quad \text{if} \quad Eu(X) \geq Eu(Y), \text{ for any concave } u(x).$$

The class of risk-averse investors is represented by the set of all concave utility functions, which contains the set $\mathcal{U}_2$. Thus, the condition in (3.3.2) is only a necessary condition for the RSD but it is not sufficient to characterize the RSD order. If the portfolio X dominates the portfolio Y in the sense of the RSD order, then a risk-averter

would never prefer Y to X. This conclusion holds for the non-satiable risk-averters as well and, therefore, the relation in (3.3.2) holds as a consequence,

$$X \succeq_{RSD} Y \quad \Longrightarrow \quad X \succeq_{SSD} Y.$$

The converse relation is not true. This can be demonstrated with the help of the example developed in section 3.3.2. If the portfolio Y pays off \$100 with probability $1/2$ and \$200 with probability $1/2$ then no risk-averse investor would prefer it to a prospect yielding \$150 with probability 1,

$$u(150) = u(100/2 + 200/2) \geq u(100)/2 + u(200)/2 = Eu(Y),$$

which is just an application of the assumption of concavity in (3.2.3). It is not possible to determine whether a risk-averse investor would prefer a prospect yielding \$150 with probability 1 or the prospect X yielding \$180 with probability 1. Those who are non-satiable would certainly prefer the larger sum but this is not universally true for all risk-averse investors because we do not assume that $u(x)$ is non-decreasing.

The condition which characterizes the RSD stochastic dominance is the following one:

$$X \succeq_{RSD} Y \quad \Longleftrightarrow \quad \begin{cases} EX = EY, \\ \displaystyle\int_{-\infty}^{x} F_X(t)dt \leq \int_{-\infty}^{x} F_Y(t)dt, \ \forall\, x \in \mathbb{R}. \end{cases} \tag{3.3.3}$$

In fact, this is the condition for the SSD order with the additional assumption that the mean payoffs should coincide.

3.3.4 Third-order stochastic dominance

We defined the coefficient of absolute risk aversion $r_A(x)$ in equation (3.2.4). Generally, its values vary for different payoffs depending on the corresponding derivatives of the utility function. Larger values of $r_A(x)$ correspond to a more pronounced risk-aversion effect.

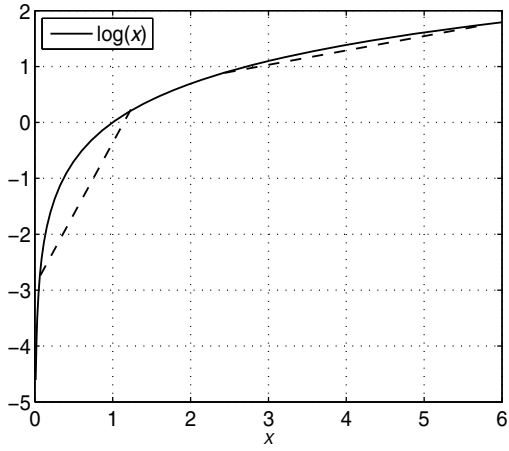

Figure 3.3: The graph of the logarithmic utility function, $u(x) = \log x$. For smaller values of x, the graph is more curved while for larger values of x, the graph is closer to a straight line and, thus, to risk neutrality.

In section 3.2.3, we noted that a negative second derivative of the utility function for all payoffs means that the investor is risk averse at any payoff level. Therefore, the closer $u''(x)$ to zero, the less risk averse the investor since the coefficient $r_A(x)$ decreases, other things held equal. The logarithmic utility function is an example of a utility function exhibiting decreasing absolute risk aversion. The larger the payoff level, the less "curved" the function is, which corresponds to a closer to zero second derivative and a less pronounced risk-aversion property. An illustration is given in Figure 3.3.

Utility functions exhibiting a decreasing absolute risk aversion are important because the investors they represent favor positive to negative skewness. This is a consequence of the decreasing risk aversion – at higher payoff levels such investors are less inclined to avoid risk in comparison to lower payoff levels at which they are much more sensitive to risk taking. Technically, a utility function with a decreasing absolute risk aversion has a non-negative third derivative, $u'''(x) \geq 0$, as this means that the second derivative is non-decreasing.

Denote by $\mathcal{U}_3$ the set of all utility functions which are non-decreasing, concave, and have a non-negative third derivative, $u'''(x) \geq 0$. Thus, $\mathcal{U}_3$ represents the class of non-satiable, risk-averse investors who prefer positive to negative skewness. A venture X is said to dominate a venture Y in the sense of *third-order stochastic dominance* (TSD), $X \succeq_{TSD} Y$, if an investor with a utility function from the set $\mathcal{U}_3$ does not prefer Y to X. In terms of the expected utility,

$$X \succeq_{TSD} Y \quad \text{if} \quad Eu(X) \geq Eu(Y), \text{ for any } u \in \mathcal{U}_3.$$

The set of utility functions $\mathcal{U}_3$ is contained in the set of non-decreasing, concave utilities, $\mathcal{U}_3 \subset \mathcal{U}_2$. Therefore, the condition (3.3.2) for SSD is only sufficient in the case of TSD,

$$X \succeq_{SSD} Y \quad \Longrightarrow \quad X \succeq_{TSD} Y.$$

The condition which characterizes the TSD stochastic dominance is

$$X \succeq_{TSD} Y \quad \Longleftrightarrow \quad E(t-X)_+^2 \leq E(t-Y)_+^2, \ \forall\, t \in \mathbb{R} \tag{3.3.4}$$

where the notation $(t-x)_+^2$ means the maximum between $t-x$ and zero raised to the second power, $(t-x)_+^2 = (\max(t-x, 0))^2$. The quantity $E(t-X)_+^2$ is known as the *second lower partial moment* of the random variable X. It measures the variability of X below a target payoff level t. Suppose that X and Y have equal means and variances. If X has a positive skewness and Y has a negative skewness, then the variability of X below any target payoff level t will be smaller than the variability of Y below the same target payoff level.

At first sight, (3.3.4) has nothing to do with (3.3.2) and it is not clear that SSD entails TSD. In fact, it is only a matter of algebraic manipulations to show that, indeed, if (3.3.2) holds, then (3.3.4) holds as well.

3.3.5 Efficient sets and the portfolio choice problem

Taking advantage of the criteria for stochastic dominance discussed in the previous sections, we can characterize the *efficient sets* of the corresponding categories of investors. The efficient set of a given class of investors is defined as the set of ventures not dominated

with respect to the corresponding stochastic dominance relation. For example, the efficient set of the non-satiable investors is the set of those ventures which are not dominated with respect to the FSD order. As a result, by construction, any venture which is not in the efficient set will be necessarily discarded by all investors in the class.

The portfolio choice problem of a given investor can be divided into two steps. The first step concerns finding the efficient set of the class of investors which the given investor belongs to. Any portfolio not belonging to the efficient set will not be selected by any of the investors in the class and is, therefore, suboptimal for the investor. Such a class may be composed of, for example, all non-satiable, risk-averse investors if the utility function of the given investor is non-decreasing and concave. In this case, the efficient set comprises all portfolios not dominated with respect to the SSD order. Note that in this step, we do not take advantage of the particular expression for the utility function of the investor.

Once we have obtained the efficient set, we proceed to the second step in which we calculate the expected utility of the investor for the portfolios in the efficient set. The portfolio which maximizes the investor's expected utility represents the optimal choice of the investor.

The difficulty of adopting this approach in practice is that it is very hard to obtain explicitly the efficient sets. That is why the problem of finding the optimal portfolio for the investor is very often replaced by a simpler one, involving only certain characteristics of the portfolios return distributions, such as the expected return and the risk. In this situation, it is critical that the simpler problem is consistent with the corresponding stochastic dominance relation in order to guarantee that its solution is among the portfolios in the efficient set. Checking the consistency reduces to choosing a risk measure which is compatible with the stochastic dominance relation.

3.3.6 Return versus payoff

Note that the expected utility theory deals with the portfolio payoff and not the portfolio return. Nevertheless, all relations defining the

stochastic dominance orders can be adopted if we consider the distribution functions of portfolio *returns* rather than portfolio *profits*. In the following, we examine the FSD and SSD orders concerning log-return distributions and the connection to the corresponding orders concerning random payoffs. The logarithmic return, or simply the log-return, is a central concept in fundamental theories in finance, such as derivative pricing and modern portfolio theory. Therefore, it makes sense to consider stochastic orders with respect to log-return distributions rather than payoff.

Suppose that P_t is a random variable describing the price of a common stock at a future time t, $t > 0$ where $t = 0$ is present time. Without loss of generality, we can assume that the stock does not pay dividends. Denote by r_t the log-return for the period $(0, t)$,

$$r_t = \log \frac{P_t}{P_0},$$

where P_0 is the price of the common stock at present and is a non-random positive quantity. The random variable P_t can be regarded as the random payoff of the common stock at time t, while r_t is the corresponding random log-return. The formula expressing the random payoff in terms of the random log-return is

$$P_t = P_0 \exp(r_t).$$

Even though log-returns and payoffs are directly linked by means of the above formulae, it turns out that, generally, stochastic dominance relations concerning two log-return distributions are not equivalent to the corresponding stochastic dominance relations concerning their payoff distributions.

Consider an investor with utility function $u(x)$ where $x > 0$ stands for payoff. In the appendix to this chapter, we demonstrate that the utility function of the investor concerning the log-return can be expressed as

$$v(y) = u(P_0 \exp(y)), \ y \in \mathbb{R} \tag{3.3.5}$$

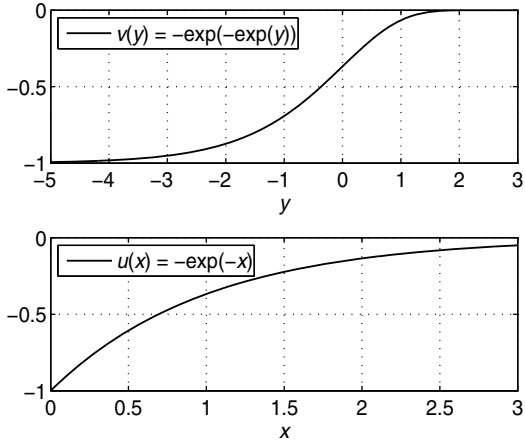

Figure 3.4: $u(x)$ represents a non-satiable and risk-averse investor on the space of payoffs and $v(y)$ is the corresponding utility on the space of log-returns. Apparently, $v(y)$ is not concave.

where y stands for the log-return of a common stock and P_0 is the price at present.[6] Equation (3.3.5) and the inverse,

$$u(x) = v(\log(x/P_0)),\ x > 0, \tag{3.3.6}$$

provide the link between utilities concerning log-returns and payoff.

It turns out that an investor who is non-satiable and risk averse with respect to payoff distributions may not be risk averse with respect to log-return distributions. The utility function $u(x)$ representing such an investor has the properties

$$u'(x) \geq 0 \quad \text{and} \quad u''(x) \leq 0,\ \forall x > 0,$$

but it does not follow that the function $v(y)$ given by (3.3.5) will satisfy them. In fact, $v(y)$ also has non-positive first derivative but the sign of the second derivative can be arbitrary. Therefore the investor is non-satiable but may not be risk-averse with respect to log-return distributions. This fact is illustrated in Figure 3.4 for the exponential utility function.

Conversely, an investor who is non-satiable and risk averse with respect to log-return distributions is also non-satiable and risk averse concerning payoff distributions. This is true because if $v(y)$ satisfies the corresponding derivative inequalities, so does $u(x)$ given by (3.3.6). Consequently, it follows that the investors who are non-satiable and risk averse on the space of log-return distributions are a subclass of those who are non-satiable and risk averse on the space of payoff distributions.

This analysis implies that the FSD order of two common stocks, for example, remains unaffected as to whether we consider their payoff distributions or their log-return distributions,

$$P_t^1 \succeq_{FSD} P_t^2 \qquad \Longleftrightarrow \qquad r_t^1 \succeq_{FSD} r_t^2,$$

where P_t^1 and P_t^2 are the payoffs of the two common stocks at time $t > t_0$, and r_t^1 and r_t^2 are the corresponding log-returns for the same period. However, such an equivalence does not hold for the SSD order. Actually, the SSD order on the space of payoff distributions implies the same order on the space of log-return distributions but not vice versa,

$$P_t^1 \succeq_{SSD} P_t^2 \qquad \Longrightarrow \qquad r_t^1 \succeq_{SSD} r_t^2.$$

In the appendix to this chapter, we demonstrate that the same conclusion holds for the TSD order and, generally, for the n-th order stochastic dominance, $n > 1$. Such kinds of relations deserve a closer scrutiny as optimal portfolio problems are usually set in terms of returns, and consistency with a stochastic dominance relation implies that the stochastic dominance relation is also set on the space of return distributions, not on the space of payoff distributions. Moreover, in this section we considered only one-period returns. In a multi-period setting, for example in the area of asset-liability management, matters get even more involved.

Note that these relations are always true if the present values of the two ventures are equal, $P_0^1 = P_0^2$. Otherwise they may be violated. Consider, for example, the FSD order of random payoffs. Suppose that P_t^1 dominates P_t^2 with respect to the FSD order, $P_t^1 \succeq_{FSD} P_t^2$. Then,

according to the characterization in terms of the c.d.f.s we obtain

$$F_{P_t^1}(x) \leq F_{P_t^2}(x), \quad \forall x \in \mathbb{R}.$$

We can represent this inequality in terms of the log-returns r_t^1 and r_t^2 in the following way:

$$P\left(r_t^1 \leq \log \frac{x}{P_0^1}\right) \leq P\left(r_t^2 \leq \log \frac{x}{P_0^2}\right), \quad \forall x \in \mathbb{R}.$$

In fact, the above inequality implies that $r_t^1 \succeq_{FSD} r_t^2$ if $P_0^1 = P_0^2$. In case the present values of the ventures differ a lot, it may happen that the c.d.f.s of the log-return distributions do not satisfy the inequality $F_{r_t^1}(y) \leq F_{r_t^2}(y)$ for all $y \in \mathbb{R}$, which means that the FSD order may not hold.

3.4 Probability Metrics and Stochastic Dominance

The conditions for stochastic dominance involving the distribution functions of the ventures X and Y represent a powerful method to determine if an entire class of investors would prefer any of the portfolios. For example, in order to verify if any non-satiable, risk-averse investor would not prefer Y to X, we have to verify if condition (3.3.2) holds. Note that a negative result does not necessarily mean that any such investor would actually prefer Y or be indifferent between X and Y. It may be the case that the inequality between the quantities in (3.3.2) is satisfied for some values of the argument, and for others, the converse inequality holds. That is, neither $X \succeq_{SSD} Y$ nor $Y \succeq_{SSD} X$ is true. Thus, only a part of the non-satiable, risk-averse investors may prefer X to Y; it now depends on the particular investor we consider.

Suppose the verification confirms that either X is preferred or the investors are indifferent between X and Y, $X \succeq_{SSD} Y$. This result is only qualitative, there are no indications whether Y would be categorically disregarded by all investors in the class, or the differences between the two portfolios are very small. Similarly, if we know that

no investors from the class prefer Y to Z, $Z \succeq_{SSD} Y$, then can we determine whether Z is more strongly preferred to Y than X is?

The only way to approach these questions is to add a quantitative element through a probability metric since only by means of a probability metric can we calculate distances between random quantities.[7] For example, we can choose a probability metric μ and we can calculate the distances $\mu(X, Y)$ and $\mu(Z, Y)$. If $\mu(X, Y) < \mu(Z, Y)$, then the return distribution of X is "closer" to the return distribution of Y than are the return distributions of Z and Y. On this ground, we can draw the conclusion that Z is more strongly preferred to Y than X is, on condition that we know in advance the relations $X \succeq_{SSD} Y$ and $Z \succeq_{SSD} Y$.

However, not any probability metric appears suitable for this calculation. This is illustrated by the following example. Suppose that Y and X are normally distributed random variables describing portfolio returns with equal means, $X \in N(a, \sigma_X^2)$ and $Y \in N(a, \sigma_Y^2)$, with $\sigma_X^2 < \sigma_Y^2$. Z is a prospect yielding a dollars with probability 1. The c.d.f.s $F_X(x)$ and $F_Y(x)$ cross only once at $x = a$ and the $F_X(x)$ is below $F_Y(x)$ to the left of the crossing point because the variance of X is assumed to be smaller than the variance of Y. Therefore, according to the condition in (3.3.3), no risk-averse investor would prefer Y to X and consequently $X \succeq_{SSD} Y$. The prospect Z provides a non-random return equal to the expected returns of X and Y, $EX = EY = a$, and, in effect, any risk-averse investor would rather choose Z from the three alternatives, $Z \succeq_{SSD} X \succeq_{SSD} Y$.

A probability metric with which we would like to quantify the second-order stochastic dominance relation should be able to indicate that, first, $\mu(X, Y) < \mu(Z, Y)$ because Z is more strongly preferred to Y and, second, $\mu(Z, X) < \mu(Z, Y)$ because Y is more strongly rejected than X with respect to Z. The assumptions in the example give us the information to order completely the three alternatives and that is why we are expecting the two inequalities should hold.

Let us choose the Kolmogorov metric,[8]

$$\boldsymbol{\rho}(X, Y) = \sup_{x \in \mathbb{R}} |F_X(x) - F_Y(x)|,$$

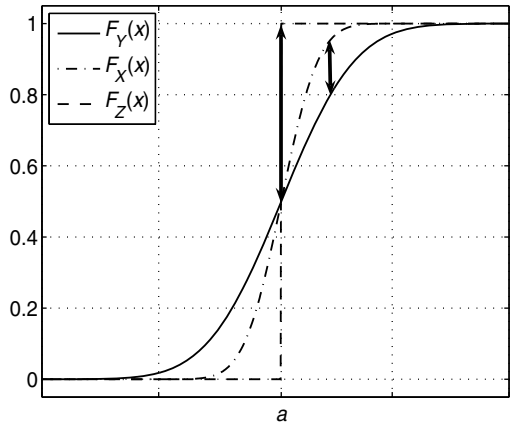

Figure 3.5: The distribution functions of two normal distributions with equal means, $EX = EY = a$ and the distribution function of $Z = a$ with probability 1. The arrows indicate the corresponding Kolmogorov distances.

for the purpose of calculating the corresponding distances. It computes the largest absolute difference between the two distribution functions. Applying the definition to the distributions in the example, we obtain that $\rho(X, Z) = \rho(Y, Z) = 1/2$ and $\rho(X, Y) < 1/2$. As a result, the Kolmogorov metric is capable of showing that Z is more strongly preferred relative to Y but cannot show that Y is more strongly rejected with respect to Z. Figure 3.5 contains a plot of the c.d.f.s of the three random variables. The arrows indicate where the largest absolute difference between the corresponding c.d.f.s is located. The arrow length equals the Kolmogorov distance.

The example shows that there are probability metrics which are not appropriate to quantify a stochastic dominance order. The task of finding a suitable metric is not a simple one because the structure of the metric should be based on the conditions defining the dominance order. Inevitably, we cannot expect that one probability metric will appear suitable for all stochastic orders; rather, a probability metric may be best suited for a selected stochastic dominance relation.

Technically, we have to impose another condition in order for the problem of quantification to have a practical meaning. The

probability metric calculating the distances between the ordered random variables should be bounded. If it explodes, then we cannot draw any conclusions. For instance, if $\mu(X, Y) = \infty$ and $\mu(Z, Y) = \infty$, then we cannot compare the investors' preferences.

Concerning the FSD order, a suitable choice for a probability metric is the Kantorovich metric,

$$\boldsymbol{\kappa}(X, Y) = \int_{-\infty}^{\infty} |F_X(x) - F_Y(x)| dx,$$

introduced in equation (2.2.5) in Chapter 2 . Note that the condition in (3.3.1) can be restated as $F_X(x) - F_Y(x) \leq 0, \ \forall x \in \mathbb{R}$. Thus, summing up all absolute differences gives an idea how "close" X is to Y, which is a natural way of measuring the distance between X and Y with respect to the FSD order. The Kantorovich metric is finite as long as the random variables have finite means. We can always count on this assumption if the random variables describe portfolio returns, for example.

The RSD order can also be quantified in a similar fashion. Consider the Zolotarev ideal metric,

$$\boldsymbol{\zeta}_2(X, Y) = \int_{-\infty}^{\infty} \left| \int_{-\infty}^{x} F_X(t) dt - \int_{-\infty}^{x} F_Y(t) dt \right| dx,$$

introduced in Chapter 4. The structure of this probability metric is directly based on the condition in (3.3.3) and it calculates in a natural way the distance between X and Y with respect to the RSD order. The requirement that $EX = EY$ in (3.3.3) combined with the additional assumption that the second moments of X and Y are finite, $EX^2 < \infty$ and $EY^2 < \infty$, represent the needed sufficient conditions for the boundedness of $\boldsymbol{\zeta}_2(X, Y)$.

3.5 Cumulative Prospect Theory

In the introduction to this chapter, we noted that expected utility theory is prescriptive in the sense that it determines what the rational behavior of economic agents should be. Empirical work in the field of

behavioral finance has identified certain situations in which people behave differently from the rational prescription of expected utility theory. Some of these instances of inconsistency are often called paradoxes, such as Allias's paradox and the Ellsberg paradox. Other instances concern particular psychological effects. For example, the so-called *framing effect* states that people's choice can vary depending on the wording of a problem – whether more emphasis is placed on the potential loss or the potential profit. Another example is the *status quo bias*. It is described as the natural preference people have for the current state. The *loss aversion effect* states that the disutility of giving up an object is greater than the utility associated with acquiring it. In a sense, forgone gains are less painful than perceived losses. For more detailed information and further examples, the reader is referred to Kahneman et al. (1991), Rabin and Thaler (2001), and Siegel and Thaler (1997).

The situations in which expected utility theory fails in describing the corresponding behavior gave rise to theories aiming at explaining observed behavior, rather than trying to prescribe rational behavior. There are a few such theories, the most prominent of which is the cumulative prospect theory proposed by Tversky and Kahneman (1992). It is built upon the following main observations. First, investors usually think about possible outcomes relative to a certain reference point rather than the final outcome. This is referred to as the status quo by behavioral finance theorists. Second, investors have different attitudes towards gains (outcomes which are larger than the reference point) and losses (outcomes which are less than the reference point), referred to by behavioral finance theorists as loss aversion. Finally, investors tend to overweight extreme events and underweight events with higher probability.

Cumulative prospect theory has been applied to a diverse range of problems, such as the asset allocation puzzle, the equity premium puzzle, the status quo bias, and various gambling and betting puzzles which appear inconsistent with standard economic rationality: see, for example, Benartzi and Thaler (1995) and Kahneman et al. (1991). However, the applications are in a simple discrete setting and only a few attempts have been made to apply the theory in a

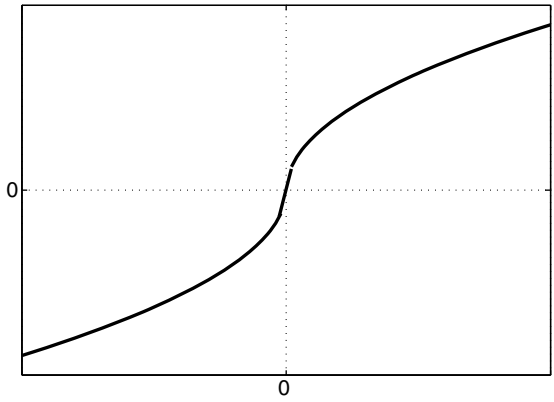

Figure 3.6: The graph of an s-shaped function.

more complicated setting: see, for example, Baucells and Heukamp (2006) and Hwang and Satchell (2003).

Cumulative prospect theory arises as an alternative to expected utility theory on the basis of the observations outlined above. The utility function, which is a basic concept in expected utility theory, is replaced by a value function. In a similar fashion, cumulative probabilities are replaced by weighted cumulative probabilities. The value function, $v(x)$, assigns values to the possible outcomes. It is non-decreasing and $v(0) = 0$ since the outcome equal to the reference point brings no value to the individual. Different functional forms for $v(x)$ have been suggested. It is often assumed that the $v(x)$ has an s-shaped form, i.e. $v(x)$ is convex for $x < 0$ and it is concave for $x > 0$ (see Kahneman and Tversky, 1979). An illustration is provided in Figure 3.6. From a financial viewpoint, the value function can be constructed for returns rather than wealth as in the classical expected utility theory. In this way, cumulative prospect theory is the natural setting for the discussion about return versus payoff based stochastic dominance in section 3.3.6. In order to account for the observed loss-aversion effect, the value function is assumed to be radially asymmetric, i.e. $v(x) \neq -v(-x)$. Furthermore, it is assumed to be steeper for losses than for gains.

The weighted cumulative probabilities are usually modeled as transformations of the c.d.f. of the prospect $F_X(x) = P(X \leq x)$ and the tail $1 - F_X(x) = P(X > x)$ depending on whether $x < 0$ or $x > 0$, respectively. The transformation for the losses is denoted by $w^-(p)$ and the one for the profits is denoted by $w^+(p)$. Both weighting functions are non-decreasing and satisfy the following conditions:

$$w^-(0) = w^+(0) = 0$$
$$w^-(1) = w^+(1) = 1.$$

Empirical studies suggest that the general shape of the weighting functions is inverse s-shaped: see, for example, Tversky and Kahneman (1992).

According to cumulative prospect theory, individuals make a choice between two risky prospects X and Y by computing the subjective expected values according to the functional

$$V(X) = \int_{-\infty}^{0} v(x)d[w^-(F_X(x))] + \int_{0}^{\infty} v(x)d[-w^+(1 - F_X(x))] \tag{3.5.1}$$

and then compare $V(X)$ and $V(Y)$: see, for example, Baucells and Heukamp (2006). The first summand is focused on losses and the second summand is related to gains. If $V(X) \geq V(Y)$, then Y is not preferred to X. If $V(X) = V(Y)$, then the individual is indifferent. Note that if the individuals do not weight the cumulative probabilities, i.e. $w^-(p) = w^+(p) = p$, then the definition of $V(X)$ reduces to

$$V(X) = Ev(X) = \int_{-\infty}^{\infty} v(x)dF_X(x).$$

The expression (3.5.1) implies that we can map all individuals to pairs (v, w) where v is the corresponding value function and w is a shorthand for both w^- and w^+. In this way, a set of individuals can be represented by $(v_j, w_j),\ j \in J$.

3.6 Summary

In this chapter, we considered the problem of choice under uncertainty as described by the classical von Neumann–Morgenstern expected utility theory. We also described the most important types of stochastic dominance relations resulting from the theory, which characterize the choices of entire classes of investors. One application of the theory of probability metrics in stochastic dominance relations is to add a quantitative element to their qualitative nature. Instead of knowing only that a venture is preferred to another venture by a whole class of investors, a probability metric is capable of showing if the differences between the two ventures are very small, or one of the two ventures is categorically discarded by the entire class.

Another major point concerning stochastic dominance relations is to take into account if probability distributions of returns or payoffs are considered. Usually, optimal portfolio problems are set in terms of returns and consistency with the SSD order is sought. In such a case, the SSD order concerns distributions of returns, rather than payoffs and this should be borne in mind when analyzing the solution.

This consideration is in line with cumulative prospect theory, which is an alternative to expected utility theory arising from the field of behavioral economics. According to it, not only do individuals value outcomes relative to a reference point, but they also subjectively weight cumulative probabilities.

3.7 Technical Appendix

In this appendix, we state the axioms of choice, which are the basis for von Neumann–Morgenstern theory, and we comment on the uniqueness of the expected utility representation of a preference order. The stochastic orders given in the chapter concern the most important classes of investors. We give examples of several others in the appendix. Finally, we briefly mention a parallel between representations of probability metrics known as *dual* and stochastic orders.

3.7.1 The axioms of choice

The axioms of choice are fundamental assumptions defining a preference order. In the following, $\mathcal{X}$ stands for the set of the probability distributions of the ventures also known as lotteries, and the notation $P_X \succeq P_Y$ means that the economic agent prefers P_X to P_Y or is indifferent between the two choices. The notation $P_X \succ P_Y$ means that P_X is strictly preferred to P_Y. The axioms of choice are the following:

Completeness	For all $P_X, P_Y \in \mathcal{X}$, either $P_X \succeq P_Y$ or $P_Y \succeq P_X$ or both are true, $P_X \sim P_Y$.
Transitivity	If $P_X \succeq P_Y$ and $P_Y \succeq P_Z$, then $P_X \succeq P_Z$, where P_X, P_Y and P_Z are three lotteries.
Archimedean axiom	If $P_X, P_Y, P_Z \in \mathcal{X}$ are such that $P_X \succ P_Y \succ P_Z$, then there is an $\alpha, \beta \in (0,1)$ such that $\alpha P_X + (1-\alpha)P_Z \succ P_Y$ and also $P_Y \succ \beta P_X + (1-\beta)P_Z$.
Independence axiom	For all $P_X, P_Y, P_Z \in \mathcal{X}$ and any $\alpha \in [0,1]$, $P_X \succeq P_Y$ if and only if $\alpha P_X + (1-\alpha)P_Z \succeq \alpha P_Y + (1-\alpha)P_Z$.

The completeness axiom states that economic agents should always be able to compare two lotteries (e.g., two portfolios). They either prefer one or the other, or are indifferent. The transitivity axiom rules out the possibility that an investor may prefer P_X to P_Y, P_Y to P_Z, and also P_Z to P_X. It states that if the first two relations hold, then necessarily the investor should prefer P_X to P_Z. The Archimedean axiom is like a "continuity" condition. It states that given any three distributions strictly preferred to each other, we can combine the most and the least preferred distribution through an $\alpha \in (0,1)$ such that the resulting distribution is strictly preferred to the middle distribution. Likewise, we can combine the most and the least preferred distribution through a $\beta \in (0,1)$ so that the middle distribution is strictly preferred to the resulting distribution. The independence axiom claims that the preference between two lotteries remains unaffected if they are both combined in the same way with a third lottery.

The basic result of von Neumann–Morgenstern is that a preference relation satisfies the four axioms of choice if and only if there is a real-valued function, $U : \mathcal{X} \to \mathbb{R}$, such that:

(a) U represents the preference order

$$P_X \succeq P_Y \qquad \Longleftrightarrow \qquad U(P_X) \geq U(P_Y)$$

for all $P_X, P_Y \in \mathcal{X}$.

(b) U has the linear property[9]

$$U(\alpha P_X + (1-\alpha)P_Y) = \alpha U(P_X) + (1-\alpha)U(P_Y)$$

for any $\alpha \in (0,1)$ and $P_X, P_Y \in \mathcal{X}$.

Moreover, the numerical representation U is unique up to a positive linear transform. That is, if U_1 and U_2 are two functions representing one and the same preference order, then $U_2 = aU_1 + b$ where $a > 0$ and b are some coefficients.

It turns out that the numerical representation has a very special form under some additional technical continuity conditions. It can be expressed as

$$U(P_X) = \int_{\mathbb{R}} u(x) dF_X(x)$$

where the function $u(x)$ is the utility function of the economic agent and $F_X(x)$ is the c.d.f. of the probability distribution P_X. Thus, the numerical representation of the preference order of an economic agent is the expected utility of X. The fact that U is known up to a positive linear transform means that the utility function of the economic agent is not determined uniquely from the preference order but is also unique up to a positive linear transform.

3.7.2 Stochastic dominance relations of order n

In the chapter, we introduced the first-, second-, and third-order stochastic dominance relations which represent non-satiable

investors, non-satiable and risk-averse investors, and non-satiable, risk-averse investors preferring positive to negative skewness. That is, including additional characteristics of the investors by imposing conditions on the utility function, we end up with more refined stochastic orders.

This method can be generalized in the n-th order stochastic dominance. Denote by $\mathcal{U}_n$ the set of all utility functions, the derivatives of which satisfy the inequalities $(-1)^{k+1}u^{(k)}(x) \geq 0,\ \ k = 1, 2, \ldots, n$ where $u^{(k)}(x)$ denotes the k-th derivative of $u(x)$. For each n, we have a set of utility functions which is a subset of $\mathcal{U}_{n-1}$,

$$\mathcal{U}_1 \subset \mathcal{U}_2 \subset \ldots \subset \mathcal{U}_n \subset \ldots$$

The classes of investors characterized by the first-, second-, and third-order stochastic dominance are $\mathcal{U}_1$, $\mathcal{U}_2$, and $\mathcal{U}_3$.

Imposing further properties on the derivatives of the utility function requires that we make more assumptions for the moments of the random variables we consider. We assume that the absolute moments $E|X|^k$ and $E|Y|^k$, $k = 1, \ldots, n$ of the random variables X and Y are finite. We say that the portfolio X dominates the portfolio Y in the sense of the n-th order stochastic dominance, $X \succeq_n Y$, if no investor with a utility function in the set $\mathcal{U}_n$ would prefer Y to X,

$$X \succeq_n Y \quad \text{if} \quad Eu(X) \geq Eu(Y),\ \forall u(x) \in \mathcal{U}_n.$$

Thus, the first-, second-, and third-order stochastic dominance appear as special cases from the n-th order stochastic dominance with $n = 1, 2, 3$.

There is an equivalent way of describing the n-th order stochastic dominance in terms of the c.d.f.s of the ventures only. The condition is

$$X \succeq_n Y \qquad \Longleftrightarrow \qquad F_X^{(n)}(x) \leq F_Y^{(n)}(x),\ \forall x \in \mathbb{R} \tag{3.7.1}$$

where $F_X^{(n)}(x)$ stands for the n-th integral of the c.d.f. of X, which can be defined recursively as

$$F_X^{(n)}(x) = \int_{-\infty}^{x} F_X^{(n-1)}(t)dt.$$

An equivalent form of the condition in (3.7.1) can be derived, which is close to the form of (3.3.4),

$$X \succeq_n Y \qquad \Longleftrightarrow \qquad E(t-X)_+^{n-1} \leq E(t-Y)_+^{n-1}, \ \forall t \in \mathbb{R} \qquad (3.7.2)$$

where $(t-x)_+^{n-1} = \max(t-x, 0)^{n-1}$. This equivalent formulation clarifies why it is necessary to assume that all absolute moments until order n are finite.

Since in the n-th order stochastic dominance we furnish the conditions on the utility function as n increases, the following relation holds,

$$X \succeq_1 Y \quad \Longrightarrow \quad X \succeq_2 Y \quad \Longrightarrow \quad \ldots \quad \Longrightarrow \quad X \succeq_n Y,$$

which generalizes the relationship between FSD, SSD, and TSD given in the chapter.

Further on, it is possible to extend the n-th order stochastic dominance to the α-order stochastic dominance in which $\alpha \geq 1$ is a real number and instead of the ordinary integrals of the c.d.f.s, fractional integrals are involved. Ortobelli et al. (2007) provide more information on extensions of stochastic dominance orderings and their relation to probability metrics and risk measures.

3.7.3 Return versus payoff and stochastic dominance

The lotteries in von Neumann–Morgenstern theory are usually interpreted as probability distributions of payoffs. That is, the domain of the utility function $u(x)$ is the positive half-line which is interpreted as the collection of all possible outcomes in terms of dollars from a given venture. Assume that the payoff distribution is actually the price distribution P_t of a financial asset at a future time t. In line with the von Neumann–Morgenstern theory, the expected utility of P_t for an investor with utility function $u(x)$ is given by

$$U(P_t) = \int_0^\infty u(x) dF_{P_t}(x) \qquad (3.7.3)$$

where $F_{P_t}(x) = P(P_t \leq x)$ is the c.d.f. of the random variable P_t. Further on, suppose that the price of the common stock at the present

time is P_0. Consider the substitution $x = P_0 \exp(y)$. Under the new variable, the c.d.f. of P_t changes to

$$F_{P_t}(P_0 \exp(y)) = P(P_t \leq P_0 \exp(y)) = P\left(\log \frac{P_t}{P_0} \leq y\right)$$

which is, in fact, the distribution function of the log-return of the financial asset $r_t = \log(P_t/P_0)$. The integration range changes from the positive half-line to the entire real line and equation (3.7.3) becomes

$$U(P_t) = \int_{-\infty}^{\infty} u(P_0 \exp(y)) dF_{r_t}(y). \tag{3.7.4}$$

On the other hand, the expected utility of the log-return distribution has the form

$$U(r_t) = \int_{-\infty}^{\infty} v(y) dF_{r_t}(y) \tag{3.7.5}$$

where $v(y)$ is the utility function of the investor on the space of log-returns which is unique up to a positive linear transform. Note that $v(y)$ is defined on the entire real line as the log-return can be any real number.

Compare equations (3.7.4) and (3.7.5). From the uniqueness of the expected utility representation, it appears that (3.7.4) is the expected utility of the log-return distribution. Therefore, the utility function $v(y)$ can be computed by means of the utility function u,

$$v(y) = a.u(P_0 \exp(y)) + b, \quad a > 0 \tag{3.7.6}$$

in which the constants a and b appear because of the uniqueness result. Conversely, the utility function $u(x)$ can be expressed via v,

$$u(x) = c.v(\log(x/P_0)) + d, \quad c > 0. \tag{3.7.7}$$

Note that the two utilities in equations (3.7.4) and (3.7.5) are identical (up to a positive linear transform) and this is not surprising. In our reasoning, the investor is one and the same. We only change the way we look at the venture, in terms of payoff or log-return, but

the venture is also fixed. As a result, we cannot expect that the utility gained by the investor will fluctuate depending on the point of view.

Because of the relationship between the functions u and v, properties imposed on the utility function u may not transfer to the function v and vice versa. We remark on what happens with the properties connected with the n-th order stochastic dominance given in this appendix. Suppose that the utility function $v(y)$ belongs to the set $\mathcal{U}_n$, i.e. it satisfies the conditions

$$(-1)^{k+1}v^{(k)}(y) \geq 0, \quad k = 1, 2, \ldots, n$$

where $v^{(k)}(y)$ denotes the k-th derivative of $v(y)$. It turns out that the function $u(x)$ given by (3.7.7) satisfies the same properties and, therefore, it also belongs to the set $\mathcal{U}_n$. This is verified directly by differentiation.

In the reverse direction, the statement holds only for $n = 1$. That is, if $u \in \mathcal{U}_n$, $n > 1$, then the function v given in (3.7.6) may not belong to $\mathcal{U}_n$, $n > 1$, and we obtain a set of functions to which $\mathcal{U}_n$ is a subset. In effect, the n-th degree stochastic dominance, $n > 1$, on the space of payoffs implies the n-th degree stochastic dominance, $n > 1$, on the space of the corresponding log-returns but not vice versa,

$$P_t^1 \succeq_n P_t^2 \qquad \Longrightarrow \qquad r_t^1 \succeq_n r_t^2.$$

where P_t^1 and P_t^2 are the payoffs of the two common stocks, for example, at time $t > 0$, and r_t^1 and r_t^2 are the corresponding log-returns for the same period.

Note that this relationship holds if we assume that the prices of the two common stocks at the present time are equal to $P_0^1 = P_0^2 = P_0$. Otherwise, as we demonstrated in the chapter, no such relationship may exist.

3.7.4 Other stochastic dominance relations

There are ways of obtaining stochastic dominance relations other than the n-th order stochastic dominance which is based on certain properties of investors' utility functions. We borrow an example from

reliability theory and adapt it for distributions describing payoffs, losses, or returns.[10] The condition defining the order relation is based on the tail behavior of the corresponding distribution.

Consider the conditional probability

$$Q_X(t, x) = P(X > t + x | X > t) \tag{3.7.8}$$

where $x \geq 0$ and suppose that X describes a random loss. Then, equation (3.7.8) calculates the probability of losing more than $t + x$ on condition that the loss is larger than t. This probability may vary depending on the level t with the additional amount of loss being fixed (x does not depend on t). For example, if $t_1 \leq t_2$, then the corresponding conditional probabilities may be related in the following way:

$$Q_X(t_1, x) \geq Q_X(t_2, x). \tag{3.7.9}$$

Thus, the deeper we go into the tail, the less likely it is to lose additional x dollars provided that the loss is larger than the selected threshold. Conversely, if the inequality is

$$Q_X(t_1, x) \leq Q_X(t_2, x), \tag{3.7.10}$$

then the further we go into the tail, the more likely it becomes to lose additional x dollars. Basically, the inequalities in (3.7.9) and (3.7.10) describe certain tail properties of the random variable X.

Denote by $\bar{F}_X(x) = 1 - F_X(x) = P(X > x)$ the tail of the random variable X. Then, according to the definition of conditional probability, equation (3.7.8) can be stated in terms of $\bar{F}_X(x)$,

$$Q_X(t, x) = \frac{\bar{F}_X(x + t)}{\bar{F}_X(t)}. \tag{3.7.11}$$

Denote by $\mathcal{Q}$ the class of all random variables for which $Q_X(t, x)$ is a *non-increasing* function of t for any $x \geq 0$, and by $\mathcal{Q}^*$ the class of all random variables for which $Q_X(t, x)$ is a *non-decreasing* function of t for any $x \geq 0$. The random variables belonging to $\mathcal{Q}$ satisfy inequality (3.7.9) and those belonging to $\mathcal{Q}^*$ satisfy inequality (3.7.10) for any $x \geq 0$.

In case the random variable X has a density $f_X(x)$, then it can be determined whether it belongs to $\mathcal{Q}$ or $\mathcal{Q}^*$ by the behavior of the function

$$h_X(t) = \frac{f_X(t)}{\bar{F}_X(t)} \tag{3.7.12}$$

which is known as the *hazard rate function* or the *failure rate function*. If $h_X(t)$ is a non-increasing function, then $X \in \mathcal{Q}$. If it is a non-decreasing function, then $X \in \mathcal{Q}^*$. In fact, the only distribution which belongs to both classes is the exponential distribution. The hazard rate function of the exponential distribution is constant with respect to t.

In the following, we introduce a stochastic dominance order assuming that the random variables describe random profits. Then, we show how the dominance order definition can be modified if the random variables describe losses or returns. Denote by $\Lambda_X(t)$ the transform

$$\Lambda_X(t) = -\log(\bar{F}_X(t)). \tag{3.7.13}$$

A positive random variable X is said to dominate another positive random variable Y with respect to the Λ transform, $X \succeq_\Lambda Y$, if the random variable $Z = \Lambda_Y(X)$ is such that $Z \in \mathcal{Q}$.

The rationale behind the Λ transform is the following. First, consider the special case $Y = X$. The random variable $Z = \Lambda_Y(X)$ has exactly the exponential distribution because $\bar{F}_Y(X)$ is uniformly distributed. If Y has a heavier tail than X, then Z has a tail which increases no more slowly than the tail of the exponential distribution and, therefore, $Z \in \mathcal{Q}$. Thus, the stochastic order $\succeq_\Lambda$ emphasizes the tail behavior of X relative to Y.

This stochastic order is interesting since it does not arise from a class of utility functions through the expected utility theory and, nevertheless, it has application in finance describing choice under uncertainty. We illustrate this by showing a relationship with SSD.

Suppose that $X \succeq_\Lambda Y$. Then, Kalashnikov and Rachev (1990) show that the following condition holds

$$\int_t^\infty \bar{F}_X(x)dx \leq \int_t^\infty \bar{F}_Y(x)dx, \quad \forall t \geq 0. \tag{3.7.14}$$

The converse statement is not true: that is, condition (3.7.14) does not ensure $X \succeq_\Lambda Y$. Equation (3.7.14) can be directly connected with SSD. In fact, if (3.7.14) holds and we assume that the expected payoffs of X and Y are equal, then

$$\int_0^t F_X(x)dx \leq \int_0^t F_Y(x)dx, \quad \forall t \geq 0.$$

This inequality means that X dominates Y with respect to RSD and, therefore, with respect to SSD. Thus, we have demonstrated that if $EX = EY$, then

$$X \succeq_\Lambda Y \quad \Longrightarrow \quad X \succeq_{RSD} Y \quad \Longrightarrow \quad X \succeq_{SSD} Y. \tag{3.7.15}$$

Suppose that the random variables describe losses. This interpretation has application in the area of operational risk management where losses are modeled as positive random variables. We modify the stochastic order in the following way. A positive random variable X is said to dominate another positive random variable Y with respect to the Λ transform, $X \succeq_{\Lambda^*} Y$, if the random variable $Z = \Lambda_Y(X)$ is such that $Z \in \mathcal{Q}^*$. In this case, the tail of X is heavier than the tail of Y.

If the random variables describe returns, then the left tail describes losses and the right tail describes profits. The random variable can be decomposed into two terms,

$$X = X_+ - X_-,$$

where $X_+ = \max(X, 0)$ stands for the profit and $X_- = \max(-X, 0)$ denotes the loss. By modifying the stochastic order, we can determine the tail of which of the two components influences the stochastic order. Consider two real valued random variables X and Y describing random returns. The order $\succeq_\Lambda$ compares the tails of the profits X_+ and Y_+, and $\succeq_{\Lambda^*}$ compares the tails of the losses X_- and Y_-.

The stochastic orders $\succeq_{\Lambda}$ and $\succeq_{\Lambda^*}$ are constructed without considering first a particular class of investors but by imposing directly a condition on the tail of the random variable. There may or may not be a corresponding set of utility functions such that if $Eu(X) \geq Eu(Y)$ for all $u(x)$ in this class, then $X \succeq_{\Lambda} Y$, for example. Nevertheless, we have demonstrated that the order $\succeq_{\Lambda}$ is consistent with SSD and is not implied by it. We can generalize by concluding that if practical problems require introducing a stochastic order on the basis of certain characteristics of the profit, the loss, or the return distribution, the stochastic order can be defined without seeking first a class of investors which can generate it. In case this question appears important, we can only search for a consistency relation with an existing stochastic order, such as the one in equation (3.7.15).

Notes

1. The equilibrium model was published in Arrow and Debreu (1954).
2. The theory of stochastic dominance was formulated in the following papers: Hadar and Russel (1969), Hanoch and Levy (1969), Rothschild and Stiglitz (1970), and Whitmore (1970).
3. Actually, the payoff was in terms of *ducats* – a gold coin used as a trade currency in Europe before World War I.
4. It is also known as the *Arrow–Pratt measure of absolute risk aversion* after the economists Kenneth Arrow and John W. Pratt. (See Pratt (1964) and Arrow (1965).)
5. The Rothschild–Stiglitz stochastic dominance order is also called *concave order*.
6. In fact, the correct relationship is a positive linear transform of the function u but this detail is immaterial for the discussion which follows.
7. Chapter 2 provides more background on probability metrics.
8. The Kolmogorov metric $\boldsymbol{\rho}(X, Y)$ is introduced in Chapter 2; see equation (2.2.2).
9. Functions satisfying this property are also called *affine*.
10. Rachev (1985) and Kalashnikov and Rachev (1990) provide more details on the application of the stochastic order discussed in this section in reliability theory.

References

Arrow, J. K. (1965), 'Aspects of theory of risk-bearing', Yrjö Johansson Foundation, Helsinki.

Arrow, J. K. and G. Debreu (1954), 'Existence of a competitive equilibrium for a competitive economy', *Econometrica* **22 (3)**, 265–290.

Baucells, M. and F. Heukamp (2006), 'Stochastic dominance and cumulative prospect theory', *Management Sciences* **52**, 1409–1423.

Benartzi, S. and R. Thaler (1995), 'Myopic loss aversion and the equity premium puzzle', *Quarterly Journal of Economics* **110**, 73–92.

Hadar, J. and W. R. Russel (1969), 'Rules for ordering uncertain prospects', *American Economic Review* **59**, 25–34.

Hanoch, G. and H. Levy (1969), 'The efficiency analysis of choices involving risk', *Review of Economic Studies* **36**, 335–346.

Hwang, S. and S. Satchell (2003), 'The magnitude of loss aversion parameters in financial markets', working paper, Cass Business School, London.

Kahneman, D. and A. Tversky (1979), 'Prospect theory: An analysis of decision under risk', *Econometrica* **47**, 263–291.

Kahneman, D., J. Knetsch and R. Thaler (1991), 'Anomalies: The endowment effect, loss aversion, and status quo bias', *Journal of Economic Perspectives* **5**, 193–206.

Kalashnikov, V. and S. T. Rachev (1990), *Mathematical Methods for Construction for Queueing Models*, Wadsworth & Brooks/Cole Advanced Books.

Ortobelli, S., S. T. Rachev, H. Shalit and F. J. Fabozzi (2007), 'Risk probability functionals and probability metrics applied to portfolio theory', working paper.

Pratt, J. W. (1964), 'Risk aversion in the small and in the large', *Econometrica* **32**, 122–136.

Rabin, M. and R. Thaler (2001), 'Anomalies: risk aversion', *Journal of Economic Perspectives* **15**, 219–232.

Rachev, S. T. (1985), *Probability Metrics and their Applications to the Problems of Stability for Stochastic Models*, Doctor of Science dissertation, Moscow, Steklov Mathematical Institute.

Rothschild, M. and J. E. Stiglitz (1970), 'Increasing risk: A definition', *Journal of Economic Theory* **2**, 225–243.

Savage, L. J. (1954), *The Foundations of Statistics*, Wiley, New York.

Siegel, J. and R. Thaler (1997), 'Anomalies: the equity premium puzzle', *Journal of Economic Perspectives* **11**, 191–200.

Tversky, A. and D. Kahneman (1992), 'Advances in prospect theory: Cumulative representation of uncertainty', *Journal of Risk and Uncertainty* **5**, 297–323.

von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

Whitmore, G. A. (1970), 'Third-degree stochastic dominance', *American Economic Review* **60**, 457–459.

Chapter 4

A Classification of Probability Distances

The goals of this chapter are the following:

- To introduce formally primary, simple, and compound probability distances.
- To provide examples and study the relationship between primary, simple, and compound distances.
- To introduce the notions of minimal probability distance, minimal norms, co-minimal functionals, and moment functions, which are needed in the study of primary, simple, and compound probability distances.
- To introduce formally the class of ideal probability metrics.

Notation introduced in this chapter:

Notation	*Description*
μ_h	A primary distance generated by a p. semidistance μ and mapping h
$\tilde{\mu}_h$	A primary h-minimal distance
$m_i P = m_i^{(p)} P$	Marginal moment of order p

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

Notation	*Description*
$\mathcal{M}_{H,p}(g)$	A primary distance generated by g, H, p
$\mathcal{M}(g)$	A primary metric generated by g
$\mathbf{\Omega}$	The discrete primary metric
$\mathbf{EN}(X, Y; H)$	The engineer's distance
$\mathbf{EN}(X, Y; p)$	The L_p-engineer's metric
$\xrightarrow{w}$	The weak convergence of laws
$\widehat{\mu}$	The minimal distance w.r.t. μ
ℓ_H	The minimal distance w.r.t. $\mathcal{L}_H$ (the Kantorovich distance)
ℓ_p	The minimal metric w.r.t. $\mathcal{L}_p$
$\boldsymbol{\sigma}$	The total variation metric
F^{-1}	The generalized inverse of the distribution function F
π	The Prokhorov metric
π_λ	The parametric version of the Prokhorov metric
π_H	The Prokhorov distance
$\boldsymbol{\theta}_H$	The Birnbaum–Orlicz distance
$\boldsymbol{\rho}_H$	The Birnbaum–Orlicz uniform distance
$\mu\nu(P_1, P_2, \alpha)$	Co-minimal metric functional w.r.t. the p. distances μ and ν
$\overline{\mu\nu}(P_1, P_2, \alpha)$	Simple semi-distance with $K_{\overline{\mu\nu}} = 2K_\mu K_\nu$
$\mu_c(m)$	The total cost of transportation of masses under the plan m
$\overset{\circ}{\mu}_c$	The minimal norm w.r.t. μ_c
$\zeta_{\mathcal{F}}$	The Zolotarev semimetric
$\zeta_{s,p}$	The Zolotarev ideal semimetric
$\zeta_{s,p,\alpha}$	The Rachev ideal semimetric
$\mathcal{M}(X, Y)$	The moment metric
$\mathbb{L}_H$	H-average compound distance
$\mathbf{KF}_H$	The Ky Fan distance
$\mathbf{K}_\lambda$	The parametric family of Ky Fan metrics
$\mathbf{\Theta}_H$	The Birnbaum–Orlicz compound distance
$\mathbf{\Theta}_p$	The Birnbaum–Orlicz compound metric
$\mathbf{R}_H$	The Birnbaum–Orlicz compound average distance
$\breve{\mu}$	The maximal distance w.r.t. μ

Notation	*Description*
$\overset{(s)}{\mu}$	μ-upper bound with marginal sum fixed
$\overset{(m,p)}{\mu}$	μ-upper bound with fixed p-th marginal moments
$\underset{(m,p)}{\mu}$	μ-lower bound with fixed p-th marginal moments
$\overline{\mu}$	μ-upper bound with fixed sum of p-th marginal moments
$\underline{\mu}$	μ-lower bound with fixed sum of p-th marginal moments

Important terms introduced in this chapter:

Term	*Concise explanation*
primary probability semimetric	A semimetric function measuring distances between random quantities in terms of the disagreement between certain characteristics of the random quantities
simple probability semimetric	A semimetric function measuring distances between random quantities in terms of their distribution function
compound probability semimetric	A semimetric function measuring distances between random quantities defined on a common probability space
minimal probability semimetric w.r.t. a given semimetric	A simple probability semimetric obtained by minimizing the given semimetric over all possible joint laws while holding the marginal distributions fixed
moment function	An axiomatically introduced function having all properties of a probability distance save for the identity property, used in obtaining upper bounds to probability distances
ideal semimetric	A probability semimetric satisfying a few properties in addition to the standard ones, which is best suited for studying rates of convergence to limit theorems

4.1 Introduction

The goal of Chapter 2 was to introduce the concept of measuring distances between random quantities and to provide examples of probability metrics. While we treated the general theory of probability metrics in detail, we did not provide much theoretical background on the distinction between different classes of probability metrics. We only noted that three classes of probability (semi-)metrics are distinguished – *primary*, *simple*, and *compound*. The goal of this chapter is to revisit these ideas but at a more advanced level.

When delving into the details of primary, simple, and compound probability metrics, we also consider a few related objects. They include co-minimal functionals, minimal norms, minimal metrics, and moment functions. In the theory, these related functionals are used to establish upper and lower bounds to given families of probability metrics. They also help specify under what conditions a given probability metric is finite.

Finally, we consider ideal probability metrics which can be simple or compound. They satisfy two properties in addition to the standard properties satisfied by any probability metric, which makes their axiomatic structure suitable for studying rates of convergence in limit theorems. Besides their theoretical properties, we discuss applications in finance.

In the appendix to this chapter, we provide more general results and proofs where necessary.

4.2 Primary Distances and Primary Metrics

The theory of probability metrics distinguishes between three categories of probability metrics. The principal criterion is contained in the answer to the question: What are the implications for X and Y provided that they have a zero distance? Suppose that X and Y stand for the random returns of two equities. Zero distance between X and Y means that the two random variables are indistinguishable

in a certain sense. This sense could be to the extent of a given set of characteristics of X and Y. For example, X is to be considered indistinguishable from Y if their expected returns and variances are the same. Therefore, a way to define the distance between them is through the distance between the corresponding characteristics (i.e., how much their expected returns and variances deviate). One example is

$$\mu(X, Y) = |EX - EY| + |\sigma^2(X) - \sigma^2(Y)|.$$

Such probability metrics are called *primary metrics*, and they imply the weakest form of sameness. Common examples of primary metrics include Engineer's metric defined in (2.2.1) and the absolute moments metric given in (2.2.11).

In this section, we define primary metrics in a general context in which the set of characteristics is introduced by means of a mapping defined on a space of probability laws taking values in $\mathbb{R}^J$. Generally, primary probability distances induce a distance in the image space of the mapping (i.e., in the space of characteristics $\mathbb{R}^J$). For example, in the case of Engineer's metric, the image space of the mapping is the real line representing the mathematical expectation of probability laws. Finally, we define a functional which under certain conditions yields primary distances. Additional examples are considered in the appendix to this chapter.

Let $h : \mathcal{P}_1 \to \mathbb{R}^J$ be a mapping, where $\mathcal{P}_1 = \mathcal{P}_1(U)$ is the set of Borel probability measures (laws) for some separable metric space (s.m.s.) (U, d) and J is some index set. This function h induces a partition of $\mathcal{P}_2 = \mathcal{P}_2(U)$ (the set of laws on U^2) into equivalence classes for the relation

$$P \overset{h}{\sim} Q \iff h(P_1) = h(Q_1) \text{ and } h(P_2) = h(Q_2)\, P_i := T_i P,\ Q_i := T_i Q \tag{4.2.1}$$

where P_i and Q_i $(i = 1, 2)$ are the ith marginals of P and Q, respectively. Let μ be a p. semidistance on $\mathcal{P}_2$ with parameter $\mathbb{K}_\mu$ (Definition 2.4.1), such that μ is constant on the equivalence

classes of $\sim$, i.e.

$$P \overset{h}{\sim} Q \iff \mu(P) = \mu(Q). \tag{4.2.2}$$

For better readability, sometimes we use hP_1 instead of $h(P_1)$ to denote the characteristics of the probability law P_1.

Definition 4.2.1. If the p. semidistance $\mu = \mu_h$ satisfies relation (4.2.2), then we call μ a *primary distance* (*with parameter* $\mathbb{K}_\mu$). If $\mathbb{K}_\mu = 1$ and μ assumes only finite values, we say that μ is a primary metric.

Obviously, by relation (4.2.2), any primary distance is completely determined by the pair of marginal characteristics (hP_1, hP_2). In the case of primary distance μ we shall write $\mu(hP_1, hP_2) := \mu(P)$ and hence μ may be viewed as a distance in the image space $h(\mathcal{P}_1) \subseteq \mathbb{R}^J$: that is, the following metric properties hold:

ID$^{(1)}$ $\quad hP_1 = hP_2 \iff \mu(hP_1, hP_2) = 0$

SYM$^{(1)}$ $\quad \mu(hP_1, hP_2) = \mu(hP_2, hP_1)$

TI$^{(1)}$ $\quad$ If the following marginal conditions are fulfilled

$$a = h(T_1P^{(1)}) = h(T_1P^{(2)}) \quad b = h(T_2P^{(2)}) = h(T_1P^{(3)})$$

$$c = h(T_2P^{(1)}) = h(T_2P^{(3)})$$

for some law $P^{(1)}, P^{(2)}, P^{(3)} \in \mathcal{P}_2$ then $\mu(a, c) \leq \mathbb{K}_\mu[\mu(a, b) + \mu(b, c)]$.

The notion of primary semidistance μ_h becomes easier to interpret assuming that a probability space $(\Omega, \mathcal{A}, \Pr)$ with property (2.6.2) is fixed (see Remark 2.6.2). In this case μ_h is a usual distance (see Definition 2.3.1) in the space

$$h(\mathfrak{X}) := \{hX := h\Pr_x, \quad \text{where } X \in \mathfrak{X}(U)\} \tag{4.2.3}$$

and thus, the metric properties of $\mu = \mu_h$ take the simplest form (cf. Definition 2.3.2):

ID$^{(1*)}$ $\quad hX = hY \iff \mu(hX, hY) = 0$

SYM$^{(2*)}$ $\quad \mu(hX, hY) = \mu(hY, hX)$,

TI$^{(3*)}$ $\quad \mu(hX, hZ) \leq \mathbb{K}_\mu[\mu(hX, hY) + \mu(hY, hZ)]$.

We noted that the Engineer's metric and the absolute moments metric are two examples of primary metrics. There is a general procedure which, under certain conditions, produces a primary metric. We illustrate this procedure in the more simple setting of random variables but it is applicable to any kind of random elements. Suppose that the random variables X and Y describe the return distribution of the common stocks of two corporations and we are interested in only three characteristics – expected return, variance, and skewness. Also, suppose that we have selected a probability metric which compares the random variables, such as the Kolmogorov metric or the L_p-metric discussed in Chapter 2. It is possible to obtain a primary metric comparing these three characteristics, which is based on the already selected probability metric. The procedure is the following. Take all random variables with expected return, variance, and skewness equal to those of X and also all random variables with the corresponding three characteristics equal to those of Y. Build all possible pairs and then from each pair construct all possible two-dimensional random vectors by varying the dependence structure. Collect all two-dimensional vectors resulting from all pairs and estimate the distance between the elements of these pairs using the selected probability metric. Compute the minimum of these distances and call it the *primary minimal distance*. Since we compute the minimum by varying everything but the three characteristics, the primary minimal distance depends only on them. It is possible to demonstrate that under certain conditions the primary minimal distance satisfies the axioms of probability semimetric; therefore it can be used to measure the dissimilarity between X and Y in terms of the three characteristics.

It is essential that the resulting minimal primary metric originates from an already selected probability metric. Thus, this procedure provides a way of obtaining a primary metric consistent in a certain sense with another probability metric. This example is illustrated in a more formal way below.

Example 4.2.1. Primary minimal distances. Each p. semidistance μ and each mapping $h : \mathcal{P}_1 \to \mathbb{R}^J$ determine a functional $\widetilde{\mu}_h : h(\mathcal{P}_1) \times$

$h(\mathcal{P}_1) \to [0, \infty]$ defined by the following equality

$$\widetilde{\mu}_h(\bar{a}_1, \bar{a}_2) := \inf\{\mu(P) : hP_i \equiv \bar{a}_i, i = 1, 2\} \tag{4.2.4}$$

where P_i are the marginals of P for any pair $(\bar{a}_1, \bar{a}_2) \in h(\mathcal{P}_1) \times h(\mathcal{P}_1)$.

As we noted, the functional (4.2.4) does not always yield a primary distance. It is a primary distance for different special functions h and spaces U. The appendix to this chapter contains other examples of primary metrics.

Definition 4.2.2. The functional $\widetilde{\mu}_h$ is called a *primary h-minimal distance* with respect to the p. semidistance μ.

4.3 Simple Distances and Metrics

Technically, simple distances are defined on the space of distribution functions. If two random variables have one and the same distribution functions, then a simple semidistance indicates that the two random variables are coincident. In a similar way, if a simple distance between two random variables is equal to zero, then the corresponding distribution functions coincide.

Important examples of simple metrics include the Kolmogorov metric defined in (2.2.2), the Lévy metric defined in (2.2.3), and the Kantorovich metric defined in (2.2.5) in Chapter 2. Some of the simple metrics defined directly as functionals of distribution functions arise as *minimal metrics*. We discuss this important construction and prove that minimal semidistances are always simple. Two other related constructions include minimal norms and co-minimal functionals. While in general they are not simple distances, they give rise to probability distances under certain conditions. They can also be used to construct lower and upper bounds respectively to simple distances.

Further on, some simple semidistances allow for alternative representations known as *dual forms*. Generally, dual forms are represented as the supremum of a functional with respect to functions belonging to some functional space. Dual forms are also referred to as

ζ-representations. In this chapter, we do not discuss in detail theories leading to dual forms. We only state the dual forms of some probability metrics in the appendix to this chapter where we also provide further examples of simple distances.

There is a link between simple and primary distances. By including additional characteristics in a primary metric, we include additional information from the distribution functions of the two random variables. If we include a sufficiently large number of characteristics, the primary metric turns into a simple metric. Generally, a very rich set of characteristics will ensure that the distribution functions coincide.

More formally, any primary distance $\mu(P)$ ($P \in \mathcal{P}_2$) is completely determined by the pair of marginal distributions $P_i = T_i P$ ($i = 1, 2$), since the equality $P_1 = P_2$ implies $hP_1 = hP_2$ (see relations (4.2.1), (4.2.2) and Definition 4.2.1). On the other hand, if the mapping h is "rich enough" then the opposite implication

$$hP_1 = hP_2 \Rightarrow P_1 = P_2$$

occurs. The simplest example of such "rich" $h : \mathcal{P}_1(U) \to \mathbb{R}^J$ is given by the equalities

$$h(P) := \{P(C), C \in \mathcal{C}, P \in \mathcal{P}_1(U)\} \tag{4.3.1}$$

where $J \equiv \mathcal{C}$ is the family of all closed non-empty subsets $C \subseteq U$.

This example allows for the following interpretation in investment management when considering random variables describing the return of two investments. Suppose that X and Y are two such random variables. The function h as defined in (4.3.1) calculates the probability that X or Y belongs to any closed interval (e.g., $P(a \le X \le b)$ for all possible choices of a and b). The relation $hX = hY$ means that all these probabilities computed for X coincide with those computed for Y (i.e., $P(a \le X \le b) = P(a \le Y \le b)$ for all possible values of a and b). This information is sufficient to conclude that the distribution function of X coincides with the distribution function of Y because the choice $a = -\infty$ and $b = x$, $x \in \mathbb{R}$ defines the distribution function. Therefore, the requirement that the probabilities of X and Y coincide on all closed intervals implies that their

distribution functions coincide. However, this does not mean that in all states of the world the return of investment X would equal the return of investment Y. Even though the two investments are identical from the point of view of their return distribution, a portfolio manager may be interested in holding both in a portfolio because of a diversification effect arising from the way X and Y depend on each other.

Another example of a sufficiently rich set of characteristics is

$$h(P) = \left\{ Pf := \int_U f \mathrm{d}P : f \in C^b(U) \right\} \quad P \in \mathcal{P}_1(U)$$

where $C^b(U)$ is the set of all bounded continuous functions on U. In this case, if we consider random variables, the function h calculates the moments $Ef(X)$ for all possible choices of f from the set of bounded continuous functions defined on the real line. Even though the practical application of this example is limited, it illustrates the connection between primary and simple distances.

Keeping in mind these two examples, we shall define the notion of "simple" distance as a particular case of primary distance with h given by equality (4.3.1).

Definition 4.3.1. The p. semidistance μ is said to be a *simple semidistance in* $\mathcal{P}_2 = \mathcal{P}_2(U)$, if for each $P \in \mathcal{P}_2$

$$\mu(P) = 0 \Leftarrow T_1 P = T_2 P.$$

If, in addition, μ is a p. semimetric, then μ will be called a *simple semimetric.* If the converse implication ($\Rightarrow$) also holds, we say that μ is *simple distance*. If, in addition, μ is a p. semimetric, then μ will be called a *simple metric*.

Since the values of the simple distance $\mu(P)$ depend only on the pair marginals P_1, P_2 we shall consider μ as a functional on $\mathcal{P}_1 \times \mathcal{P}_1$ and we shall use the notation

$$\mu(P_1, P_2) := \mu(P_1 \times P_2) \quad (P_1, P_2 \in \mathcal{P}_1)$$

where $P_1 \times P_2$ means the measure product of laws P_1 and P_2. In this case the metric properties of μ take the form (cf. Definition 2.4.1) (for each $P_1, P_2, P_3 \in \mathcal{P}_1$):

$\mathbf{ID}^{(2)}$ $\quad P_1 = P_2 \iff \mu(P_1, P_2) = 0$
$\mathbf{SYM}^{(2)}$ $\quad \mu(P_1, P_2) = \mu(P_2, P_1)$
$\mathbf{TI}^{(2)}$ $\quad \mu(P_1, P_2) \leq \mathbb{K}_\mu(\mu(P_1, P_2) + \mu(P_2, P_3))$.

Hence, the space $\mathcal{P}_1$ of laws P with a simple distance μ is a distance space (see Definition 2.3.2). Clearly each primary distance is a simple semidistance in $\mathcal{P}_1$. The Kolmogorov metric $\boldsymbol{\rho}$ (2.2.2), the Lévy metric $\mathbf{L}$ (2.2.3), and the $\boldsymbol{\theta}_p$-metrics (2.2.6) are simple metrics in $\mathcal{P}(\mathbb{R})$.

Let us consider a few more examples of simple metrics which we shall use later on.

Example 4.3.1. Minimal distances. We described the primary minimal distance as a procedure for obtaining primary distances starting from a given probability distance. There is a similar procedure which yields simple distances from a given probability distance. In contrast to the primary minimal distance, this procedure always yields a simple distance. We will illustrate it in the context of two random variables X and Y describing the return distribution of the common stocks of two corporations.

Denote the distribution functions of the two random variables by F_X and F_Y, respectively. Construct the set of all random variables with the same distribution function as F_X and also the set of all random variables with the same distribution function as F_Y. Take a member from the first set and one from the second set and build all possible bivariate random vectors by varying the dependence structure between the two random variables. To evaluate the probability distance, we start with all those pairs and then calculate the minimum. This minimum is called the *minimal distance*. It depends only on the distribution functions F_X and F_Y, and satisfies the axioms of probability semidistances. Therefore, it is a simple semidistance.

If a simple distance indicates that F_X and F_Y coincide, this does not necessarily mean that in all states of the world the common stocks

have equal returns – they only have equal distribution functions. As we noted, a portfolio manager may include both in a portfolio if there is a diversification effect (e.g., if X is negatively correlated to Y).

The minimal semidistance is defined formally in the following way.

Definition 4.3.2. For a given p. semidistance μ on $\mathcal{P}_2$ the functional $\widehat{\mu}$ on $\mathcal{P}_1 \times \mathcal{P}_1$ defined by the equality

$$\widehat{\mu}(P_1, P_2) := \inf\{\mu(P); T_i P = P_i, i = 1, 2\} \qquad P_1, P_2 \in \mathcal{P}_1 \qquad (4.3.2)$$

is said to be (simple) *minimal* (w.r.t. μ) *distance*.

As we showed in section 2.6.2, for a "rich enough" probability space, the space $\mathcal{P}_2$ of all laws on U^2 coincides with the set of joint distributions $\Pr_{X,Y}$ of U-valued random variables. Thus, it is $\mu(P) = \mu(\Pr_{X,Y})$ for some $X, Y \in \mathfrak{X}(U)$ and as a result equation (4.3.2) can be rewritten as follows:

$$\widehat{\mu}(P_1, P_2) = \inf\{\mu(X, Y) : \Pr_X = P_1, \Pr_Y = P_2\}.$$

The last is the Zolotarev definition of a minimal metric (Zolotarev, 1976).

In the next theorem we shall consider the conditions on U that guarantee $\widehat{\mu}$ to be a simple metric. We use the notation $\xrightarrow{\text{w}}$ to mean "weak convergence of laws" (see, for example, Billingsley, 1968).

Theorem 4.3.1. Let U be a u.m. s.m.s. (see Definition 2.6.2) and let μ be a p. semidistance with parameter $\mathbb{K}_\mu$. Then $\widehat{\mu}$ is a simple semidistance with parameter $\mathbb{K}_{\widehat{\mu}} = \mathbb{K}_\mu$. Moreover, if μ is a p. distance satisfying the following "continuity" condition

$$\left.\begin{array}{l} P^{(n)} \in \mathcal{P}_2 \quad P^{(n)} \xrightarrow{\text{w}} P \in \mathcal{P}_2 \\ \mu(P^{(n)}) \to 0 \end{array}\right\} \Rightarrow \mu(P) = 0$$

then $\widehat{\mu}$ is a simple distance with parameter $\mathbb{K}_{\widehat{\mu}} = \mathbb{K}_\mu$.

Remark 4.3.1. The continuity condition is not restrictive; in fact, all p. distances we are going to use satisfy this condition.

Remark 4.3.2. Clearly, if μ is a p. semimetric then, by the above theorem, $\widehat{\mu}$ a simple semimetric.

Proof. **ID**$^{(2)}$: If $P_1 \in \mathcal{P}_1$ then we let $X \in \mathfrak{X}(U)$ have the distribution P_1. Then, by **ID**$^{(*)}$ (Definition 2.4.2),

$$\widehat{\mu}(P_1, P_1) \leq \mu(\mathrm{Pr}_{(X,X)}) = 0.$$

Suppose now that μ is a p. distance and the continuity condition holds. If $\widehat{\mu}(P_1, P_2) = 0$ then there exists a sequence of laws $P^{(n)} \in \mathcal{P}_2$ with fixed marginals $T_i P^{(n)} = P_i$ $(i = 1, 2)$ such that $\mu(P^{(n)}) \to 0$ as $n \to \infty$. Since P_i is a tight measure then the sequence $\{P^{(n)}, n \geq 1\}$ is uniformly tight, i.e., for any $\varepsilon > 0$ there exists a compact $K_\varepsilon \subseteq U^2$ such that $P^{(n)}(K_\varepsilon) \geq 1 - \varepsilon$ for all $n \geq 1$ (cf. Dudley (1989), Section 11.5). Using Prokhorov compactness criteria (see, for instance, Billingsley (1968), Theorem 6.1) we choose a subsequence $P^{(n')}$ that weakly tends to a law $P \in \mathcal{P}_2$, hence, $T_i P = P_i$ and $\mu(P) = 0$. Since μ is a p. distance, P is concentrated on the diagonal $x = y$ and thus $P_1 = P_2$ as desired.

SYM$^{(2)}$: Obvious.

TI$^{(2)}$: Let $P_1, P_2, P_3 \in \mathcal{P}_1$. For any $\varepsilon > 0$ define a law $P_{12} \in \mathcal{P}_2$ with marginals $T_i P_{12} = P_i$ $(i = 1, 2)$ and a law $P_{23} \in \mathcal{P}_2$ with $T_i P_{23} = P_{i+1}$ $(i = 1, 2)$ such that $\widehat{\mu}(P_1, P_2) \geq \mu(P_{12}) - \varepsilon$ and $\widehat{\mu}(P_2, P_3) \geq \mu(P_{23}) - \varepsilon$. Since U is a u.m. s.m.s. then there exist Markov kernels $P'(A|z)$ and $P''(A|z)$ defined by the equalities

$$P_{12}(A_1 \times A_2) := \int_{A_2} P'(A_1|z) P_2(\mathrm{d}z) \tag{4.3.3}$$

$$P_{23}(A_2 \times A_3) := \int_{A_2} P''(A_3|z) P_2(\mathrm{d}z) \tag{4.3.4}$$

for all $A_1, A_2, A_3 \in \mathcal{B}_1$ (see Corollary 2.6.2). Then define a set function Q on the algebra $\mathcal{A}$ of finite unions of Borel rectangles $A_1 \times A_2 \times A_3$

by the equation

$$Q(A_1 \times A_2 \times A_3) := \int_{A_2} P'(A_1|z)P''(A_3|z)P_2(\mathrm{d}z). \tag{4.3.5}$$

It is easily checked that Q is countably additive on $\mathcal{A}$ and therefore extends to a law on U^3. We use "Q" to represent this extension also. The law Q has the projections $T_{12}Q = P_{12}$, $T_{23}Q = P_{23}$. Since μ is a p. semidistance with parameter $\mathbb{K} = \mathbb{K}_\mu$ we have

$$\begin{aligned}\mu(P_1, P_3) &\leq \mu(T_{13}Q) \leq \mathbb{K}[\mu(P_{12}) + \mu(P_{13})] \\ &\leq \mathbb{K}[\widehat{\mu}(P_1, P_2) + \widehat{\mu}(P_2, P_3)] + 2\mathbb{K}\varepsilon.\end{aligned}$$

Letting $\varepsilon \to 0$ we complete the proof of $\mathbf{TI}^{(2)}$. □

We will demonstrate in the next chapters that all simple distances in the examples in the appendix to this chapter are actually simple minimal $\widehat{\mu}$ distances with respect to p. distances μ that will be introduced in section 4.4 (see further examples 4.7.10 to 4.7.12).

Example 4.3.2. Co-minimal metrics. There is an extension of the construct of minimal metrics which also produces simple metrics. The extension involves not one but several preselected probability distances which restrict the set of pairs of probability laws we consider for computing the minimal distance. For example, suppose that X and Y are two random variables describing the return distribution of a portfolio and a benchmark, respectively. We may want to compute the distance between X and Y through the minimal metric construct but considering only those pairs of random variables which are coupled in a particular way. For instance, we would like to consider only those for which the tracking error is below a given number. We may want to impose this restriction because without it the minimal distance may indicate that X is too close to Y because it may be attained at an unrealistic dependence model. In effect, the set of bivariate random variables (U, V) we would like to consider is

$$\{(U, V) : F_U = F_X, F_V = F_Y, \sigma(X - Y) \leq \alpha\}$$

where $\alpha > 0$ is the upper bound on the tracking error. As a result, we compute

$$\mu\sigma(X, Y, \alpha) = \inf\{\mu(U, V), F_U = F_X, F_V = F_Y, \sigma(X - Y) \leq \alpha\}$$

where $\mu(U, V)$ is chosen in advance as in the construct of the minimal metric. The functional $\mu\sigma(X, Y, \alpha)$ is called *co-minimal functional* with respect to the two probability distances μ and σ.

Following the main idea of obtaining primary and simple distances by means of minimization procedures of certain types (see Definitions 4.2.2 and 4.3.2), we define formally the notion of "co-minimal distance." For a given compound, semidistances μ and ν with parameters $\mathbb{K}_\mu$ and $\mathbb{K}_\nu$, respectively, and for each $\alpha > 0$ denote

$$\mu\nu(P_1, P_2, \alpha) = \inf\{\mu(P) : P \in \mathcal{P}_2, T_1P = P_1, T_2P = P_2, \nu(P) \leq \alpha\}$$
$$P_1, P_2 \in \mathcal{P}_1 \quad (4.3.6)$$

(see equations (4.7.14) and (4.7.16)).

Definition 4.3.3. The functional $\mu\nu(P_1, P_2, \alpha)$ $(P_1, P_2 \in \mathcal{P}_1, \alpha > 0)$ will be called the *co-minimal (metric) functional w.r.t. the p. distances μ and ν* (see Figure 4.2)

The co-minimal functional is not a probability semidistance itself. Nevertheless, it induces a probability semidistance. The details are given in the appendix to this chapter. We will only note that there is a relationship between the minimal distance and the co-minimal distance which, in this example, is the inequality

$$\widehat{\mu}(X, Y) \leq \mu\sigma(X, Y, \alpha).$$

Intuitively, the more α increases, the less restrictive the additional constraint becomes. Therefore, at the limit $\widehat{\mu}(X, Y) = \mu\sigma(X, Y, \infty)$.

Example 4.3.3. Minimal norms. We considered a number of different examples of probability distances in Chapter 2. Note that some of the simple metrics directly depend on the difference between the probability laws $P_1 - P_2$, such as the Kolmogorov metric (2.2.2), the

Kantorovich metric (2.2.5), and the L_p-metrics between distribution functions (2.2.6). If we consider random variables, the difference $P_1 - P_2$ translates into the difference between the corresponding distribution functions $F_X(x) - F_Y(x)$. There is a set of such functionals which are called *minimal norms* and denoted by $\mathring{\mu}$ but not all probability semidistances which directly depend on the difference $P_1 - P_2$ are minimal norms.

We introduce the minimal norms in this section by means of their dual representation. The formal definition of minimal norms is given in the appendix to this chapter. The dual representation allows for an interesting interpretation in the field of behavioral finance concerning making choice under uncertainty. We discussed expected utility theory and cumulative prospect theory in Chapter 3.

The minimal norm $\mathring{\mu}_c(P_1, P_2)$ of two probability laws $P_1, P_2 \in \mathcal{P}_1$ is defined as

$$\mathring{\mu}_c(P_1, P_2) = \sup_{f \in \mathcal{F}} \left| \int_U f \mathrm{d}(P_1 - P_2) \right| \tag{4.3.7}$$

where $f : U \to \mathbb{R}$ is a function satisfying the "Lipschitz" condition $|f(x) - f(y)| \leq c(x, y)$ in which $c(x, y) : U^2 \to \mathbb{R}^+$ is a given continuous, symmetric, and non-negative function. The Lipschitz condition can be regarded as a growth condition on f. Thus, we compute the supremum by varying the function f in the set $\mathcal{F}$ of all functions satisfying the growth condition. The function c is specified in advance and this is reflected by the subscript in the notation of the minimal norm.

From the point of view of cumulative prospect theory, the function f appearing in the definition can be interpreted as a value function of an individual. It indicates how much "utility" the individual gains from a given outcome. Suppose that X and Y are two random variables describing the return distribution of two investments. In this setting, we can interpret $\int_{\mathbb{R}} f \mathrm{d}F_X$ as the expected value gained by the individual with a value function f. In effect, the functional $\int_{\mathbb{R}} f \mathrm{d}(F_X - F_Y)$ computes the difference between the expected values of the two investments gained by the same individual. Therefore, the

minimal norm in (4.3.7) can be interpreted as the largest difference running through all the investors. Intuitively, if the largest difference is zero, then all investors are indifferent between the two opportunities. If the set of investors is large enough, this would indicate that the distribution functions F_X and F_Y coincide.

Using the theory of probability metrics, it is possible to show that there exists a relationship between minimal norms and minimal metrics. In the context of the example above, the following inequality holds true,

$$\mathring{\mu}_c(X, Y) \le \widehat{\mu}_c(Y, X) \le \mu_c(Y, X),$$

where $\mu_c = Ec(X, Y)$. In the appendix to this chapter, this relationship is considered in the general case.

It is important to note that generally the minimal norm provides a lower bound for the minimal metric in contrast to the co-minimal distance which provides an upper bound. In fact, the minimal norm is not always a simple semidistance. It depends on how rich the set $\mathcal{F}$ is. In the appendix to this chapter, we provide a sufficient condition.

4.4 Compound Distances and Moment Functions

We continue the classification of probability distances. Recall some basic examples of p. metrics on a s.m.s. (U, d):

(a) The *moment metric* (see Example 4.7.1):

$$\mathcal{M}(X, Y) = |Ed(X, a) - Ed(Y, a)| \quad X, Y \in \mathfrak{X}(U)$$

($\mathcal{M}$ is a primary metric in the space $\mathfrak{X}(U)$ of U-valued r.v.s).
(b) The Kantorovich metric (see Example 4.7.6):

$$\boldsymbol{\kappa}(X, Y) = \sup\{|Ef(X) - Ef(Y)| : f : U \to \mathbb{R} \text{ bounded}, \\ |f(x) - f(y)| \le d(x, y) \quad \forall x \text{ and } y \in U\}$$

($\boldsymbol{\kappa}$ is a simple metric in $\mathfrak{X}(U)$).

(c) The L_1-metric (see (2.4.3)):

$$\mathcal{L}_1(X, Y) = Ed(X, Y) \qquad X, Y \in \mathfrak{X}(U).$$

The $\mathcal{L}_1$-metric is a p. metric in $\mathfrak{X}(U)$ (Definition 2.4.2). Since the value of $\mathcal{L}_1(X, Y)$ depends on the joint distribution of the pair (X, Y), we shall call $\mathcal{L}_1$ a compound metric.

Definition 4.4.1. A *compound distance* (resp., metric) is any probability distance μ (resp., metric). See Definitions 2.4.1 and 2.4.2.

Remark 4.4.1. In many papers on probability metrics, "compound" metric stands for a metric which is not simple. However, all "non-simple" metrics that have been used in these papers are in fact "compound" in the sense of Definition 4.4.1. The problem of classification of p. metrics which are neither compound (in the sense of Definition 4.4.1) nor simple is open.

We noted that the coincidence of distribution functions is stronger than the coincidence of certain characteristics, such as absolute moments. There is a stronger form of identity than coincidence of distribution functions, which is actually the strongest possible. Consider the case in which no matter what happens, the returns of common stock 1 and common stock 2 are identical. Hence, we can describe the two random variables as being coincident in each state of the world. As a consequence, their distribution functions are the same because the probabilities of all events of the return of common stock 1 are exactly equal to the corresponding events of the return of common stock 2. This identity is also known as *almost everywhere identity* because it considers all states of the world which happen with non-zero probability. The compound metrics imply the almost everywhere identity.

Since a compound metric μ may take infinite values, we have to determine a concept of μ-boundedness. This concept allows construction of upper bounds by means of special functionals called *moment functions*. We will illustrate this concept with the L_2-metric defined in equation (2.2.9) in Chapter 2.

We noted that the L_2-metric is closely related to tracking error. Assume that an approximate model for the daily return distribution of a portfolio and a benchmark are two zero-mean random variables X and Y with standard deviations σ_X and σ_Y, respectively. Under these assumptions, the tracking error equals the L_2-metric,

$$\mathcal{L}_2(X, Y) = (E(X - Y)^2)^{1/2}.$$

If we want to compute and use the tracking error in practice, we have to make sure it is a finite number. The sample estimate of tracking error is always a finite number, as any other sample estimate, but it may become infinite depending on the assumed model for the return distributions. Using the properties of variance, we obtain the following bounds for the tracking error,

$$|\sigma_X - \sigma_Y| \leq \mathcal{L}_2(X, Y) \leq \sigma_X + \sigma_Y.$$

Both bounds are obtained by varying the correlation between the portfolio and benchmark returns. The upper bound appears when X and Y are perfectly positively correlated and the lower bound when they are perfectly negatively correlated. The upper bound appears to be the sum of the standard deviations of X and Y. Therefore, if the assumed models are such that $\sigma_X < \infty$ and $\sigma_Y < \infty$, the tracking error is always finite.

Note that we can view the upper bound as a sum of two moments. In the general case, the upper bound can also be a sum of two moments which is why it is called a moment function. Moments functions have an axiomatic construction which differs from the notion of simple distance in the "identity" property only (cf. Definition 4.3.1 and $\mathbf{ID}^{(2)}$, $\mathbf{TI}^{(2)}$).

Definition 4.4.2. A mapping $\mathbb{M} : \mathcal{P}_1 \times \mathcal{P}_1 \to [0, \infty]$ is said to be a *moment function* (with parameter $\mathbb{K}_{\mathbb{M}} \geq 1$) if it possesses the following properties for all $P_1, P_2, P_3 \in \mathcal{P}_1$.

$\mathbf{SYM}^{(4)}$ $\mathbb{M}(P_1, P_2) = \mathbb{M}(P_2, P_1)$,
$\mathbf{TI}^{(4)}$ $\mathbb{M}(P_1, P_3) \leq \mathbb{K}_{\mathbb{M}}[\mathbb{M}(P_1, P_2) + \mathbb{M}(P_2, P_3)]$.

As we noted, we use moment functions as upper bounds for p. distances μ. As a more general example, we now consider μ to be the p. average distance (see equalities (4.7.49) and (4.7.50)):

$$\mathcal{L}_p(P) := \left[\int_{U\times U} d^p(x,y)P(\mathrm{d}x,\mathrm{d}y)\right]^{p'} \quad p>0$$
$$\times\, p' := \min(1,1/p) \quad P\in\mathcal{P}_2. \tag{4.4.1}$$

For any $p>0$ and $a\in U$ define the moment function:

$$\Lambda_{p,a}(P_1,P_2) := \left[\int_U d^p(x,a)P_1(\mathrm{d}x)\right]^{p'} + \left[\int_U d^p(x,a)P_2(\mathrm{d}x)\right]^{p'} \tag{4.4.2}$$

Taking advantage of the Minkovski inequality, we get our first (rough) upper bound for the value $\mathcal{L}_p(P)$ under the convention that the marginals $T_iP = P_i$ $(i=1,2)$ are known:

$$\mathcal{L}_p(P) \le \Lambda_{p,a}(P_1,P_2). \tag{4.4.3}$$

Obviously, by the inequality (4.4.3), we can get a more refined estimate:

$$\mathcal{L}_p(P) \le \Lambda_p(P_1,P_2) \tag{4.4.4}$$

where

$$\Lambda_p(P_1,P_2) := \inf_{a\in U}\Lambda_{p,a}(P_1,P_2). \tag{4.4.5}$$

Further, we shall consider the following question.

Problem 4.4.1. What is the best possible inequality of the type

$$\mathcal{L}_p(P) \le \check{\mathcal{L}}_p(P_1,P_2), \tag{4.4.6}$$

where $\check{\mathcal{L}}_p$ is a functional that depends on the marginals $P_i = T_iP$ $(i=1,2)$ only?

Remark 4.4.2. Suppose (X,Y) is a pair of *dependent* random variables taking on values in s.m.s. (U,d). Knowing only the marginal distributions $P_1 = \mathrm{Pr}_X$ and $P_2 = \mathrm{Pr}_Y$, what is the best possible

improvement of the "triangle inequality" bound

$$\mathcal{L}_1(X, Y) := Ed(X, Y) \leq Ed(X, a) + Ed(Y, a). \tag{4.4.7}$$

The answer is simple: The best possible upper bound for $Ed(X, Y)$ is given by

$$\check{\mathcal{L}}_1(P_1, P_2) := \sup\{\mathcal{L}_1(X_1, X_2) : \Pr_{X_i} = P_i, i = 1, 2\}. \tag{4.4.8}$$

More difficult is to determine dual and explicit representations for $\check{\mathcal{L}}_1$ similar to those of the minimal metric $\widehat{\mathcal{L}}_1$ (the Kantorovich metric).

More generally, for any compound semidistance $\mu(P)$ ($P \in \mathcal{P}_2$) let us define the functional

$$\check{\mu}(P_1, P_2) := \sup\{\mu(P) : T_i P = P_i, i = 1, 2\} P_1, P_2 \in \mathcal{P}_1. \tag{4.4.9}$$

Definition 4.4.3. The functional $\check{\mu} : \mathcal{P}_1 \times \mathcal{P}_1 \to [0, \infty]$ will be called *maximal distance* w.r.t. the given compound semidistance μ.

Conceptually, the idea of maximal distances is close to the idea of minimal distances. According to the definitions, we compute the probability semidistance at all bivariate laws with one-dimensional projections equal to P_1 and P_2. However, instead of computing the infimum, we calculate the supremum. While similar on a conceptual level, the maximal distance does not have metric properties in contrast to the minimal distance. The appendix to this chapter contains further properties of maximal distances.

There are examples of probability semidistances for which the minimal and the maximal distances can be computed explicitly. Suppose that the probability semidistance can be represented in the following special form,

$$\mu_c(P) = \int_{U^2} c(x, y) P(\mathrm{d}x, \mathrm{d}y), P \in \mathcal{P}_2,$$

where $c(x, y) : U^2 \to \mathbb{R}^+$ is a symmetric, non-negative function. If the function c can be represented as $c(x, y) = \psi(x - y)$ where ψ is a convex, non-negative function, then the minimal and the maximal

distances allow for the following explicit representations,

$$\widehat{\mu}(P_1, P_2) = \int_0^1 \psi(F_1^{-1}(t) - F_2^{-1}(t))\mathrm{d}t$$
$$\breve{\mu}(P_1, P_2) = \int_0^1 \psi(F_1^{-1}(t) - F_2^{-1}(1-t))\mathrm{d}t,$$

where F_i^{-1} denotes the generalized inverse of the distribution function F_i, $i = 1, 2$. This fact can be used to derive explicit forms of minimal and maximal distances for the average compound distances $\mathcal{L}_H(P)$ given in equation (4.7.10).

Minimal and maximal distances are only one possible way to obtain lower and upper bounds of probability semidistances. There are other general constructs, which are described in the appendix to this chapter. In this section, we provide an illustration of only one of them-the *μ-lower and μ-upper bound with fixed marginal moments*. The general idea is again to compute infimum and supremum but making the constraints more loose: that is, instead of holding the marginal distribution functions fixed, we hold two marginal moments fixed. In the general setting, we compute,

$$\underset{(m)}{\mu}(a, b) = \inf\left\{\mu(P) : \int_U f\mathrm{d}P_1 = a, \int_U f\mathrm{d}P_2 = b\right\}$$

and

$$\overset{(m)}{\mu}(a, b) = \sup\left\{\mu(P) : \int_U f\mathrm{d}P_1 = a, \int_U f\mathrm{d}P_2 = b\right\}$$

where P_1 and P_2 are the marginal laws of $P \in \mathcal{P}_2$ and f defines the moment. From general arguments, it follows that

$$\underset{(m)}{\mu}(a, b) \leq \widehat{\mu}(P_1, P_2) \leq \mu(P) \leq \breve{\mu}(P_1, P_2) \leq \overset{(m)}{\mu}(a, b)$$

where P_1 and P_2 are the marginal laws of $P \in \mathcal{P}_2$. For some special cases of μ, these two bounds allow for explicit representations. For example, consider the average compound metric $\mathcal{L}_2(X, Y) =$

$(E(X-Y)^2)^{1/2}$ and the following bounds,

$$\underset{(m,p)}{\mu}(a,b) = \inf\{\mathcal{L}_2(X,Y) : (E(X-u)^p)^{1/p} = a, (E(Y-u)^p)^{1/p} = b\}$$

and

$$\overset{(m,p)}{\mu}(a,b) = \sup\{\mathcal{L}_2(X,Y) : (E(X-u)^p)^{1/p} = a, (E(Y-u)^p)^{1/p} = b\}$$

where $p \geq 1$. In this case, it can be proved that

$$\underset{(m,p)}{\mu}(a,b) = |a-b| \quad \text{and} \quad \overset{(m,p)}{\mu}(a,b) = a+b.$$

This result is consistent with the example we gave at the beginning of the section with tracking error which appears as a special case when $p = 2$.

Note that the μ-lower bound with marginal moments fixed resembles the minimal primary distance. In fact, the concept of minimal primary distance is more general and the μ-lower bound with marginal moments fixed is a special case.

In the appendix to this section, we provide other examples of lower and upper bounds and how they are related to one another.

4.5 Ideal Probability Metrics

We noted that an important application of probability metrics is in establishing the rate of convergence in limit theorems. In the literature, there are many results which state the convergence rate in terms of different simple probability metrics, such as the Kolmogorov metric, the total variation metric, the uniform metric between densities, the Kantorovich metric, etc.[1] In fact, it turned out that probability metrics with special structure have to be introduced in order for exact estimates of the convergence rate to be obtained in limit theorems. These metrics are called *ideal metrics* and their special structure is dictated by the particular problem under study – different additional axioms are added depending on the limit problem. In this respect, they are called ideal because they solve the problem in the

best possible way due to their special structure. In this section, we describe the notion of ideal probability metrics used to obtain exact convergence rates in the generalized central limit theorem (CLT) which we described in Chapter 1. It appears that the additional axioms have an interesting interpretation from the point of view of finance.

In Chapter 2, we introduced the axiomatic definition of probability metrics. We briefly repeat the definition. A probability metric $\mu(X, Y)$ is a functional which measures the "closeness" between the random variables X and Y, satisfying the following three properties:

Property 1. $\mu(X, Y) \geq 0$ for any X, Y and $\mu(X, X) = 0$
Property 2. $\mu(X, Y) = \mu(Y, X)$ for any X, Y
Property 3. $\mu(X, Y) \leq \mu(X, Z) + \mu(Z, Y)$ for any X, Y, Z

The three properties are called the *identity axiom*, the *symmetry axiom*, and the *triangle inequality*, respectively.

The ideal probability metrics are probability metrics which satisfy two additional properties which make them uniquely positioned to study problems related to the generalized CLT. The two additional properties are the homogeneity property and the regularity property.

Homogeneity property The homogeneity property is

Property 4. $\mu(cX, cY) = |c|^r \mu(X, Y)$ for any X, Y and constants $c \in \mathbb{R}$ and $r \in \mathbb{R}$.

Basically, the homogeneity property states that if we scale the two random variables by one and the same constant, the distance between the scaled quantities ($\mu(cX, cY)$) is proportional to the initial distance ($\mu(X, Y)$) by $|c|^r$. In particular, if $r = 1$, then the distance between the scaled quantities changes linearly with c.

The homogeneity property has the following financial interpretation. If X and Y are random variables describing the random return

of two portfolios, then converting proportionally into cash, for example, 30% of the two portfolios results in returns scaled down to $0.3X$ and $0.3Y$. Since the returns of the two portfolios appear scaled by the same factor, it is reasonable to assume that the distance between the two scales down proportionally.

Regularity property The regularity property is

Property 5. $\mu(X + Z, Y + Z) \leq \mu(Y, X)$ for any X, Y and Z independent of X and Y.

The regularity property states that if we add to the initial random variables X and Y one and the same random variable Z independent of X and Y, then the distance decreases.

The regularity property has the following financial interpretation. Suppose that X and Y are random variables describing the random values of two common stock portfolios and Z describes the random price of a common stock. Then buying one share of stock Z per portfolio results in two new portfolios with random wealth $X + Z$ and $Y + Z$. Because of the common factor in the two new portfolios, we can expect that the distance between $X + Z$ and $Y + Z$ is smaller than the one between X and Y.

4.5.1 Interpretation and examples of ideal probability metrics

Any functional satisfying all five properties is called an ideal probability metric of order r.

There are examples of both compound and simple ideal probability metrics. For instance, the p-average compound metric $\mathcal{L}_p(X, Y)$ defined in equation (4.7.49) in the appendix to this chapter and the Birnbaum–Orlicz metric $\mathbf{\Theta}_p(X, Y)$ defined in equation (4.7.59) in the appendix to this chapter are ideal compound probability metrics of order 1 and $1/p$, respectively. In fact, almost all known examples of ideal probability metrics of order $r > 1$ are simple metrics.

Other examples of simple ideal probability distances include:

(a) The uniform metric between densities $\ell(X, Y)$ defined as

$$\ell(X, Y) = \max_{x \in \mathbb{R}} |f_X(x) - f_Y(x)| \tag{4.5.1}$$

is an ideal metric of order -1.

(b) The L_p-metrics between distribution functions $\boldsymbol{\theta}_p(X, Y)$ defined in equation (4.7.40) in the appendix to this chapter is an ideal probability metric of order $1/p$, $p \geq 1$.

(c) The Kolmogorov metric $\boldsymbol{\rho}(X, Y)$ defined in equation (4.7.41) in the appendix to this chapter is an ideal metric of order 0. This can also be inferred from the relationship $\boldsymbol{\rho}(X, Y) = \boldsymbol{\theta}_\infty(X, Y)$.

(d) The L_p-metrics between inverse distribution functions $\ell_p(X, Y)$ defined in equation (4.7.23) in the appendix to this chapter is an ideal metric of order 1.

(e) The total variation metric $\boldsymbol{\sigma}(X, Y)$ defined in equation (4.7.25) in the appendix to this chapter is an ideal probability metric of order 0.

Let us illustrate the order of ideality, or the homogeneity order, by the ideal metrics $\ell(X, Y)$ and $\boldsymbol{\sigma}(X, Y)$ which are both based on measuring distances between density functions. The left part of Figure 4.1 shows the densities $f_X(x)$ and $f_Y(x)$ of two random variables X and Y. At the bottom of the figure, we can see the absolute difference between the two densities $|f_X(x) - f_Y(x)|$ as a function of x. The upper-right plot shows the densities of the scaled random variables $0.5X$ and $0.5Y$. Note that they are more peaked at the means of X and Y. The lower-right plot shows the absolute difference $|f_{X/2}(x) - f_{Y/2}(x)|$ as a function of x.

The uniform distance between the two densities is equal to the maximum absolute difference between them – see the definition in (4.5.1). On Figure 4.1 we can see that the maximum between the densities of the scaled random variables is clearly larger than the

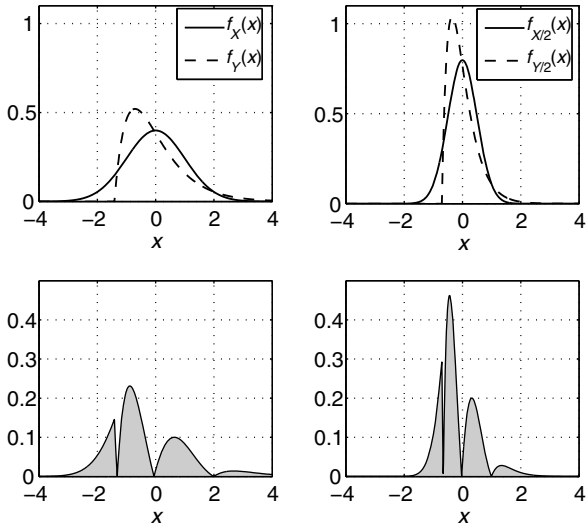

Figure 4.1: The left part shows the densities of X and Y and the absolute difference between them. The right part shows the same information but for the scaled random variables $0.5X$ and $0.5Y$.

maximum of the non-scaled counterparts. Actually, it is exactly twice as large,

$$\ell(X/2, Y/2) = 2\ell(X, Y)$$

because the metric $\ell(X, Y)$ is ideal of order -1.

The total variation metric $\boldsymbol{\sigma}(X, Y)$ can be expressed as

$$\boldsymbol{\sigma}(X, Y) = \frac{1}{2} \int_{\mathbb{R}} |f_X(x) - f_Y(x)| dx$$

provided that X and Y have densities $f_X(x)$ and $f_Y(x)$. Since the total variation metric is ideal of order zero,

$$\boldsymbol{\sigma}(X/2, Y/2) = \boldsymbol{\sigma}(X, Y),$$

then it follows that the surface closed between the two graphs is not changed by the scaling. Therefore, the shaded areas on Figure 4.1 are exactly the same.

Suppose that X and Y are random variables describing the return of two portfolios. In line with the interpretation of the homogeneity property, if we start converting those portfolios into cash, then their returns appear scaled by a smaller and smaller factor. Our expectations are that the portfolios should appear more and more alike: that is, when decreasing the scaling factor, the ideal metric should indicate that the distance between the two portfolios decreases. We verified that the metrics $\ell(X, Y)$ and $\sigma(X, Y)$ indicate otherwise. Therefore, from the perspective of applications in finance, it makes more sense to consider ideal metrics of order greater than zero, $r > 0$.

Besides the ideal metrics we have listed above, there are others which allow for interesting interpretations.

Example 4.5.1. The Zolotarev ideal metric. The general form of the Zolotarev ideal metric is

$$\zeta_s(X, Y) = \int_{-\infty}^{\infty} \left| F_{s,X}(x) - F_{s,Y}(x) \right| dx \tag{4.5.2}$$

where $s = 1, 2, \ldots$ and

$$F_{s,X}(x) = \int_{-\infty}^{x} \frac{(x-t)^{s-1}}{(s-1)!} dF_X(t) \tag{4.5.3}$$

The Zolotarev metric $\zeta_s(X, Y)$ is ideal of order $r = s$. Zolotarev (1997) provides more information.

Example 4.5.2. The Rachev metric. The general form of the Rachev metric is

$$\zeta_{s,p,\alpha}(X, Y) = \left(\int_{-\infty}^{\infty} \left| F_{s,X}(x) - F_{s,Y}(x) \right|^p |x|^{\alpha p'} dx \right)^{1/p'} \tag{4.5.4}$$

where $p' = \max(1, p), \alpha \geq 0, p \in [0, \infty]$, and $F_{s,X}(x)$ is defined in equation (4.5.3). If $\alpha = 0$, then the Rachev metric $\zeta_{s,p,0}(X, Y)$ is ideal of

order $r = (s-1)p/p' + 1/p'$. The Zolotarev metric in (4.5.2) is a special case of the Rachev metric with $\alpha = 0$ and $p = 1$.

Note that $\zeta_{s,p,\alpha}(X, Y)$ can be represented in terms of lower partial moments,

$$\zeta_{s,p,\alpha}(X, Y) = \frac{1}{(s-1)!} \left(\int_{-\infty}^{\infty} \left| E(t-X)_+^s - E(t-X)_+^s \right|^p |t|^{\alpha p'} dt \right)^{1/p'}.$$

In financial theory, the lower partial moments are used to characterize preferences of difference classes of investors. For example, the lower partial moment of order 2 characterizes the investors, preferences who are non-satiable, risk averse, and prefer positively skewed distributions. Suppose that X and Y describe the return distribution of two portfolios. X is preferred to Y by this class of investors if $EX = EY$ and

$$E(t-X)_+^2 \le E(t-Y)_+^2, \quad \forall t \in \mathbb{R}.$$

The Rachev ideal metric $\zeta_{2,p,0}(X, Y)$ quantifies such a preference order in a natural way – if X is preferred to Y, then we can calculate the distance by $\zeta_{2,p,0}(X, Y)$ and check whether X significantly dominates Y.

Example 4.5.3. The Kolmogorov–Rachev metrics. The Kolmogorov–Rachev metrics arise from other ideal metrics by a process known as *smoothing*. Suppose the metric μ is ideal of order $0 \le r \le 1$. Consider the metric defined as

$$\mu_s(X, Y) = \sup_{h \in \mathbb{R}} |h|^s \mu(X + hZ, X + hZ) \tag{4.5.5}$$

where Z is independent of X and Y and is a symmetric random variable $Z \stackrel{d}{=} -Z$. The metric $\mu_s(X, Y)$ defined in this way is ideal of order $r = s$. Note that while (4.5.5) always defines an ideal metric of order s, this does not mean that the metric is finite. The finiteness of μ_s should be studied for every choice of the metric μ.

For example, suppose that $\mu(X, Y)$ is the total variation metric $\sigma(X, Y)$ defined in (4.7.25) in the appendix to this chapter and Z has the standard normal distribution, $Z \in N(0, 1)$. If we assume that

X and Y have densities, we calculate that

$$\begin{aligned}\boldsymbol{\sigma}_s(X,Y) &= \sup_{h\in\mathbb{R}} |h|^s \boldsymbol{\sigma}(X+hZ, X+hZ) \\ &= \sup_{h\in\mathbb{R}} |h|^s \frac{1}{2}\int_{\mathbb{R}} |f_X(x) - f_Y(x)| \frac{f_Z(x/h)}{h} dx \\ &= \sup_{h\in\mathbb{R}} |h|^s \frac{1}{2}\int_{\mathbb{R}} |f_X(x) - f_Y(x)| \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{x^2}{2h^2}} dx\end{aligned} \tag{4.5.6}$$

in which we use the explicit form of the standard normal density, $f_Z(u) = \exp(-u^2/2)/\sqrt{2\pi},\ u \in \mathbb{R}$. Note that the absolute difference between the two densities of X and Y is averaged with respect to the standard normal density. This is why the Kolmogorov–Rachev metrics are also called *smoothing metrics*.

The Kolmogorov–Rachev metrics are applied in estimating the convergence rate in the Generalized CLT and other limit theorems. Rachev and Rüschendorf (1998) and Rachev (1991) provide more background and further details on the application in limit theorems.

4.5.2 Conditions for boundedness of ideal probability metrics

In the following, we specify the exact conditions which need to be satisfied in order for the ideal metrics considered to be finite. We briefly mention a few general conditions.

Suppose that the probability metric $\mu(X, Y)$ is a simple ideal metric of order r, $r > 1$. The finiteness of $\mu(X, Y)$ guarantees equality of all integer moments up to order r,

$$\mu(X,Y) < \infty \quad \Longrightarrow \quad E(X^k - Y^k) = 0, \quad k = 1, 2, \ldots, n < r.$$

Conversely, if all moments $k = 1, 2, \ldots, n < r$ agree and, in addition to this, the absolute moments of order r are finite, then metric $\mu(X, Y)$, which can be the Zolotarev metric, the Rachev metric, or the

Kolmogorov–Rachev metric, is finite,

$$\begin{aligned} &EX^k = EY^k \\ &E|X|^r < \infty \quad \Longrightarrow \quad \mu(X, Y) < \infty \\ &E|Y|^r < \infty \end{aligned}$$

where $k = 1, 2, \ldots, n < r$. In fact this result seems to be universal for all known ideal metrics of order $r > 1$.

The conditions which guarantee finiteness of the ideal metric μ are very important when investigating the problem of convergence in distribution of random variables in the context of the metric μ.[2] Consider a sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ and a random variable X which satisfy the conditions,

$$EX_n^k = EX^k, \ \forall n, \ k = 1, 2, \ldots, n < r$$

and

$$E|X|^r < \infty, \ E|X_n|^r < \infty, \ \forall n.$$

For all known ideal metrics $\mu(X, Y)$ of order $r > 0$, given the above moment assumptions, the following holds: $\mu(X_n, X) \to 0$ if and only if X_n converges to X in distribution and the absolute moment of order r converge,

$$\mu(X_n, X) \to 0 \quad \text{if and only if} \quad X_n \xrightarrow{d} X \quad \text{and} \quad E|X_n|^r \to E|X^r|.$$

This abstract result has the following interpretation. Suppose that X and Y describe the returns of two portfolios. Choose an ideal metric μ of order $3 < r < 4$, for example. The convergence result above means that if $\mu(X, Y) \approx 0$, then both portfolios have very similar distribution functions and also they have very similar means, volatilities and skewness.

Note that, generally, the distribution functions of two portfolios being "close" to each other does not necessarily mean that their moments will be approximately the same. The ideal metrics have this

nice property that they guarantee convergence of certain moments. Rachev (1991) provides an extensive review of the properties of ideal metrics and their applications to limit theorems in probability theory.

4.6 Summary

In this chapter, we described in detail the characterization of probability distances in terms of the three major classes of primary, simple, and compound distances. We discussed constructions of minimal functionals which have a very important place in the theory as they lead to the notions of minimal metrics, minimal norms, and co-minimal functionals used to derive lower bounds to probability distances. Also, we considered the notion of moment functions which are used to construct upper bounds of probability distances and, as a consequence, are related to the problem of specifying the conditions under which given probability distances are finite. Finally, we considered the class of ideal probability metrics and provided applications in finance.

4.7 Technical Appendix

Particular examples of primary, simple, and compound probability distances are provided in this appendix.

4.7.1 Examples of primary distances

In this section, we provide additional examples of primary semimetrics. Before starting with the examples, we would like to stress that as far as the minimal functional $\widetilde{\mu}_h$ introduced in section 4.2 is concerned, the following problem arises.

In general it is not true that the metric properties of a p. distance μ imply that $\widetilde{\mu}_h$ is a distance. The following two examples

illustrate this fact:

(a) Let $U = \mathbb{R}$, $d(x, y) = |x - y|$. Consider the p. metric

$$\mu(X, Y) = \mathcal{X}_0(X, Y) = \Pr(X \neq Y) \quad X, Y \in \mathfrak{X}(\mathbb{R})$$

and the mapping $h : \mathfrak{X}(\mathbb{R}) \to [0, \infty]$ given by $hX = E|X|$. Then

$$\widetilde{\mu}_h(a, b) = \inf\{\Pr(X \neq Y) : E|X| = a,\ E|Y| = b\} = 0$$

for all $a \geq 0$ and $b \geq 0$. Hence in this case the metric properties of μ imply only semimetric properties for $\widetilde{\mu}_h$.

(b) Now let μ be defined as in (a) but $h : \mathfrak{X}(\mathbb{R}) \to [0, \infty] \times [0, \infty]$ be defined by $hX = (E|X|, EX^2)$. Then

$$\begin{aligned} &\mu_h((a_1, a_2), (b_1, b_2)) \\ &= \inf\{\Pr(X \neq Y) : E|X| = a_1, EX^2 = a_2, E|Y| = b_1, EY^2 = b_2\} \end{aligned} \tag{4.7.1}$$

where $\widetilde{\mu}_h$ is not even p. semidistance since the triangle inequality $TI^{(3*)}$ is not valid.

With respect to this, the following problem arises, which is not yet completely resolved in the literature on probability metrics: *under which condition on the space U, p. distance μ on $\mathfrak{X}(U)$ and transformation $h : \mathfrak{X}(U) \to R^l$ the primary h-minimal distance $\widetilde{\mu}_h$ is a primary p. distance in $h(\mathfrak{X})$?*

As we shall see later on, all further examples 4.7.1 to 4.7.4 of primary distances are special cases of primary h-minimal distances.

Example 4.7.1. Let $H \in \mathcal{H}$ (see Example 2.3.1) and $\overline{0}$ be a fixed point of a s.m.s. (U, d). For each $P \in \mathcal{P}_2$ with marginals $P_i = T_i P$, let $m_1 P$, $m_2 P$ denote the "marginal moments of order $p > 0$",

$$m_i P := m_i^{(p)} P := \left(\int_U d^p(x, \overline{0}) P_i(\mathrm{d}x) \right)^{p'} \quad p > 0 \quad p' := \min(1, 1/p).$$

Then

$$\mathcal{M}_{H,p}(P) := \mathcal{M}_{H,p}(m_1 P, m_2 P) := H(|m_1 P - m_2 P|) \tag{4.7.2}$$

is a primary distance. One can also consider $\mathcal{M}_{H,p}$ as a distance in the space

$$m^{(p)}(\mathcal{P}_1) := \left\{ m^{(p)} := \left(\int_U d^p(x,a) P(\mathrm{d}x) \right)^{p'} < \infty, P \in \mathcal{P}_1(U) \right\} \tag{4.7.3}$$

of moments $m^{(p)}P$ of order $p > 0$. If $H(t) = t$ then

$$\mathcal{M}(P) := \mathcal{M}_{H,1}(P) = \left| \int_U d(x,\overline{0})(P_1 - P_2)(\mathrm{d}x) \right|$$

is a primary metric in $m^{(p)}(\mathcal{P}_1)$.

Example 4.7.2. Let $g : [0,\infty] \to \mathbb{R}$ and $H \in \mathcal{H}$ Then

$$\mathcal{M}(g)_{H,p}(m_1P, m_2P) := H(|g(m_1P) - g(m_2P)|) \tag{4.7.4}$$

is a primary distance in $g \circ m(\mathcal{P}_1)$ and

$$\mathcal{M}(g)(m_1P, m_2P) := |g(m_1P) - g(m_2P)| \tag{4.7.5}$$

is a primary metric.

If U is a Banach space with norm $\|\cdot\|$ then we define the primary distance $\mathcal{M}_{H,p}(g)$ as follows

$$\mathcal{M}_{H,p}(g)(m^{(p)}X, m^{(p)}Y) := H(|m^{(p)}P - m^{(p)}Y|) \tag{4.7.6}$$

where (cf. (2.2.10)) $m^{(p)}X$ is the *'p-th moment (norm) of X'*

$$m^{(p)}X := \{E\|X\|^p\}^{p'}.$$

By equation (4.7.5), $\mathcal{M}_{H,p}(g)$ may be viewed as a distance (see Definition 2.3.2) in the space

$$\begin{aligned} g \circ m(\mathfrak{X}) := \{ g \circ m(X) &:= g(\{E\|X\|^p\}^{p'}), X \in \mathfrak{X} \}\ p' \\ &= \min(1, p^{-1}),\ \mathfrak{X} = \mathfrak{X}(U) \end{aligned} \tag{4.7.7}$$

of moments $g \circ m(X)$. If U is the real line $\mathbb{R}$ and $g(t) = H(t) = t(t \geq 0)$ then $\mathcal{M}_{H,p}(m^{(p)}X, m^{(p)}Y)$ is the usual deviation between moments $m^{(p)}X$ and $m^{(p)}Y$ (see (2.2.11)).

Example 4.7.3. Let J be an index set (with arbitrary cardinality), g_i $(i \in J)$ be real functions on $[0, \infty]$ and for each $P \in \mathcal{P}_1(U)$ define the set

$$hP := \{g_i(mP), i \in J\} \tag{4.7.8}$$

Further, for each $P \in \mathcal{P}_2(U)$ let us consider hP_1 and hP_2 where P_i's are the marginals of P. Then

$$\mathbf{\Omega}(hP_1, hP_2) = \begin{cases} 0 & \text{if } hP_1 \equiv hP_2 \\ 1 & \text{otherwise} \end{cases} \tag{4.7.9}$$

is a primary metric.

Example 4.7.4. Let U be the n-dimensional Euclidean space $\mathbb{R}^n$, $H \in \mathcal{H}$. Define the *'engineer distance'*

$$\mathbf{EN}(X, Y; H) := H\left(\left|\sum_{i=1}^{n}(EX_i - EY_i)\right|\right) \tag{4.7.10}$$

where $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$ belong to the subset $\widetilde{\mathfrak{X}}(\mathbb{R}^n) \subseteq \mathfrak{X}(\mathbb{R}^n)$ of all n-dimensional random vectors that have integrable components. Then $\mathbf{EN}(\cdot, \cdot; H)$ is a p. semidistance in $\widetilde{\mathfrak{X}}(\mathbb{R}^n)$. Analogously, the *'$L_p$-engineer metric'*

$$\mathbf{EN}(X, Y, p) := \left[\sum_{i=1}^{n} |EX - EY|^p\right]^{\min(1,1/p)}, p > 0 \tag{4.7.11}$$

is a primary metric in $\widetilde{\mathfrak{X}}(\mathbb{R}^n)$. In the case $p = 1$ and $n = 1$, the metric $\mathbf{EN}(\cdot, \cdot; p)$ coincides with the engineer metric in $\mathfrak{X}(\mathbb{R})$ (see (2.2.1)).

4.7.2 Examples of simple distances

In the introduction to this chapter, we noted that there are simple distances which have alternative representations known as dual forms. In this section, we provide additional examples of simple distances together with their dual forms.

Example 4.7.5. Co-minimal metrics. As we have seen in section 4.2 each primary distance $\mu(P) = \mu(h(T_1P), h(T_2P))$ $(P \in \mathcal{P}_2)$ determines a semidistance (see Definition 2.3.2) in the space of equivalence classes

$$\{P \in \mathcal{P}_2 : h(T_1P) = a, h(T_2P) = b\} \quad a, b \in \mathbb{R}^J. \tag{4.7.12}$$

Analogously, the minimal distance

$$\begin{aligned}\widehat{\mu}(P) :=& \widehat{\mu}(T_1P, T_2P)\\ :=& \inf\{\mu(\widetilde{P}) : \widetilde{P} \in \mathcal{P}_2(U), \widetilde{P} \text{ and } P \text{ have one and the same marginals,}\\ & T_i\widetilde{P} = T_iP, i = 1, 2\}, P \in \mathcal{P}_2(U)\end{aligned}$$

may be viewed as a semidistance in the space of classes of equivalence

$$\{P \in \mathcal{P}_2 : T_1P = P_1, T_2P = P_2\} \quad P_1, P_2 \in \mathcal{P}_1 \tag{4.7.13}$$

The partitioning (4.7.13) is more refined than equation (4.7.12) and hence each primary semidistance is a simple semidistance. Thus

$$\begin{aligned}&\{\textit{the class of primary distances} \text{ (Definition 4.2.1)}\}\\ &\quad \subset \{\textit{the class of simple semidistances} \text{ (Definition 4.3.1)}\}\\ &\quad \subset \{\textit{the class of all p. semidistances} \text{ (Definition 2.4.1)}\}.\end{aligned}$$

A basic problem in TPM, which is still open, is to find a good classification of the set of all p. semidistances. Does there exist a 'Mendeleyev periodic table' of p. semidistances?

One can get a classification of probability semidistances considering more and more refined partitions of $\mathcal{P}_2$. For instance, one can

use a partition finer than equation (4.7.13), generated by

$$\{P \in \mathcal{PC}_t \subset \mathcal{P}_2 : T_1P = P_1, T_2P = P_2\}, \quad t \in T \tag{4.7.14}$$

where P_1 and P_2 are laws in $\mathcal{P}_1$ and $\mathcal{PC}_t$ ($t \in T$) are subsets of $\mathcal{P}_2$, whose union covers $\mathcal{P}_2$. As an example of the set $\mathcal{PC}_t$ one could consider

$$\mathcal{PC}_t = \left\{P \in \mathcal{P}_2 : \int_{U^2} f_i \mathrm{d}P \le b_i, i \in J\right\} \quad t = (J, \overline{b}, \overline{f}) \tag{4.7.15}$$

where J is an index set, $\overline{b} := (b_i, i \in J)$ is a set of reals and $\overline{f} = \{f_i, i \in J\}$ is a family of bounded continuous functions on U^2 (Kemperman, 1983; Levin and Rachev, 1990).

Another useful example of a set $\mathcal{PC}_t$ is constructed using a given probability metric $\nu(P)$ ($P \in \mathcal{P}_2$) and has the form

$$\mathcal{PC}_t = \{P \in \mathcal{P}_2 : \nu(P) \le t\} \tag{4.7.16}$$

where $t \in [0, \infty]$ is a fixed number.

Related is the following question. Under which conditions is the functional

$$\mu(P_1, P_2; \mathcal{PC}_t) := \inf\{\mu(P) : P \in \mathcal{P}_2, T_iP = P_i (i = 1, 2), P \in \mathcal{PC}_t\} \quad (P_1, P_2 \in \mathcal{P}_1)$$

a simple semidistance (resp., semimetric) w.r.t. the given p. distance (resp. metric) μ? TPM does not provide a complete solution as this question is still open.

Further, we shall examine this problem in the special case of (4.7.16) (see Theorem 4.7.1). Analogously, one can investigate the case of $\mathcal{PC}_t = \{P \in \mathcal{P}_2 : \nu_i(P) \le \alpha_i, i = 1, 2 \dots\}$ ($t = (\alpha_1, \alpha_2, \dots)$) for fixed p. metrics ν_i, and $\alpha_i \in [0, \infty]$.

As we will see in the next theorem, the functional $\mu\nu(\cdot, \cdot, \alpha)$ has some metric properties but nevertheless it is not a p. distance. However, $\mu\nu(\cdot, \cdot, \alpha)$ induces p. semidistances as follows.

Let $\mu\nu$ be the so-called *co-minimal distance*

$$\mu\nu(P_1, P_2) = \inf\{\alpha > 0; \mu\nu(P_1, P_2, \alpha) < \alpha\} \tag{4.7.17}$$

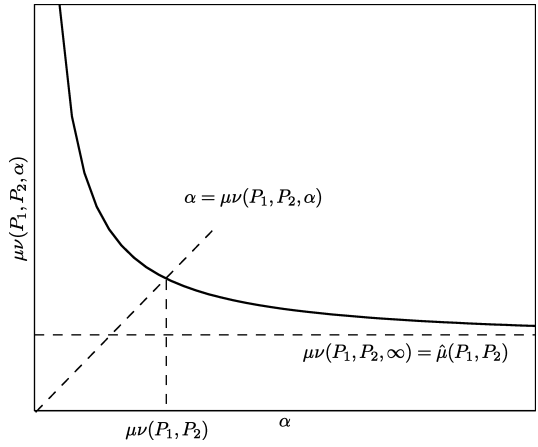

Figure 4.2: Co-minimal distance $\mu\nu(P_1, P_2)$.

(see Figure 4.2) and let

$$\overline{\mu\nu}(P_1, P_2) = \lim_{\alpha \to 0} \sup \alpha\mu\nu(P_1, P_2, \alpha).$$

Then the following theorem is true.

Theorem 4.7.1. Let U be an u.m. s.m.s. and μ be a p. distance satisfying the 'continuity' condition in Theorem 4.3.1. Then, for any p. distance ν,

(a) $\mu\nu(\cdot, \cdot, \alpha)$ satisfies the following metric properties:

$$\begin{aligned}
&\mathbf{ID}^{(3)}: && \mu\nu(P_1, P_2, \alpha) = 0 \iff P_1 = P_2 \\
&\mathbf{SYM}^{(3)}: && \mu\nu(P_1, P_2, \alpha) = \mu\nu(P_2, P_1, \alpha) \\
&\mathbf{TI}^{(3)}: && \mu\nu(P_1, P_3, \mathbb{K}_\nu(\alpha + \beta)) \le \mathbb{K}_\mu(\mu\nu(P_1, P_2, \alpha) + \\
& && \quad \mu\nu(P_2, P_3, \beta))
\end{aligned}$$

for any $P_1, P_2, P_3 \in \mathcal{P}_1$, $\alpha \ge 0$, $\beta \ge 0$.

(b) $\mu\nu$ is a simple distance with parameter $\mathbb{K}_{\mu\nu} = \max[\mathbb{K}_\mu, \mathbb{K}_\nu]$. In particular, if μ and ν are p. metrics then $\mu\nu$ is a simple metric.

(c) $\overline{\mu\nu}$ is a simple semidistance with parameter $\mathbb{K}_{\overline{\mu\nu}} = 2\mathbb{K}_\mu\mathbb{K}_\nu$.

Proof. (a) By Theorem 4.3.1, and Figure 4.2, $\mu\nu(P_1, P_2, \alpha) = 0 \Rightarrow \widehat{\mu}(P_1, P_2) = 0 \to P_1 = P_2$ as well as if $P_1 \in \mathcal{P}_1$ and X is a r.v. with distribution P_1 then $\mu\nu(P_1, P_2, \alpha) \leq \mu(\mathrm{Pr}_{X,X}) = 0$. So, $\mathbf{ID}^{(3)}$ is valid. Let us prove $\mathbf{TI}^{(3)}$. For each P_1, P_2, $P_3 \in \mathcal{P}_1$ $\alpha \geq 0$, $\beta \geq 0$ and $\varepsilon \geq 0$ define laws $P_{12} \in \mathcal{P}_2$ and $P_{23} \in \mathcal{P}_2$ such that $T_i P_{12} = P_i$, $T_i P_{23} = P_{i+1}$ $(i = 1, 2)$, $\nu(P_{12}) \leq \alpha$, $\nu(P_{23}) \leq \beta$ and $\mu\nu(P_1, P_2, \alpha) \geq \mu(P_{12}) - \varepsilon$, $\mu\nu(P_2, P_3, \alpha) \geq \mu(P_{23}) - \varepsilon$. Define a law $Q \in \mathcal{P}_3$ by equation (4.3.5). Then Q has bivariate marginals $T_{12}Q = P_{12}$ and $T_{23}Q = P_{23}$, hence, $\nu(T_{13}Q) \leq \mathbb{K}_\nu[\nu(P_{12}) + \nu(P_{23})] \leq \mathbb{K}_\nu(\alpha + \beta)$ and

$$\begin{aligned}\mu\nu(P_1, P_3, \mathbb{K}_\nu(\alpha + \beta)) &\leq \mu(T_{13}Q) \leq \mathbb{K}[\mu(P_{12}) + \mu(P_{23})] \\ &\leq \mathbb{K}_\mu[\mu\nu(P_1, P_2, \alpha) + \mu\nu(P_2, P_3, \beta) + 2\varepsilon].\end{aligned}$$

Letting $\varepsilon \to 0$, we get $\mathbf{TI}^{(3)}$.
(b) If $\mu\nu(P_1, P_2) < \alpha$ and $\mu\nu(P_2, P_3) < \beta$, then there exists P_{12}; (resp, P_{23}) with marginals P_1 and P_2 (resp. P_2 and P_3) such that $\mu(P_{12}) < \alpha$, $\nu(P_{12}) < \alpha$, $\mu(P_{23}) < \beta$. In a similar way, as in (a) we conclude that $\mu\nu(P_1, P_3, \mathbb{K}_\nu(\alpha + \beta)) < \mathbb{K}_\mu(\alpha + \beta)$, thus, $\mu\nu(P_1, P_2) < \max(\mathbb{K}_\mu, \mathbb{K}_\nu)(\alpha + \beta)$.
(c) Follows from (a) with $\alpha = \beta$. □

Example 4.7.6. (Kantorovich metric and Kantorovich distance). In section 2.2, we introduced the Kantorovich metric $\boldsymbol{\kappa}$ and its 'dual' representation

$$\begin{aligned}\boldsymbol{\kappa}(P_1, P_2) &= \int_{-\infty}^{+\infty} |F_1(x) - F_2(x)| \mathrm{d}x \\ &= \sup\left\{\left|\int_{\mathbb{R}} f \mathrm{d}(P_1 - P_2)\right| : f : \mathbb{R} \to \mathbb{R}, f' \text{ exists a.e. and } |f'| < 1 \text{ a.e.}\right\}\end{aligned}$$

where P_is are laws on $\mathbb{R}$ with d.f.s F_i and finite first absolute moment. From the above representation it also follows that

$$\boldsymbol{\kappa}(P_1, P_2) = \sup\left\{\left|\int_{\mathbb{R}} f \mathrm{d}(P_1 - P_2)\right| : f : \mathbb{R} \to \mathbb{R}, f \text{ is } (1,1)\text{-Lipschitz, i.e., } |f(x) - f(y)| \leq |x - y| \forall x, y \in \mathbb{R}\right\}$$

In this example we shall extend the definition of the above simple p. metric of the set $\mathcal{P}_1(U)$ of all laws on a s.m.s. (U, d). For any $\alpha \in (0, \infty)$ and $\beta \in [0, 1]$ define the Lipschitz functions class

$$\mathrm{Lip}_{\alpha\beta} := \{f : U \to \mathbb{R} : |f(x) - f(y)| \leq \alpha \mathrm{d}^\beta(x, y) \; \forall x, y \in U\} \quad (4.7.18)$$

with the convention

$$\mathrm{d}^0(x, y) := \begin{cases} 1 & \text{if} \quad x \neq y \\ 0 & \text{if} \quad x = y. \end{cases} \quad (4.7.19)$$

Denote the set of all bounded functions $f \in \mathrm{Lip}_{\alpha\beta}(U)$ by $\mathrm{Lip}^b_{\alpha\beta}(U)$. Let $\mathcal{G}_H(U)$ be the class of all pairs (f, g) of functions that belong to the set

$$\mathrm{Lip}^b(U) := \bigcup_{\alpha > 0} \mathrm{Lip}_{\alpha,1}(U) \quad (4.7.20)$$

and satisfy the inequality

$$f(x) + g(y) \leq H(d(x, y)) \; \forall x, y \in U \quad (4.7.21)$$

where H is a convex function from $\mathcal{H}$. Recall that $H \in \mathcal{H}$ if H is a non-decreasing continuous function from $[0, \infty)$ onto $[0, \infty)$, vanishes at the origin and $K_H := \sup_{t>0} H(2t)/H(t) < \infty$. For any two laws P_1 and P_2 on a s.m.s. (U, d) define

$$\ell_H(P_1, P_2) := \sup \left\{ \int_U f \mathrm{d}P_1 + \int_U g \mathrm{d}P_2 : (f, g) \in \mathcal{G}_H(U) \right\}. \quad (4.7.22)$$

We shall prove further that ℓ_H is a simple distance with $\mathbb{K}_{\ell_H} = \mathbb{K}_H$ in the space of all laws P with finite 'H-moment', $\int H(\mathrm{d}(x, a))P(\mathrm{d}x) < \infty$. The proof is based on the representation of ℓ_H as a minimal distance $\ell_H = \widehat{\mathcal{L}}_H$ with respect to a p. distance (with $\mathbb{K}_{\mathcal{L}_H} = \mathbb{K}_H$) $\ell_H(P) = \int_{U^2} H(\mathrm{d}(x, y))P(\mathrm{d}x, \mathrm{d}y)$ and then an appeal to Theorem 4.3.1 proves that ℓ_H is a simple p. distance if (U, d) is a universally measurable s.m.s. In the case $H(t) = t^p$ $(1 < p < \infty)$ define

$$\ell_p(P_1, P_2) := \ell_H(P_1, P_2)^{1/p} \quad 1 < p < \infty. \quad (4.7.23)$$

In addition, for $p \in [0, 1]$ and $p = \infty$, denote

$$\ell_p(P_1, P_2) := \sup\left\{\left|\int_U f \mathrm{d}(P_1 - P_2)\right| : f \in \mathrm{Lip}^b_{1,p}(U)\right\} \times p \in (0, 1]\, P_1, P_2 \in \mathcal{P}_1(U) \tag{4.7.24}$$

$$\ell_0(P_1 - P_2) := \left\{\left|\int_U f \mathrm{d}(P_1 - P_2)\right| : f \in \mathrm{Lip}_{1,0}(U)\right\} \tag{4.7.25}$$

$$= \boldsymbol{\sigma}(P_1, P_2) := \sup_{A \in \mathcal{B}_1} |P_1(A) - P_2(A)|$$

$$\ell_\infty(P_1, P_2) := \inf\{\varepsilon > 0 : P_1(A) \le P_2(A^\varepsilon)\ \forall A \in \mathcal{B}_1\} \tag{4.7.26}$$

where, as above, $\mathcal{B}_1 = \mathcal{B}(U)$ is the Borel σ-algebra on a s.m.s. (U, d), and $A^\varepsilon := \{x : d(x, A) < \varepsilon\}$.

For any $0 \le p \le 1$, $p = \infty$, ℓ_p is a simple metric in $\mathcal{P}_1(U)$ which follows immediately from the definition. To prove that ℓ_p is a p. metric (taking possibly infinite values) one can use the equality

$$\sup_{A \in \mathcal{B}_1}[P_1(A) - P_2(A^\varepsilon)] = \sup_{A \in \mathcal{B}_1}[P_2(A) - P_1(A^\varepsilon)].$$

The equality $\ell_0 = \boldsymbol{\sigma}$ in equation (4.7.25) follows from the fact that both metrics are minimal with respect to one and the same p. distance $\mathcal{L}_0(P) = P((x, y) : x \ne y)$. We shall prove also that $\ell_H = \widehat{\mathcal{L}}_H$, as a minimal distance w.r.t. ℓ_H defined above, admits the Birnbaum–Orlicz representation (see Example 2.3.2)

$$\ell_H(P_1, P_2) = \ell_H(F_1, F_2) := \int_0^1 H(|F_1^{-1}(t) - F_2^{-1}(t)|)\mathrm{d}t \tag{4.7.27}$$

in the case of $U = \mathbb{R}$ and $d(x, y) = |x - y|$. In equation (4.7.27),

$$F_i^{-1}(t) := \sup\{x : F_i(x) \le t\} \tag{4.7.28}$$

is the (generalized) *inverse* of the d.f. F_i determined by P_i $(i = 1, 2)$. Letting $H(t) = t$ we claim that

$$\begin{aligned}\ell_1(P_1, P_2) &= \int_0^1 |F_1^{-1}(t) - F_2^{-1}(t)| \mathrm{d}t \\ &= \boldsymbol{\kappa}(P_1, P_2) := \int_{-\infty}^{\infty} |F_1(x) - F_2(x)| \mathrm{d}x \; P_i \in \mathcal{P}(\mathbb{R}) \; i = 1, 2\end{aligned} \tag{4.7.29}$$

Remark 4.7.1. Here and in the sequel, for any simple semidistance μ on $\mathcal{P}(\mathbb{R}^n)$ we shall use the following notations interchangeably:

$$\mu = \mu(P_1, P_2) \; \forall P_1, P_2 \in \mathcal{P}(\mathbb{R}^n)$$
$$\mu = \mu(X_1, X_2) := \mu(\mathrm{Pr}_{X_1}, \mathrm{Pr}_{X_2}) \; \forall X_1, X_2 \in \mathcal{X}(\mathbb{R}^n)$$
$$\mu = \mu(F_1, F_2) := \mu(P_1, P_2) \; \forall F_1, F_2 \in \mathcal{F}(\mathbb{R}^n)$$

where Pr_{X_i} is the distribution of X_i, F_i is the d.f. of P_i and $\mathcal{F}(\mathbb{R}^n)$ stands for the class of d.f.s on $\mathbb{R}^n$.

The ℓ_1-metric (4.7.29) is known as the *average metric* in $\mathcal{F}(\mathbb{R})$ as well as the *first difference pseudomoment*, and it is also denoted by $\boldsymbol{\kappa}$ (see Zolotarev 1976). A great contribution in the investigation of ℓ_1-metric properties was made by Kantorovich 1942, 1948, Kantorovich and Akilov 1984, 4, Chap. VIII). That is the reason the metric ℓ_1 is called the Kantorovich metric. Considering ℓ_H as a generalization ℓ_1, we shall call ℓ_H the *Kantorovich distance.*

Example 4.7.7. (Prokhorov metric and Prokhorov distance). Prokhorov 1956 introduced his famous metric

$$\begin{aligned}\boldsymbol{\pi}(P_1, P_2) := &\inf\{\varepsilon > 0 : P_1(C) \le P_2(C^\varepsilon) + \varepsilon, P_2(C) \le P_1(C^\varepsilon) \\ &+ \varepsilon \; \forall C \in \mathcal{C}\}\end{aligned} \tag{4.7.30}$$

where $\mathcal{C} := \mathcal{C}(U)$ is the set of all non-empty closed subsets of a Polish space U and

$$C^{\varepsilon} := \{x : d(x, C) < \varepsilon\}. \tag{4.7.31}$$

The metric $\boldsymbol{\pi}$ admits the following representations: for any laws P_1 and P_2 on a s.m.s. (U, d)

$$\begin{aligned}\boldsymbol{\pi}(P_1, P_2) &= \inf\{\varepsilon > 0 : P_1(C) \le P_2(C^{\varepsilon}) + \varepsilon \text{ for any } C \in \mathcal{C}\} \\ &= \inf\{\varepsilon > 0 : P_1(C) \le P_2(C^{\varepsilon]}) + \varepsilon \text{ for any } C \in \mathcal{C}\} \\ &= \inf\{\varepsilon > 0 : P_1(A) \le P_2(A^{\varepsilon}) + \varepsilon, \text{ for any } A \in \mathcal{B}_1\}\end{aligned} \tag{4.7.32}$$

where

$$C^{\varepsilon]} = \{x : d(x, C) \le \varepsilon\} \tag{4.7.33}$$

is the ε-closed neighborhood of C (see, for example, Theorem 8.1, Dudley 1976).

Let us introduce a *parametric version of the Prokhorov metric*

$$\boldsymbol{\pi}_{\lambda}(P_1, P_2) := \inf\{\varepsilon > 0 : P_1(C) \le P_2(C^{\lambda\varepsilon}) + \varepsilon \text{ for any } C \in \mathcal{C}\}. \tag{4.7.34}$$

The next lemma gives the main relationship between the Prokhorov-type metrics and the metrics ℓ_0 and ℓ_{∞} denned by equalities (4.7.25) and (4.7.26).

Lemma 4.7.1. For any $P_1, P_2 \in \mathcal{P}_1(U)$

$$\lim_{\lambda \to 0} \boldsymbol{\pi}_{\lambda}(P_1, P_2) = \boldsymbol{\sigma}(P_1, P_2) = \ell_0(P_1, P_2) \tag{4.7.35}$$

$$\lim_{\lambda \to 0} \lambda \boldsymbol{\pi}_{\lambda}(P_1, P_2) = \ell_{\infty}(P_1, P_2).$$

Proof. For any fixed $\varepsilon > 0$ the function $A_{\varepsilon}(\lambda) := \sup\{P_1(C) - P_2(C^{\lambda\varepsilon}) : C \in \mathcal{C}\}$, $\lambda \ge 0$ is non-increasing on $\varepsilon > 0$, hence

$$\boldsymbol{\pi}_{\lambda}(P_1, P_2) = \inf\{\varepsilon > 0 : A_{\varepsilon}(\lambda) \le \varepsilon\} = \max_{\varepsilon > 0} \min(\varepsilon, A_{\varepsilon}(\lambda)).$$

For any fixed $\varepsilon > 0$, $A_\varepsilon(\cdot)$ is non-increasing and

$$\begin{aligned}\lim_{\lambda\to 0} A_\varepsilon(\lambda) &= A_\varepsilon(0) = \sup_{C\in\mathcal{C}}(P_1(C) - P_2(C)) = \sup_{A\in\mathcal{B}(U)}(P_1(A) - P_2(A)) \\ &= \sup_{A\in\mathcal{B}(U)} |P_1(A) - P_2(A)| =: \boldsymbol{\sigma}(P_1, P_2)\end{aligned}$$

Thus

$$\begin{aligned}\lim_{\lambda\to 0} \boldsymbol{\pi}_\lambda(P_1, P_2) &= \max_{\varepsilon>0} \min\left(\varepsilon, \lim_{\lambda\to 0} A_\varepsilon(\lambda)\right) \\ &= \max_{\varepsilon>0} \min(\varepsilon, \boldsymbol{\sigma}(P_1, P_2)) = \boldsymbol{\sigma}(P_1, P_2).\end{aligned}$$

Analogously, as $\lambda \to \infty$

$$\begin{aligned}\lambda\boldsymbol{\pi}_\lambda(P_1, P_2) &= \inf\{\lambda\varepsilon > 0 : A_\varepsilon(\lambda) \le \varepsilon\} \\ &= \inf\{\varepsilon > 0 : A_\varepsilon(1) \le \varepsilon/\lambda\} \to \inf\{\varepsilon > 0 : A_\varepsilon(1) \le 0\} \\ &= \ell_\infty(P_1, P_2)\end{aligned}$$

□

As a generalization of $\boldsymbol{\pi}_\lambda$ we define the *Prokhorov distance*

$$\boldsymbol{\pi}_H(P_1, P_2) := \inf\{H(\varepsilon) > 0 : P_1(A^\varepsilon) \le P_2(A) + H(\varepsilon),\ \forall A \in \mathcal{B}_1\} \tag{4.7.36}$$

for any strictly increasing function $H \in \mathcal{H}$.

From equation (4.7.36),

$$\boldsymbol{\pi}(P_1, P_2) = \inf\{\delta > 0 : P_1(A) \le P_2(A^{H^{-1}(\delta)}) + \delta \text{ for any } A \in \mathcal{B}_1\} \tag{4.7.37}$$

and it is easy to check that $\boldsymbol{\pi}_H$ is a simple distance with $\mathbb{K}_{\pi_H} = \mathbb{K}_H$ (see condition (2.3.3)). The metric $\boldsymbol{\pi}_\lambda$ is a special case of $\boldsymbol{\pi}_H$ with $H(t) = t/\lambda$.

Example 4.7.8. (Birnbaum–Orlicz distance ($\boldsymbol{\theta}_H$) and $\boldsymbol{\theta}_p$-metric in $\mathcal{P}(\mathbb{R})$). Let $U = \mathbb{R}$, $\mathrm{d}(x, y) = |x - y|$. Following Example 2.3.2 we define

the *Birnbaum–Orlicz average distance*

$$\boldsymbol{\theta}_H(F_1, F_2) := \int_{-\infty}^{+\infty} H(|F_1(t) - F_2(t)|)\mathrm{d}t \quad H \in \mathcal{H} \quad F_i \in \mathcal{F}(\mathbb{R}) \quad i = 1, 2 \tag{4.7.38}$$

and the *Birnbaum–Orlicz uniform distance*

$$\boldsymbol{\rho}_H(F_1, F_2) := H(\boldsymbol{\rho}(F_1, F_2)) = \sup_{x\in\mathbb{R}} H(|F_1(x) - F_2(x)|). \tag{4.7.39}$$

The $\boldsymbol{\theta}_p$*-metric* ($p > 0$)

$$\boldsymbol{\theta}_p(F_1, F_2) := \left\{ \int_{-\infty}^{\infty} |F_1(t) - F_2(t)|^p \mathrm{d}t \right\}^{p'} \quad p' := \min(1, 1/p) \tag{4.7.40}$$

is a special case of $\boldsymbol{\theta}_H$ with appropriate normalization that makes $\boldsymbol{\theta}_p$ p. metric taking finite and infinite values in the distribution functions space $\mathcal{F} := \mathcal{F}(\mathbb{R})$. In case $p = \infty$ we denote $\boldsymbol{\theta}_\infty$ to be the Kolmogorov metric

$$\boldsymbol{\theta}_\infty(F_1, F_2) := \boldsymbol{\rho}(F_1, F_2) := \sup_{x\in\mathbb{R}} |F_1(x) - F_2(x)|. \tag{4.7.41}$$

In the case $p = 0$ we put

$$\boldsymbol{\theta}_0(F_1, F_2) := \int_{-\infty}^{\infty} I\{t : F_1(t) \neq F_2(t)\}\mathrm{d}t = \mathrm{Leb}(F_1 \neq F_2)$$

Here as in the following, $I(A)$ is the indicator of the set A.

Example 4.7.9. Minimal norms. We noted that each co-minimal distance $\mu\nu$ is greater than the minimal distance $\widehat{\mu}$ (see Figure 4.3). We now consider examples of simple metrics $\overset{\circ}{\mu}$ corresponding to given p. distances $\widehat{\mu}$ that have (like $\mu\nu$) a 'minimal' structure but $\overset{\circ}{\mu} \leq \widehat{\mu}$.

Let $\mathcal{M}_k$ be the set of all finite non-negative measures on the Borel σ-algebra $\mathcal{B}_k = \mathcal{B}(U^k)$ (U is a s.m.s.). Let $\mathcal{M}_0$ denote the space of all finite signed measures ν on $\mathcal{B}_1$ with total mass $m(U) = 0$. Denote

by $\mathcal{CS}(U^2)$ the set of all continuous, symmetric, and non-negative functions on U^2. Define the functionals

$$\mu_c(m) := \int_{U^2} c(x,y)m(\mathrm{d}x,\mathrm{d}y),\ m \in \mathcal{M}_2 \quad c \in \mathcal{CS}(U^2) \tag{4.7.42}$$

and

$$\overset{\circ}{\mu}_c(\nu) := \inf\{\mu_c(m) : T_1 m - T_2 m = \nu\} \quad \nu \in \mathcal{M}_0 \tag{4.7.43}$$

where $T_i m$ means the i-th marginal measure of m.

Lemma 4.7.2. For any $c \in \mathcal{CS}(U^2)$ the functional $\overset{\circ}{\mu}_c$ is a seminorm in the space $\mathcal{M}_0$.

Proof. Obviously, $\overset{\circ}{\mu}_c \geq 0$. For any positive constant a we have

$$\begin{aligned}\overset{\circ}{\mu}_c(a\nu) &= \inf\{\mu_c(m) : T_1(1/a)m - T_2(1/a)m = \nu\}\\ &= a\inf\{\mu_c((1/a)m) : T_1(1/a)m - T_2(1/a)m = \nu\}\\ &= a\overset{\circ}{\mu}_c(\nu).\end{aligned}$$

If $a \leq 0$ and $\widetilde{m}(A \times B) := m(B \times A)$ $(A, B \in \mathcal{B}_1)$ then by the symmetry of c we get

$$\begin{aligned}\mu_c(a\nu) &= \inf\{\mu_c(m) : T_2(-1/a)m - T_1(-1/a)m = \nu\}\\ &= \inf\{\mu_c(\widetilde{m}) : T_1(-1/a)\widetilde{m} - T_2(-1/a)\widetilde{m} = \nu\}\\ &= |a|\overset{\circ}{\mu}_c(\nu).\end{aligned}$$

Let us prove now that $\overset{\circ}{\mu}_c$ is a sub-additive function. Let ν_1, $\nu_2 \in \mathcal{M}_0$. For m_1, $m_2 \in \mathcal{M}_2$ with $T_1 m_i - T_2 m_i = \nu_i$ $(i = 1, 2)$, let $m = m_1 + m_2$. Then we have $\mu_c(m) = \mu_c(m_1) + \mu_c(m_2)$ and $T_1 m - T_2 m = \nu_1 + \nu_2$, hence, $\overset{\circ}{\mu}_c(\nu_1 + \nu_2) \leq \overset{\circ}{\mu}_c(\nu_1) + \overset{\circ}{\mu}_c(\nu_2)$. □

In the next theorem we give a sufficient condition for

$$\overset{\circ}{\mu}_c(P_1, P_2) := \overset{\circ}{\mu}_c(P_1 - P_2) \quad P_1, P_2 \in \mathcal{P}_1 \tag{4.7.44}$$

to be a simple metric in $\mathcal{P}_1$. In the proof we shall make use of *Zolotarev's semimetric* $\zeta_{\mathcal{F}}$. Namely, for a given class $\mathcal{F}$ of bounded

continuous function $f : U \to \mathbb{R}$, we define

$$\zeta_{\mathcal{F}}(P_1, P_2) = \sup_{f \in \mathcal{F}} \left| \int_U f \mathrm{d}(P_1 - P_2) \right| \quad P_i \in \mathcal{P}(U)$$

Clearly, $\zeta_{\mathcal{F}}$ is a simple semimetric. Moreover, if the class $\mathcal{F}$ is 'rich enough' to preserve the implication $\zeta_{\mathcal{F}}(P_1, P_2) = 0 \iff P_1 = P_2$, we have that $\zeta_{\mathcal{F}}$ is a simple metric.

Theorem 4.7.2.
(i) For any $c \in \mathcal{CS}(U^2)$, $\overset{\circ}{\mu}_c(P_1, P_2)$ defined by equality (4.7.44) is a semimetric in $\mathcal{P}_1$.
(ii) Further, if the class $\mathcal{F}_c := \{f : U \to \mathbb{R}, |f(x) - f(y)| \leq c(x, y)\ \forall x, y \in U\}$ contains the class $\mathcal{G}$ of all functions

$$f(x) := f_{k,C}(x) := \max\{0, 1/k - d(x, C)\} \quad x \in U$$

(k is an integer greater than some fixed k_0, C is a closed non-empty set) then $\overset{\circ}{\mu}_c$ is a simple metric in $\mathcal{P}_1$.

Proof.
(i) The proof follows immediately from Lemma 4.7.2 and the definition of semimetric (see Definition 2.3.1).
(ii) For any $m \in \mathcal{M}_2$ such that $T_1 m - T_2 m = P_1 - P_2$ and for any $f \in \mathcal{F}_c$ we have

$$\begin{aligned} \left| \int_U f \mathrm{d}(P_1 - P_2) \right| &= \left| \int_{U^2} f(x) - f(y) m(\mathrm{d}x, \mathrm{d}y) \right| \\ &\leq \int_{U^2} |f(x) - f(y)| m(\mathrm{d}x, \mathrm{d}y) \leq \mu_c(m); \end{aligned}$$

hence, the Zolatarev's metric $\zeta_{\mathcal{F}_c}(P_1, P_2)$ is a lower bound for $\overset{\circ}{\mu}_c(P_1, P_2)$. On the other hand, by assumption, $\zeta_{\mathcal{F}_c} \geq \zeta_{\mathcal{G}}$. Thus assuming $\overset{\circ}{\mu}_c(P_1, P_2) = 0$ we get $0 \leq \zeta_{\mathcal{G}}(P_1, P_2) \leq \zeta_{\mathcal{F}_c}(P_1, P_2) \leq \overset{\circ}{\mu}_c(P_1, P_2) = 0$. Next, for any closed non-empty set C we have

$$P_1(C) < k \int_U f_{k,C} \mathrm{d}P_1 \leq k\zeta_{\mathcal{G}}(P_1, P_2) + k \int_U f_{k,C} \mathrm{d}P_2 \leq P_2(C^{1/k}).$$

Letting $k \to \infty$ we get $P_1(C) \le P_2(C)$ and hence, by the symmetry, $P_1 = P_2$. □

Remark 4.7.2. Obviously $\mathcal{F}_d \supseteq \mathcal{G}$ and hence $\mathring{\mu}_d$ is a simple metric in $\mathcal{P}_1$. However, if $p > 1$ then $\mathring{\mu}_{d^p}$ is not a metric in $\mathcal{P}_1$ as is shown in the following example. Let $U = [0, 1]$, $d(x, y) = |x - y|$. Let P_1 be a law concentrated on the origin and P_2 a law concentrated on 1. For any $n = 1, 2, \dots$ consider a measure $m^{(n)} \in \mathcal{M}_2$ with total mass $m^{(n)}(U^2) = 2n + 1$ and

$$m^{(n)}\left(\left\{\frac{i}{n}, \frac{i}{n}\right\}\right) = 1 \quad i = 0, \dots, n$$
$$m^{(n)}\left(\left\{\frac{i}{n}, \frac{(i+1)}{n}\right\}\right) = 1 \quad i = 0, \dots, n-1$$

Then obviously, $T_1 m^{(n)} - T_2 m^{(n)} = P_1 - P_2$ and

$$\int_{U \times U} |x - y|^p m^{(n)}(\mathrm{d}x, \mathrm{d}y) = \sum_{i=0}^{n-1} \left(\frac{1}{n}\right)^p = n^{1-p}$$

Hence, if $p > 1$ then

$$\mathring{\mu}_d(P_1, P_2) \le \inf_{n>0} \int_{U^2} |x - y|^p m^{(n)}(\mathrm{d}x, \mathrm{d}y) = 0$$

and thus $\mathring{\mu}_{d^p}(P_1, P_2) = 0$.

Definition 4.7.1. The simple semimetric $\mathring{\mu}_c$ (see equality (4.7.44)) is said to be the *minimal norm w.r.t. the functional* μ_c.

Obviously, for any $c \in \mathcal{CS}$,

$$\mathring{\mu}_c(P_1, P_2) \le \widehat{\mu}_c(P_1, P_2) := \inf\{\mu_c(P) : P \in \mathcal{P}_2, T_i P = P_i, i = 1, 2\}$$
$$P_1, P_2 \in \mathcal{P}_1. \quad (4.7.45)$$

Hence, for each minimal metric of the type $\widehat{\mu}_c$ we can construct an estimate from below by means of $\mathring{\mu}_c$, but what is more important,

$\overset{\circ}{\mu}_c$ is a *simple semimetric even though μ_c is not a probability semidistance.* For instance, let $c_h(x,y) := d(x,y)h(\max(d(x,a),d(y,a)))$, where h is a non-decreasing non-negative continuous function on $[\alpha,\infty)$ for some $\alpha > 0$. Then, as in Theorem 4.7.2, we conclude that $\zeta_{c_h} \leq \overset{\circ}{\mu}_{c_h}$ and $\zeta_{c_h}(P_1,P_2) = 0 \Rightarrow P_1 = P_2$. Thus $\overset{\circ}{\mu}_{c_h}$ is a simple metric, while if $h(t) = t^p$, $p > 1$, then μ_{c_h} is not a p. distance. Further, we shall prove that $\overset{\circ}{\mu}$ admits the dual formula: for any laws P_1 and P_2 on a s.m.s. (U,d), with finite $\int d(x,a)h(d(x,a))(P_1+P_2)(\mathrm{d}x)$,

$$\overset{\circ}{\mu}_{c_h}(P_1,P_2) = \sup\Big\{\Big|\int_U f\mathrm{d}(P_1-P_2)\Big| : f: U \to \mathbb{R}, \\ |f(x)-f(y)| \leq c_h(x,y) \quad \forall x,y \in U\Big\}. \quad (4.7.46)$$

From equality (4.7.46) it follows that if $U = \mathbb{R}$ and $d(x,y) = |x-y|$, then $\overset{\circ}{\mu}_c$ may be represented explicitly as an average metric with weight $h(\cdot - a)$ between d.f.s

$$\overset{\circ}{\mu}_{c_h}(P_1,P_2) = \overset{\circ}{\mu}_{c_h}(F_1,F_2) := \int_{-\infty}^{\infty} |F_1(x)-F_2(x)|h(|x-a|)\mathrm{d}x \quad (4.7.47)$$

where F_i is the d.f. of P_i.

4.7.3 Examples of compound distances

In this section, we provide examples of compound distances. We also discuss the links with some of the simple distances considered in section 4.7.2 as they arise as minimal distances with respect to compound distances.

Example 4.7.10. (Average compound distances). Let (U,d) be a s.m.s. and $H \in \mathcal{H}$ (see Example 2.3.1). Then

$$\mathcal{L}_H(P) := \int_{U^2} H(d(x,y))P(\mathrm{d}x,\mathrm{d}y) \quad P \in \mathcal{P}_2 \quad (4.7.48)$$

is a compound distance with parameter K_H (see (2.3.3)), and will be called *H-average compound distance*. If $H(t) = t^p$, $p > 0$ and $p' = \min(1, 1/p)$ then

$$\mathcal{L}_p(P) := [\mathcal{L}_H(P)]^{p'} \quad P \in \mathcal{P}_2 \tag{4.7.49}$$

is a compound metric in

$$\mathcal{P}_2^{(p)} := \left\{ P \in \mathcal{P}_2 : \int_{U^2} d^p(x,a)[P(\mathrm{d}x, \mathrm{d}y) + P(\mathrm{d}y, \mathrm{d}x)] < \infty \right\}.$$

In the space

$$\mathfrak{X}^{(p)}(U) := \{X \in \mathfrak{X}(U) : Ed^p(X,a) < \infty\}$$

the metric $\mathcal{L}_p$ is the usual *p-average metric*

$$\mathcal{L}_p(X,Y) := [Ed^p(X,Y)]^{p'} \quad 0 < p < \infty. \tag{4.7.50}$$

In the limit cases $p = 0, p = \infty$ we shall define the compound metrics

$$\mathcal{L}_0(P) := P\left(\bigcup_{x \neq y} (x,y) \right) \quad P \in \mathcal{P}_2 \tag{4.7.51}$$

and

$$\mathcal{L}_\infty(P) := \inf\{\varepsilon > 0 : P(d(x,y) > \varepsilon) = 0\} \quad P \in \mathcal{P}_2 \tag{4.7.52}$$

that on $\mathfrak{X}$ have the forms

$$\mathcal{L}_0(X,Y) := EI\{X \neq Y\} = \Pr(X \neq Y) \quad X, Y \in \mathfrak{X} \tag{4.7.53}$$

and

$$\mathcal{L}_\infty(X,Y) := \operatorname{ess\,sup} d(X,Y) := \inf\{\varepsilon > 0 : \Pr(d(X,Y) > \varepsilon) = 0\}. \tag{4.7.54}$$

Example 4.7.11. (Ky Fan distance and Ky Fan metric). The Ky Fan metric **K** in $\mathfrak{X}(\mathbb{R})$ was defined by Equality (2.2.7) and we shall extend that definition considering the space $\mathcal{P}_2(U)$ for a s.m.s. (U, d). We define

the Ky Fan metric in $\mathcal{P}_2(U)$ as follows:

$$\mathbf{K}(\mathbf{P}) := \inf\{\varepsilon > 0 : P(d(x,y) > \varepsilon) < \varepsilon\} \quad P \in \mathcal{P}_2$$

and on $\mathfrak{X}(U)$ by $\mathbf{K}(X,Y) = \mathbf{K}(\mathrm{Pr}_{X,Y})$. In this way $\mathbf{K}$ takes the form of the *distance in probability* in $\mathfrak{X} = \mathfrak{X}(U)$

$$\mathbf{K}(X,Y) := \inf\{\varepsilon > 0 : \Pr(d(X,Y) > \varepsilon) < \varepsilon\} \quad X, Y \in \mathfrak{X}. \tag{4.7.55}$$

A possible extension of the metric structure of $\mathbf{K}$ is the *Ky Fan distance*:

$$\mathbf{KF}_H(P) := \inf\{\varepsilon > 0 : P(H(d(x,y)) > \varepsilon) < \varepsilon\} \tag{4.7.56}$$

for each $H \in \mathcal{H}$. It is easy to check that $\mathbf{KF}_H$ is a compound distance with parameter $\mathbb{K}_{\mathbf{KF}} := K_H$ (see (2.3.3)). A particular case of the Ky Fan distance is the *parametric family of Ky Fan metrics*

$$\mathbf{K}_\lambda(P) := \inf\{\varepsilon > 0 : P(d(x,y) > \lambda\varepsilon) < \varepsilon\}. \tag{4.7.57}$$

For each $\lambda > 0$

$$\mathbf{K}_\lambda(X,Y) := \inf\{\varepsilon > 0 : \Pr(d(X,Y) > \lambda\varepsilon) < \varepsilon\} \quad X, Y \in \mathfrak{X}$$

metrizes the convergence 'in probability' in $\mathfrak{X}(U)$, i.e.

$$\mathbf{K}_\lambda(X_n, Y) \to 0 \iff \Pr(d(X_n, Y) > \varepsilon) \to 0 \text{ for any } \varepsilon > 0.$$

In the limit cases,

$$\lim_{\lambda\to 0} \mathbf{K}_\lambda = \mathcal{L}_0 \quad \lim_{\lambda\to\infty} \lambda\mathbf{K}_\lambda = \mathcal{L}_\infty \tag{4.7.58}$$

we get, however, average compound metrics (see equalities (4.7.51)–(4.7.54)) that induce convergence, stronger than convergence in probability, i.e.,

$$\mathcal{L}_0(X_n, Y) \to 0 \underset{\not\Leftarrow}{\Rightarrow} X_n \to Y \text{ 'in probability'}$$

and

$$\mathcal{L}_\infty(X_n, Y) \to 0 \underset{\not\Leftarrow}{\Rightarrow} X_n \to Y \text{ 'in probability'}$$

Example 4.7.12. (Birnbaum–Orlicz compound distances). Let $U = \mathbb{R}$, $d(x, y) = |x - y|$. For each $p \in [0, \infty]$ consider the following compound metrics in $\mathfrak{X}(\mathbb{R})$:

$$\mathbf{\Theta}_p(X_1, X_2) := \left[\int_{-\infty}^{\infty} \tau^p(t; X_1, X_2)\mathrm{d}t\right]^{p'} \quad 0 < p < \infty\, p' := \min(1, 1/p) \tag{4.7.59}$$

$$\mathbf{\Theta}_\infty(X_1, X_2) := \sup_{t\in\mathbb{R}} \tau(t; X_1, X_2) \tag{4.7.60}$$

$$\mathbf{\Theta}_0(X_12, X_2) := \int_{-\infty}^{\infty} I\{t : \tau(t; X_1, X_2) \neq 0\}\mathrm{d}t$$

where

$$\tau(t; X_1, X_2) := \Pr(X_1 \leq t < X_2) + \Pr(X_2 \leq t < X_1). \tag{4.7.61}$$

If $H \in \mathcal{H}$ then

$$\mathbf{\Theta}_H(X_1, X_2) := \int_{-\infty}^{\infty} H(\tau(t; X_1, X_2))\mathrm{d}t \tag{4.7.62}$$

is a compound distance with $\mathbb{K}_{\mathbf{\Theta}_H} = K_H$. The functional $\mathbf{\Theta}_H$ will be called a *Birnbaum–Orlicz compound average distance* and

$$\mathbf{R}_H(X_1, X_2) := H(\mathbf{\Theta}_\infty(X_1, X_2)) = \sup_{t\in\mathbb{R}} H(\tau(t; X_1, X_2)) \tag{4.7.63}$$

will be called a *Birnbaum–Orlicz compound uniform distance*.

Each example 3.3.*i*. ($i = 1, 2, 3$) is closely related to the corresponding example 3.2.*i*. In fact, it can be demonstrated that ℓ_H is a minimal distance (see Definition 4.3.2) w.r.t. $\mathcal{L}_H$ for any convex $H \in \mathcal{H}$, i.e.,

$$\ell_H = \widehat{\mathcal{L}}_H. \tag{4.7.64}$$

Analogously, the simple metrics ℓ_p (see (4.7.23)–(4.7.26)), the Prokhorov metric $\boldsymbol{\pi}_\lambda$ (see (4.7.34)), and the Prokhorov distance $\boldsymbol{\pi}_H$

(see (4.7.36)) are minimal with respect to the $\mathcal{L}_p$-metric, Ky Fan metric $\mathbf{K}_\lambda$ and Ky Fan distance $\mathbf{KF}_H$, i.e.,

$$\ell_p = \widehat{\mathcal{L}}_p \ (p \in [0, \infty]) \quad \boldsymbol{\pi}_\lambda = \widehat{\mathbf{K}}_\lambda \ (\lambda > 0) \quad \boldsymbol{\pi}_H = \widehat{\mathbf{KF}}_H. \tag{4.7.65}$$

Finally, the Birnbaum–Orlicz metric and distance $\boldsymbol{\theta}_p$ and $\boldsymbol{\theta}_H$ (see equations (4.7.40) and (4.7.38)) and the Birnbaum–Orlicz uniform distance $\boldsymbol{\rho}_H$ (see equation (4.7.39)) are minimal with respect to their 'compound versions' $\boldsymbol{\Theta}_p$, $\boldsymbol{\Theta}_H$ and $\mathbf{R}_H$, i.e.,

$$\boldsymbol{\theta}_p = \widehat{\boldsymbol{\Theta}}_p \ (p \in [0, \infty]) \quad \boldsymbol{\theta}_H = \widehat{\boldsymbol{\Theta}}_H \quad \rho_H = \widehat{\mathbf{R}}_H. \tag{4.7.66}$$

The equalities (4.7.64) to (4.7.66) represent the main relationships between simple and compound distances (resp., metrics) and serve as a framework for TPM.

Analogous relationships exist between primary and compound distances. For example, the primary distance

$$\mathcal{M}_{H,1}(\alpha, \beta) := H(|\alpha - \beta|) \tag{4.7.67}$$

(see (4.7.2)) is a primary minimal distance (see Definition 4.2.2) w.r.t. the p. distance $H(\mathcal{L}_1)$ ($H \in \mathcal{H}$), i.e.,

$$\begin{aligned} &\mathcal{M}_{H,1}(\alpha, \beta) := \\ &\inf \left\{ H(\mathcal{L}_1(P)) : \int_{U^2} d(x, a) P(\mathrm{d}x, \mathrm{d}y) = \alpha, \int_{U^2} d(a, y) P(\mathrm{d}x, \mathrm{d}y) = \beta \right\}. \end{aligned} \tag{4.7.68}$$

4.7.4 Examples of moment functions

In this section, we provide examples of functionals which can be used to bound compound semidistances. Finally, we summarize the relationship between all bounds.

Note that, by definition, a maximal distance need not be a distance. We prove the following theorem.

Theorem 4.7.3. If (U, d) is an u.m. s.m.s. and μ is a compound distance with parameter $\mathbb{K}_\mu$ then $\check{\mu}$ is a moment function and $\mathbb{K}_{\check{\mu}} = K_\mu$.

Moreover, the following stronger version of the $\mathbf{TI}^{(4)}$ is valid:

$$\check{\mu}(P_1, P_3) \leq \mathbb{K}_\mu[\widehat{\mu}(P_1, P_2) + \check{\mu}(P_2, P_3)] \quad P_1, P_2, P_3 \in \mathcal{P}_1 \tag{4.7.69}$$

where $\widehat{\mu}$ is the minimal metric w.r.t. μ.

Proof. We shall prove inequality (4.7.69) only. For each $\varepsilon > 0$ define laws $P_{12}, P_{13} \in \mathcal{P}_2$ such that

$$T_1P_{12} = P_1 \quad T_2P_{12} = P_2 \quad T_1P_{13} = P_1 \quad T_2P_{13} = P_3$$

and

$$\widehat{\mu}(P_1, P_2) \geq \mu(P_{12}) - \varepsilon, \quad \check{\mu}(P_1, P_3) \leq \mu(P_{13}) + \varepsilon.$$

As in Theorem 4.3.1, let us define a law $Q \in \mathcal{P}_3$ (cf. (4.3.5)) having marginals $T_{12}Q = P_{12}, T_{13}Q = P_{13}$. By Definitions 2.4.1, 4.3.2 and 4.4.3 we have

$$\begin{aligned}\check{\mu}(P_1, P_3) &\leq \mu(T_{13}Q) + \varepsilon \leq \mathbb{K}_\mu[\mu(P_{12}) + \mu(P_{23})] + \varepsilon \\ &\leq \mathbb{K}_\mu[\widehat{\mu}(P_1, P_2) + \varepsilon + \check{\mu}(P_2, P_3)] + \varepsilon.\end{aligned}$$

Letting $\varepsilon \to 0$, we get equation (4.7.69). □

Definition 4.7.2. The moment functions $\check{\mu}$ will be called *a maximal distance with parameter* $\mathbb{K}_{\check{\mu}} = \mathbb{K}_\mu$ and if $\mathbb{K}_\mu = 1$, then $\check{\mu}$ will be called *maximal metric*.

As before, we note that a maximal distance (resp. metric) may fail to be distance (resp. metric). (The **ID** property may fail.)

Corollary 4.7.1. If (U, d) is an u.m. s.m.s. and μ is a compound metric on $\mathcal{P}_2$ then

$$|\check{\mu}(P_1, P_3) - \check{\mu}(P_2, P_3)| \leq \widehat{\mu}(P_1, P_2) \tag{4.7.70}$$

for all $P_1, P_2, P_3 \in \mathcal{P}_1$.

Remark 4.7.3. By the triangle inequality $\mathbf{TI}^{(4)}$ we have

$$|\check{\mu}(P_1, P_3) - \check{\mu}(P_2, P_3)| \leq \check{\mu}(P_1, P_2). \tag{4.7.71}$$

Inequality (4.7.70) thus gives us refinement of the triangle inequality for maximal metrics.

We shall further investigate the following problem, which is related to a description of the minimal and maximal distances.

Problem 4.7.1. If c is a non-negative continuous function on U^2 and

$$\mu_c(P) := \int_{U^2} c(x, y)P(\mathrm{d}x, \mathrm{d}x) \quad P \in \mathcal{P}_2 \tag{4.7.72}$$

then what are the best possible inequalities of the type

$$\phi(P_1, P_2) \leq \mu_c(P) \leq \psi(P_1, P_2) \tag{4.7.73}$$

when the marginals $T_iP = P_i$, $i = 1, 2$ are fixed?

If $c(x, y) = H(d(x, y)), H \in \mathcal{H}$ then $\mu_c = \mathcal{L}_H$ (see equation (4.7.48)) and the best possible lower and upper bounds for $\mathcal{L}_H(P)$ (with fixed $P_i = T_iP$ $(i = 1, 2)$) are given by the minimal distance $\phi(P_1, P_2) = \widehat{\mathcal{L}}(P_1, P_2)$ and the maximal distance $\psi(P_1, P_2) = \check{\mathcal{L}}_H(P_1, P_2)$.

Remark 4.7.4. In particular, for any convex non-negative function ψ on $\mathbb{R}$ and $c(x, y) = \psi(x - y)$ $(x, y \in \mathbb{R})$, the functionals of $\widehat{\mathcal{L}}_H$ and $\check{\mathcal{L}}_H$ have the following explicit forms:

$$\widehat{\mathcal{L}}_H(P_1, P_2) := \int_0^1 H(F_1^{-1}(t) - F_2^{-1}(t))\mathrm{d}t$$
$$\check{\mathcal{L}}_H(P_1, P_2) := \int_0^1 H(F_1^{-1}(t) - F_2^{-1}(1 - t))\mathrm{d}t$$

where F_i^{-1} is the generalized inverse function (4.7.28) w.r.t. the d.f. F_i.

Another example of a moment function that is an upper bound for $\mathcal{L}_H$ ($H \in \mathcal{H}$) is given by

$$\Lambda_{H,\mathbf{0}}(P_1, P_2) := K_H \int_U H(d(x, \mathbf{0}))(P_1 + P_2)(\mathrm{d}x) \tag{4.7.74}$$

where $\mathbf{0}$ is a fixed point of U. In fact, since $H \in \mathcal{H}$ then $H(d(x, y)) \leq K_H[H(d(x, \mathbf{0})) + H(d(y, \mathbf{0}))]$ for all $x, y \in U$ and hence

$$\mathcal{L}_H(P) \leq \overline{\Lambda}_{H,\mathbf{0}}(P_1, P_2). \tag{4.7.75}$$

One can easily improve inequality (4.7.75) by the following inequality

$$\mathcal{L}_H(P) \leq \overline{\Lambda}_H(P_1, P_2) := \inf_{a \in U} \overline{\lambda}_{H,a}(P_1, P_2). \tag{4.7.76}$$

The upper bounds $\overline{\Lambda}_{H,a}$, $\overline{\Lambda}_H$ of $\mathcal{L}_H$ depend on the sum $P_1 + P_2$ only. Hence, if P is an unknown law in $\mathcal{P}_2$ and we know only the sum of marginals $P_1 + P_2 = T_1P + T_2P$, then the best improvement of inequality (4.7.76) is given by

$$\mathcal{L}_H(P) \leq \mathcal{L}_H^{(s)}(P_1 + P_2) \tag{4.7.77}$$

where

$$\mathcal{L}_H^{(s)}(P_1 + P_2) := \sup\{\mathcal{L}_H(P) : T_1P + T_2P = P_1 + P_2\}. \tag{4.7.78}$$

Remark 4.7.5. Following Remark 4.4, we have that if (X, Y) is a pair of dependent U-valued r.v.s, and we know only the sum of distributions $\mathrm{Pr}_X + \mathrm{Pr}_Y$, then $\mathcal{L}_1^{(s)}(\mathrm{Pr}_X + \mathrm{Pr}_Y)$ is the best possible improvement of the triangle inequality (4.4.7). It can be demonstrated that in the particular case $U = \mathbb{R}$, $\mathrm{d}(x, y) = |x - y|$, and $p \geq 1$

$$\mathcal{L}_p^{(s)}(P_1 + P_2) = \left(\int_0^1 |V^{-1}(t) - V^{-1}(1 - t)|^p \mathrm{d}t \right)^{1/p}$$

where V^{-1} is the generalized inverse (see equation (4.7.28)) of $V(t) = \frac{1}{2}(F_1(t) + F_2(t))$, $t \in \mathbb{R}$ and F_i, is the d.f. of P_i ($i = 1, 2$). For additional details, see Rachev (1991).

For more general cases we shall use the following definition.

Definition 4.7.3. For any compound distance μ, the functional

$$\overset{(s)}{\mu}(P_1, P_2) := \sup\{\mu(P) : T_1P + T_2P = P_1 + P_2\}$$

will be called the *μ-upper bound with marginal sum fixed.*

Let us consider another possible improvement of Minkovski's inequality (4.4.3). Suppose we need to estimate from above (in the best possible way) the value $\mathcal{L}(X, Y)$ ($p > 0$), having available only the moments

$$m_p(X) := [Ed^p(X, \mathbf{0})]^{p'} \quad p' := \min(1, 1/p) \tag{4.7.79}$$

and $m_p(Y)$. Then the problem consists in evaluating the quantity

$$\psi_p(a_1, a_2) := \sup\left\{\mathcal{L}_p(P) : P \in \mathcal{P}_2(U), \left(\int_U d^p(x, \mathbf{0})T_iP(\mathrm{d}x)\right)^{p'} = a_i, i=1, 2\right\}$$
$$p' = \min(1, 1/p)$$

for each $a_i \geq 0$ and $a_2 \geq 0$.

Obviously, ψ_p is a moment function. It is possible to obtain an explicit representation of $\psi_p(a_1, a_2)$. For additional details, see Rachev (1991).

Definition 4.7.4. For any p. distance μ, the function

$$\overset{(m,p)}{\mu}(a_1, a_2) := \sup\left\{\mu(P) : P \in \mathcal{P}_2(U), \left(\int_U d^p(x, \mathbf{0})T_iP(\mathrm{d}x)\right)^{p'} = a_i, i=1, 2\right\}$$

where $a_1 \geq 0, a_2 \geq 0, p > 0$ is said to be the μ-upper bound with fixed pth marginal moments a_1 *and* a_2.

Hence, $\overset{(m,1)}{\mathcal{L}}(a_1, a_2)$ is the best possible improvement of the triangle inequality (4.4.7) when we know only the 'marginal' moments

$$a_1 = Ed(X, \mathbf{0}) \quad a_2 = Ed(Y, \mathbf{0}).$$

We shall investigate improvements of inequalities of the type

$$Ed(X, \mathbf{0}) - Ed(Y, \mathbf{0}) \leq Ed(X, Y) \leq Ed(X, \mathbf{0}) + Ed(Y, \mathbf{0})$$

for dependent r.v.s X and Y. We make the following definition.

Definition 4.7.5. For any p. distance μ,
(i) the functional

$$\underset{(m,p)}{\mu}(a_1, a_2) := \inf\left\{\mu(P) : P \in \mathcal{P}_2(U), \left[\int_U d^p(x, \mathbf{0}) T_i P(\mathrm{d}x)\right]^{p'} = a_i, i{=}1, 2\right\}$$

where $a_1 \geq 0$, $a_2 \geq 0$, $p > 0$ is said to be the *μ-lower bound with fixed marginal pth moments* a_1 and a_2;
(ii) the functional

$$\overline{\mu}(a_1 + a_2; m, p) := \sup\left\{\mu(P) : P \in \mathcal{P}_2(U), \left[\int_U d^p(x, \mathbf{0}) T_1 P(\mathrm{d}x)\right]^{p'} + \left[\int_U d^p(x, \mathbf{0}) T_2 P(\mathrm{d}x)\right]^{p'} = a_1 + a_2\right\}$$

where $a_1 \geq 0$, $a_2 \geq 0$, $p > 0$ is said to be the *μ-upper bound with fixed sum of marginal pth moments* $a_1 + a_2$;

(iii) the functional

$$\underline{\mu}(a_1 - a_2; m, p) := \inf\left\{\mu(P) : P \in \mathcal{P}_2(U), \left[\int_U d^p(x, \mathbf{0}) T_1 P(\mathrm{d}x)\right]^{p'} - \left[\int_U d^p(x, \mathbf{0}) T_2 P(\mathrm{d}x)\right]^{p'} = a_1 - a_2\right\}$$

where $a_1 \geq 0$, $a_2 \geq 0$, $p > 0$ is said to be the *μ-lower bound with fixed difference* of marginal p. moments $a_1 - a_2$.

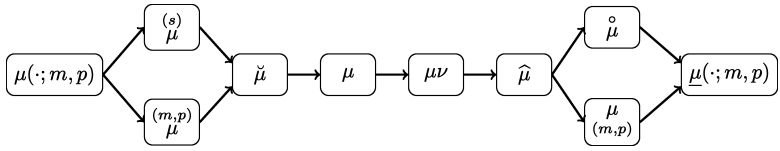

Figure 4.3: Lower and upper bounds for $\mu(P)$ ($P \in \mathcal{P}_2$) of a compound distance μ when different kinds of marginal characteristics of P are fixed. The arrow $\to$ indicates inequality of the type $\leq$.

Knowing explicit formulae for $\overset{(m,p)}{\mu}$ and $\underset{(m,p)}{\mu}$ (see Rachev (1991)), we can easily determine $\overline{\mu}(a_1 + a_2; m, p)$ and $\overline{\mu}(a_1 - a_2; m, p)$ by using the representations

$$\overline{\mu}(a; m, p) = \sup\left\{\overset{(m,p)}{\mu}(a_1, a_2) : a_1 \geq 0, a_2 \geq 0, a_1 + a_2 = a\right\}$$

and

$$\underline{\mu}(a; m, p) = \inf\left\{\underset{(m,p)}{\mu}(a_1, a_2) : a_1 \geq 0, a_2 \geq 0, a_1 - a_2 = a\right\}$$

Let us summarize the bounds for μ we have obtained up to now. For any compound distance μ (see Figure 4.3), the maximal distance $\breve{\mu}$ (see Definition 4.7.2) is not greater than the moment distance

$$\overset{(m,p)}{\mu}(a_1, a_2) := \sup\left\{\mu(P_1, P_2) : \left[\int_U d^p(x, \mathbf{0}) P_i(\mathrm{d}x)\right]^{p'} = a_i, i = 1, 2\right\}. \tag{4.7.80}$$

As we have seen, all compound distances μ can be estimated from above by means of $\breve{\mu}$, $\overset{(s)}{\mu}$, $\overset{(m,p)}{\mu}$, $\mu(\cdot; m, p)$ and in addition, the following inequality holds

$$\mu \leq \breve{\mu} \leq \overset{(s)}{\mu} \leq \overline{\mu}(\cdot; m, p), \quad \breve{\mu} \leq \overset{(m,p)}{\mu}. \tag{4.7.81}$$

The p. distance μ can be estimated from below by means of the minimal metric $\widehat{\mu}$ (see Definition 4.3.2), the co-minimal metric $\mu\nu$ (see Definition 4.3.3), the primary minimal distance $\widetilde{\mu}_h$ (see

Definition 4.2.2), as well as for such μ as $\mu = \mu_c$ (see equation (4.7.45)) by means of minimal norms $\overset{\circ}{\mu}$ (see Definition 4.7.1).

Thus

$$\underline{\mu}(\cdot; m, p) \le \widetilde{\mu}_h \le \widehat{\mu} \le \mu v \le \mu, \quad \overset{\circ}{\mu}_c \le \mu_c \tag{4.7.82}$$

and moreover, we can compute the values of $\widetilde{\mu}_h$ by using the values of the minimal distances μ, since

$$\widetilde{\mu}_h(a_1, a_2) = (\widetilde{\widehat{\mu}})_h(a_1, a_2) := \inf\{\widehat{\mu}(P_1, P_2) : hP_i = a_i, i = 1, 2\}. \tag{4.7.83}$$

Also, if $c(x, y) = H(d(x, y))$, $H \in \mathcal{H}$, then μ_c is a p. distance and

$$\overset{\circ}{\mu}_c \le \widehat{\mu}_c \le \mu. \tag{4.7.84}$$

The inequalities (4.7.80)–(4.7.84) are represented in the scheme on Figure 4.3. The double arrows are interpreted in the following way. The functional $\overset{(s)}{\mu}$ dominates $\breve{\mu}$, but $\overset{(s)}{\mu}$ and $\overset{(m,p)}{\mu}$ are not generally comparable.

As an example illustrating the list of bounds in Figure 4.3 let us consider the case $p = 1$ and $\mu(X, Y) = Ed(X, Y)$. Then for a fixed point $\mathbf{0} \in U$

$$(*) \quad \mu(a_1 + a_2; m, 1) = \sup\{Ed(X, Y) : Ed(X, \mathbf{0}) + Ed(Y, \mathbf{0}) = a_1 + a_2\}$$

$$a_1 + a_2 \ge 0 \tag{4.7.85}$$

$$(**) \quad \overset{(m,1)}{\mu}(a_1, a_2) = \sup\{Ed(X, Y) : Ed(X, \mathbf{0}) = a_1, Ed(Y, \mathbf{0}) = a_2\}$$

$$a_1 \ge 0 \quad a_2 \ge 0 \tag{4.7.86}$$

$$(***) \quad \overset{(s)}{\mu}(P_1 + P_2) = \sup\{Ed(X, Y) : \mathrm{Pr}_X + \mathrm{Pr}_Y = P_1 + P_2\}$$

$$P_1, P_2 \in \mathcal{P}_1 \tag{4.7.87}$$

$$(****) \quad \breve{\mu}(P_1, P_2) = \sup\{Ed(X, Y) : \mathrm{Pr}_X = P_1, \mathrm{Pr}_Y = P_2\}$$

$$P_1, P_2 \in \mathcal{P}_1 \tag{4.7.88}$$

and each of these functionals gives the best possible refinement of the inequality

$$Ed(X, Y) \leq Ed(X, \mathbf{0}) + Ed(Y, \mathbf{0})$$

under the respective conditions

$$(*) \qquad Ed(X, \mathbf{0}) + Ed(Y, \mathbf{0}) = a_1 + a_2$$

$$(**) \qquad Ed(X, \mathbf{0}) = a_1 \quad Ed(Y, \mathbf{0}) = a_2$$

$$(***) \qquad \Pr_X + \Pr_Y = P_1 + P_2$$

$$(****) \qquad \Pr_X = P_1, \quad Pr_Y = P_2.$$

Analogously, the functionals

(i) $$\underline{\mu}(a_1 - a_2; m, 1) = \inf\{Ed(X, Y) : Ed(X, \mathbf{0}) - Ed(Y, \mathbf{0}) = a_1 - a_2\}$$

$$a_1, a_2 \in \mathbb{R} \tag{4.7.89}$$

(ii) $$\underset{(m,1)}{\mu}(a_1, a_2) = \inf\{Ed(X, Y) : Ed(X, \mathbf{0}) = a_1, Ed(Y, \mathbf{0}) = a_2\}$$

$$a_1 \geq 0 \quad a_2 \geq 0 \tag{4.7.90}$$

(iii) $$\mathring{\mu}(P_1, P_2) = \inf\{\alpha Ed(X, Y) : \text{ for some } \alpha > 0, X \in \mathfrak{X}, Y \in \mathfrak{X}$$

$$\text{such that } \alpha(\Pr_X - \Pr_Y) = P_1 - P_2 \quad P_1, P_2 \in \mathcal{P}_1 \tag{4.7.91}$$

(iv) $$\widehat{\mu}(P_1, P_2) = \inf\{Ed(X, Y) : \Pr_X = P_1, \Pr_Y = P_2\}$$

$$P_1, P_2 \in \mathcal{P}_1 \tag{4.7.92}$$

(v) $$\mu\nu(P_1, P_2) = \inf\{Ed(X, Y) : \Pr_X = P_1, \Pr_Y = P_2, \nu(X, Y) < \alpha\}$$

$$(P_1, P_2 \in \mathcal{P}_1, \nu \text{ is a p. distance in } \mathfrak{X}(U)) \tag{4.7.93}$$

describe the best possible refinement of the inequality

$$Ed(X, Y) \geq Ed(X, \mathbf{0}) - Ed(Y, \mathbf{0})$$

under the respective conditions,

(i) $Ed(X, \mathbf{0}) - Ed(Y, \mathbf{0}) = a_1 - a_2$
(ii) $Ed(X, \mathbf{0}) = a_1 \;\; Ed(Y, \mathbf{0}) = a_2$
(iii) $\alpha(Pr_X - \mathrm{Pr}_Y) = P_1 - P_2$ for some $\alpha > 0$
(iv) $\mathrm{Pr}_X = P_1 \;\; \mathrm{Pr}_Y = P_2$
(v) $\mathrm{Pr}_X = P_1 \;\; \mathrm{Pr}_Y = P_2 \;\; \nu(X, Y) < \alpha$.

Remark 4.7.6. If $\mu(X, Y) = Ed(X, Y)$, then $\overset{\circ}{\mu} = \widehat{\mu}$, hence, in this case,

$$\widehat{\mu}(P_1, P_2) = \inf\{Ed(X, Y) : \mathrm{Pr}_X - \mathrm{Pr}_Y = P_1 - P_2\}. \tag{4.7.94}$$

Notes

1. See Chapter 2 for definitions and discussion.
2. Technically, it is said that the metric μ *metrizes* the convergence in distribution if a sequence of random variables $X_1, \ldots, X_n, \ldots$ converges in distribution to the random variable X, if and only if $\mu(X_n, X) \to 0$ as $n \to \infty$.

References

Billingsley, P. (1968), *Convergence of Probability Measures*, John Wiley, New York.

Dudley, R. M. (1976), *Probabilities and Metrics: Convergence of Laws on Metric Spaces, With a View to Statistical Testing*, Aarhus University Mathematics Institute Lecture Notes Series no. 45, Aarhus.

Dudley, R. M. (1989), *Real Analysis and Probability*, Wadsworth & Brooks-Cole, Pacific Grove, California.

Kantorovich, L. V. (1942), 'On the transfer of masses', *Dokl. Akad. Nauk. USSR* **37**, 227–229.

Kantorovich, L. V. (1948), 'On a problem of Monge', *Usp. Mat. Nauk* **3**, 225–226 (in Russian).

Kantorovich, L. V. and G. P. Akilov (1984), *Functional Analysis*, Nauka, Moscow (in Russian).

Kemperman, J. H. B (1983), 'On the role of duality in the theory of moments. Semi-infinite programming and applications' (*Lect. Notes Economic Math. Syst.*, **215**), Springer, Berlin, pp. 63–92.

Levin, V. L. and S. T. Rachev (1990), 'New duality theorems for marginal problems with some applications in stochastics' (*Lect. Notes Math.*, **1412**), Springer, Berlin, pp. 137–171.

Prokhorov, Yu. V. (1956), 'Convergence of random processes and limit theorems in probability theory', *Theory Prob. Appl.* **1**, 157–214.

Rachev, S. and L. Rüschendorf (1998), *Mass Transportation Problems, V1, V2*, Springer-Verlag, NY.

Rachev, S. T. (1991), *Probability Metrics and the Stability of Stochastic Models*, Wiley, New York.

Zolotarev, V. M. (1976), 'The effect of stability for characterization of probability distributions', *Zap. Nauchn. Sem. LOMI* **61**, 38–55 (in Russian).

Zolotarev, V. M. (1997), *Modern Theory of Summation of Random Variable*, Brill Academic Publishers, Boston.

Chapter 5

Risk and Uncertainty

The goals of this chapter are the following:

- To introduce measures of dispersion quantifying the notion of uncertainty of a random variable describing financial return.
- To describe links between measures of dispersion and probability metrics.
- To introduce the notion of risk measure and consider as examples the well-known value-at-risk (VaR) and the more general class of coherent risk measures.
- To consider links between risk measures and dispersion measures and consistency between risk measures and stochastic orders.
- To demonstrate that all deviation measures arise from probability quasi-metrics.

Notation introduced in this chapter:

Notation	*Description*
σ_X	The standard deviation of a random variable X
MAD_X	The mean absolute deviation of a random variable X
$\sigma_X^{+/-}$	The positive/negative semi-standard deviation of a random variable X

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

Notation	*Description*
$D(X)$	The dispersion measure of a random variable X
$VaR_\epsilon(X)$	The VaR of a random variable X at tail probability ϵ
$\rho(X)$	A general risk measure of a random variable X

Important terms introduced in this chapter:

Term	*Concise explanation*
dispersion measure	A statistic calculating whether there is high or low variability around the mean of the distribution.
deviation measure	A sub-additive and translation invariant dispersion measure.
risk measure	A general functional quantifying the notion of risk of a random variable describing financial return.
coherent risk measure	A risk measure which is monotonic, positively homogeneous, sub-additive and translation antivariant.

5.1 Introduction

There has been a major debate on the differences and common features of risk and uncertainty. Both notions are related but they do not coincide. Risk is often argued to be a *subjective* phenomenon involving *exposure* and *uncertainty*.[1] That is, generally, risk may arise whenever there is uncertainty.

While risk is an essential factor in all human decision making, in this chapter we consider it only in the context of investment management. In our context, exposure is identified with monetary loss. Thus, investment risk is related to the uncertain monetary loss to which a manager may expose a client. Subjectivity appears because two managers may define the same investment as having different risk – it is a question of personal predisposition.

A major activity in many financial institutions is to recognize the sources of risk, then manage and control them. This is possible only

if risk is quantified. If we can measure the risk of a portfolio, then we can identify the financial assets which constitute the main risk contributors, rebalance the portfolio, and, in this way, minimize the potential loss by minimizing the portfolio risk. Even though the recognition that risk involves exposure and uncertainty is illuminating, it appears insufficient in order for risk to be quantified. It merely shows that both uncertainty and monetary loss are essential characteristics. For example, if an asset will surely lose 30% of its value tomorrow, then it is not risky even though money will be lost. Uncertainty alone is not synonymous with risk either. If the price of an asset will certainly increase between 5% and 10% tomorrow, then there is uncertainty but no risk as there is no monetary loss. As a result, risk is qualified as an *asymmetric phenomenon* in the sense that it is related to loss only.

Concerning uncertainty, it is our assumption that it is an intrinsic feature of the future values of traded assets on the market. If we consider two time instants, the present and a future one, then the inherent uncertainty materializes as a probability distribution of future prices or returns: that is, these are random variables as of the present instant. Investment managers do not know the probabilistic law exactly but can infer it, to a degree, from the available data – they approximate the unknown law by assuming a parametric model and by calibrating its parameters. Uncertainty relates to the probable deviations from the expected price or return where the probable deviations are described by the unknown law. Therefore, a measure of uncertainty should be capable of quantifying the probable positive and negative deviations. In this aspect, any uncertainty measure is symmetric. As an extreme case, consider a variable characterized by no uncertainty whatsoever. It follows that this variable is non-random and we know its future value with certainty.

A classical example of an uncertainty measure is variance. It equals the average squared deviation from the mean of a distribution – it captures both the upside and the downside deviations from the mean of the distribution. Another measure is the standard deviation, which is the square root of the variance. It is more understandable as it is measured in the same units as the random variable. For instance, if

the random variable describes prices, then the standard deviation is measured in dollars; if the random variable describes percentage return, then the standard deviation is measured in percentage points. There are many other measures of uncertainty besides standard deviation and we will discuss them in this chapter.

Besides the essential features of risk discussed above, there are other characteristics. For example, investment risk may be *relative*. In benchmark tracking problems, it is reasonable to demand a smaller risk of the strategy relative to a benchmark: that is, smaller potential loss but relative to the loss of the benchmark. If there are multiple benchmarks, then there are multiple relative risks to take into account and strategy construction becomes a multi-dimensional, or a multi-criterion, problem.

Depending on the sources of risk, a financial institution may face *market*, *credit*, or *operational risk*.[2] Market risk describes the portfolio exposure to the moves of certain market variables. There are four standard market risk variables – equities, interest rates, exchange rates, and commodities. A financial instrument is dependent on those market factors and its price fluctuates as the underlying market factors move. Credit risk arises due to a debtor's failure to satisfy the terms of a borrowing arrangement. Operational risk is defined as the risk of loss resulting from inadequate or failed internal processes, people, and systems. Its contribution to total portfolio risk varies from firm to firm and its management falls under the responsibility of internal auditors.

Apparently, a true functional definition of investment risk is out of reach. Nevertheless, financial institutions have made a lot of effort to model it. Generally, a risk model consists of two parts. First, probabilistic models are constructed for the underlying sources of risk, such as market or credit risk factors, and the portfolio loss distribution is described by means of the probabilistic models. Second, risk is quantified by means of a *risk measure* which associates a real number to the portfolio loss distribution. It is important to recognize that both steps are crucial. Non-realistic probabilistic models may compromise the risk estimate just as an inappropriate choice for the risk measure may do.

Due to the lack of a functional definition of risk, no perfect risk measure exists. A risk measure captures only some of the characteristics of risk and, in this sense, every risk measure is incomplete. Nonetheless, we believe that it is reasonable to search for risk measures which are ideal for the particular problem under investigation.

In this chapter, we provide several examples of widely used dispersion measures that quantify the notion of uncertainty. A few of their basic features can be summarized into axioms leading to an axiomatic construction of dispersion measures and deviation measures, which are convex dispersion measures. The notion of a probability metric is related to the notion of dispersion. In fact, we demonstrate that probability metrics can be used to generate dispersion measures.

Measures of dispersion are inadequate for quantifying risk. We discuss in detail *value-at-risk* (VaR), its properties, estimation methods, and why it fails to be a true risk measure.

An axiomatic construction of risk measures is possible by setting key characteristics as axioms. We describe this approach in the section devoted to *coherent risk measures* and illustrate the defining axioms depending on whether the random variable describes return or payoff. It turns out that the "coherent" properties very much depend on the interpretation of the random variable. If a risk measure is coherent for return distributions, it may not be coherent for payoff distributions.

Finally, we stress the importance of consistency of a true risk measure with the second-order stochastic dominance as it concerns risk-averse investors.

5.2 Measures of Dispersion

Measures of dispersion can be constructed by means of different descriptive statistics. They calculate how observations in a dataset are distributed, whether there is high or low variability around the mean of the distribution. Intuitively, if we consider a non-random quantity, then it is equal to its mean with probability 1 and there is no fluctuation whatsoever around the mean.

In this section, we provide several descriptive statistics widely used in practice and we give a generalization which axiomatically describes measures of dispersion.

5.2.1 Standard deviation

Standard deviation is, perhaps, the most widely used measure of uncertainty. It is calculated as the square root of variance, which itself can be regarded as a measure of uncertainty. The standard deviation is usually denoted by σ_X,[3] where X stands for the random variable we consider:

$$\sigma_X = \sqrt{E(X - EX)^2} \tag{5.2.1}$$

in which E stands for mathematical expectation. For a discrete distribution, equation (5.2.1) changes to

$$\sigma_X = \left(\sum_{k=1}^{n} (x_k - EX)^2 p_k \right)^{1/2}$$

where $x_k,\ k = 1, \ldots, n$ are the outcomes, $p_k,\ k = 1, \ldots, n$ are the probabilities of the outcomes, and

$$EX = \sum_{k=1}^{n} x_k p_k$$

is the mathematical expectation. The standard deviation is always a non-negative number; if it is equal to zero, then the random variable is equal to its mean with probability 1 and, therefore, it is non-random. This conclusion holds for an arbitrary distribution.

In order to see why the standard deviation can measure uncertainty, consider the following simple example. Suppose that X describes the outcomes in a game in which one wins \$1 or \$3 with probabilities equal to $1/2$. The mathematical expectation of X, the expected win, is \$2,

$$EX = 1(1/2) + 3(1/2) = 2.$$

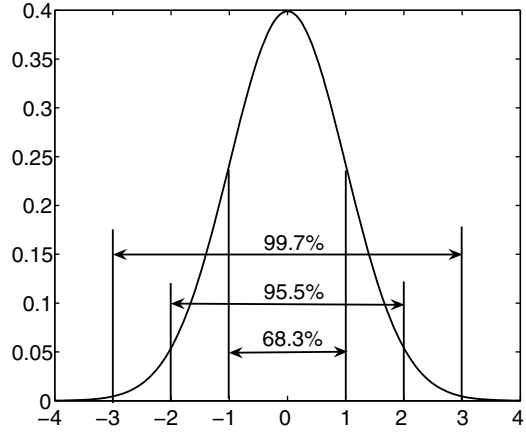

Figure 5.1: The standard normal density and the probabilities of the intervals $EX \pm \sigma_X$, $EX \pm 2\sigma_X$, and $EX \pm 3\sigma_X$, where $X \in N(0, 1)$, as a percentage of the total mass.

The standard deviation equals \$1,

$$\sigma_X = \left((1-2)^2 \frac{1}{2} + (3-2)^2 \frac{1}{2} \right)^{1/2} = 1.$$

In this equation, both the positive and the negative deviations from the mean are taken into account. In fact, all possible values of the random variable X are within the limits $EX \pm \sigma_X$. That is why it is also said that the standard deviation is a measure of statistical dispersion, i.e. how widely spread the values in a dataset are.

The interval $EX \pm \sigma_X$ covers all the possible values of X only in a few isolated examples. Suppose that X has the normal distribution with mean equal to a, $X \in N(a, \sigma_X)$. Then, the probability of the interval $a \pm \sigma_X$ is 0.683. That is, when sampling from the corresponding distribution, 68.3% of the simulations will be in the interval $(a - \sigma_X, a + \sigma_X)$. The probabilities of the intervals $a \pm 2\sigma_X$ and $a \pm 3\sigma_X$ are 0.955 and 0.997, respectively. Figure 5.1 provides an illustration for the standard normal case.

The probabilities in this example are specific for the normal distribution only. Actually, in the general case when the distribution of

Table 5.1 The values $p_k = 1 - 1/k^2$ provide a lower bound for the probability $P(X \in EX \pm k\sigma_X)$ when the distribution of X is unknown.

k	1.4	2	3	4	5	6	7
p_k	0.5	0.75	0.889	0.94	0.96	0.97	0.98

the random variable X is unknown, we can obtain bounds on the probabilities by means of *Chebyshev's inequality*,

$$P(|X - EX| > x) \leq \frac{\sigma_X^2}{x^2}, \tag{5.2.2}$$

provided that the random variable X has a finite second moment, $E|X|^2 < \infty$. With the help of Chebyshev's inequality, we calculate that the probability of the interval $EX \pm k\sigma_X$, $k = 1, 2, \ldots$ exceeds $1 - 1/k^2$,

$$P(|X - EX| \leq k\sigma_X) \geq 1 - 1/k^2.$$

If we choose $k = 2$, we compute that $P(X \in EX \pm 2\sigma_X)$ is at least 0.75. Table 5.1 contains the corresponding bounds on the probabilities computed for several choices of k.

5.2.2 Mean absolute deviation

Even though the standard deviation is widely used, it does not provide the only way to measure uncertainty. In fact, there are important cases where it is inappropriate – there are distributions for which the standard deviation is infinite. An example of an uncertainty measure also often used, which may be finite when the standard deviation does not exist, is the *mean absolute deviation* (MAD). This measure is defined as the average deviation in absolute terms around the mean of the distribution,

$$MAD_X = E|X - EX|, \tag{5.2.3}$$

where X is a random variable with finite mean. For a discrete distribution, equation (5.2.3) becomes

$$MAD_X = \sum_{k=1}^{n} |x_k - EX| p_k,$$

where $x_k,\ k = 1, \ldots, n$, are the outcomes and $p_k,\ k = 1, \ldots, n$, are the corresponding probabilities. It is clear from the definition that both the positive and the negative deviations are taken into account in the MAD formula. Similar to the standard deviation, the MAD is a non-negative number and if it is equal to zero, then X is equal to its mean with probability 1.

The analysis made for the standard deviation can be repeated for the MAD without any modification. Therefore, the MAD and the standard deviation are merely two alternative measures estimating the uncertainty of a random variable. There are distributions, for which one of the quantities can be expressed from the other. For example, if X has a normal distribution, $X \in N(a, \sigma_X^2)$, then

$$MAD_X = \sigma_X \sqrt{\frac{2}{\pi}}.$$

Thus, for the normal distribution case, the MAD is just a scaled standard deviation.

5.2.3 Semi-standard deviation

The *semi-standard deviation* is a measure of dispersion which differs from the standard deviation and the MAD in that it takes into account only the positive or only the negative deviations from the mean. Therefore, it is not symmetric. The positive and the negative semi-standard deviations are defined as,

$$\begin{aligned} \sigma_X^+ &= (E(X - EX)_+^2)^{1/2} \\ \sigma_X^- &= (E(X - EX)_-^2)^{1/2} \end{aligned} \tag{5.2.4}$$

where

$(x - EX)_+^2$ equals the squared difference between the outcome x and the mean EX if the difference is positive, $(x - EX)_+^2 = \max(x - EX, 0)^2$

$(x - EX)_-^2$ equals the squared difference between the outcome x and the mean EX if the difference is negative, $(x - EX)_-^2 = \min(x - EX, 0)^2$.

Thus, σ_X^+ takes into account only the positive deviations from the mean and it may be called an *upside dispersion measure*. Similarly, σ_X^- takes into account only the negative deviations from the mean and it may be called a *downside dispersion measure*.

As with the standard deviation, both σ_X^- and σ_X^+ are non-negative numbers which are equal to zero if and only if the random variable equals its mean with probability 1.

If the random variable is symmetric around the mean, then the upside and the downside semi-standard deviations are equal. For example, if X has a normal distribution, $X \in N(a, \sigma_X^2)$, then both quantities are equal and can be expressed by means of the standard deviation,

$$\sigma_X^- = \sigma_X^+ = \frac{\sigma_X}{\sqrt{2}}.$$

If the distribution of X is skewed,[4] then $\sigma_X^- \neq \sigma_X^+$. Positive skewness corresponds to larger positive semi-standard deviation, $\sigma_X^- < \sigma_X^+$. Similarly, negative skewness corresponds to larger negative semi-standard deviation, $\sigma_X^- > \sigma_X^+$.

5.2.4 Axiomatic description

Besides the examples considered in section 5.2, measures of dispersion also include *interquartile range* and can be based on *central absolute moments*. The interquartile range is defined as the difference between the 75% and the 25% quantile. The central absolute moment of order k is defined as

$$m_k = E|X - EX|^k$$

and an example of a dispersion measure based on it is

$$(m_k)^{1/k} = (E|X - EX|^k)^{1/k}.$$

The common properties of the dispersion measures we have considered can be synthesized into axioms. In this way, a dispersion measure is called any functional satisfying the axioms. Rachev et al. (2008) provide the following set of general axioms. We denote the dispersion measure of a random variable X by $D(X)$.

Positive shift	$D(X + C) \leq D(X)$ for all X and constants $C \geq 0$.
Positive homogeneity	$D(0) = 0$ and $D(\lambda X) = \lambda D(X)$ for all X and all $\lambda > 0$.
Positivity	$D(X) \geq 0$ for all X, with $D(X) > 0$ for non-constant X.

According to the positive shift property, adding a positive constant does not increase the dispersion of a random variable. According to the positive homogeneity and the positivity properties, the dispersion measure D is equal to zero only if the random variable is a constant. This property is very natural for any measure of dispersion. Recall that it holds for the standard deviation, MAD, and semi-standard deviation – all examples we considered in the previous sections.

An example of a dispersion measure satisfying these properties is the *colog measure* defined by

$$\text{colog}(X) = E(X \log X) - E(X)E(\log X).$$

where X is a positive random variable. The colog measure is sensitive to additive shifts and has applications in finance as it is consistent with the preference relations of risk-averse investors.

5.2.5 Deviation measures

Rockafellar et al. (2006) provide an axiomatic description of dispersion measures which arises as a special case of our approach in

section 5.2.4. The axioms of Rockafellar et al. (2006) define convex dispersion measures called *deviation measures*. An interesting link exists between deviation measures and risk measures, which we illustrate in section 5.5 in this chapter.

Besides the axioms given in section 5.2.4, the deviation measures satisfy the property

Sub-additivity $D(X + Y) \leq D(X) + D(Y)$ for all X and Y.

and the positive shift property is replaced by

Translation invariance $D(X + C) = D(X)$ for all X and constants $C \in \mathbb{R}$.

As a consequence of the translation invariance axiom, the deviation measure is influenced only by the difference $X - EX$. If $X = EX$ in all states of the world, then the deviation measure is a constant and, therefore, it is equal to zero because of the positivity axiom. Conversely, if $D(X) = 0$, then $X = EX$ in all states of the world. The positive homogeneity and the sub-additivity axioms establish the convexity property of $D(X)$.

Apparently not all deviation measures are symmetric: that is, it is possible to have $D(X) \neq D(-X)$ if the random variable X is not symmetric. This is not a drawback of the construction. Quite the opposite: this is an advantage because an investment manager is more attentive to the negative deviations from the mean. Examples of asymmetric deviation measures include the semi-standard deviation, σ_X^- defined in equation (5.2.4). Deviation measures which depend only on the negative deviations from the mean are called downside deviation measures. As a matter of fact, symmetric deviation measures can easily be constructed. The quantity $\tilde{D}(X)$ is a symmetric deviation measure if we define it as

$$\tilde{D}(X) := \frac{1}{2}(D(X) + D(-X)),$$

where $D(X)$ is an arbitrary deviation measure.

A downside deviation measure possesses several of the characteristics of a risk measure but it is not a risk measure. Here is an example. Suppose that we have initially in our portfolio a common stock, X, with a current market value of \$95 and an expected return of 0.5% in a month. Let us choose one particular deviation measure, D_1, and compute $D_1(r_X) = 20\%$, where r_X stands for the portfolio return. Assume that we add to our portfolio a risk-free government bond, B, worth \$95 with a face value of \$100 and a one-month maturity. The return on the bond equals $r_B = \$5/\$95 = 5.26\%$ and is non-random. Our portfolio now consists of equal dollar amounts of the common stock and the bond and its return equals $r_p = r_X/2 + r_B/2$. Using the positive homogeneity and the translation invariance axioms from the definition, we obtain $D_1(r_p) = D_1(r_X)/2 = 10\%$. Indeed, the uncertainty of the portfolio return r_p decreases twice, since the share of the risky stock decreases twice – this is what the deviation measure is informing us about. Intuitively, the risk of r_p decreases more than twice if compared to r_X because half of the new portfolio earns a sure profit of 5.26%. This effect is due to the translation invariance which makes the deviation measure insensitive to non-random profit.

Examples of deviation measures include the standard deviation, the MAD, and the semi-standard deviation.

5.3 Probability Metrics and Dispersion Measures

Probability metrics were introduced in Chapter 2. They are functionals which are constructed to measure distances between random quantities. Thus, every probability metric involves two random variables X and Y, and the distance between them is denoted by $\mu(X, Y)$ where μ stands for the probability metric.

Suppose that μ is a compound probability metric.[5] In this case, if $\mu(X, Y) = 0$, it follows that the two random variables are coincident in all states of the world. Therefore, the quantity $\mu(X, Y)$ can be interpreted as a measure of relative deviation between X and Y. A positive distance, $\mu(X, Y) > 0$, means that the two variables fluctuate

with respect to each other, and zero distance, $\mu(X, Y) = 0$, implies that there is no deviation of any of them relative to the other.

This idea is closely related to the notion of dispersion but it is much more profound because we obtain the notion of dispersion measures as a special case by considering the distance between X and the mean of X, $\mu(X, EX)$. In fact, the functional $\mu(X, EX)$ provides a very general notion of a dispersion measure as it arises as a special case from a probability metric which represents the only general way of measuring distances between random quantities. In the appendix to this chapter, we demonstrate how the family of symmetric deviation measures arises from probability metrics. Stoyanov et al. (2008) consider similar questions and provide a more general treatment.

5.4 Measures of Risk

As we noted in the introduction, risk is related to uncertainty but it is not synonymous with it. Therefore, a risk measure may share some of the features of a dispersion measure but is, generally, a different object.

From a historical perspective, Markowitz (1952) was the first to recognize the relationship between risk and reward and introduced standard deviation as a proxy for risk. The standard deviation is not a good choice for a risk measure because it penalizes symmetrically both the negative and the positive deviations from the mean. It is an uncertainty measure and cannot account for the asymmetric nature of risk: that is, risk concerns losses only. The deficiencies of the standard deviation as a risk measure were acknowledged by Markowitz, who was the first to suggest the semi-standard deviation as a substitute (Markowitz, 1959). In section 5.2.5, we gave an example illustrating why the semi-standard deviation, as well as any other deviation measure, cannot be a true risk measures.

In this section, we provide several examples of risk measures. We consider VaR, and we comment on its properties and different calculation methods. Where possible, the definitions and equations

are geometrically interpreted, making the ideas more intuitive and understandable. We also consider the more general family of coherent risk measures, which includes the *average value-at-risk* (AVaR) and the *spectral risk measures* as particular representatives. Finally, we address the question of consistency of a risk measure with a stochastic dominance order and remark on the relationship between risk measures and uncertainty measures.

5.4.1 Value-at-risk

A risk measure which has been widely accepted since the 1990s is value-at-risk (VaR). In the late 1980s, it was integrated by JP Morgan on a firmwide level into its risk-management system. In this system, they developed a service called RiskMetrics which was later spun off into a separate company called the RiskMetrics Group. It is usually thought that JP Morgan invented the VaR measure. In fact, similar ideas had been used by large financial institutions in computing their exposure to market risk. The contribution of JP Morgan was that the notion of VaR was introduced to a wider audience.

In the mid-1990s, the VaR measure was approved by regulators as a valid approach to calculating capital reserves needed to cover market risk. The Basel Committee on Banking Supervision released several amendments to the requirements for banking institutions allowing them to use their own internal systems for risk estimation. In this way, capital reserves, which financial institutions are required to keep, could be based on the VaR numbers computed by an in-house risk management system. Generally, regulators demand that the capital reserve equal the VaR number multiplied by a factor between 3 and 4. Thus, regulators link the capital reserves for market risk directly to the risk measure.

VaR is defined as the minimum level of loss at a given, sufficiently high, confidence level for a predefined time horizon. The recommended confidence levels are 95% and 99%. Suppose that we hold a portfolio with a 1-day 99% VaR equal to \$1 million. This means that over the horizon of 1 day, the portfolio may lose more than \$1 million with probability equal to 1%.

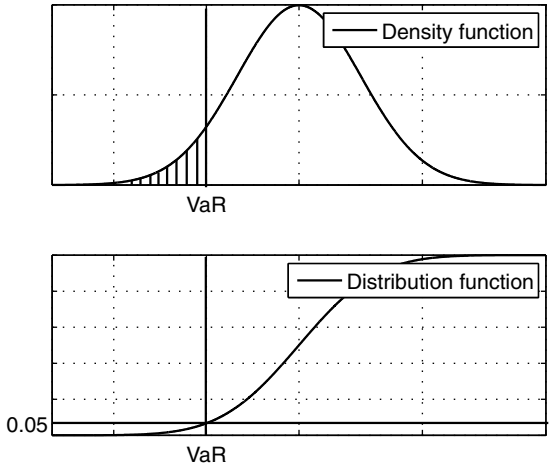

Figure 5.2: VaR at 95% confidence level of a random variable X. The top plot shows the density of X, the marked area equals the tail probability, and the bottom plot shows the distribution function.

The same example can be constructed for percentage returns. Suppose that the current value of a portfolio we hold is \$10 million. If the 1-day 99% VaR of the return distribution is 2%, then over the time horizon of 1 day, we lose more than 2% (\$200,000) of the portfolio present value with probability equal to 1%.

Denote by $(1-\epsilon)\,100\%$ the confidence level parameter of the VaR. As we explained, losses larger than the VaR occur with probability ϵ. The probability ϵ, we call *tail probability*. Depending on the interpretation of the random variable, VaR can be defined in different ways. Formally, VaR at confidence level $(1-\epsilon)\,100\%$ (tail probability ϵ) is defined as the negative of the lower ϵ-quantile of the return distribution,

$$VaR_\epsilon(X) = -\inf_x\{x|P(X \le x) \ge \epsilon\} = -F_X^{-1}(\epsilon) \tag{5.4.1}$$

where $\epsilon \in (0,1)$ and $F_X^{-1}(\epsilon)$ is the inverse of the distribution function. If the random variable X describes random returns, then the VaR number is given in terms of a return figure. The definition of VaR is illustrated in Figure 5.2.

If X describes random payoffs, then VaR is a threshold in dollar terms below which the portfolio value falls with probability ϵ,

$$VaR_\epsilon(X) = \inf_x\{x|P(X \leq x) \geq \epsilon\} = F_X^{-1}(\epsilon) \tag{5.4.2}$$

where $\epsilon \in (0, 1)$ and $F_X^{-1}(\epsilon)$ is the inverse of the distribution function of the random payoff. VaR can also be expressed as a distance to the current value when considering the profit distribution. The random profit is defined as $X - P_0$ where X is the payoff and P_0 is the current value. VaR of the random profit equals

$$VaR_\epsilon(X - P_0) = -\inf_x\{x|P(X - P_0 \leq x) \geq \epsilon\} = P_0 - VaR_\epsilon(X)$$

in which $VaR_\epsilon(X)$ is defined according to (5.4.2), since X is interpreted as a random payoff. In this case, the definition of VaR is essentially given by equation (5.4.1).

According to the definition in equation (5.4.1), VaR may become a negative number. If $VaR_\epsilon(X)$ is a negative number, then this means that at tail probability ϵ we do not observe losses but profits. Losses happen with even smaller probability than ϵ. If for any tail probability $VaR_\epsilon(X)$ is a negative number, then no losses can occur and, therefore, the random variable X bears no risk as no exposure is associated with it. In this chapter, we assume that random variables describe either random returns or random profits and we adopt the definition in equation (5.4.1).

We illustrate one aspect in which VaR differs from the deviation measures and all uncertainty measures. As a consequence of the definition, if we add to the random variable X a non-random profit C, the resulting VaR can be expressed by VaR of the initial variable in the following way:

$$VaR_\epsilon(X + C) = VaR_\epsilon(X) - C. \tag{5.4.3}$$

Thus, adding a non-random profit decreases the risk of the portfolio. Furthermore, scaling the return distribution by a positive constant λ scales the VaR by the same constant,

$$VaR_\epsilon(\lambda X) = \lambda VaR_\epsilon(X). \tag{5.4.4}$$

It turns out that these properties characterize not only VaR. They are identified as key features of a risk measure. We will come back to them in section 5.4.4.

Consider again the example developed in section 5.2.5. Initially, the portfolio we hold consists of a common stock with random monthly return r_X. We rebalance the portfolio so that it becomes an equally weighted portfolio of the stock and a bond with a non-random monthly return of 5.26%, $r_B = 5.26\%$. Thus, the portfolio return can be expressed as

$$r_p = r_X(1/2) + r_B(1/2) = r_X/2 + 0.0526/2.$$

Using equations (5.4.3) and (5.4.4), we calculate that if $VaR_\epsilon(r_X) = 12\%$, then $VaR_\epsilon(r_p) \approx 3.365\%$ which is by far less than 6% – half of the initial risk. Recall from section 5.2.5 that any deviation measure would indicate that the dispersion (or the uncertainty) of the portfolio return r_p would be twice as small as the uncertainty of r_X.

A very important remark has to be made with respect to the performance of VaR and, as it turns out, of any other risk measure. It is heavily dependent on the assumed probability distribution of the variable X. An unrealistic hypothesis may result in the underestimation or overestimation of true risk. If we use VaR to build reserves in order to cover losses in times of crises, then underestimation may be fatal and overestimation may lead to inefficient use of capital. An inaccurate model is even more dangerous in an optimal portfolio problem in which we minimize risk subject to some constraints, as it may adversely influence the optimal weights and therefore not reduce the true risk.

Even though VaR has been largely adopted by financial institutions and approved by regulators, it turns out that VaR has important deficiencies. While it provides an intuitive description of how much a portfolio may lose, generally, it should be abandoned as a risk measure. The most important drawback is that, in some cases, the reasonable diversification effect that every portfolio manager should expect to see in a risk measure is not present: that is, a portfolio's VaR

may be greater than the sum of the VaRs of the constituents,

$$VaR_\epsilon(X+Y) > VaR_\epsilon(X) + VaR_\epsilon(Y), \tag{5.4.5}$$

in which X and Y stand for the random payoff of the instruments in the portfolio. This shows that VaR cannot be a true risk measure.

We give a simple example which shows that VaR may satisfy (5.4.5). Suppose that X denotes a bond which either defaults with probability 4.5% and we lose \$50 or does not default, in which case the loss is equal to zero. Let Y be the same bond but assume that the defaults of the two bonds are independent events. The VaR of the two bonds at 95% confidence level (5% tail probability) is equal to zero:

$$VaR_{0.05}(X) = VaR_{0.05}(Y) = 0.$$

Being the 5% quantile of the payoff distribution in this case, VaR fails to recognize losses occurring with probability smaller than 5%. A portfolio of the two bonds has the following payoff profile: it loses \$100 with probability of about 0.2%, loses \$50 with probability of about 8.6%, and the loss is zero with probability 91.2%. Thus, the corresponding 95% VaR of the portfolio equals \$50 and clearly,

$$\$50 = VaR_{0.05}(X+Y) > VaR_{0.05}(X) + VaR_{0.05}(Y) = 0.$$

What are the consequences of using a risk measure which may satisfy property (5.4.5)? It is going to mislead portfolio managers that there is no diversification effect in the portfolio and they may make the irrational decision to concentrate it only into a few positions. As a consequence, the portfolio risk actually increases.

Besides being sometimes incapable of recognizing the diversification effect, another drawback is that VaR is not very informative about losses beyond the VaR level. It only reports that losses larger than the VaR level occur with probability equal to ϵ but it does not provide any information about the likely magnitude of such losses, for example.

Nonetheless, VaR is not a useless concept to be abandoned altogether. For example, it can be used in risk-reporting only as a

characteristic of the portfolio return (payoff) distribution, since it has a straightforward interpretation. The criticism of VaR is focused on its wide application by practitioners as a true risk measure which, in view of the deficiencies described above, is not well grounded and should be reconsidered.

5.4.2 Computing portfolio VaR in practice

In this section, we provide three approaches for portfolio VaR calculation which are used in practice. We assume that the portfolio contains common stocks, which is only to make the description easier to grasp; this is not a restriction of any of the approaches.

Suppose that a portfolio contains n common stocks and we are interested in calculating the daily VaR at 99% confidence level. Denote the random daily returns of the stocks by $X_1, \ldots, X_n$ and by $w_1, \ldots, w_n$ the weight of each stock in the portfolio. Thus, the portfolio return r_p can be calculated as

$$r_p = w_1 X_1 + w_2 X_2 + \ldots + w_n X_n.$$

The portfolio VaR is derived from the distribution of r_p. The three approaches vary in the assumptions they make.

The approach of RiskMetrics

The approach of the RiskMetrics Group is centered on the assumption that the stock returns have a multivariate normal distribution. Under this assumption, the distribution of the portfolio return is also normal. Therefore, in order to calculate the portfolio VaR, we only have to calculate the expected return of r_p and the standard deviation of r_p. The 99% VaR will appear as the negative of the 1% quantile of the $N(Er_p, \sigma_{r_p}^2)$ distribution.

The portfolio expected return can be directly expressed through the expected returns of the stocks,

$$Er_p = w_1 EX_1 + w_2 EX_2 + \ldots + w_n EX_n = \sum_{k=1}^{n} w_k EX_k, \tag{5.4.6}$$

where E denotes mathematical expectation. Similarly, the variance of the portfolio return $\sigma^2_{r_p}$ can be computed through the variances of the stock returns and their covariances,

$$\sigma^2_{r_p} = w_1^2\sigma^2_{X_1} + w_2^2\sigma^2_{X_2} + \ldots + w_n^2\sigma^2_{X_n} + \sum_{i \neq j} w_i w_j \text{cov}(X_i, X_j),$$

in which the last term appears because we have to sum up the covariances between all pairs of stock returns. There is a more compact way of writing down the expression for $\sigma^2_{r_p}$ using matrix notation,

$$\sigma^2_{r_p} = w'\Sigma w, \tag{5.4.7}$$

in which $w = (w_1, \ldots, w_n)$ is the vector of portfolio weights and Σ is the covariance matrix of stock returns,

$$\Sigma = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma^2_{X_2} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma^2_{X_n} \end{pmatrix},$$

in which σ_{ij}, $i \neq j$, is the covariance between X_i and X_j, $\sigma_{ij} = \text{cov}(X_i, X_j)$. As a result, we obtain that the portfolio return has a normal distribution with mean given by equation (5.4.6) and variance given by equation (5.4.7).

The standard deviation is the scale parameter of the normal distribution and the mean is the location parameter. Due to the normal distribution properties, if $r_p \in N(Er_p, \sigma^2_{r_p})$, then

$$\frac{r_p - Er_p}{\sigma_{r_p}} \in N(0, 1).$$

Thus, because of the properties (5.4.3) and (5.4.4) of the VaR, the 99% portfolio VaR can be represented as

$$VaR_{0.01}(r_p) = q_{0.99}\sigma_{r_p} - Er_p \tag{5.4.8}$$

where the standard deviation of the portfolio return σ_{r_p} is computed from equation (5.4.7), the expected portfolio return Er_p is given in (5.4.6), and $q_{0.99}$ is the 99% quantile of the standard normal distribution.

Note that $q_{0.99}$ is a quantity independent of the portfolio composition; it is merely a constant which can be calculated in advance. The parameters which depend on the portfolio weights are the standard deviation of portfolio returns σ_{r_p} and the expected portfolio return. As a consequence, VaR under the assumption of normality is symmetric even though, by definition, VaR is centered on the left tail of the distribution: that is, VaR is asymmetric by construction. This result appears because the normal distribution is symmetric around the mean.

The approach of RiskMetrics can be extended for other types of distributions. Lamantia et al. (2006*a*) and Lamantia et al. (2006*b*) provide such extensions and comparisons for the Student's t and stable distributions.

The historical method

The *historical method* does not impose any distributional assumptions; the distribution of portfolio returns is constructed from historical data. Hence, sometimes the historical simulation method is called a *non-parametric method*. For example, the 99% daily VaR of the portfolio return is computed as the negative of the empirical 1% quantile of the observed daily portfolio returns. The observations are collected from a predetermined time window such as the most recent business year.

While the historical method seems to be more general as it is free of any distributional hypotheses, it has a number of major drawbacks.

(a) It assumes that the past trends will continue in the future. This is not a realistic assumption because extreme events may be experienced in the future, for instance, which have not happened in the past.
(b) It treats the observations as independent and identically distributed (i.i.d.), which is not realistic. The daily returns data

exhibit clustering of the volatility phenomenon, autocorrelations and so on, which are sometimes a significant deviation from the i.i.d. assumption.

(c) It is not reliable for estimation of VaR at very high confidence levels. A sample of one year of daily data contains 250 observations, which is a rather small sample for the purpose of the 99% VaR estimation.

The hybrid method

The *hybrid method* is a modification of the historical method in which the observations are not regarded as i.i.d. but certain weights are assigned to them depending on how close they are to the present. The weights are determined using the *exponential smoothing* algorithm. The exponential smoothing accentuates the most recent observations and seeks to take into account time-varying volatility phenomena.

The algorithm of the hybrid approach consists of the following steps.

(a) Exponentially declining weights are attached to historical returns, starting from the current time and going back in time. Let $r_{t-k+1}, \ldots, r_{t-1}, r_t$ be a sequence of k observed returns on a given asset, where t is the current time. The i-th observation is assigned a weight

$$\theta_i = c^* \lambda^{t-i},$$

where $0 < \lambda < 1$, and $c = \frac{1-\lambda}{1-\lambda^k}$ is a constant chosen such that the sum of all weights is equal to one, $\sum \theta_i = 1$.

(b) Similarly to the historical simulation method, the hypothetical future returns are obtained from the past returns and sorted in increasing order.

(c) The VaR measure is computed from the empirical cumulative distribution function (c.d.f.), in which each observation has a probability equal to the weight θ_i.

Generally, the hybrid approach is appropriate for VaR estimation of heavy-tailed time series. It overcomes, to some degree, the first

and the second deficiency of the historical method but it is also not reliable for VaR estimation of very high confidence levels.

The Monte Carlo method

In contrast to the historical method, the *Monte Carlo method* requires specification of a statistical model for the stock returns. The statistical model is multivariate, hypothesizing both the behavior of the stock returns on a stand-alone basis and their dependence. For instance, the multivariate normal distribution assumes normal distributions for the stock returns viewed on a stand-alone basis and describes the dependencies by means of the covariance matrix. The multivariate model can also be constructed by specifying explicitly the one-dimensional distributions of the stock returns, and their dependence through a copula function.

The Monte Carlo method consists of the following basic steps:

Step 1. *Selection of a statistical model.* The statistical model should be capable of explaining a number of observed phenomena in the data, such as heavy tails, clustering of the volatility, etc., which we think influence the portfolio risk.

Step 2. *Estimation of the statistical model parameters.* A sample of observed stocks returns is used from a predetermined time window: for instance, the most recent 250 daily returns.

Step 3. *Generation of scenarios from the fitted model.* Independent scenarios are drawn from the fitted model. Each scenario is a vector of stock returns which depend on each other according to the presumed dependence structure of the statistical model.

Step 4. *Calculation of portfolio risk.* Compute portfolio risk on the basis of the portfolio return scenarios obtained from the previous step.

The Monte Carlo method is a very general numerical approach to risk estimation. It does not require any closed-form expressions and, by choosing a flexible statistical model, accurate risk numbers can be obtained. A disadvantage is that the computed portfolio VaR is

Table 5.2 The 99% VaR of the standard normal distribution computed from a sample of scenarios. The 95% confidence interval is calculated from 100 repetitions of the experiment. The true value is $VaR_{0.01}(X) = 2.326$.

Number of scenarios	*99% VaR*	*95% confidence interval*
500	2.067	[1.7515, 2.3825]
1,000	2.406	[2.1455, 2.6665]
5,000	2.286	[2.1875, 2.3845]
10,000	2.297	[2.2261, 2.3682]
20,000	2.282	[2.2305, 2.3335]
50,000	2.342	[2.3085, 2.3755]
100,000	2.314	[2.2925, 2.3355]

dependent on the generated sample of scenarios and will fluctuate a little if we regenerate the sample. This side effect can be reduced by generating a larger sample. An illustration is provided in the following example.

Suppose that the daily portfolio return distribution is standard normal and, therefore, at Step 4 of the algorithm we have scenarios from the standard normal distribution. Under the assumption of normality, we can use the approach of RiskMetrics and compute the 99% daily VaR directly from formula (5.4.8). Nevertheless, we will use the Monte Carlo method to gain more insight into the deviations of the VaR based on scenarios from the VaR computed according to formula (5.4.8).

In order to investigate how the fluctuations of the 99% VaR change about the theoretical value, we generate samples of different sizes: 500, 1,000, 5,000, 10,000, 20,000, 50,000, and 100,000 scenarios. The 99% VaR is computed from these samples and the numbers are stored. We repeat the experiment 100 times. In the end, we have 100 VaR numbers for each sample size. We expect that as the sample size increases, the VaR values will fluctuate less about the theoretical value, which is $VaR_{0.01}(X) = 2.326$, $X \in N(0, 1)$.

Table 5.2 contains the result of the experiment. From the 100 VaR numbers, we calculate the 95% confidence interval for the true value

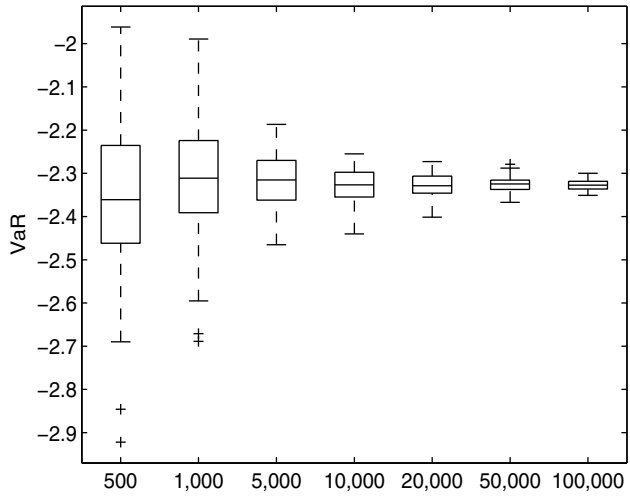

Figure 5.3: Boxplot diagrams of the fluctuation of the 99% VaR of the standard normal distribution based on scenarios. The horizontal axis shows the number of scenarios and the boxplots are computed from 100 independent samples.

given in the third column. The confidence intervals cover the theoretical value 2.326 and also we notice that the length of the confidence interval decreases as the sample size increases. This effect is best illustrated with the help of the boxplot diagrams[6] shown in Figure 5.3. A sample of 100,000 scenarios results in VaR numbers which are tightly packed around the true value, while a sample of only 500 scenarios may give a very inaccurate estimate.

This simple experiment shows that the number of scenarios in the Monte Carlo method has to be carefully chosen. The approach we used to determine the fluctuations of the VaR based on scenarios is a statistical method called *parametric bootstrap*. Bootstrap methods in general are powerful statistical methods which are used to compute confidence intervals when the problem is not analytically tractable but the calculations may be quite computationally intensive.

The true merits of the Monte Carlo method can only be realized when the portfolio contains complicated instruments such as

derivatives. In this case, it is no longer possible to use a closed-form expression for the portfolio VaR (and any risk measure in general) because the distribution of the portfolio return (or payoff) becomes quite arbitrary. The Monte Carlo method provides the general framework to generate scenarios for the risk-driving factors, then revaluates the financial instruments in the portfolio under each scenario, and, finally, estimates portfolio risk on the basis of the computed portfolio returns (or payoffs) in each state of the world.

While it may seem a straightforward approach, the practical implementation is a very challenging endeavor from both software development and financial modeling points of view. The portfolios of large financial institutions often contain products which require yield curve modeling, development of fundamental and statistical factor models, and, on top of that, a probabilistic model capable of describing the heavy tails of the risk-driving factor returns, the autocorrelation, clustering of the volatility, and the dependence between these factors. Processing large portfolios is related to manipulation of colossal data structures, which requires excellent skills of software developers in order to be efficiently performed.

5.4.3 Back-testing of VaR

If we adopt VaR for analysis of portfolio exposure, then a reasonable question is whether the VaR calculated according to any of the methods discussed in the previous section is realistic. Suppose that we calculate the 99% daily portfolio VaR. This means that according to our assumption for the portfolio return (payoff) distribution, the portfolio loses more than the 99% daily VaR with 1% probability. The question is whether this estimate is correct: that is, does the portfolio really lose more than this amount with 1% probability? This question can be answered by back-testing of VaR.

Generally, the procedure consists of the following steps.

Step 1. Choose a time window for the back-testing. Usually the time window is the most recent one or two years.

Step 2. For each day in the time window, calculate the VaR number.

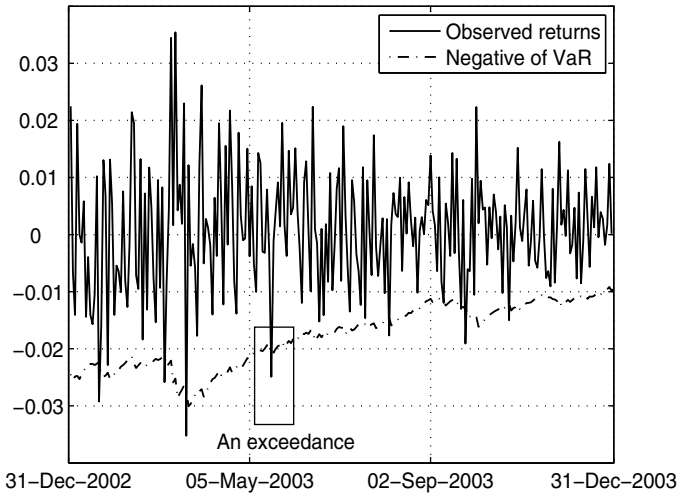

Figure 5.4: The observed daily returns of S&P 500 index between December 31, 2002 and December 31, 2003 and the negative of VaR. The marked observation is an example of an exceedance.

Step 3. Check if the loss on a given day is below or above the VaR number computed the day before. If the observed loss is larger, then we say that there is a case of an *exceedance*. Figure 5.4 provides an example.

Step 4. Count the number of exceedances. Check if there are too many or too few of them by verifying if the number of exceedances belong to the corresponding 95% confidence interval.

If in Step 4 we find out that there are too large a number of exceedances, then the VaR numbers produced by the model are too optimistic. Losses exceeding the corresponding VaR happen too frequently. If capital reserves are determined on the basis of VaR, then there is a risk of being incapable of covering large losses. Conversely, if we find out that there are too small a number of exceedances, then the VaR numbers are too pessimistic. This is also an undesirable

situation. Note that the actual size of the exceedances is immaterial, we only count them.

The confidence interval for the number of exceedances is constructed on the basis of the indicator-type events "we observe an exceedance," "we do not observe an exceedance" on a given day. If we consider the 99% VaR, then the probability of the first event, according to the model, is 1%. Let us associate a number with each of the events similar to a coin-tossing experiment. If we observe an exceedance on a given day, then we say that the number 1 has occurred, otherwise 0 has occurred. If the back-testing time window is two years, then we have a sequence of 500 zeros and ones and the expected number of exceedances is 5. Thus, finding the 95% confidence interval for the number of exceedances reduces to finding an interval around 5, such that the probability of the number of ones belonging to this interval is 95%.

If we assume that the corresponding events are independent, then there is a complete analogue of this problem in terms of coin tossing. We toss independently 500 times an unfair coin with probability of success equal to 1%. What is the range of the number of success events with 95% probability? In order to find the 95% confidence interval, we can resort to the normal approximation to the binomial distribution. The formula is

$$\text{left bound} = N\epsilon - F^{-1}(1 - 0.05/2)\sqrt{N\epsilon(1-\epsilon)}$$

$$\text{right bound} = N\epsilon + F^{-1}(1 - 0.05/2)\sqrt{N\epsilon(1-\epsilon)}$$

where N is the number of indicator-type events, ϵ is the tail probability of the VaR, and $F^{-1}(t)$ is the inverse distribution function of the standard normal distribution. In the example, $N = 500, \epsilon = 0.01$, and the 95% confidence interval for the number of exceedances is $[0, 9]$. Similarly, if we are back-testing the 95% VaR, under the same circumstances the confidence interval is $[15, 34]$.

Note that the statistical test based on the back-testing of VaR at a certain tail probability cannot answer the question if the distributional assumptions for the risk-driving factors are correct in general. For instance, if the portfolio contains only common stocks, then we

presume a probabilistic model for stock returns. By back-testing the 99% daily VaR of portfolio return, we verify if the probabilistic model is adequate for the 1% quantile of the portfolio return distribution: that is, we are back-testing if a certain point in the left tail of the portfolio return distribution is sufficiently accurately modeled. This should not be confused with statistical tests such as the Kolmogorov test or the Kolmogorov–Smirnov test, which concern accepting or rejecting a given distributional hypothesis.

5.4.4 Coherent risk measures

Even though VaR has an intuitive interpretation and has been widely adopted as a risk measure, it does not always satisfy the important property that the VaR of a portfolio should not exceed the sum of the VaRs of the portfolio positions. This means that VaR is not always capable of representing the diversification effect.

This fact raises an important question. Can we find a set of desirable properties that a risk measure should satisfy? An answer is given by Artzner et al. (1998). They provide an axiomatic definition of a functional which they call a *coherent risk measure*. The axioms follow with remarks given below each axiom.[7] We denote the risk measure by the functional $\rho(X)$, assigning a real-valued number to a random variable. Usually, the random variable X is interpreted as a random payoff and the motivation for the axioms in Artzner et al. (1998) follows this interpretation. In the remarks below each axiom, we provide an alternative interpretation which holds if X is interpreted as a random return.

The monotonicity property

Monotonicity $\rho(Y) \leq \rho(X)$, if $Y \geq X$ in almost sure sense.

Monotonicity states that if investment A has random return (payoff) Y which is not less than the return (payoff) X of investment B at a given horizon in all states of the world, then the risk of A is not greater than the risk of B. This is quite intuitive but it really does matter whether the random variables represent random return

or profit because an inequality in almost sure sense between random returns may not translate into the same inequality between the corresponding random profits and vice versa.

Suppose that X and Y describe the random percentage returns on two investments A and B and let $Y = X + 3\%$. Apparently, $Y > X$ in all states of the world. The corresponding payoffs are obtained according to the equations

$$\text{Payoff}(X) = I_A(1 + X)$$
$$\text{Payoff}(Y) = I_B(1 + Y) = I_B(1 + X + 3\%)$$

where I_A is the initial investment in opportunity A and I_B is the initial investment in opportunity B. If the initial investment I_A is much larger than I_B, then Payoff(X) > Payoff(Y) irrespective of the inequality $Y > X$. In effect, investment A may seem less risky than investment B in terms of payoff but in terms of return, the converse may hold.

The positive homogeneity property

Positive homogeneity $\rho(0) = 0$, $\rho(\lambda X) = \lambda\rho(X)$, for all X and all $\lambda > 0$.

The positive homogeneity property states that scaling the return (payoff) of the portfolio by a positive factor scales the risk by the same factor. The interpretation for payoffs is obvious – if the investment in a position doubles, so does the risk of the position. We give a simple example illustrating this property when X stands for a random percentage return.

Suppose that today the value of a portfolio is I_0 and we add a certain amount of cash C. The value of our portfolio becomes $I_0 + C$. The value tomorrow is random and equals $I_1 + C$ in which I_1 is the random payoff. The return of the portfolio equals

$$X = \frac{I_1 + C - I_0 - C}{I_0 + C} = \frac{I_1 - I_0}{I_0}\left(\frac{I_0}{I_0 + C}\right)$$
$$= h\frac{I_1 - I_0}{I_0} = hY$$

where $h = I_0/(I_0 + C)$ is a positive constant. The axiom positive homogeneity property implies that $\rho(X) = h\rho(Y)$: that is, the risk of the new portfolio will be the risk of the portfolio without the cash but scaled by h.

The sub-additivity property

Sub-additivity $\rho(X + Y) \leq \rho(X) + \rho(Y)$, for all X and Y.

If X and Y describe random payoffs, then the sub-additivity property states that the risk of the portfolio is not greater than the sum of the risks of the two random payoffs.

The positive homogeneity property and the sub-additivity property imply that the functional is convex

$$\begin{aligned}\rho(\lambda X + (1-\lambda)Y) &\leq \rho(\lambda X) + \rho((1-\lambda)Y) \\ &= \lambda\rho(X) + (1-\lambda)\rho(Y)\end{aligned}$$

where $\lambda \in [0, 1]$. If X and Y describe random returns, then the random quantity $\lambda X + (1 - \lambda)Y$ stands for the return of a portfolio composed of two financial instruments with returns X and Y having weights λ and $1 - \lambda$, respectively. Therefore, the convexity property states that the risk of a portfolio is not greater than the sum of the risks of its constituents, meaning that it is the convexity property which is behind the diversification effect that we expect in the case of X and Y denoting random returns.

The invariance property

Invariance $\rho(X + C) = \rho(X) - C$, for all X and $C \in \mathbb{R}$.

The invariance property has various labels. Originally, it was called *translation invariance* while in other texts it is called *cash invariance*.[8] If X describes a random payoff, then the invariance property suggests that adding cash to a position reduces its risk by the amount of cash added. This is motivated by the idea that the risk measure can be used to determine capital requirements. As a consequence, the risk measure $\rho(X)$ can be interpreted as the minimal amount of

cash necessary to make the position free of any capital requirements,

$$\rho(X + \rho(X)) = 0.$$

The invariance property has a different interpretation when X describes random return. Suppose that the random variable X describes the return of a common stock and we build a long-only portfolio by adding a government bond yielding a risk-free rate r_B. The portfolio return equals $wX + (1 - w)r_B$, where $w \in [0, 1]$ is the weight of the common stock in the portfolio. Note that the quantity $(1 - w)r_B$ is non-random by assumption. The invariance property states that the risk of the portfolio can be decomposed as

$$\begin{aligned} \rho(wX + (1 - w)r_B) &= \rho(wX) - (1 - w)r_B \\ &= w\rho(X) - (1 - w)r_B \end{aligned} \tag{5.4.9}$$

where the second equality appears because of the positive homogeneity property. In effect, the risk measure admits the following interpretation: Assume that the constructed portfolio is equally weighted, i.e. $w = 1/2$, then the risk measure equals the level of the risk-free rate such that the risk of the equally weighted portfolio consisting of the risky asset and the risk-free asset is zero. The investment in the risk-free asset will be, effectively, the reserve investment.

Alternative interpretations are also possible. Suppose that the present value of the position with random percentage return X is I_0. Assume that we can find a government security earning return r_B^* at the horizon of interest. Then we can ask the question in the opposite direction: How much should we reallocate from I_0 and invest in the government security in order to hedge the risk $\rho(X)$? The needed capital C should satisfy the equation

$$\frac{I_0 - C}{I_0}\rho(X) - \frac{C}{I_0}r_B^* = 0$$

which is merely a re-statement of equation (5.4.9) with the additional requirement that the risk of the resulting portfolio should be zero.

The solution is

$$C = I_0 \frac{\rho(X)}{\rho(X) + r_B^*}.$$

Note that if in the invariance property the constant is non-negative, $C \geq 0$, then it follows that $\rho(X + C) \leq \rho(X)$. This result is in agreement with the monotonicity property as $X + C \geq X$. In fact, the invariance property can be regarded as an extension of the monotonicity property when the only difference between X and Y is in their means.

According to the discussion in the previous section, VaR is not a coherent risk measure because it may violate the sub-additivity property.

An example of a coherent risk measure is the average VaR (AVaR), defined as the average of the VaRs which are larger than the VaR at a given tail probability ϵ. The accepted notation is $AVaR_\epsilon(X)$, in which ϵ stands for the tail probability level. A larger family of coherent risk measures is the family of spectral risk measures, which includes the AVaR as a representative. The spectral risk measures are defined as weighted averages of VaRs. The AVaR and the spectral risk measures will be considered in detail in Chapter 6.

5.5 Risk Measures and Dispersion Measures

In the introduction to this chapter, we remarked that there is a certain relationship between risk and uncertainty. While the two notions are different, without uncertainty there is no risk. Having this in mind, it is not surprising that there are similarities between the axioms behind the deviation measures in section 5.2.5 and the axioms behind the coherent risk measures in section 5.4.4. Both classes, the deviation measures and the coherent risk measures, are not the only classes capable of quantifying statistical dispersion and risk respectively.[9] Nevertheless, they describe basic features of uncertainty and risk and, in effect, we may expect that a relationship between them exists.[10]

Inspecting the defining axioms, we conclude that the common properties are the sub-additivity property and the positive homogeneity property. The specific features are the monotonicity property and the invariance property of the coherent risk measures and the translation invariance and positivity of deviation measures. The link between them concerns a subclass of the coherent risk measures called *strictly expectation-bounded risk measures* and a subclass of the deviation measures called *lower-range-dominated deviation measures*.

A coherent risk measure $\rho(X)$ is called *strictly expectation bounded* if it satisfies the condition

$$\rho(X) > -EX \tag{5.5.1}$$

for all non-constant X, in which EX is the mathematical expectation of X. If X describes the portfolio return distribution, then the inequality in (5.5.1) means that the risk of the portfolio is always greater than the negative of the expected portfolio return. A coherent risk measure satisfying this condition is the AVaR, for example.

A deviation measure $D(X)$ is called *lower range dominated* if it satisfies the condition

$$D(X) \leq EX \tag{5.5.2}$$

for all non-negative random variables, $X \geq 0$. A deviation measure which is lower range dominated is, for example, the downside semi-standard deviation σ_X^- defined in (5.2.4).

The relationship between the two subclasses is a one-to-one correspondence between them established through the equations

$$D(X) = \rho(X - EX) \tag{5.5.3}$$

and

$$\rho(X) = D(X) - E(X). \tag{5.5.4}$$

That is, if $\rho(X)$ is a strictly expectation-bounded coherent risk measure, then through the formula in (5.5.3) we obtain the corresponding lower-range-dominated deviation measure and, conversely,

through the formula in (5.5.4), we obtain the corresponding strictly expectation-bounded coherent risk measure.

In effect, there is a deviation measure behind each strictly expectation-bounded coherent risk measure. Consider the AVaR, for instance. Since it satisfies the property in (5.5.1), according to the relationship discussed above, the quantity

$$D_\epsilon(X) = AVaR_\epsilon(X - EX)$$

represents the deviation measure underlying the AVaR risk measure at tail probability ϵ. In fact, the quantity $D_\epsilon(X)$, as well as any other lower-range-dominated deviation measure, is obtained by computing the risk of the centered random variable. The definition of AVaR and different calculation methods are provided in Chapter 6.

5.6 Risk Measures and Stochastic Orders

In section 3.3 of Chapter 3, we considered stochastic dominance relations. The second-order stochastic dominance (SSD), for example, states that X dominates Y with respect to SSD when all risk-averse investors prefer X to Y. Suppose that we estimate the risk of X and Y through a risk measure ρ. If all risk-averse investors prefer X to Y, then does it follow that $\rho(X) \leq \rho(Y)$? This question describes the issue of consistency of a risk measure with the SSD order. Intuitively, a realistic risk measure should be consistent with the SSD order, since there is no reason to assume that an investment with higher risk as estimated by the risk measure will be preferred by all risk-averse investors.

Note that the monotonicity property of the coherent risk measures implies consistency with first-order stochastic dominance (FSD). The condition that $X \geq Y$ in all states of the world translates into the following inequality in terms of the c.d.f.s:

$$F_X(x) \leq F_Y(x), \quad \forall x \in \mathbb{R},$$

which, in fact, characterizes the FSD order.[11] As a result, if all non-satiable investors prefer X to Y, then any coherent risk measure will indicate that the risk of X is below the risk of Y.

Concerning the more important SSD order, the consistency question is more involved. The defining axioms of the coherent risk measures cannot guarantee consistency with the SSD order. Therefore, if we want to use a coherent risk measure in practice, we have to verify separately the consistency with the SSD order.

DeGiorgi (2005) shows that the AVaR, and spectral risk measures in general, are consistent with the SSD order. Note that if the AVaR, for example, is used to measure the risk of portfolio return distributions, then the corresponding SSD order concerns random variables describing returns. Similarly, if the AVaR is applied to random variables describing payoff, then the SSD order concerns random payoffs. SSD orders involving returns do not coincide with SSD orders involving payoffs (see section 3.3.6 in Chapter 3 for further details).

5.7 Summary

In this chapter, we described different approaches to quantifying risk and uncertainty. We discussed in detail the following dispersion measures: standard deviation, mean absolute deviation, upside and downside semi-standard deviations, an axiomatic description of dispersion measures, and the family of deviation measures. We also discussed in detail the following risk measures: value-at-risk and the family of coherent risk measures.

We emphasized that a realistic statistical model for risk estimation includes two essential components: (1) a realistic statistical model for the financial asset return distributions and their dependence, capable of accounting for empirical phenomena and (2) a true risk measure capable of describing the essential characteristics of risk.

We explored a link between risk measures and dispersion measures through two subclasses of coherent risk measures and

deviation measures. Behind every such coherent risk measure, we can find a corresponding deviation measure and vice versa. The intuitive connection between risk and uncertainty materializes quantitatively in a particular form.

Finally, we emphasized the importance of consistency of risk measures with the SSD order. In the appendix to this chapter, we demonstrate a relationship between probability quasi-metrics and deviation measures.

5.8 Technical Appendix

In this appendix, we provide an example of a class of risk measures more general than the coherent risk measures described in the chapter. Then we demonstrate that all symmetric deviation measures are generated from probability metrics.

5.8.1 Convex risk measures

In the chapter, we noted that the sub-additivity and the positive homogeneity properties of coherent risk measures guarantee that they are convex. The convexity property is the essential feature describing the diversification effect when the random variables are interpreted as portfolio returns. Thus, it is possible to postulate convexity directly and obtain the larger class of *convex risk measures*.

A risk measure ρ is said to be a convex risk measure if it satisfies the following properties:

Monotonicity	$\rho(Y) \leq \rho(X)$, if $Y \geq X$ in almost sure sense.
Convexity	$\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$, for all X, Y and $\lambda \in [0,1]$.
Invariance	$\rho(X + C) = \rho(X) - C$, for all X and $C \in \mathbb{R}$.

The remarks from section 5.4.4 concerning the interpretation of the axioms of coherent risk measures depending on whether X describes payoff or return are valid for the convex risk measures as well. The

convex risk measures are more general than the coherent risk measures because every coherent risk measure is convex but not vice versa. The convexity property does not imply positive homogeneity. Föllmer and Schied (2002) provide more details on convex risk measures and their relationship with preference relations.

5.8.2 Probability metrics and deviation measures

In this section, we demonstrate that the symmetric deviation measures[12] arise from probability metrics equipped with two additional properties – *translation invariance* and *positive homogeneity*. In fact, not only the symmetric but all deviation measures can be described with the general method of probability metrics by extending the framework.

We briefly repeat the definition of a probability semimetric given in Chapter 2. The probability semimetric is denoted by $\mu(X, Y)$, in which X and Y are random variables. The properties which $\mu(X, Y)$ should satisfy are the following:

Property 1. $\mu(X, Y) \geq 0$ for any X, Y and $\mu(X, Y) = 0$ if $X = Y$ in almost sure sense.
Property 2. $\mu(X, Y) = \mu(Y, X)$ for any X, Y.
Property 3. $\mu(X, Y) \leq \mu(X, Z) + \mu(Z, Y)$ for any X, Y, Z.

A probability metric is called *translation invariant* and *positively homogeneous* if, besides properties 1, 2, and 3, it satisfies also

Property 4. $\mu(X + Z, Y + Z) = \mu(Y, X)$ for any X, Y, Z.
Property 5. $\mu(aX, aY) = a\mu(X, Y)$ for any X, Y and $a > 0$.

Property 4 is the translation invariance axiom and Property 5 is the positive homogeneity axiom.

Note that translation invariance and positive homogeneity have a different meaning depending on whether probability metrics or dispersion measures are concerned. To avoid confusion, we enumerate the axioms of symmetric deviation measures given in section 5.2.5

of this chapter. A symmetric deviation measure $D(X)$ satisfies the following axioms:

Property 1*. $D(X + C) = D(X)$ for all X and constants $C \in \mathbb{R}$.
Property 2*. $D(X) = D(-X)$ for all X.
Property 3*. $D(0) = 0$ and $D(\lambda X) = \lambda D(X)$ for all X and all $\lambda > 0$.
Property 4*. $D(X) \geq 0$ for all X, with $D(X) > 0$ for non-constant X.
Property 5*. $D(X + Y) \leq D(X) + D(Y)$ for all X and Y.

We will demonstrate that the functional

$$\mu_D(X, Y) = D(X - Y) \tag{5.8.1}$$

is a probability semimetric satisfying properties 1 through 5 if D satisfies properties 1* through 5*. Furthermore, the functional

$$D_\mu(X) = \mu(X - EX, 0) \tag{5.8.2}$$

is a symmetric deviation measure if μ is a probability metric satisfying properties 2 through 5.

Demonstration of equation (5.8.1)
We show that properties 1 through 5 hold for μ_D defined in equation (5.8.1).

Property 1. $\mu_D(X, Y) \geq 0$ follows from the non-negativity of D, Property 4*. Further on, if $X = Y$ in almost sure sense, then $X - Y = 0$ in almost sure sense and $\mu_D(X, Y) = D(0) = 0$ from Property 3*.
Property 2. A direct consequence of Property 2*.
Property 3. Follows from Property 5*:

$$\begin{aligned}\mu(X, Y) &= D(X - Y) = D(X - Z + (Z - Y)) \\ &\leq D(X - Z) + D(Z - Y) = \mu(X, Z) + \mu(Z, Y)\end{aligned}$$

Property 4. A direct consequence of the definition in (5.8.1).
Property 5. Follows from Property 3*.

Demonstration of equation (5.8.2)
We show that properties 1* through 5* hold for D_μ defined in equation (5.8.2).

Property 1*. A direct consequence of the definition in (5.8.2).
Property 2*. Follows from Property 4 and Property 2.

$$\begin{aligned} D_\mu(-X) &= \mu(-X + EX, 0) = \mu(0, X - EX) \\ &= \mu(X - EX, 0) = D_\mu(X) \end{aligned}$$

Property 3*. Follows from Property 1 and Property 5. $D_\mu(0) = \mu(0, 0) = 0$ and

$$D_\mu(\lambda X) = \lambda\mu(X - EX, 0) = \lambda D_\mu(X)$$

Property 4*. Follows because μ is a probability metric. If $D_\mu(X) = 0$, then $X - EX$ is equal to zero almost surely, which means that X is a constant in all states of the world.
Property 5*. Arises from Property 3 and Property 4.

$$\begin{aligned} D(X + Y) &= \mu(X - EX + Y - EY, 0) = \mu(X - EX, -Y + EY) \\ &\leq \mu(X - EX, 0) + \mu(0, -Y + EY) \\ &= \mu(X - EX, 0) + \mu(Y - EY, 0) \\ &= D(X) + D(Y) \end{aligned}$$

Conclusion
Equation (5.8.2) shows that all symmetric deviation measures arise from the translation invariant, positively homogeneous probability metrics.

Note that because of the properties of the deviation measures, μ_D is a semimetric and cannot become a metric. This is because D is not sensitive to additive shifts and this property is inherited by μ_D,

$$\mu_D(X + a, Y + b) = \mu_D(X, Y),$$

where a and b are constants. In effect, $\mu_D(X, Y) = 0$ implies that the two random variables differ by a constant, $X = Y + c$ in all states of the world.

Due to the translation invariance property, equation (5.8.2) can be equivalently re-stated as

$$D_\mu(X) = \mu(X, EX). \tag{5.8.3}$$

In fact, as we remarked in the chapter, equation (5.8.3) represents a very natural generic way of defining measures of dispersion. Starting from equation (5.8.3) and replacing the translation invariance property by the regularity property of ideal probability metrics given in section 4.5 of Chapter 4, the sub-additivity property (Property 5*) of $D_\mu(X)$ breaks down and a property similar to the positive shift property given in the chapter holds instead of Property 1*,

$$D_\mu(X + C) = \mu(X + C, EX + C) \leq \mu(X, EX) = D_\mu(X)$$

for all constants C. In fact, this property is more general than the positive shift property as it holds for arbitrary constants.

5.8.3 Deviation measures and probability quasi-metrics

In this section, we demonstrate that the deviation measures arise from probability quasi-metrics[13] equipped with the two additional properties of translation invariance and positive homogeneity given in section 5.8.2.

Theorem 5.8.1. The functional μ_D defined as

$$\mu_D(X, Y) = D(X - Y)$$

is a positively homogeneous, translation-invariant probability quasi-semimetric if D is a deviation measure. Furthermore, the functional D_μ defined as

$$D_\mu(X) = \mu(X - EX, 0)$$

is a deviation measure if μ is a a positively homogeneous, translation-invariant probability quasi-metric.

The proof can be found in Stoyanov et al. (2008) but basically it repeats the steps considered in section 5.8.2. The result in the theorem

is more general than the one given in section 5.8.2 and includes it as a special case.

Notes

1. Holton (2004) provides a thorough analysis of the notion of risk. Knight (1921) started the debate about risk and uncertainty.
2. This distinction is made by the Basel Committee on Banking Supervision. The Basel Committee consists of representatives from central banks and regulatory authorities of the G10 countries. It has issued two banking supervision accords, Basel I and Basel II, with the purpose of ensuring that financial institutions retain enough capital as a protection against unexpected losses. In the two accords, a distinction is made between market, credit, and operational risk, and a simple methodology is provided for their quantification.
3. At times, we will use the notation $\sigma(X)$ instead of σ_X to accentuate that the standard deviation is a functional of the underlying distribution.
4. Symmetric random variables are described through their distribution function: that is, X is symmetric (around zero) if X has the same distribution function as $-X$, $X \stackrel{d}{=} -X$, where the notation $\stackrel{d}{=}$ means equality in distribution. If the mean of the distribution is not zero, then the condition of symmetry changes to $X - EX \stackrel{d}{=} -(X - EX)$, and we say that X is symmetric around the mean.
5. Chapter 4 provides more details on primary, simple, and compound probability metrics.
6. A boxplot, or a box-and-whiskers diagram, is a convenient way of depicting several statistical characteristics of the sample. The size of the box equals the difference between the third and the first *quartile* (75% quantile – 25% quantile), also known as the *interquartile range*. The line in the box corresponds to the median of the data (50% quantile). The lines extending out of the box are called *whiskers* and each of them is long up to 1.5 times the interquartile range. All observations outside the whiskers are labeled outliers and are depicted by a plus sign.
7. Further remarks on this and other axiomatic constructions can be found in Pflug and Roemisch (2007) and Heyde et al. (2009)

8. This label can be found in Föllmer and Schied (2002).
9. The appendix to this chapter contains an example of a class of risk measures which is more general than the coherent risk measures. This is the class of *convex risk measures*.
10. The relationship is studied in Rockafellar et al. (2006).
11. Section 3.3 of Chapter 3 provides more details.
12. Deviation measures are described in section 5.2.5.
13. Probability quasi-semimetrics satisfy all properties of probability semimetrics save for the symmetry property: see Chapter 2 for the corresponding definition.

References

Artzner, P., F. Delbaen, J.-M. Eber and D. Heath (1998), 'Coherent measures of risk', *Mathematical Finance* **6**, 203–228.

DeGiorgi, E. (2005), 'Reward–risk portfolio selection and stochastic dominance', *Journal of Banking and Finance* **29 (4)**, 895–926.

Föllmer, H. and A. Schied (2002), *Stochastic Finance: An Introduction in Discrete Time, second revised and extended revision*, Walter de Gruyter, Berlin.

Heyde, C., S. Kou and X. Peng (2009), 'What is a good risk measure: bridging the gaps between data, coherent risk measure, and insurance risk measures', Columbia University.

Holton, Glyn A. (2004), 'Defining risk', *Financial Analysts Journal* **60 (6)**, 19–25.

Knight, F. H. (1921), *Risk, Uncertainty, and Profit*, Houghton Mifflin, Boston and New York.

Lamantia, F., S. Ortobelli and S. Rachev (2006*a*), 'An empirical comparison among var models and time rules with elliptical and stable distributed returns', *Investment Management and Financial Innovations* **3**, 8–29.

Lamantia, F., S. Ortobelli and S. Rachev (2006*b*), 'Var, cvar and time rules with elliptical and asymmetric stable distributed returns', *Investment Management and Financial Innovations* **4**, 19–39.

Markowitz, H. M. (1952), 'Portfolio selection', *Journal of Finance* **7 (1)**, 77–91.

Markowitz, H. M. (1959), *Portfolio Selection: Efficient Diversification of Investments*, John Wiley & Sons, New York.

Pflug, G. and W. Roemisch (2007), *Modeling, Measuring and Managing Risk*, World Scientific, Hackensack, NJ.

Rachev, S. T., S. Ortobelli, S. Stoyanov, F. J. Fabozzi and A. Biglova (2008), 'Desirable properties of an ideal risk measure in portfolio theory', *International Journal of Theoretical and Applied Finance* **11 (1)**, 19–54.

Rockafellar, R. T., S. Uryasev and M. Zabarankin (2006), 'Generalized deviations in risk analysis', *Finance and Stochastics* **10**, 51–74.

Stoyanov, S., S. Rachev and F. Fabozzi (2008), 'Probability metrics with applications in finance', *Journal of Statistical Theory and Practice* **2 (2)**, 253–277.

Chapter 6

Average Value-at-Risk

The goals of this chapter are the following:

- To introduce formally the *average value-at-risk* (AVaR) measure.
- To provide closed-form expressions for AVaR for a few distributions widely used for modeling financial returns.
- To consider the problem of estimating AVaR from a sample.
- To discuss spectral risk measures and distortion risk measures which include AVaR as a special case.
- To explain the important place of AVaR as a building block because any other distortion risk measure can be represented as an appropriate weighted average of AVaRs.

Notation introduced in this chapter:

Notation	*Description*
$AVaR_\epsilon(X)$	The AVaR of a random variable X at tail probability ϵ
$\rho_\phi(X)$	The spectral risk measure of a random variable X with a risk-aversion function ϕ
$\rho_H(X)$	The distortion risk measure of a random variable X
$m_\epsilon^n(X)$	The tail moment of order n of a random variable X at tail probability ϵ

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

Notation	*Description*
$\mu_\epsilon^n(X)$	The absolute central tail moment of order n of a random variable X at tail probability ϵ
$MTL_\epsilon(X)$	The median tail loss of a random variable X at tail probability ϵ
$AVaR_\epsilon^{(n)}(X)$	The AVaR of order n of a random variable X at tail probability ϵ
$ETL_\epsilon(X)$	The expected tail loss of a random variable X at tail probability ϵ
$\mathcal{RV}_\alpha$	The set of monotonic functions regularly varying at infinity with index α

Important terms introduced in this chapter:

Term	*Concise explanation*
heavy-tailed distribution	A probability distribution the tails of which decay faster than the tails of the exponential distribution.
robust estimator	An estimator which is not excessively influenced by small departures from the model assumptions (e.g. by presence of outliers in the data).

6.1 Introduction

The value-at-risk (VaR) measure has been adopted as a standard risk measure in the financial industry. Nonetheless, it has a number of deficiencies recognized by financial professionals. In Chapter 5, we remarked that there is one very important property which does not hold for VaR. This is the sub-additivity property which ensures that the VaR measure cannot always account for diversification. There are cases in which the portfolio VaR is larger than the sum of the VaRs of the portfolio constituents. This shows that VaR cannot be used as a true risk measure.

AVaR is a risk measure which is a superior alternative to VaR. Not only does it lack the deficiencies of VaR, but it also has an

intuitive interpretation. There are convenient ways for computing and estimating AVaR which allows its application in optimal portfolio problems. Moreover, it satisfies all axioms of coherent risk measures and it is consistent with the preference relations of risk-averse investors.

In this chapter, we explore in detail the properties of AVaR and illustrate its superiority to VaR. We develop new geometric interpretations of AVaR and the various calculation methods. We also provide closed-form expressions for the AVaR of the normal distribution, Student's t distribution, and a practical formula for Lévy stable distributions. Finally, we describe different estimation methods and remark on potential pitfalls.

Besides AVaR, we consider a more general family of risk measures satisfying the axioms of coherent risk measures. This is the class of spectral risk measures which contains AVaR as a special case. In contrast to AVaR, spectral risk measures in general are harder to work with. There are subtle conditions which have to be satisfied in order for spectral risk measures to be a practical concept. Such conditions are stated in the appendix to this chapter.

At the end of the chapter, we note an interesting link between probability metrics and risk measures. Having selected a risk measure, it is possible to find a probability metric which ensures that random variables closer to each other with respect to the probability metric have similar risk profiles.

6.2 Average Value-at-Risk

In section 5.4.1 of Chapter 5, we noted that a disadvantage of VaR is that it does not give any information about the severity of losses beyond the VaR level. Consider the following example. Suppose that X and Y describe the random returns of two financial instruments with densities and distribution functions such as the ones in Figure 6.1. The expected returns are 3% and 1%, respectively. The standard deviations of X and Y are equal to 10%.[1] The cumulative distribution functions (c.d.f.s) $F_X(x)$ and $F_Y(x)$ cross at

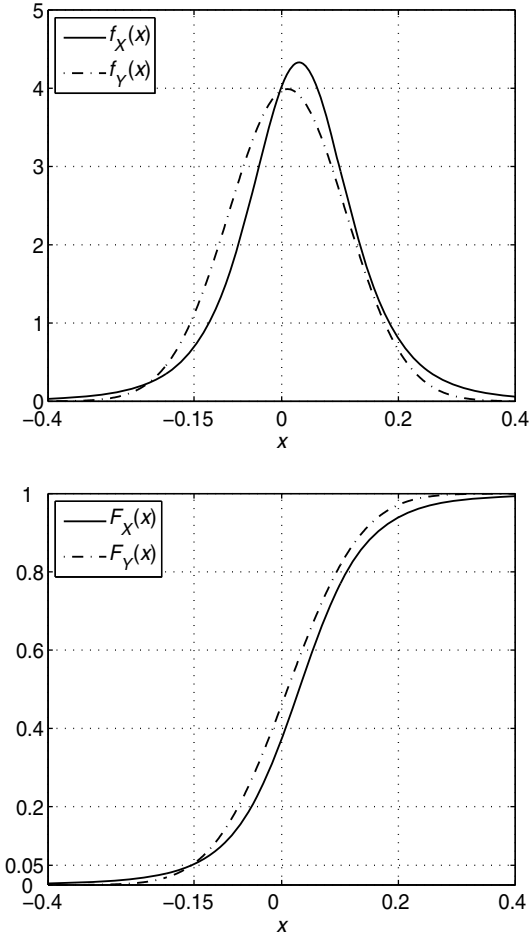

Figure 6.1: The top plot shows the densities of X and Y and the bottom plot shows their c.d.f.s. The 95% VaRs of X and Y are equal to 0.15 but X has a thicker tail and is more risky.

$x = -0.15$ and $F_X(-0.15) = F_Y(-0.15) = 0.05$. According to the definition of VaR in equation (5.4.1), the 95% VaRs of both X and Y are equal to 15%. That is, the two financial instruments lose more than 15% of their present values with probability of 5%. In effect, we

may conclude that their risks are equal because their 95% VaRs are equal.

This conclusion is wrong because we pay no attention to the losses which are larger than the 95% VaR level. It is visible in Figure 6.1 that the left tail of X is heavier than the left tail of Y.[2] Therefore, it is more likely that the losses of X will be larger than the losses of Y, on condition that they are larger than 15%. Thus, looking only at the losses occurring with probability smaller than 5%, the random return X is riskier than Y. Note that both X and Y have equal standard deviations. If we base the analysis on the standard deviation and the expected return, we would conclude that not only is the uncertainty of X equal to the uncertainty of Y but X is actually preferable because of the higher expected return. In fact, we realize that it is exactly the opposite, which shows how important it is to ground the reasoning on a proper risk measure.

The disadvantage of VaR, that it is not informative about the magnitude of the losses larger than the VaR level, is not present in the risk measure known as *average value-at-risk*. In the literature, it is also called *conditional value-at-risk*[3] or *expected shortfall* but we will use average value-at-risk (AVaR) as it best describes the quantity it refers to.

The AVaR at tail probability ϵ is defined as the average of the VaRs which are larger than the VaR at tail probability ϵ. Therefore, by construction, the AVaR is focused on the losses in the tail which are larger than the corresponding VaR level. The average of the VaRs is computed through the integral

$$AVaR_\epsilon(X) := \frac{1}{\epsilon}\int_0^\epsilon VaR_p(X)dp \tag{6.2.1}$$

where $VaR_p(X)$ is defined in equation (5.4.1) in Chapter 5. As a matter of fact, the AVaR is not well defined for all real-valued random variables but only for those with finite mean: that is, $AVaR_\epsilon(X) < \infty$ if $E|X| < \infty$. This should not be disturbing because random variables with infinite mathematical expectation have limited application in the field of finance. For example, if such a random variable is used

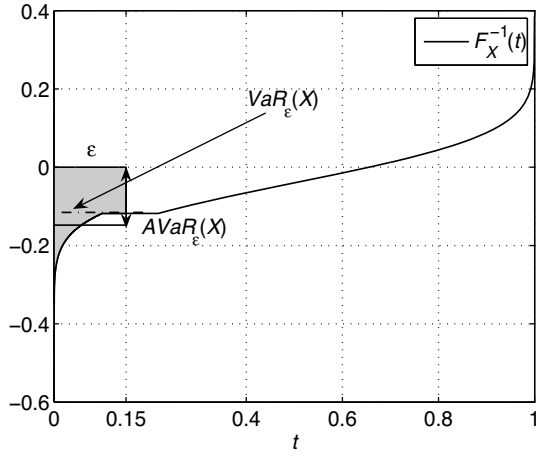

Figure 6.2: Geometrically, $AVaR_\epsilon(X)$ is the height for which the area of the drawn rectangle equals the shaded area closed between the graph of the inverse c.d.f. and the horizontal axis for $t \in [0, \epsilon]$. The $VaR_\epsilon(X)$ value is shown by a dash-dotted line.

for a model of stock returns, then it is assumed that the common stock has infinite expected return, which is not realistic.

The AVaR satisfies all the axioms of coherent risk measures. One consequence is that, unlike VaR, it is convex for all possible portfolios, which means that it always accounts for the diversification effect.

A geometric interpretation of the definition in equation (6.2.1) is provided in Figure 6.2. In this figure, the inverse c.d.f. of a random variable X is plotted. The shaded area is closed between the graph of $F_X^{-1}(t)$ and the horizontal axis for $t \in [0, \epsilon]$ where ϵ denotes the selected tail probability. $AVaR_\epsilon(X)$ is the value for which the area of the drawn rectangle, equal to $\epsilon \times AVaR_\epsilon(X)$, coincides with the shaded area, which is computed by the integral in equation (6.2.1). The $VaR_\epsilon(X)$ value is always smaller than $AVaR_\epsilon(X)$. In Figure 6.2, $VaR_\epsilon(X)$ is shown by a dash-dotted line and is indicated by an arrow.

Let us revisit the example developed at the beginning of this section. We concluded that even though the VaRs at 5% tail probability of both random variables are equal, X is riskier than Y because the

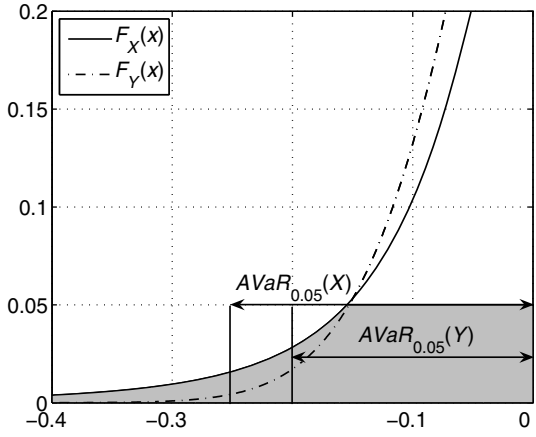

Figure 6.3: The AVaRs of the return distributions from Figure 6.1 in line with the geometric intuition. Even though the 95% VaRs are equal, the AVaRs at 5% tail probability differ, $AVaR_{0.05}(X) > AVaR_{0.05}(Y)$.

left tail of X is heavier than the left tail of Y: that is, the distribution of X is more likely to produce larger losses than the distribution of Y on condition that the losses are beyond the VaR at the 5% tail probability. We apply the geometric interpretation illustrated in Figure 6.2 to this example. First, notice that the shaded area in Figure 6.2, which concerns the graph of the inverse of the c.d.f., can also be identified through the graph of the c.d.f. This is done in Figure 6.3, which shows a magnified section of the left tails of the c.d.f.s plotted in Figure 6.1. The shaded area appears as the intersection of the area closed below the graph of the distribution function and the horizontal axis, and the area below a horizontal line shifted at the tail probability above the horizontal axis. In Figure 6.3, we show the area for $F_X(x)$ at 5% tail probability. The corresponding area for $F_Y(x)$ is smaller because $F_Y(x) \leq F_X(x)$ to the left of the crossing point of the two c.d.f.s, which is exactly at 5% tail probability.

In line with the geometric interpretation, the $AVaR_{0.05}(X)$ is a number, such that if we draw a rectangle with height 0.05 and width equal to $AVaR_{0.05}(X)$, the area of the rectangle $(0.05 \times AVaR_{0.05}(X))$ equals

the shaded area in Figure 6.3. The same exercise for $AVaR_{0.05}(Y)$ shows that $AVaR_{0.05}(Y) < AVaR_{0.05}(X)$ because the corresponding shaded area is smaller and both rectangles share a common height of 0.05.

Besides the definition in equation (6.2.1), AVaR can be represented through a minimization formula,[4]

$$AVaR_\epsilon(X) = \min_{\theta \in \mathbb{R}} \left(\theta + \frac{1}{\epsilon} E(-X - \theta)_+ \right) \tag{6.2.2}$$

where $(x)_+$ denotes the maximum between x and zero, $(x)_+ = \max(x, 0)$ and X describes the portfolio return distribution. It turns out that this formula has an important application in optimal portfolio problems based on AVaR as a risk measure. In the appendix to this chapter, we provide an illuminating geometric interpretation of equation (6.2.2) which shows the connection to definition of AVaR.

How can we compute the AVaR for a given return distribution? Throughout this section, we assume that the return distribution function is a continuous function (i.e., there are no point masses). Under this condition, after some algebra and using the fact that VaR is the negative of a certain quantile, we obtain that the AVaR can be represented in terms of a conditional expectation,

$$\begin{aligned} AVaR_\epsilon(X) &= -\frac{1}{\epsilon} \int_0^\epsilon F_X^{-1}(t)dt \\ &= -E(X|X < -VaR_\epsilon(X)), \end{aligned} \tag{6.2.3}$$

which is called *expected tail loss* (ETL) and is denoted by $ETL_\epsilon(X)$. The conditional expectation implies that the AVaR equals the average loss provided that the loss is larger than the VaR level. In fact, the average of VaRs in equation (6.2.1) equals the average of losses in equation (6.2.3) only if the c.d.f. of X is continuous at $x = VaR_\epsilon(X)$. If there is a discontinuity, or a point mass, the relationship is more involved. The general formula is given in the appendix to this chapter.

Equation (6.2.3) implies that AVaR is related to the conditional loss distribution. In fact, under certain conditions, it is the mathematical expectation of the conditional loss distribution, which represents

only one characteristic of it. In section 6.10.1 in the appendix to this chapter, we introduce several sets of characteristics of the conditional loss distribution, which provide a more complete picture of it. Also, in section 6.10.2, we introduce the more general concept of higher-order AVaR.

For some continuous distributions, it is possible to calculate explicitly the AVaR through equation (6.2.3). We provide the closed-form expressions for the normal distribution and Student's t distribution. In the next section, we give a semi-explicit formula for the class of stable distributions.

(a) *The normal distribution*
Suppose that X is distributed according to a normal distribution with standard deviation σ_X and mathematical expectation EX. The AVaR of X at tail probability ϵ equals

$$AVaR_\epsilon(X) = \frac{\sigma_X}{\epsilon\sqrt{2\pi}} \exp\left(-\frac{(VaR_\epsilon(Y))^2}{2}\right) - EX \tag{6.2.4}$$

where Y has the standard normal distribution, $Y \in N(0,1)$.

(b) *The Student's t distribution*
Suppose that X has Student's t distribution with ν degrees of freedom, $X \in t(\nu)$. The AVaR of X at tail probability ϵ equals

$$AVaR_\epsilon(X) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \dfrac{\sqrt{\nu}}{(\nu-1)\epsilon\sqrt{\pi}} \left(1 + \dfrac{(VaR_\epsilon(X))^2}{\nu}\right)^{\frac{1-\nu}{2}}, & \nu > 1 \\ \infty, & \nu = 1 \end{cases}$$

where the notation $\Gamma(x)$ stands for the gamma function. It is not surprising that for $\nu = 1$ the AVaR explodes because the Student's t distribution with one degree of freedom, also known as the *Cauchy distribution*, has infinite mathematical expectation.[5]

Note that equation (6.2.4) can be represented in a more compact way,

$$AVaR_\epsilon(X) = \sigma_X C_\epsilon - EX, \tag{6.2.5}$$

where C_ϵ is a constant which depends only on the tail probability ϵ. Therefore, the AVaR of the normal distribution has the same structure as the normal VaR given in (5.4.8) in Chapter 5 – the difference between the properly scaled standard deviation and the mathematical expectation. In effect, similar to the normal VaR, the normal AVaR properties are dictated by the standard deviation. Even though AVaR is focused on the extreme losses only, due to the limitations of the normal assumption, it is symmetric.

Exactly the same conclusion holds for the AVaR of Student's t distribution. The true merits of AVaR become apparent if the underlying distributional model is skewed.

6.2.1 AVaR for stable distributions

A major criticism for assuming the normal distribution as a model is that very often there are outliers in the data. These outliers cannot be explained by the normal distribution. In contrast to the normal distribution, the stable Paretian distributions are heavy-tailed and they have the potential to describe the heavy tails and the asymmetry of the empirical data.

The class of the stable distributions is defined by means of their characteristic functions.[6] With very few exceptions, no closed-form expressions are known for their densities and distribution functions. A random variable X is said to have a stable distribution if there are parameters $0 < \alpha \leq 2,\ \sigma > 0,\ -1 \leq \beta \leq 1,\ \mu \in \mathbb{R}$ such that its characteristic function $\varphi_X(t) = Ee^{itX}$ has the following form:

$$\varphi_X(t) = \begin{cases} \exp\{-\sigma^\alpha |t|^\alpha \left(1 - i\beta \frac{t}{|t|} \tan\left(\frac{\pi\alpha}{2}\right)\right) + i\mu t\}, & \alpha \neq 1 \\ \exp\{-\sigma|t| \left(1 + i\beta \frac{2}{\pi} \frac{t}{|t|} \ln(|t|)\right) + i\mu t\}, & \alpha = 1 \end{cases} \tag{6.2.6}$$

where $\frac{t}{|t|} = 0$ if $t = 0$. Zolotarev (1986) and Samorodnitsky and Taqqu (1994) provide further details on the properties of stable distributions.

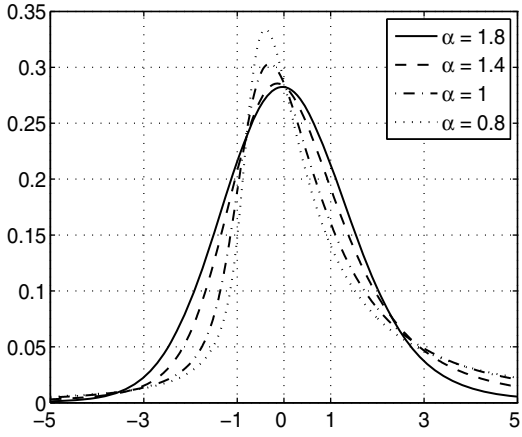

Figure 6.4: The density functions of stable laws with parameters $\alpha = 1.8,\ 1.4,\ 1,$ and 0.8, $\beta = 0.6$, $\sigma = 1$, $\mu = 0$.

The parameters appearing in equation (6.2.6) are the following:

α is called the *index of stability* or the *tail exponent*
β is a skewness parameter
σ is a scale parameter
μ is a location parameter

Since stable distributions are uniquely determined by these four parameters, the common notation is $S_\alpha(\sigma, \beta, \mu)$.

Figure 6.4 shows several stable densities with different tail exponents and $\beta = 0.6$. All densities are asymmetric but the skewness is more pronounced when the tail exponent is lower. Figure 6.5 shows several stable densities with different tail exponents and $\beta = 0$. All densities are symmetric.

The parameter α determines how heavy the tails of the distribution are. That is why it is also called the tail exponent. The lower the tail exponent, the heavier the tails. If $\alpha = 2$, then we obtain the normal distribution. Figure 6.5 illustrates the increase of the tail thickness as α decreases. Thicker tails indicate that the extreme events become more frequent. Due to the important effect of the parameter α on the

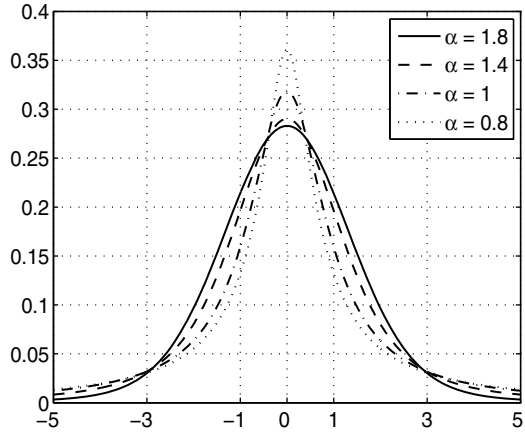

Figure 6.5: The density functions of stable laws with parameters $\alpha = 1.8,\ 1.4,\ 1,$ and 0.8, $\beta = 0$, $\sigma = 1$, $\mu = 0$.

properties of the stable distributions, they are often called *α-stable* or *alpha stable*.

Apart from the appealing feature that the probabilistic properties of only the stable distributions are close to the probabilistic properties of sums of i.i.d. random variables, there is another important characteristic which is the stability property. According to the stability property, appropriately centered and normalized sums of i.i.d. α-stable random variables are again α-stable. This property is unique to the class of stable laws.

Working with the class of stable distributions in practice is difficult because there are no closed-form expressions for their densities and distribution functions. Thus, practical work relies on numerical methods.

Stoyanov et al. (2006) give an account of the approaches to estimating AVaR of stable distributions. It turns out that there is a formula which is not exactly a closed-form expression, such as the ones for the normal and Student's t AVaR stated in the chapter, but is suitable for numerical work. It involves numerical integration but the integrand is nicely behaved and the integration range is a bounded

interval. Numerical integration can be performed by standard toolboxes in many software packages, such as MATLAB. Moreover, there are libraries freely available on the Internet. Therefore, numerical integration itself is not a severe restriction for applying a formula in practice. Since the formula involves numerical integration, we call it a *semi-analytic expression*.

Suppose that the random variable X has a stable distribution with tail exponent α, skewness parameter β, scale parameter σ, and location parameter μ, $X \in S_\alpha(\sigma, \beta, \mu)$. If $\alpha \leq 1$, then $AVaR_\epsilon(X) = \infty$. The reason is that stable distributions with $\alpha \leq 1$ have infinite mathematical expectation and the AVaR is unbounded.

If $\alpha > 1$ and $VaR_\epsilon(X) \neq 0$, then the AVaR can be represented as

$$AVaR_\epsilon(X) = \sigma A_{\epsilon,\alpha,\beta} - \mu$$

where the term $A_{\epsilon,\alpha,\beta}$ does not depend on the scale and the location parameters. In fact, this representation is a consequence of the positive homogeneity and the invariance property of AVaR. Concerning the term $A_{\epsilon,\alpha,\beta}$,

$$A_{\epsilon,\alpha,\beta} = \frac{\alpha}{1-\alpha} \frac{|VaR_\epsilon(X)|}{\pi\epsilon} \int_{-\bar{\theta}_0}^{\pi/2} g(\theta) \exp\left(-|VaR_\epsilon(X)|^{\frac{\alpha}{\alpha-1}} v(\theta)\right) d\theta$$

where

$$g(\theta) = \frac{\sin(\alpha(\bar{\theta}_0 + \theta) - 2\theta)}{\sin\alpha(\bar{\theta}_0 + \theta)} - \frac{\alpha\cos^2\theta}{\sin^2\alpha(\bar{\theta}_0 + \theta)},$$

$$v(\theta) = (\cos\alpha\bar{\theta}_0)^{\frac{1}{\alpha-1}} \left(\frac{\cos\theta}{\sin\alpha(\bar{\theta}_0 + \theta)}\right)^{\frac{\alpha}{\alpha-1}} \frac{\cos(\alpha\bar{\theta}_0 + (\alpha-1)\theta)}{\cos\theta},$$

in which $\bar{\theta}_0 = \frac{1}{\alpha}\arctan\left(\bar{\beta}\tan\frac{\pi\alpha}{2}\right)$, $\bar{\beta} = -\text{sign}(VaR_\epsilon(X))\beta$, and $VaR_\epsilon(X)$ is the VaR of the stable distribution at tail probability ϵ.

If $VaR_\epsilon(X) = 0$, then the AVaR admits a very simple expression,

$$AVaR_\epsilon(X) = \frac{2\Gamma\left(\frac{\alpha-1}{\alpha}\right)}{(\pi - 2\theta_0)} \frac{\cos\theta_0}{(\cos\alpha\theta_0)^{1/\alpha}}.$$

in which $\Gamma(x)$ is the gamma function and $\theta_0 = \frac{1}{\alpha}\arctan(\beta\tan\frac{\pi\alpha}{2})$.

It is possible to find expressions suitable for numerical work for VaR and AVaR of the much more general class of infinitely divisible distributions to which stable distributions belong. For more information and how this approach can be applied to some classes of tempered stable distributions, see Kim et al. (2009). Tempered stable distributions have been successfully applied as a model for stock returns in option pricing theory.

Besides the class of stable distributions, asymmetric versions of the classical Student's t distributions have been suggested as a model for stock returns (see, for example, Rachev and Mittnik (2000)). An approach for numerical calculation of VaR and AVaR for an asymmetric Student's t model is developed in Dokov et al. (2008).

6.3 AVaR Estimation from a Sample

Suppose that we have a sample of observed portfolio returns and we are not aware of their distribution. Provided that we do not impose any distributional model, the AVaR of portfolio return can be estimated from the sample of observed portfolio returns. Denote the observed portfolio returns by $r_1, r_2, \ldots, r_n$ at time instants $t_1, t_2, \ldots, t_n$. The numbers in the sample are given in order of observation.

Denote the sorted sample by $r_{(1)} \leq r_{(2)} \leq, \ldots, \leq r_{(n)}$. Thus, $r_{(1)}$ equals the smallest observed portfolio return and $r_{(n)}$ is the largest. The AVaR of portfolio returns at tail probability ϵ is estimated according to the formula[7]

$$\widehat{AVaR}_\epsilon(r) = -\frac{1}{\epsilon}\left(\frac{1}{n}\sum_{k=1}^{\lceil n\epsilon\rceil-1} r_{(k)} + \left(\epsilon - \frac{\lceil n\epsilon\rceil - 1}{n}\right) r_{(\lceil n\epsilon\rceil)}\right) \tag{6.3.1}$$

where the notation $\lceil x\rceil$ stands for the smallest integer larger than x.[8] The "hat" above AVaR denotes that the number calculated by equation (6.3.1) is an estimate of the true value because it is based on a sample. This is a standard notation in statistics.

We demonstrate how equation (6.3.1) is applied in the following example. Suppose that the sorted sample of portfolio returns is -1.37%, -0.98%, -0.38%, -0.26%, 0.19%, 0.31%, and 1.91%, and our goal is to calculate the portfolio AVaR at 30% tail probability. In this case, the sample contains 7 observations and $\lceil n\epsilon \rceil = \lceil 7 \times 0.3 \rceil = 3$. According to equation (6.3.1), we calculate

$$\begin{aligned}\widehat{AVaR}_{0.3}(r) &= -\frac{1}{0.3}\left(\frac{1}{7}(-1.37\% - 0.98\%) + (0.3 - 2/7)(-0.38\%)\right) \\ &= 1.137\%.\end{aligned}$$

Formula (6.3.1) can be applied not only to a sample of empirical observations. We may want to work with a statistical model for which no closed-form expressions for AVaR are known. Then we can simply sample from the distribution and apply formula (6.3.1) to the generated simulations.

Besides formula (6.3.1), there is another method for calculation of AVaR. It is based on the minimization formula (6.2.2) in which we replace the mathematical expectation by the sample average,

$$\widehat{AVaR}_{\epsilon}(r) = \min_{\theta \in \mathbb{R}} \left(\theta + \frac{1}{n\epsilon}\sum_{i=1}^{n} \max(-r_i - \theta, 0)\right). \tag{6.3.2}$$

Even though it is not obvious, equations (6.3.1) and (6.3.2) are completely equivalent.

The minimization formula in equation (6.3.2) is appealing because it can be calculated through the methods of linear programming. It can be restated as a linear optimization problem by introducing auxiliary variables $d_1, \ldots, d_n$, one for each observation in the sample,

$$\begin{aligned} \min_{\theta, d} \quad & \theta + \frac{1}{n\epsilon}\sum_{k=1}^{n} d_k \\ \text{subject to} \quad & -r_k - \theta \le d_k, k = 1, n \\ & d_k \ge 0, k = 1, n \\ & \theta \in \mathbb{R}. \end{aligned} \tag{6.3.3}$$

The linear problem (6.3.3) is obtained from (6.3.2) through standard methods in mathematical programming. We briefly demonstrate the equivalence between them. Let us fix the value of θ to θ^*. Then the following choice of the auxiliary variables yields the minimum in (6.3.3). If $-r_k - \theta^* < 0$, then $d_k = 0$. Conversely, if it turns out that $-r_k - \theta^* \geq 0$, then $-r_k - \theta^* = d_k$. In this way, the sum in the objective function becomes equal to the sum of maxima in equation (6.3.2).

Applying (6.3.3) to the sample in the example above, we obtain the optimization problem,

$$\begin{aligned}
\min_{\theta,d} \quad & \theta + \frac{1}{7 \times 0.3} \sum_{k=1}^{7} d_k \\
\text{subject to} \quad & 0.98\% - \theta \leq d_1 \\
& -0.31\% - \theta \leq d_2 \\
& -1.91\% - \theta \leq d_3 \\
& 1.37\% - \theta \leq d_4 \\
& 0.38\% - \theta \leq d_5 \\
& 0.26\% - \theta \leq d_6 \\
& -0.19\% - \theta \leq d_7 \\
& d_k \geq 0, k = 1, 7 \\
& \theta \in \mathbb{R}.
\end{aligned}$$

The solution to this optimization problem is the number 1.137% which is attained for $\theta = 0.38\%$. In fact, this value of θ coincides with the VaR at 30% tail probability and this is not by chance but a feature of the problem which is demonstrated in the appendix to this chapter. We verify that the solution of the problem is indeed the number 1.137% by calculating the objective in equation (6.3.2) for $\theta = 0.38\%$,

$$AVaR_\epsilon(r) = 0.38\% + \frac{0.98\% - 0.38\% + 1.37\% - 0.38\%}{7 \times 0.3} = 1.137\%.$$

Thus, we obtain the number calculated through equation (6.3.1).

6.4 Computing Portfolio AVaR in Practice

The ideas behind the approaches of VaR estimation can be applied to AVaR. We revisit the four methods from section 5.4.2 of Chapter 5, focusing on the implications for AVaR. We assume that there are n common stocks with random returns described by the random variables $X_1, \ldots, X_n$. Thus, the portfolio return is represented by

$$r_p = w_1 X_1 + \ldots + w_n X_n$$

where $w_1, \ldots, w_n$ are the weights of the common stocks in the portfolio.

6.4.1 The multivariate normal assumption

We noted in section 5.4.2 of Chapter 5 that if the stock returns are assumed to have a multivariate normal distribution, then the portfolio return has a normal distribution with variance $w'\Sigma w$, where w is the vector of weights and Σ is the covariance matrix between stock returns. The mean of the normal distribution is

$$Er_p = \sum_{k=1}^{n} w_k EX_k$$

where E stands for the mathematical expectation. Thus, under this assumption the AVaR of portfolio return at tail probability ϵ can be expressed in closed-form through equation (6.2.4),

$$\begin{aligned} AVaR_\epsilon(r_p) &= \frac{\sqrt{w'\Sigma w}}{\epsilon\sqrt{2\pi}} \exp\left(-\frac{(VaR_\epsilon(Y))^2}{2}\right) - Er_p \\ &= C_\epsilon \sqrt{w'\Sigma w} - Er_p \end{aligned} \tag{6.4.1}$$

where C_ϵ is a constant independent of the portfolio composition and can be calculated in advance. In effect, due to the limitations of the multivariate normal assumption, the portfolio AVaR appears symmetric and is representable as the difference between the properly scaled standard deviation of the random portfolio return and portfolio expected return.

6.4.2 The historical method

As we noted in section 5.4.2 of Chapter 5, the historical method is not related to any distributional assumptions. We use the historically observed portfolio returns as a model for the future returns and apply formula (6.3.1) or (6.3.2).

The historical method has several drawbacks as mentioned in section 5.4.2. We emphasize that it is very inaccurate for low tail probabilities, such as 1% or 5%. Even with one year of daily returns, which amounts to 250 observations, in order to estimate the AVaR at 1% probability, we have to use the three smallest observations, which is quite insufficient. What makes the estimation problem even worse is that these observations are in the tail of the distribution: that is, they are the *smallest* ones in the sample. The implication is that when the sample changes, the estimated AVaR may change a lot because the smallest observations tend to fluctuate a lot.

6.4.3 The hybrid method

According to the hybrid method described in section 5.4.2 of Chapter 5, different weights are assigned to the observations by which the more recent observations get a higher weight. The rationale is that the observations far back in the past have less impact on the portfolio risk at the present time.

The hybrid method can be adapted for AVaR estimation. The weights assigned to the observations are interpreted as probabilities and, thus, the portfolio AVaR can be estimated from the resulting discrete distribution according to the formula

$$\widehat{AVaR}_\epsilon(r) = -\frac{1}{\epsilon}\left(\sum_{j=1}^{k_\epsilon} p_j r_{(j)} + \left(\epsilon - \sum_{j=1}^{k_\epsilon} p_j\right) r_{(k_{\epsilon+1})}\right) \tag{6.4.2}$$

where $r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(k_m)}$ denotes the sorted sample of portfolio returns or payoffs and $p_1, p_2, \ldots, p_{k_m}$ stand for the probabilities of the sorted observations: that is, p_1 is the probability of $r_{(1)}$. The number

k_ϵ in equation (6.4.2) is an integer satisfying the inequalities,

$$\sum_{j=1}^{k_\epsilon} p_j \leq \epsilon < \sum_{j=1}^{k_\epsilon+1} p_j.$$

Equation (6.4.2) follows directly from the definition of AVaR[9] under the assumption that the underlying distribution is discrete without the additional simplification that the outcomes are equally probable. In the appendix to this chapter, we demonstrate the connection between equation (6.4.2) and the definition of AVaR in equation (6.2.1).

6.4.4 The Monte Carlo method

The basic steps of the Monte Carlo method are described in section 5.4.2 of Chapter 5. They are applied without modification. Essentially, we assume and estimate a multivariate statistical model for the stock returns distribution. Then we sample from it, and we calculate scenarios for portfolio return. On the basis of these scenarios, we estimate portfolio AVaR using equation (6.3.1), in which $r_1, \ldots, r_n$ stands for the vector of generated scenarios.

Similar to the case of VaR, an artifact of the Monte Carlo method is the variability of the risk estimate. Since the estimate of portfolio AVaR is obtained from a generated sample of scenarios, by regenerating the sample, we will obtain a slightly different value. We illustrate the variability issue by a simulation example, similar to the one developed for VaR in section 6.3.7.

Suppose that the portfolio daily return distribution is the standard normal law, $r_p \in N(0, 1)$. By the closed-form expression in equation (6.2.4), we calculate that the AVaR of the portfolio at 1% tail probability equals

$$AVaR_{0.01}(r_p) = \frac{1}{0.01\sqrt{2\pi}} \exp\left(-\frac{2.326^2}{2}\right) = 2.665.$$

In order to investigate how the fluctuations of the 99% AVaR change about the theoretical value, we generate samples of different sizes:

Table 6.1 The 99% AVaR of the standard normal distribution computed from a sample of scenarios. The 95% confidence interval is calculated from 100 repetitions of the experiment. The true value is $AVaR_{0.01}(X) = 2.665$.

Number of scenarios	*AVaR at 99%*	*95% confidence interval*
500	2.646	[2.2060, 2.9663]
1,000	2.771	[2.3810, 2.9644]
5,000	2.737	[2.5266, 2.7868]
10,000	2.740	[2.5698, 2.7651]
20,000	2.659	[2.5955, 2.7365]
50,000	2.678	[2.6208, 2.7116]
100,000	2.669	[2.6365, 2.6872]

500, 1,000, 5,000, 10,000, 20,000, 50,000, and 100,000 scenarios. The 99% AVaR is computed from these samples using equation 6.3.1 and the numbers are stored. We repeat the experiment 100 times. In the end, we have 100 AVaR numbers for each sample size. We expect that as the sample size increases, the AVaR values will fluctuate less about the theoretical value, which is $AVaR_{0.01}(X) = 2.665$, $X \in N(0, 1)$.

Table 6.1 contains the result of the experiment. From the 100 AVaR numbers, we calculate the 95% confidence interval reported in the third column. The confidence intervals cover the theoretical value 2.665 and also we notice that the length of the confidence interval decreases as the sample size increases. This effect is illustrated in Figure 6.6 with boxplot diagrams. A sample of 100,000 scenarios results in AVaR numbers which are tightly packed around the true value while a sample of only 500 scenarios may give a very inaccurate estimate.

By comparing Table 6.1 to Table 5.2 in section 5.4.2 of Chapter 5, we notice that the lengths of the 95% confidence intervals for AVaR are larger than the corresponding confidence intervals for VaR. This result is not surprising. Given that both quantities are at the same tail probability of 1%, the AVaR has larger variability than the VaR for a fixed number of scenarios because the AVaR is the average of terms

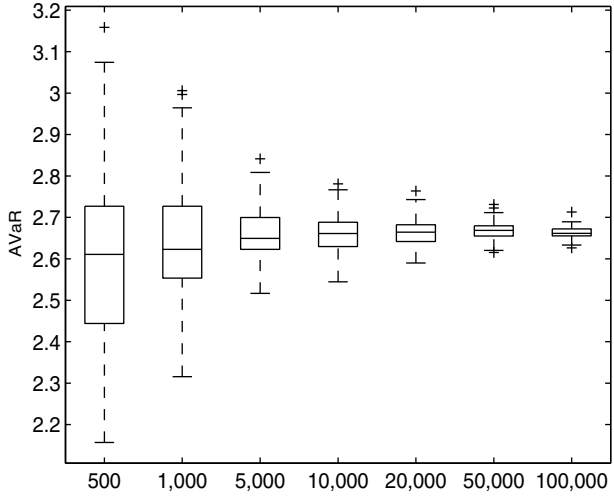

Figure 6.6: Boxplot diagrams of the fluctuation of the AVaR at 1% tail probability of the standard normal distribution based on scenarios. The horizontal axis shows the number of scenarios and the boxplots are computed from 100 independent samples.

fluctuating more than the 1% VaR. This effect is more pronounced the more heavy tailed the distribution is.

6.4.5 Kernel methods

Both the Monte Carlo and the historical method rely on the natural sample AVaR estimator in equation (6.3.1). An advantage of this estimator is its computational simplicity. A disadvantage appears when considering the question of differentiability of portfolio AVaR with respect to portfolio weights.[10] It can be demonstrated that the sample AVaR in (6.3.1) is piece-wise linear with respect to portfolio weights and, therefore, is not everywhere differentiable. As a consequence, in order to be able to calculate derivatives, we need a smooth approximation to portfolio AVaR.

The reason for lack of differentiability can be explained on an intuitive level by looking at the way the formula in (6.3.1) is derived. It

is derived by using the empirical quantile function in the definition of AVaR and that function is a step function. If it were smooth, then portfolio AVaR would be everywhere differentiable with respect to portfolio weights.

If we view the empirical c.d.f. as a step approximation to the theoretical c.d.f., then one way to deal with this issue is to consider a smooth approximation to the theoretical c.d.f. Such a smooth approximation can be constructed with the help of *kernel functions*. Since our discussion is about computing portfolio AVaR, we consider the multivariate case directly.

A kernel function, or simply a *kernel*, is a function $\mathcal{K}(u) : \mathbb{R}^n \to \mathbb{R}^+$ satisfying the following property:

$$\int_{\mathbb{R}^n} \mathcal{K}(u)du = 1.$$

Therefore, a kernel is nothing more than a density function of an n-dimensional random variable. Since we use kernel functions to compute AVaR, we need to impose one additional assumption,

$$\int_{\mathbb{R}^n} u\mathcal{K}(u)du < \infty,$$

which is analogous to the condition that the random variable should have a finite mean in order for its AVaR to be finite.

We consider kernel functions equipped with the property that the kernel of any linear combination of the n-dimensional random variable remain the same parametric family. This technical requirement has a very straightforward interpretation. If we assume an n-dimensional kernel for the stock returns and we calculate, for example, the c.d.f. of portfolio returns, then we obtain a kernel approximation of the portfolio returns c.d.f. in which the kernel belongs to the same parametric family.

One example of a kernel satisfying all conditions is the density of the multivariate normal distribution and, in general, the densities of any multivariate elliptical distribution with a finite mean. In fact, in this case the kernel can be expressed as $\mathcal{K}(u) = K_1(u'u)$, in which

$K_1 : \mathbb{R}^+ \to \mathbb{R}^+$. In the multivariate normal case, for example, $K_1(t) = (2\pi)^{-d/2} e^{-t/2}$.

We choose the multivariate normal density as a kernel function, even though all results can be re-stated for any of the elliptical kernels. Denote by $X^1, \ldots, X^N$ a sample of N observations on a n-dimensional random vector X. We assume that the random vector describes the returns of n stocks. The kernel density estimator of the density function $f_X(x)$, $x \in \mathbb{R}^n$, has the form

$$\widehat{f}_X(x) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\det(H)} \mathcal{K}(H^{-1}(x - X^k)),$$

where H is a matrix called a *bandwidth matrix*. If we assume a multivariate normal kernel, then the kernel density estimator equals

$$\widehat{f}_{w'X}(y) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{\sqrt{2\pi h_w}} \exp\left(\frac{(y - w'X^k)^2}{2h_w} \right),$$

where $h_w = w'\Sigma w$, in which $\Sigma = H'H$. The smoothness of the kernel estimate is inherited by the corresponding smoothness of the kernel itself. As a result, the kernel estimator is infinitely many times differentiable.

Starting from the expression of $\widehat{f}_{w'X}(y)$, we can compute a kernel estimator of the portfolio returns c.d.f.

$$\widehat{F}_{w'X}(y) = \frac{1}{N} \sum_{k=1}^{N} \Phi\left(\frac{y - r_k}{\sqrt{h_w}} \right),$$

where Φ is the c.d.f. of the standard normal distribution.

A comparison between a kernel estimate of a c.d.f. and the sample c.d.f. is shown in Figure 6.7. The empirical c.d.f. is a step function with jumps, while the kernel estimate is a smooth function. Looking at Figure 6.7, the effect of the kernel function can be explained in the following way. The point masses at the observations are "spread out" in a small interval around the observed value. In what way the probability mass gets spread out is determined by the form of the kernel function. The size of the interval depends on the bandwidth

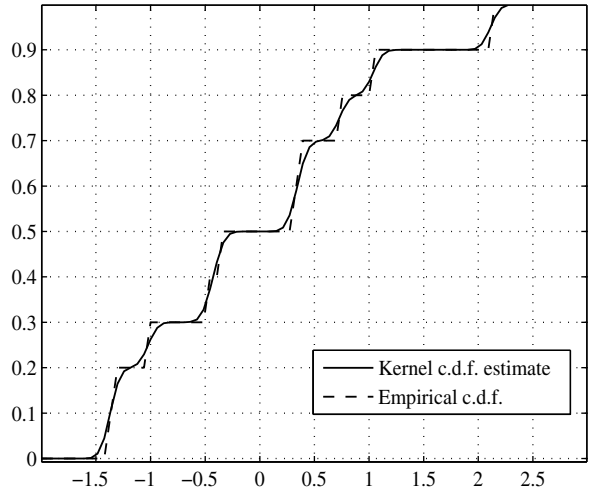

Figure 6.7: A kernel estimate of the c.d.f. compared to the empirical c.d.f.

parameter. The smaller the parameter is, the smaller the interval is and, at the limit, the kernel estimate coincides with the empirical c.d.f.

We can construct an estimator of AVaR through the kernel estimator of the portfolio returns c.d.f. We provide the formula without proof in the text, but the proof, as well as additional information about the estimator, can be found in the appendix to this chapter. The kernel estimator of AVaR equals

$$AVaR_{\epsilon}^{H}(w'X) = -\frac{1}{N\epsilon}\sum_{k=1}^{N}\left(r_k\Phi\left(\frac{q_{\epsilon}^{w}-r_k}{\sqrt{h_w}}\right) - \sqrt{h_w}g\left(\frac{q_{\epsilon}^{w}-r_k}{\sqrt{h_w}}\right)\right), \tag{6.4.3}$$

where Φ and g are the c.d.f. and the density of the standard normal distribution, $r_k = w'X^k$ is the k-th scenario for the portfolio return, $h_w = (Hw)'(Hw)$, and q_{ϵ}^{w} is the solution of the equation $\widehat{F}_{w'X}(y) = \epsilon$. The partial derivatives of $AVaR_{\epsilon}^{H}(w'X)$ of any order with respect to portfolio weights exist.

We emphasize that kernel methods can be used to compute VaR as well. Actually, the kernel estimator of VaR, $VaR_\epsilon^H(w'X)$, appears in (6.4.3) implicitly. It can be computed as

$$VaR_\epsilon^H(w'X) = q_\epsilon^w : \quad \widehat{F}_{w'X}(q_\epsilon^w) = \epsilon. \tag{6.4.4}$$

In order to compare the kernel estimators for VaR and AVaR described in this chapter and the natural sample estimators, we construct a numerical experiment. We generate a number of return scenarios from a two-dimensional distribution and we compute the resulting scenarios for all possible long-only portfolios, $w_1 + w_2 = 1$, $w_1 \geq 0, \;\; w_2 \geq 0$. Then, we calculate $VaR_\epsilon^H(w'X)$ and $AVaR_\epsilon^H(w'X)$ and also the corresponding sample estimators, and we compare them plotted as functions of w_1.

The plots are shown in Figures 6.8 and 6.9. The lack of differentiability of empirical VaR is evident from the top plot on Figure 6.8. The bottom plot contains a section of the top plot zoomed in. This section shows more clearly what happens with $VaR_\epsilon^H(w'X)$ near a point where the sample VaR is not differentiable. It is evident that the kernel estimate is a smooth function of portfolio weights.

In a similar way, Figure 6.9 compares $AVaR_\epsilon^H(w'X)$ and sample AVaR as functions of w_1. The top plot indicates that the kernel estimate is larger than the sample estimate for all portfolios. In the appendix to this chapter, we demonstrate this inequality in a general setting. The bottom plot in Figure 6.9 shows the region where both functions attain their minimums. We can see that empirical AVaR is a convex piece-wise linear function of portfolio weights. In contrast, $AVaR_\epsilon^H(w'X)$ is a smooth function, exactly as theory suggests.

The theory of kernel methods in statistics suggests that the choice of a kernel function is not so important.[11] The bandwidth matrix estimation appears to be a much more significant factor.

There are a number of methods in the literature about bandwidth matrix estimation:

(a) $H = hI$. Setting the smoothing parameter to be constant for every variable implies the amount of smoothing in each

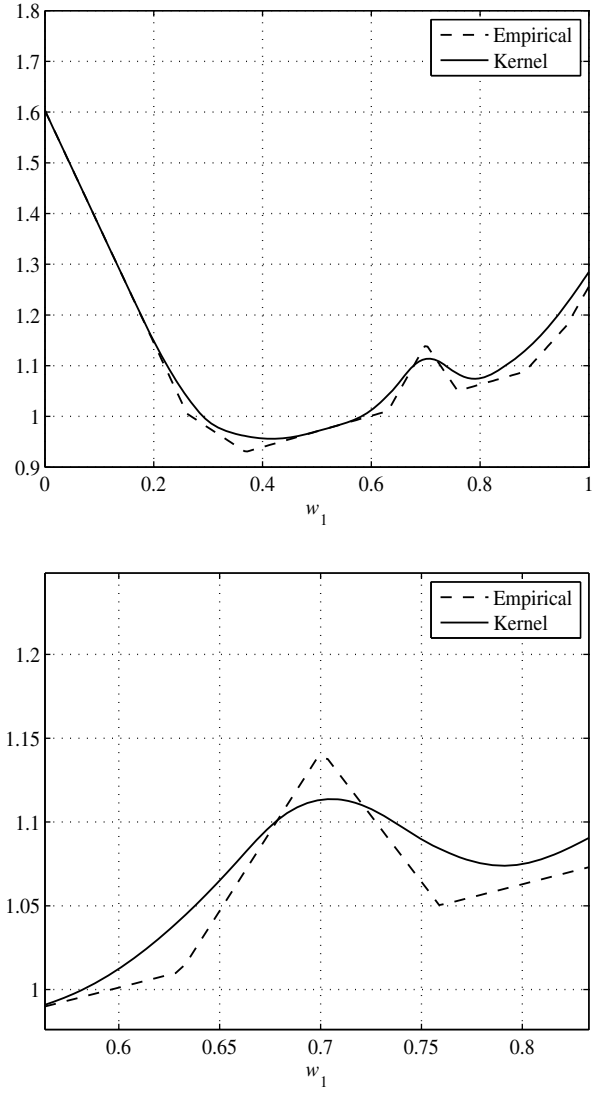

Figure 6.8: The portfolio empirical and the kernel VaRs at 1% tail probability as functions of w_1, $w_1 + w_2 = 1$. The bottom plot shows a zoomed section of the top plot.

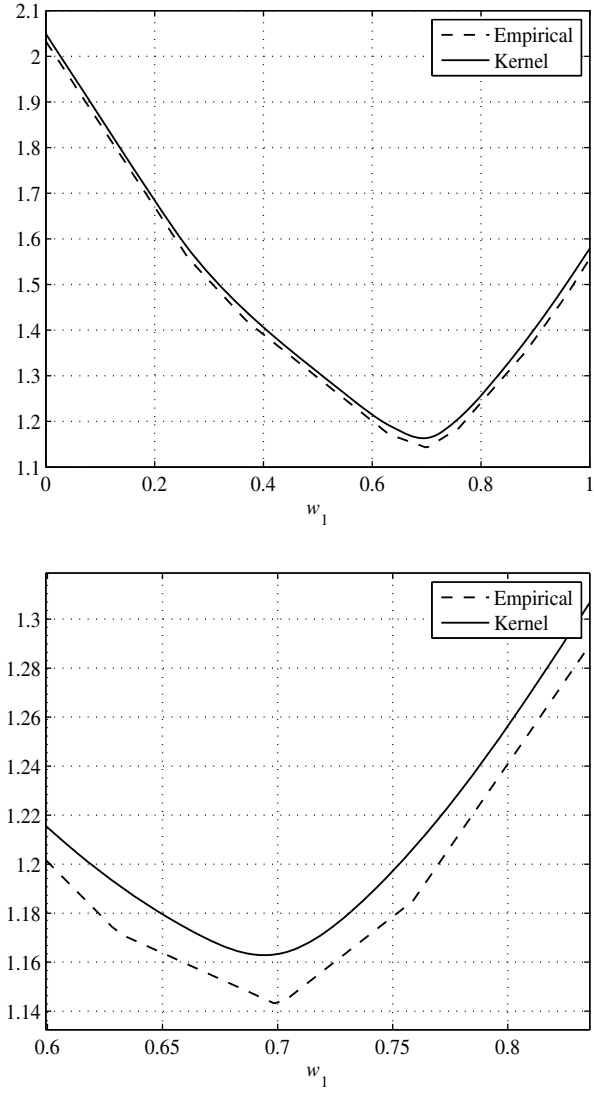

Figure 6.9: The portfolio empirical and the kernel AVaRs at 1% tail probability as functions of w_1, $w_1 + w_2 = 1$. The bottom plot shows a zoomed section of the top plot.

direction is the same. This is sensible only if the scales of all variables are roughly constant.

(b) $H = \text{diag}(h_1, \ldots, h_n)$. This parametrization allows different amounts of smoothing in each coordinate direction. This approach would be a "practical" version of the approach in item 1; if s_j is the scaling constant for the j-th variable, the approach in item 1 is equivalent to using $H = h \times \text{diag}(s_1, \ldots, s_d)$.

(c) $H = h\hat{\Sigma}^{1/2}$, where $\hat{\Sigma}$ is an estimate of the covariance matrix. This is the multivariate generalization of coordinate-wise scaling, since it is equivalent to linearly transforming the data to have unit estimated covariance (often called sphering the data), using a constant bandwidth $H = hI$, and then transforming back to the original scale. The idea is to use a kernel that has the same general shape as the density.

Arguments from statistics can be used to provide a guideline for the scaling coefficient h. For example, if the sample is generated from a normal distribution and we use a normal kernel, then an optimal choice for the bandwidth matrix is given by[12]

$$H = \left(\frac{4}{d+2}\right)^{1/(d+4)} \hat{\Sigma}^{1/2} n^{-1/(d+4)}.$$

6.5 Back-testing of AVaR

Suppose that we have selected a method for calculating the daily AVaR of a portfolio. A reasonable question is how we can verify if the estimates of daily AVaR are realistic.

In section 5.4.2 of Chapter 5, we considered the same issue in the context of VaR and the solution was to carry out a back-testing of VaR. Essentially, VaR back-testing consists of computing the portfolio VaR for each day back in time using the information available up to that day only. In this way, we have the VaR numbers back in time as if we had used exactly the same methodology in the past. On the basis of the VaR numbers and the realized portfolio returns, we can use

statistical methods to assess whether the forecasted loss at the VaR tail probability is consistent with the observed losses. If there are too many observed losses larger than the forecasted VaR, then the model is too optimistic. Conversely, if there are too few losses larger than the forecasted VaR, then the model is too pessimistic.

Note that in the case of VaR back-testing, we are simply counting the cases in which there is an exceedance: that is, when the size of the observed loss is larger than the predicted VaR. The magnitude of the exceedance is immaterial for the statistical test.

Unlike VaR, back-testing of AVaR is not straightforward and is a much more challenging task. By definition, the AVaR at tail probability ϵ is the average of VaRs larger than the VaR at tail probability ϵ. Thus, the most direct approach to test AVaR would be to perform VaR back-tests at all tail probabilities smaller than ϵ. If all these VaRs are correctly modeled, then so is the corresponding AVaR.

One general issue with this approach is that it is impossible to perform in practice. Suppose that we consider the AVaR at tail probability of 1%, for example. Back-testing VaRs deeper in the tail of the distribution can be infeasible because the back-testing time window is too short. The lower the tail probability, the larger time window we need in order for the VaR test to be conclusive. Another general issue is that this approach is too demanding. Even if the VaR back-testing fails at some tail probability ϵ_1 below ϵ, this does not necessarily mean that the AVaR is incorrectly modeled because the test failure may be due to purely statistical reasons and not to incorrect modeling.

These arguments illustrate why AVaR back-testing is a difficult problem – we need the information about the entire tail of the return distribution describing the losses larger than the VaR at tail probability ϵ and there may be too few observations from the tail upon which to base the analysis. For example, in one business year, there are typically 250 trading days. Therefore, a one-year back-testing results in 250 daily portfolio returns, which means that if $\epsilon = 1\%$, then there are only 2 observations available from the losses larger than the VaR at 1% tail probability.

As a result, in order to be able to back-test AVaR, we can assume a certain "structure" of the tail of the return distribution which

would compensate for the lack of observations. There are two general approaches:

(a) Use the tails of the Lévy stable distributions[13] as a proxy for the tail of the loss distribution and take advantage of the practical semi-analytic formula for the AVaR given in the appendix to this chapter to construct a statistical test.
(b) Make the weaker assumption that the loss distribution belongs to the domain of attraction of a max-stable distribution. Thus, the behavior of the large losses can be approximately described by the limit max-stable distribution and a statistical test can be based on it.

The rationale of the first approach is that, generally, the Lévy stable distribution provides a good fit to the stock returns data and, thus, the stable tail may turn out to be a reasonable approximation. Moreover, from the generalized Central Limit Theorem we know that stable distributions have domains of attraction which makes them an appealing candidate for an approximate model.

The second approach is based on a weaker assumption. The family of max-stable distributions arises as the limit distribution of properly scaled and centered maxima of i.i.d. random variables. If the random variable describes portfolio losses, then the limit max-stable distribution can be used as a model for the large losses (i.e., the ones in the tail). Unfortunately, as a result of the weaker assumption, estimators of poor quality have to be used to estimate the parameters of the limit max-stable distribution, such as the Hill estimator. This represents the basic trade-off in this approach.

6.6 Spectral Risk Measures

By definition, the AVaR at tail probability ϵ is the average of the VaRs larger than the VaR at tail probability ϵ. It appears possible to obtain a larger family of coherent risk measures by considering the weighted average of the VaRs instead of simple average. Thus, the

AVaR becomes just one representative of this larger family, which is known as *spectral risk measures*. Acerbi (2004) provides a detailed description of spectral risk measures.

Spectral risk measures are defined as,[14]

$$\rho_\phi(X) = \int_0^1 VaR_p(X)\phi(p)dp, \tag{6.6.1}$$

where $\phi(p), p \in [0, 1]$ is the weighting function also known as *the risk spectrum* or *risk-aversion function*. It has the following interpretation. Consider a small interval $[p_1, p_2]$ of tail probabilities with length $p_2 - p_1 = \Delta p$. The weight corresponding to this interval is approximately equal to $\phi(p_1) \times \Delta p$. Thus, the VaRs at tail probabilities belonging to this interval have approximately the weight $\phi(p_1) \times \Delta p$.

The risk-aversion function should possess some properties in order for $\rho_\phi(X)$ to be a coherent risk measure. It should be:

Positive	$\phi(p) \geq 0, p \in [0, 1]$.
Non-increasing	Larger losses are multiplied by larger weights, $\phi(p_1) \geq \phi(p_2), p_1 \leq p_2$.
Normed	All weights should sum up to 1, $\int_0^1 \phi(p)dp = 1$.

If we compare equations (6.6.1) and (6.2.1), we notice that the AVaR at tail probability ϵ arises from a spectral risk measure with a constant risk-aversion function for all tail probabilities below ϵ. The left plot in Figure 6.10 illustrates a typical risk-aversion function. The right plot shows the graph of the risk-aversion function yielding the AVaR at tail probability ϵ.

It is possible to obtain formulae through which we can estimate the spectral risk measures from a sample of observations. They are essentially counterparts of (6.3.1) and (6.3.2): see Acerbi and Simonetti (2002) for further details.

In section 5.4.2 of Chapter 5 and section 6.4 of this chapter, we emphasized that if a sample is used to estimate VaR and AVaR, then there is certain variability of the estimates. We illustrated it through a Monte Carlo example for the standard normal distribution.

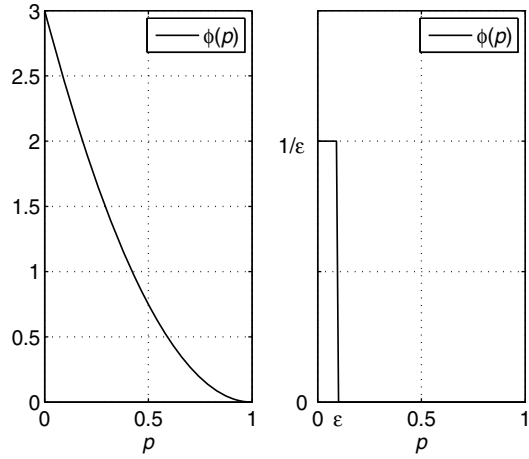

Figure 6.10: Examples of risk-aversion functions. The right plot shows the risk-aversion function yielding the AVaR at tail probability ϵ.

Comparing the results, we concluded that the variability of AVaR is larger than the VaR at the same tail probability because in the AVaR, we average terms with larger variability. The heavier the tail, the more pronounced this effect becomes.

When spectral risk measures are estimated from a sample, the variability of the estimate may become a big issue. Note that due to the non-increasing property of the risk-aversion function, the larger losses, which are deeper in the tail of the return distribution, are multiplied by a larger weight. The larger losses (VaRs at lower tail probability) have higher variability and the multiplication by a larger weight further increases the variability of the weighted average. Therefore, a larger number of scenarios may turn out to be necessary to achieve given stability of the estimate for spectral risk measures than for AVaR. Ultimately, this is dependent on the choice of the risk-aversion function and the assumed distribution of portfolio return.

In fact, the distributional assumption for the random variable X is very important because it may lead to unbounded spectral risk measures for some choices of the risk-aversion function. An infinite risk

measure is not informative for decision makers and an unfortunate combination of a distributional model and a risk-aversion function cannot be identified by looking at the sample estimate of $\rho_\phi(X)$. In practice, when $\rho_\phi(X)$ is divergent in theory, we will observe high variability of the risk estimates when regenerating the simulations and also non-decreasing variability of the risk estimates as we increase the number of simulations. We can regard these effects as symptoms for a bad combination of a statistical model and a risk-aversion function. The appendix to this chapter contains guidelines for avoiding inappropriate choices of a risk-aversion function depending on certain information about the probability distribution of X.

We would like to stress that this problem does not exist for AVaR because a finite mean of X guarantees that the AVaR is well defined on all tail probability levels. The problem for the spectral measures of risk arises from the non-increasing property of the risk-aversion function. Larger losses are multiplied by larger weights, which may result in an unbounded weighted average.

6.7 Risk Measures and Probability Metrics

In Chapter 4, we introduced the notion of probability metrics and remarked that they provide the only way of measuring distances between random quantities. It turns out that a small distance between random quantities does not necessarily imply that selected characteristics of those quantities will be close to each other. For example, a probability metric may indicate that two distributions are close to each other and, still, the standard deviations of the two distributions may be arbitrarily different. As a very extreme case, one of the distributions may even have an infinite standard deviation. Thus, if we want small distances measured by a probability metric to imply similar characteristics, the probability metric should be carefully chosen.

A risk measure can be viewed as calculating a particular characteristic of a random variable. Furthermore, there are problems in finance in which the goal is to find a random variable closest to

another random variable. For instance, such is the benchmark tracking problem which is at the heart of passive portfolio construction strategies. Essentially, we are trying to construct a portfolio tracking the performance a given benchmark. In some sense, this can be regarded as finding a portfolio return distribution which is closest to the return distribution of the benchmark. Usually, the distance is measured through the tracking error, which is the standard deviation of the active return.

Suppose that we have found the portfolio tracking the benchmark most closely with respect to the tracking error. Can we be sure that the risk of the portfolio is close to the risk of the benchmark? Generally, the answer is affirmative only if we use the standard deviation as a risk measure. Active return is refined as the difference between the portfolio return r_p and the benchmark return r_b, $r_p - r_b$. The conclusion that smaller tracking error implies that the standard deviation of r_p is close to the standard deviation of r_b is based on the inequality,

$$|\sigma(r_p) - \sigma(r_b)| \leq \sigma(r_p - r_b).$$

The right part corresponds to the tracking error and, therefore, smaller tracking error results in $\sigma(r_p)$ being closer to $\sigma(r_b)$.

In order to guarantee that small distance between portfolio return distributions corresponds to similar risks, we have to find a suitable probability metric. Technically, for a given risk measure we need to find a probability metric with respect to which the risk measure is a continuous functional,

$$|\rho(X) - \rho(Y)| \leq \mu(X, Y),$$

where ρ is the risk measure and μ stands for the probability metric. We continue with examples of how this can be done for VaR, AVaR, and the spectral risk measures.[15]

VaR

Suppose that X and Y describe the return distributions of two portfolios. The absolute difference between the VaRs of the two portfolios

at any tail probability can be bounded by,

$$\begin{aligned} |VaR_\epsilon(X) - VaR_\epsilon(Y)| &\leq \max_{p\in(0,1)} |VaR_p(X) - VaR_p(Y)| \\ &= \max_{p\in(0,1)} |F_Y^{-1}(p) - F_X^{-1}(p)| \\ &= \mathbf{W}(X, Y), \end{aligned}$$

where $\mathbf{W}(X, Y)$ is the uniform metric between inverse distribution functions defined in Chapter 2. If the distance between X and Y is small, as measured by the metric $\mathbf{W}(X, Y)$, then the VaR of X is close to the VaR of Y at any tail probability level ϵ.

AVaR

Suppose that X and Y describe the return distributions of two portfolios. The absolute difference between the AVaRs of the two portfolios at any tail probability can be bounded by

$$\begin{aligned} |AVaR_\epsilon(X) - AVaR_\epsilon(Y)| &\leq \tfrac{1}{\epsilon} \int_0^\epsilon |F_X^{-1}(p) - F_Y^{-1}(p)| dp \\ &\leq \int_0^1 |F_X^{-1}(p) - F_Y^{-1}(p)| dp \\ &= \kappa(X, Y), \end{aligned}$$

where $\kappa(X, Y)$ is the Kantorovich metric defined in Chapter 2. If the distance between X and Y is small, as measured by the metric $\kappa(X, Y)$, then the AVaR of X is close to the AVaR of Y at any tail probability level ϵ. Note that the quantity

$$\kappa_\epsilon(X, Y) = \frac{1}{\epsilon} \int_0^\epsilon |F_X^{-1}(p) - F_Y^{-1}(p)| dp$$

can also be used to bound the absolute difference between the AVaRs. It is a probability semimetric giving the best possible upper bound on the absolute difference between the AVaRs.

Spectral risk measures

Suppose that X and Y describe the return distributions of two portfolios. The absolute difference between the spectral risk measures of

the two portfolios for a given risk-aversion function can be bounded by,

$$|\rho_\phi(X) - \rho_\phi(Y)| \leq \int_0^1 |F_X^{-1}(p) - F_Y^{-1}(p)|\phi(p)dp$$
$$= \kappa_\phi(X, Y),$$

where $\kappa_\phi(X, Y)$ is a weighted Kantorovich metric. If the distance between X and Y is small, as measured by the metric $\kappa_\phi(X, Y)$, then the risk of X is close to the risk of Y as measured by the spectral risk measure ρ_ϕ.

6.8 Risk Measures Based on Distortion Functionals

We introduced spectral risk measures as functionals computing a weighted average of the VaRs at all possible tail probabilities. The weighting function should satisfy certain conditions in order for spectral risk measures to be coherent. Basically, the weights should be positive and the deeper we go into the left tail of the return distribution, the larger the weight should be. In a certain sense, the spectral risk measure is a *pessimistic* model about the average loss.

In a similar but more general way, we can introduce a pessimistic model about the average loss through the functional

$$\rho_H(X) = -\int_0^1 F_X^{-1}(p)dH(p), \tag{6.8.1}$$

where $H(p) : [0, 1] \to [0, 1]$ is a bounded right continuous increasing function, and F_X^{-1} is the inverse c.d.f. of the random variable X describing the return on a common stock. Functionals such as (6.8.1) are also known as *distortion functionals*.[16] From a more general viewpoint, we can think of H as a distribution function of a non-negative probability measure on the interval $[0, 1]$.

The pessimistic model about the average loss of X as introduced in (6.8.1) can be a coherent risk measure if we assume that $H(p)$ is

a concave function. In this case we say that $\rho_H(X)$ is a *distortion risk measure*.[17]

Spectral risk measures as defined in (6.6.1) arise from distortion risk measures if $H(p)$ is differentiable. Under this assumption, the risk-aversion function $\phi(p) = H'(p)$. It can be directly verified that $\phi(p) = H'(p)$ is a decreasing function due to the assumed concavity of $H(p)$ and satisfies all properties of risk-aversion functions stated in section 6.6.

Among all distortion risk measures, AVaR has a very special place. There is a representation result, according to which AVaR plays the role of a building block. Every distortion risk measure $\rho_H(X)$ can be viewed as a weighted average of AVaRs,

$$\rho_H(X) = \int_0^1 AVaR_p(X) dM(p), \tag{6.8.2}$$

where M is a monotonically increasing function and satisfies $M(0) = 0$, and $M(1) = 1$. The proof of this representation result can be found in Pflug and Roemisch (2007), for example.

6.9 Summary

In this chapter, we considered in detail the AVaR risk measure. We noted the advantages of AVaR, described a number of methods for its calculation and estimation, and remarked some potential pitfalls including estimates variability and problems in AVaR back-testing. We illustrated geometrically many of the formulae for AVaR calculation, which makes them more intuitive and easier to understand.

Besides the AVaR, we considered a more general family of coherent risk measures – the spectral risk measures. The AVaR is a spectral risk measure with a specific risk-aversion function. We emphasized the importance of proper selection of the risk-aversion function to avoid explosion of the risk measure.

We discussed the more general concept of distortion risk measures, which include spectral risk measures as a special case. AVaR plays the important role of a building block, since any distortion risk measure

can be represented as an average of AVaRs at different tail probability levels.

Finally, we demonstrated a connection between the theory of probability metrics and risk measures. Basically, by choosing an appropriate probability metric we can guarantee that if two portfolio return distributions are close to each other, their risk profiles are also similar.

6.10 Technical Appendix

In this appendix, we start with a more general view that better describes the conditional loss distribution in terms of certain characteristics in which AVaR appears as a special case. We continue with the notion of higher-order AVaR, generating a family of coherent risk measures. Next, we provide an intuitive geometric interpretation of the minimization formula for the AVaR calculation. We also provide a semi-analytic expression for the AVaR of stable distributions and compare the expected tail loss measure to AVaR. Finally, we comment on the proper choice of a risk-aversion function in spectral risk measures which does not result in an infinite risk measure.

6.10.1 Characteristics of conditional loss distributions

In the chapter, we defined AVaR as a risk measure and showed how it can be calculated in practice. While it is an intuitive and easy-to-use coherent risk measure, AVaR represents the average of the losses larger than the VaR at tail probability ϵ, which is only one characteristic of the distribution of extreme losses. We remarked that if the distribution function is continuous, then AVaR coincides with ETL, which is the mathematical expectation of the conditional loss distribution. Besides the mathematical expectation, there are other important characteristics of the conditional loss distribution. For example, AVaR does not provide any information about how dispersed the conditional losses are around the AVaR value. In this section, we state a couple of families of useful characteristics in which AVaR appears as one example.

Consider the following tail moment of order n at tail probability ϵ,

$$m_\epsilon^n(X) = \frac{1}{\epsilon}\int_0^\epsilon (F_X^{-1}(t))^n dt, \tag{6.10.1}$$

where $n = 1, 2, \ldots$, $F_X^{-1}(t)$ is the inverse c.d.f. of the random variable X. If the distribution function of X is continuous, then the tail moment of order n can be represented through the following conditional expectation,

$$m_\epsilon^n(X) = E(X^n | X < VaR_\epsilon(X)), \tag{6.10.2}$$

where $n = 1, 2, \ldots$ In the general case, if the c.d.f. has a jump at $VaR_\epsilon(X)$, a link exists between the conditional expectation and equation (6.10.1), which is similar to formula (6.10.12) for AVaR. In fact, AVaR appears as the negative of the tail moment of order 1, $AVaR_\epsilon(X) = -m_\epsilon^1(X)$.

The higher-order tail moments provide additional information about the conditional distribution of the extreme losses. We can make a parallel with the way the moments of a random variable are used to describe certain properties of it. In our case, it is the conditional distribution that we are interested in.

In addition to the moments $m_\epsilon^n(X)$, we introduce the central tail moments of order n at tail probability ϵ,

$$M_\epsilon^n(X) = \frac{1}{\epsilon}\int_0^\epsilon (F_X^{-1}(t) - m_\epsilon^1(X))^n dt, \tag{6.10.3}$$

where $m_\epsilon^1(X)$ is the tail moment of order 1. If the distribution function is continuous, then the central moments can be expressed in terms of the conditional expectation,

$$M_\epsilon^n(X) = E((X - m_\epsilon^1(X))^n | X < VaR_\epsilon(X)).$$

The tail variance of the conditional distribution appears as $M_\epsilon^2(X)$ and the tail standard deviation equals

$$(M_\epsilon^2(X))^{1/2} = \left(\frac{1}{\epsilon}\int_0^\epsilon (F_X^{-1}(t) - m_\epsilon^1(X))^2 dt\right)^{1/2}.$$

There is a formula expressing the tail variance in terms of the tail moments introduced in (6.10.2),

$$\begin{aligned} M_\epsilon^2(X) &= m_\epsilon^2(X) - (m_\epsilon^1(X))^2 \\ &= m_\epsilon^2(X) - (AVaR_\epsilon(X))^2. \end{aligned}$$

This formula is similar to the representation of variance in terms of the first two moments,

$$\sigma_X^2 = EX^2 - (EX)^2.$$

The tail standard deviation can be used to describe the dispersion of conditional losses around AVaR as it satisfies the general properties of dispersion measures given in section 5.2.4 of Chapter 5. It can be viewed as complementary to AVaR in the sense that if there are two portfolios with equal AVaRs of their return distributions but different tail standard deviations, the portfolio with the smaller standard deviation is preferable.

Another central tail moment which can be interpreted is $M_\epsilon^3(X)$. After proper normalization, it can be employed to measure the skewness of the conditional loss distribution. In fact, if the tail probability is sufficiently small, the tail skewness will be quite significant. In the same fashion, by normalizing the central tail moment of order 4, we obtain a measure of kurtosis of the conditional loss distribution.
In a similar way, we introduce the absolute central tail moments of order n at tail probability ϵ,

$$\mu_\epsilon^n(X) = \frac{1}{\epsilon}\int_0^\epsilon |F_X^{-1}(t) - m_\epsilon^1(X)|^n dt. \tag{6.10.4}$$

The tail moments $\mu_\epsilon^n(X)$ raised to the power of $1/n$, $(\mu_\epsilon^n(X))^{1/n}$, can be applied as measures of dispersion of the conditional loss distribution if the distribution is such that they are finite.

In the chapter, we remarked that the tail of the random variable can be so heavy that AVaR becomes infinite. Even if it is theoretically finite, it can be hard to estimate because the heavy tail will result in the AVaR estimator having a large variability. Thus, under certain conditions it may turn out to be practical to employ a robust estimator

instead. The *median tail loss* (MTL), defined as the median of the conditional loss distribution, is a robust alternative to AVaR. It has the advantage of always being finite no matter the tail behavior of the random variable. Formally, it is defined as

$$MTL_\epsilon(X) = -F_X^{-1}(1/2|X < -VaR_\epsilon(X)), \tag{6.10.5}$$

where $F_X^{-1}(p|X < -VaR_\epsilon(X))$ stands for the inverse distribution function of the c.d.f. of the conditional loss distribution

$$\begin{aligned} F_X(x|X < -VaR_\epsilon(X)) &= P(X \leq x|X < -VaR_\epsilon(X)) \\ &= \begin{cases} P(X \leq x)/\epsilon, & x < -VaR_\epsilon(X) \\ 1, & x \geq -VaR_\epsilon(X). \end{cases} \end{aligned}$$

In effect, MTL, as well as any other quantile of the conditional loss distribution, can be directly calculated as a quantile of the distribution of X,

$$\begin{aligned} MTL_\epsilon(X) &= -F_X^{-1}(\epsilon/2) \\ &= VaR_{\epsilon/2}(X), \end{aligned} \tag{6.10.6}$$

where $F_X^{-1}(p)$ is the inverse c.d.f. of X and ϵ is the tail probability of the corresponding VaR in equation (6.10.5). Thus, MTL shares the properties of VaR. Equation (6.10.7) shows that MTL is not a coherent risk measure even though it is a robust alternative to AVaR, which is a coherent risk measure.

In the universe of the three families of moments that we introduced, AVaR is one special case providing only limited information. It may be the only coherent risk measure among them, but the other moments can be employed in addition to AVaR in order to gain more insight into the conditional loss distribution. Furthermore, it could appear that other reasonable risk measures can be based on some of the moments. Thus, we believe that they all should be considered in financial applications.

6.10.2 Higher-order AVaR

By definition, AVaR is the average of VaRs larger than the VaR at tail probability ϵ. In the same fashion, we can pose the question of what happens if we average all AVaRs larger than the AVaR at tail probability ϵ. In fact, this quantity is an average of coherent risk measures and, therefore, is a coherent risk measure itself since it satisfies all defining properties of coherent risk measures given in section 5.4.4 of Chapter 5. We call it *AVaR of order 1* and denote it by $AVaR_\epsilon^{(1)}(X)$ because it is a derived quantity from AVaR. In this section, we consider similar derived quantities from AVaR, which we call *higher-order AVaRs*.

Formally, the AVaR of order 1 is represented in the following way:

$$AVaR_\epsilon^{(1)}(X) = \frac{1}{\epsilon}\int_0^\epsilon AVaR_p(X)dp$$

where $AVaR_p(X)$ is the AVaR at tail probability p. Replacing AVaR by the definition given in equation (6.2.1), we obtain

$$\begin{aligned} AVaR_\epsilon^{(1)}(X) &= -\frac{1}{\epsilon}\int_0^\epsilon \left(\int_0^1 F_X^{-1}(y)g_p(y)dy\right)dp \\ &= -\frac{1}{\epsilon}\int_0^1 F_X^{-1}(y)\left(\int_0^\epsilon g_p(y)dp\right)dy \end{aligned}$$

where

$$g_p(y) = \begin{cases} 1/p, & y \in [0,p] \\ 0, & y > p. \end{cases}$$

and after certain algebraic manipulations, we get the expression

$$\begin{aligned} AVaR_\epsilon^{(1)}(X) &= -\frac{1}{\epsilon}\int_0^\epsilon F_X^{-1}(y)\log\frac{\epsilon}{y}dy \\ &= \int_0^\epsilon VaR_y(X)\phi_\epsilon(y)dy. \end{aligned} \tag{6.10.7}$$

In effect, the AVaR of order 1 can be expressed as a weighted average of VaRs larger than the VaR at tail probability ϵ with a weighting

function $\phi_\epsilon(y)$ equal to

$$\phi_\epsilon(y) = \begin{cases} \frac{1}{\epsilon}\log\frac{\epsilon}{y}, & 0 \le y \le \epsilon \\ 0, & \epsilon < y \le 1. \end{cases}$$

The AVaR of order 1 can be viewed as a spectral risk measure with $\phi_\epsilon(y)$ being the risk-aversion function.

Similarly, we define the higher-order AVaR through the recursive equation

$$AVaR_\epsilon^{(n)}(X) = \frac{1}{\epsilon}\int_0^\epsilon AVaR_p^{(n-1)}(X)dp \tag{6.10.8}$$

where $AVaR_p^{(0)}(X) = AVaR_p(X)$ and $n = 1, 2\ldots$ Thus, the AVaR of order 2 equals the average of AVaRs of order 1 which are larger than the AVaR of order 1 at tail probability ϵ. The AVaR of order n appears as an average of AVaRs of order $n-1$.

The quantity $AVaR_\epsilon^{(n)}(X)$ is a coherent risk measure because it is an average of coherent risk measures. This is a consequence of the recursive definition in (6.10.8). It is possible to show that AVaR of order n admits the representation

$$AVaR_\epsilon^{(n)}(X) = \frac{1}{\epsilon}\int_0^\epsilon VaR_y(X)\frac{1}{n!}\left(\log\frac{\epsilon}{y}\right)^n dy \tag{6.10.9}$$

and $AVaR_\epsilon^{(n)}(X)$ can be viewed as a spectral risk measure with a risk-aversion function equal to

$$\phi_\epsilon^{(n)}(y) = \begin{cases} \frac{1}{\epsilon n!}\left(\log\frac{\epsilon}{y}\right)^n, & 0 \le y \le \epsilon \\ 0, & \epsilon < y \le 1. \end{cases}$$

As a simple consequence of the definition, the sequence of higher-order AVaRs is monotonic,

$$AVaR_\epsilon(X) \le AVaR_\epsilon^{(1)}(X) \le \ldots \le AVaR_\epsilon^{(n)}(X) \le \ldots$$

In the chapter, we remarked that if the random variable X has a finite mean, $E|X| < \infty$, then AVaR is also finite. This is not true for spectral

risk measures and the higher-order AVaR in particular. In line with the general theory developed in section 6.10.6 in this appendix, $AVaR_\epsilon^{(n)}(X)$ is finite if all moments of X exist. For example, if the random variable X has an exponential tail, then $AVaR_\epsilon^{(n)}(X) < \infty$ for any $n < \infty$.

6.10.3 The minimization formula for AVaR

In this section, we provide a geometric interpretation of the minimization formula (6.2.2) for AVaR. We restate equation (6.2.2) in the following equivalent form,

$$AVaR_\epsilon(X) = \frac{1}{\epsilon} \min_{\theta \in \mathbb{R}} (\epsilon\theta + E(-X - \theta)_+) \tag{6.10.10}$$

where $(x)_+ = \max(x, 0)$. Note the similarity between equation (6.10.10) and the definition of AVaR in (6.2.1). Instead of the integral of the quantile function in the definition of AVaR, a minimization formula appears in (6.10.10). We interpreted the integral of the inverse c.d.f. as the shaded area in Figure 6.2. Similarly, we will find the area corresponding to the objective function in the minimization formula and we will demonstrate that as θ changes, there is a minimal area which coincides with the area corresponding to the shaded area in Figure 6.2. Moreover, the minimal area is attained for $\theta = VaR_\epsilon(X)$ when the c.d.f. of X is continuous at $VaR_\epsilon(X)$. In fact, all illustrations in this section are based on the assumption that X has a continuous distribution function.

Consider first the expectation in equation (6.10.10). Assuming that X has a continuous c.d.f., we obtain an expression for the expectation involving the inverse c.d.f.,

$$\begin{aligned} E(-X - \theta)_+ &= \int_{\mathbb{R}} \max(-x - \theta, 0) dF_X(x) \\ &= \int_0^1 \max(-F_X^{-1}(t) - \theta, 0) dt \\ &= -\int_0^1 \min(F_X^{-1}(t) + \theta, 0) dt. \end{aligned}$$

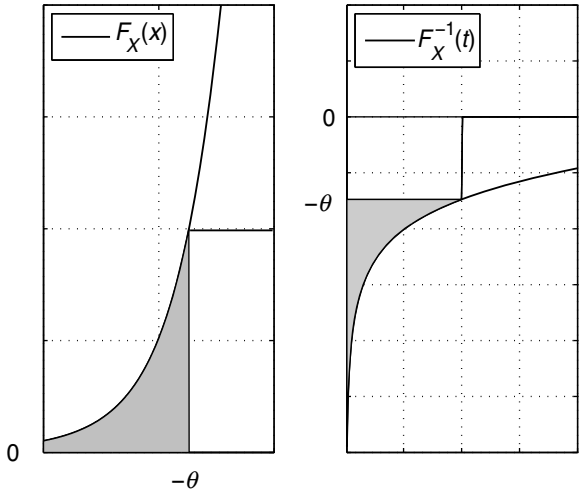

Figure 6.11: The shaded area is equal to the expectation $E(-X-\theta)_+$ in which X has a continuous distribution function.

This representation implies that the expectation $E(-X-\theta)_+$ equals the area closed between the graph of the inverse c.d.f. and a line parallel to the horizontal axis passing through the point $(0, -\theta)$. This is the shaded area on the right plot in Figure 6.11. The same area can be represented in terms of the c.d.f. This is done on the left plot in Figure 6.11.

Let us get back to equation (6.10.10). The tail probability ϵ is fixed. The product $\epsilon \times \theta$ equals the area of a rectangle with sides equal to ϵ and θ. This area is added to $E(-X-\theta)_+$. Figure 6.12 shows the two areas together. The shaded areas on the top and the bottom plots equal $\epsilon \times AVaR_\epsilon(X)$. The top plot shows the case in which $-\theta < -VaR_\epsilon(X)$. Comparing the plot to Figure 6.11, we find out that adding the marked area to the shaded area, we obtain the total area corresponding to the objective in the minimization formula, $\epsilon\theta + E(-X-\theta)_+$. If $-\theta > -VaR_\epsilon(X)$, then we obtain a similar case shown on the bottom plot. Again, adding the marked area to the shaded area, we obtain the the total area computed by the objective in the minimization formula. By varying θ, the total area

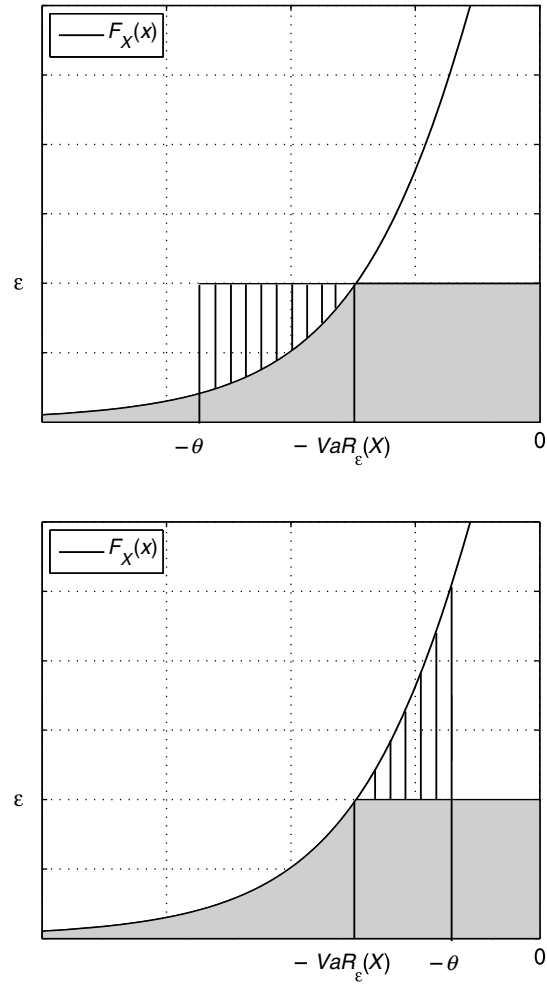

Figure 6.12: The marked area is in addition to the shaded one. The marked area is equal to zero if $\theta = VaR_\epsilon(X)$.

changes but it always remains larger than the shaded area unless $\theta = VaR_\epsilon(X)$.

Thus, when $\theta = VaR_\epsilon(X)$ the minimum area is attained which equals exactly $\epsilon \times AVaR_\epsilon(X)$. According to equation (6.10.10), we

have to divide the minimal area by ϵ in order to obtain the AVaR. As a result, we have demonstrated that the minimization formula in equation (6.2.2) calculates the AVaR.

6.10.4 ETL vs AVaR

The expected tail loss and the average value-at-risk are two related concepts. In the chapter, we remarked that ETL and AVaR coincide if the portfolio return distribution is continuous at the corresponding VaR level. However, if there is a discontinuity, or a point mass, then the two notions diverge. Still, the AVaR can be expressed through the ETL and the VaR at the same tail probability. In this section, we illustrate this relationship and show why the AVaR is more appealing. Moreover, it will throw light on why equation (6.3.1) should be used when considering a sample of observations.

The ETL at tail probability ϵ is defined as the average loss provided that the loss exceeds the VaR at tail probability ϵ,

$$ETL_\epsilon(X) = -E(X|X < -VaR_\epsilon(X)). \tag{6.10.11}$$

As a consequence of the definition, the ETL can be expressed in terms of the c.d.f. and the inverse c.d.f. Suppose, additionally, that the c.d.f. of X has a jump at $-VaR_\epsilon(X)$. In this case, the loss $VaR_\epsilon(X)$ occurs with probability equal to the size of the jump and, because of the strict inequality in (6.10.11), it will not be included in the average.

Figure 6.13 shows the graphs of the c.d.f. and the inverse c.d.f. of a random variable with a point mass at $-VaR_\epsilon(X)$. If ϵ splits the jump of the c.d.f. as on the left plot in Figure 6.13, then the ETL at tail probability ϵ equals

$$\begin{aligned} ETL_\epsilon(X) &= -E(X|X < -VaR_\epsilon(X)) \\ &= -E(X|X < -VaR_{\epsilon_0}(X)) \\ &= ETL_{\epsilon_0}(X). \end{aligned}$$

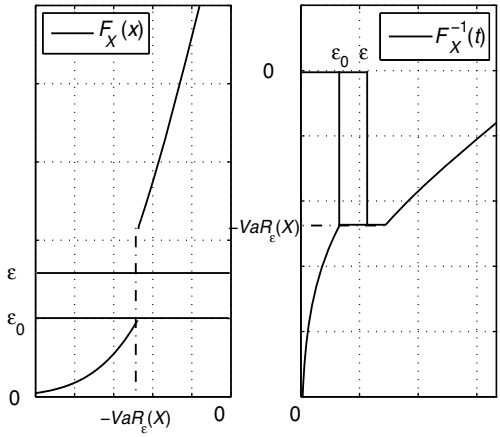

Figure 6.13: The c.d.f. and the inverse c.d.f. of a random variable X with a point mass at $-VaR_\epsilon(X)$. The tail probability ϵ splits the jump of the c.d.f.

In terms of the inverse c.d.f., the quantity $ETL_{\epsilon_0}(X)$ can be represented as

$$ETL_{\epsilon_0}(X) = -\frac{1}{\epsilon_0}\int_0^{\epsilon_0} F_X^{-1}(t)dt.$$

The relationship between AVaR and ETL follows directly from the definition of AVaR.[18] Suppose that the c.d.f. of the random variable X is as on the left plot in Figure 6.13. Then,

$$\begin{aligned} AVaR_\epsilon(X) &= -\frac{1}{\epsilon}\int_0^{\epsilon} F_X^{-1}(t)dt \\ &= -\frac{1}{\epsilon}\left(\int_0^{\epsilon_0} F_X^{-1}(t)dt + \int_{\epsilon_0}^{\epsilon} F_X^{-1}(t)dt\right) \\ &= -\frac{1}{\epsilon}\int_0^{\epsilon_0} F_X^{-1}(t)dt + \frac{\epsilon-\epsilon_0}{\epsilon}VaR_\epsilon(X). \end{aligned}$$

where the last inequality holds because the inverse c.d.f. is flat in the interval $[\epsilon_0, \epsilon]$ and the integral is merely the surface of the rectangle shown on the right plot in Figure 6.13. The integral in the first summand can be related to the ETL at tail probability ϵ and, finally, we

arrive at the expression

$$AVaR_\epsilon(X) = \frac{\epsilon_0}{\epsilon} ETL_\epsilon(X) + \frac{\epsilon - \epsilon_0}{\epsilon} VaR_\epsilon(X). \tag{6.10.12}$$

Equation (6.10.12) shows that $AVaR_\epsilon(X)$ can be represented as a weighted average between the ETL and the VaR at the same tail probability as the coefficients in front of the two summands are positive and sum up to 1. In the special case in which there is no jump, or if $\epsilon = \epsilon_1$, then AVaR equals ETL.

Why is equation (6.10.12) important if in all statistical models we assume that the random variables describing return or payoff distribution have densities? Under this assumption, not only are the corresponding c.d.f.s continuous but they are also smooth. Equation (6.10.12) is important because if the estimate of AVaR is based on the Monte Carlo method, then we use a sample of scenarios which approximate the nicely behaved hypothesized distribution. Even though we are approximating a smooth distribution function, the sample c.d.f. of the scenarios is completely discrete, with jumps at the scenarios the size of which equals the $1/n$, where n stands for the number of scenarios.

In fact, equation (6.3.1) given in the chapter is actually equation (6.10.12) restated for a discrete random variable. The outcomes are the available scenarios which are equally probable. Consider a sample of observations or scenarios $r_1, \ldots, r_n$ and denote by $r_{(1)} \leq r_{(2)} \leq \ldots \leq r_{(n)}$ the ordered sample. The natural estimator of the ETL at tail probability ϵ is

$$\widehat{ETL}_\epsilon(r) = -\frac{1}{\lceil n\epsilon \rceil - 1} \sum_{k=1}^{\lceil n\epsilon \rceil - 1} r_{(k)} \tag{6.10.13}$$

where $\lceil x \rceil$ is the smallest integer larger than x. Formula (6.10.13) means that we average $\lceil n\epsilon \rceil - 1$ of the $\lceil n\epsilon \rceil$ smallest observations which is, in fact, the definition of the conditional expectation in (6.10.11) for a discrete distribution. The VaR at tail probability ϵ is equal to the negative of the empirical quantile,

$$\widehat{VaR}_\epsilon(r) = -r_{(\lceil n\epsilon \rceil)}. \tag{6.10.14}$$

It remains to determine the coefficients in (6.10.12). Having in mind that the observations in the sample are equally probable, we calculate that

$$\epsilon_0 = \frac{\lceil n\epsilon \rceil - 1}{n}.$$

Plugging ϵ_0, (6.10.14), and (6.10.13) into equation (6.10.12), we obtain (6.3.1) which is the sample AVaR.

Similarly, equation (6.4.2) also arises from (6.10.12). The assumption is that the underlying random variable has a discrete distribution but the outcomes are not equally probable. Thus, the corresponding equation for the average loss on condition that the loss is larger than the VaR at tail probability ϵ is given by

$$\widehat{ETL}_\epsilon(r) = -\frac{1}{\epsilon_0} \sum_{j=1}^{k_\epsilon} p_j r_{(j)} \tag{6.10.15}$$

where $\epsilon_0 = \sum_{j=1}^{k_\epsilon} p_j$ and k_ϵ is the integer satisfying the inequalities,

$$\sum_{j=1}^{k_\epsilon} p_j \leq \epsilon < \sum_{j=1}^{k_\epsilon+1} p_j.$$

The sum $\sum_{j=1}^{k_\epsilon} p_j$ stands for the cumulative probability of the losses larger than the the VaR at tail probability ϵ. Note that equation (6.10.15) turns into equation (6.10.13) when the outcomes are equally probable. With these remarks, we have demonstrated the connection between equations (6.3.1), (6.4.2), and (6.10.12).

The differences between ETL and AVaR are not without any practical importance. In fact, ETL is not a coherent risk measure. Furthermore, the sample ETL in (6.10.13) is not a smooth function of the tail probability while the sample AVaR is smooth. This is illustrated in Figure 6.14. The top plot shows the graph of the sample ETL and AVaR with the tail probability varying between 1% and 10%. The sample contains 100 independent observations on a standard normal distribution, $X \in N(0, 1)$. The bottom plots shows the same but

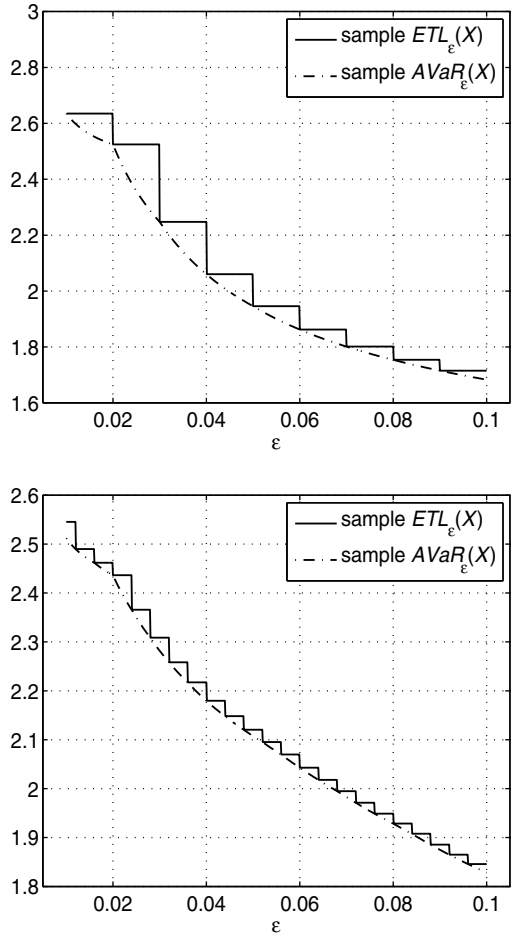

Figure 6.14: The graphs of the sample ETL and AVaR with tail probability varying between 1% and 10%. The top plot is produced from a sample of 100 observations and the bottom plot from a sample of 250 observations. In both cases, $X \in N(0, 1)$.

the sample is larger. It contains 250 independent observations on a standard normal distribution.

Both plots demonstrate that the sample ETL is a step function of the tail probability while the AVaR is a smooth function of it. This is

not surprising because, as ϵ increases, new observations appear in the sum in (6.10.13), producing the jumps in the graph of the sample ETL. In contrast, the AVaR changes gradually as it is a weighted average of the ETL and the VaR at the same tail probability. Note that, as the sample size increases, the jumps in the graph of the sample ETL diminish. In a sample of 5,000 scenarios, both quantities almost overlap. This is because the standard normal distribution has a smooth c.d.f. and the sample c.d.f. constructed from a larger sample better approximates the theoretical c.d.f. In this case, as the sample size approaches infinity, the AVaR becomes indistinguishable from the ETL at the same tail probability.[19]

6.10.5 Kernel-based estimation of AVaR

In this section, we provide more details on the the results in section 6.4.5. We start with a proof of the kernel-based estimator of AVaR.

Proposition 6.10.1. Adopting the multivariate normal kernel, the kernel estimator of the c.d.f. equals

$$\widehat{F}_{w'X}(y) = \frac{1}{N}\sum_{k=1}^{N} \Phi\left(\frac{y - r_k}{\sqrt{h_w}}\right), \tag{6.10.16}$$

and the kernel estimator of the *AVaR* equals

$$AVaR_{\epsilon}^{H}(w'X) = -\frac{1}{N\epsilon}\sum_{k=1}^{N}\left(r_k\Phi\left(\frac{q_{\epsilon}^{w} - r_k}{\sqrt{h_w}}\right) - \sqrt{h_w}g\left(\frac{q_{\epsilon}^{w} - r_k}{\sqrt{h_w}}\right)\right), \tag{6.10.17}$$

where Φ and g are the c.d.f. and the density of the standard normal distribution, $r_k = w'X^k$ is the k-th scenario for the portfolio return, $h_w = (Hw)'(Hw)$, and q_{ϵ}^{w} is the solution of the equation $\widehat{F}_{w'X}(y) = \epsilon$. Furthermore, the partial derivatives of $AVaR_{\epsilon}^{H}(w'X)$ of any order exist.

Proof. The proof of (6.10.16) is obvious: we integrate the corresponding expression for the density. The estimator of AVaR is obtained starting from the conditional expectation and applying the substitution $y = u\sigma + r_k$:

$$\begin{aligned} AVaR_\epsilon^H(w'X) &= -\frac{1}{N\epsilon}\sum_{k=1}^{N}\int_{-\infty}^{q_\epsilon^w}\frac{y}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-r_k)^2}{2\sigma^2}\right)dy \\ &= -\frac{1}{N\epsilon}\sum_{k=1}^{N}\int_{-\infty}^{(q_\epsilon^w-r_k)/\sigma}\frac{u\sigma+r_k}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)du \\ &= -\frac{1}{N\epsilon}\sum_{k=1}^{N} I_k. \end{aligned}$$

We write σ instead of $\sqrt{h_w}$ to simplify the expressions. The resulting integrals I_k are easy to calculate:

$$\begin{aligned} I_k &= r_k\int_{-\infty}^{(q_\epsilon^w-r_k)/\sigma}\frac{e^{-u^2/2}}{\sqrt{2\pi}}du + \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{(q_\epsilon^w-r_k)/\sigma}ue^{-u^2/2}du \\ &= r_k\Phi\left(\frac{q_\epsilon^w-r_k}{\sigma}\right) - \frac{\sigma}{\sqrt{2\pi}}\exp\left(-\frac{(q_\epsilon^w-r_k)^2}{2\sigma^2}\right) \\ &= r_k\Phi\left(\frac{q_\epsilon^w-r_k}{\sigma}\right) - \sigma g\left(\frac{q_\epsilon^w-r_k}{\sigma}\right) \end{aligned}$$

in which $\Phi(x)$ is the cdf of the standard normal distribution and $g(x) = \Phi'(x)$ is the density. Concerning the differentiability property, it holds because the density function of the standard normal is analytic. □

One can argue that it would be fine to start with the portfolio return scenarios and reduce the problem to a univariate one. This would be acceptable if we needed only an approximation to AVaR without necessarily keeping the positive homogeneity and the sub-additivity properties. However, we need these properties as they guarantee convexity of the risk measure. In the following, we check the positive homogeneity property, which indicates why it is important to choose the bandwidth in the multivariate setting.

Proposition 6.10.2. The kernel estimator of the *AVaR* in (6.10.17) satisfies the positive homogeneity property,

$$AVaR_\epsilon^H(aw'X) = aAVaR_\epsilon^H(w'X),$$

where $a > 0$. Furthermore, the kernel approximation is a convex function of portfolio weights.

Proof. This is straightforward to verify. Notice that $h_{aw} = a^2 h_w$ which leads to $\widehat{F}_{aw'X}(y) = \widehat{F}_{w'X}(y/a)$, taking advantage of the definition in (6.10.16) and hence $q_\epsilon^{aw} = aq_\epsilon^w$. In effect,

$$\begin{aligned} AVaR_\epsilon^H(aw'X) &= -\frac{1}{n\epsilon}\sum_{k=1}^{n}\left(ar_k\Phi\left(\frac{q_\epsilon^w - r_k}{\sqrt{h_w}}\right) - a\sqrt{h_w}f\left(\frac{q_\epsilon^w - r_k}{\sqrt{h_w}}\right)\right) \\ &= aAVaR_\epsilon^H(w'X) \end{aligned}$$

The second statement holds from the general properties of AVaR as the kernel approximation can be viewed as the AVaR of a linear combination of random variables with a multivariate density function equal to the kernel estimate. Therefore, the convexity of $AVaR_\epsilon^H(w'X)$ follows as a consequence of the convexity of AVaR in general. □

It can be demonstrated that the kernel estimator of AVaR is always larger than the empirical and approaches it asymptotically.

Proposition 6.10.3. The following relations hold true:

$$\widehat{AVaR}_\epsilon(w'X) \leq AVaR_\epsilon^H(w'X)$$

and

$$\lim_{h_w \to 0} AVaR_\epsilon^H(w'X) = \widehat{AVaR}_\epsilon(w'X)$$

where $\widehat{AVaR}_\epsilon$ denotes the empirical estimator defined in (6.3.1). As a consequence of the limit relation, the kernel estimator converges almost surely to $AVaR_\epsilon(w'X)$ as $n \to \infty$.

Proof. The inequality follows due to the following relationship:

$$\int_{-\infty}^{q_\epsilon^w} \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-r_k)^2}{2\sigma^2}\right) dy \leq \int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-r_k)^2}{2\sigma^2}\right) dy = r_k$$

The limit is a consequence of

$$\int_{-\infty}^{q_\epsilon^w} \frac{y}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-r_k)^2}{2\sigma^2}\right) dy \longrightarrow \begin{cases} r_k & \text{if } r_k < q_\epsilon^w \\ 0 & \text{if } r_k \geq q_\epsilon^w \end{cases}$$

since the normal c.d.f. approaches the Dirac delta function at r_k. □

6.10.6 Remarks on spectral risk measures

In the chapter, we remarked that by selecting a particular risk-aversion function, we can obtain an infinite risk measure for some return distributions. The AVaR can also become infinite but all distributions for which this happens are not reasonable as a model for financial assets returns because they have infinite mathematical expectation. This is not the case with the spectral risk measures. There are plausible statistical models which, if combined with an inappropriate risk-aversion function, result in an infinite spectral risk measure.

In this section, we provide conditions which guarantee that if a risk-aversion function satisfies them, then it generates a finite spectral risk measure. These conditions can be divided into two groups depending on what kind of information about the random variable is used. The first group of conditions is based on information about existence of certain moments and the second group contains more precise conditions based on the tail behavior of the random variable. This section is based on Stoyanov (2005).

Moment-based conditions

Moment-based conditions are related to the existence of a certain norm of the risk-aversion function. We take advantage of the norms

behind the classical Lebesgue spaces of functions denoted by

$$L^p([0,1]) := \left\{ f : ||f||_p = \int_0^1 |f(t)|^p dt < \infty \right\}$$

where $||\cdot||_p$ denotes the corresponding norm. If $p = \infty$, then the norm is the essential supremum, $||f||_\infty = \text{ess sup}_{t\in[0,1]}|f(t)|$. If the function f is continuous and bounded, then $||f||_\infty$ is simply the maximum of the absolute value of the function.

The sufficient conditions for the finiteness of the spectral risk measure involve the quantity

$$I_\phi(X) = \int_0^1 |F_X^{-1}(p)\phi(p)|dp \tag{6.10.18}$$

which is, essentially, the definition of the spectral risk measure but the integrand is taken in absolute value. Therefore,

$$|\rho_\phi(X)| \leq I_\phi(X)$$

and, as a consequence, if the quantity $I_\phi(X)$ is finite, so is the spectral risk measure $\rho_\phi(X)$. Formally, this is a sufficient condition for the absolute convergence of the integral behind the definition of spectral risk measures.

Moment-based conditions are summarized by the following inequalities,

$$C \cdot E|X| \leq I_\phi(X) \leq (E|X|^s)^{1/s} \, ||\phi||_r \tag{6.10.19}$$

where $0 \leq C < \infty$ is a constant and $1/s + 1/r = 1$ with $r, s > 1$. Further on, if $r = 1$ or $s = 1$, the second inequality[20] in (6.10.19) changes to

$$\begin{aligned} I_\phi(X) &\leq \text{sup}_{u\in[0,1]}|F_X^{-1}(u)|, \quad \text{if} \quad r = 1 \\ I_\phi(X) &\leq E|X| \cdot ||\phi||_\infty, \quad \text{if} \quad s = 1. \end{aligned} \tag{6.10.20}$$

As a consequence of equation (6.10.19), it follows that if the absolute moment of order s exists, $E|X|^s < \infty$, $s > 1$, then $\phi \in L^r([0,1])$ is a sufficient condition for $\rho_\phi(X) < \infty$. The $AVaR_\epsilon(X)$ has a special place among $\rho_\phi(X)$ because if $AVaR_\epsilon(X) = \infty$, then $E|X| = \infty$ and

$\rho_\phi(X)$ is not absolutely convergent for any choice of ϕ. In the reverse direction, if there exists $\phi \in L^1([0,1])$ such that $I_\phi(X) < \infty$, then $AVaR_\epsilon(X) < \infty$.

The limit cases in inequalities (6.10.20) show that if X has a bounded support, then all possible risk spectra are meaningful. In addition, if we consider the space of all essentially bounded risk spectra, then the existence of $E|X|$ is a necessary and sufficient condition for the absolute convergence of $\rho_\phi(X)$.

Conditions based on the tail behavior of X

More precise sufficient conditions can be derived assuming a particular tail behavior of the distribution function of X. A fairly general assumption for the tail behavior is *regular variation*. A monotonic function $f(x)$ is said to be *regularly varying* at infinity with index α, $f \in \mathcal{RV}_\alpha$, if

$$\lim_{x\to\infty} \frac{f(tx)}{f(x)} = t^\alpha. \tag{6.10.21}$$

Examples of random variables with regularly varying distribution functions include stable distributions, Student's t distribution, and Pareto distribution. Thus, it is natural to look for sufficient conditions for the convergence of $\rho_\phi(X)$ in the general setting of regularly varying tails. A set of such conditions is provided below.

Suppose that $\rho_\phi(X)$ is the spectral measure of risk of a random variable X such that $E|X| < \infty$ and $P(-X > u) \in \mathcal{RV}_{-\alpha}$. Let the inverse of the risk spectrum $\phi^{-1} \in \mathcal{RV}_{-\delta}$, if existing. Then

$$\begin{aligned} \rho_\phi(X) &= \infty, \quad \text{if} \quad 1 < \delta \leq \alpha/(\alpha-1) \\ \rho_\phi(X) &< \infty, \quad \text{if} \quad \delta > \alpha/(\alpha-1) \end{aligned}$$

The inverse of the risk-aversion function ϕ^{-1} exists if we assume that ϕ is smooth because by assumption ϕ is a monotonic function.

In some cases, we may not know explicitly the inverse of the risk-aversion function, or the inverse may not be regularly varying. Then, the next sufficient condition can be adopted. It is based on comparing the risk-aversion function to a power function.

Suppose that the same condition as above holds, the random variable X is such that $E|X| < \infty$ and $P(-X > u) \in \mathcal{RV}_{-\alpha}$. If the condition

$$\lim_{x \to 0} \phi(x) x^{\beta} = C$$

is satisfied with $0 < \beta < \frac{\alpha - 1}{\alpha}$ and $0 \leq C < \infty$, then $\rho_{\phi}(X) < \infty$. If $\frac{\alpha - 1}{\alpha} \leq \beta < 1$ and $0 < C < \infty$, then $\rho_{\phi}(X) = \infty$.

This condition emphasizes that it is the behavior of the risk-aversion function $\phi(t)$ close to $t = 0$ that matters. This is reasonable because in this range, the risk-aversion function defines the weights of the very extreme losses and if the weights increase very quickly as $t \to 0$, then the risk measure may explode.

In fact, these conditions are more specific than assuming that a certain norm of the risk-aversion function is finite. It is possible to derive them because of the hypothesized tail behavior of the distribution function of X which is a stronger assumption than the existence of certain moments.

Notes

1. In fact, $X = 0.05\sqrt{3}Z + 0.03$ where Z has Student's t distribution with 4 degrees of freedom and Y has a normal distribution with standard deviation equal to 0.1 and mathematical expectation equal to 0.01. The coefficient of Z is chosen so that the standard deviation of X is also equal to 0.1.
2. By comparing the c.d.f.s, we notice that the c.d.f. of X is "above" the c.d.f. of Y to the left of the crossing point, $F_X(x) \geq F_Y(x), \ x \leq -0.15$.
3. This term is adopted in Rockafellar and Uryasev (2002).
4. Equation (6.2.2) was first studied by Pflug (2000). A proof that equation (6.2.1) is indeed the AVaR can be found in Rockafellar and Uryasev 2002.
5. As we remarked, $AVaR_{\epsilon}(X)$ can be infinite only if the mathematical expectation of X is infinite. Nevertheless, if this turns out to be an issue, one can use instead of AVaR the median of the loss distribution,

provided that the loss is larger than $VaR_\epsilon(X)$, as a robust version of AVaR. The median of the conditional loss is always finite and, therefore, the issue disappears but at the cost of violating the coherence axioms. Section 6.10.1 in the appendix to this chapter provides more details.

6. A characteristic function provides a third possibility (besides the cumulative distribution function and the probability density function) to uniquely define a probability distribution. It is a mapping from the set of real numbers $\mathbb{R}$ into the set of complex numbers $\mathbb{C}$ denoted by $\varphi_X(t) = Ee^{itX}$ which represents the so-called "Fourier transform" of the distribution of the random variable X. Knowing the characteristic function $\varphi_X(t)$ is mathematically equivalent to knowing the probability density function $f_X(x)$ or the cumulative distribution function $F_X(x)$.
7. This formula is a simple consequence of the definition of AVaR for discrete distributions: see the appendix to this chapter. A detailed derivation is provided by Rockafellar and Uryasev (2002).
8. For example, $\lceil 3.1 \rceil = \lceil 3.8 \rceil = 4$.
9. A formal proof can be found in Rockafellar and Uryasev (2002). The reasoning in Rockafellar and Uryasev (2002) is based on the assumption that the random variable describes losses, while in equation (6.4.2), the random variable describes the portfolio return or payoff.
10. This question is important in investigating which stocks are the largest contributors to portfolio AVaR but is not within the scope of this chapter.
11. See, for example, van der Vaart (1998) and Simonoff (1996).
12. See, for example, Simonoff (1996) for more information about bandwidth selection and kernel methods in general.
13. Section 6.2.1 of this chapter provides more details on the class of stable distributions and its application as a model in finance.
14. In fact, the formal definition is more involved. See Acerbi (2004) for further details.
15. The examples are based on Stoyanov et al. (2008).
16. For additional information on distortion functionals, see Pflug and Roemisch (2007).
17. For additional information about application of distortion risk measures in optimal portfolio problems, see Sereda et al. (2009).

18. Formal derivation of this relationship can be found, for example, in Rockafellar and Uryasev (2002).
19. In fact, this is a consequence of the celebrated Glivenko–Cantelli theorem claiming that the sample c.d.f. converges almost surely to the true c.d.f.
20. As a matter of fact, the right-hand side inequalities of both cases can be unified as a consequence of the norm relationship $||fg||_1 \leq ||f||_r ||g||_s$ where $f \in L^r$ and $g \in L^s$ and r and s are conjugate exponents: that is, $1/s + 1/r = 1$ and $1 \leq r, s \leq \infty$. See, for example, Rudin (1970).

References

Acerbi, C. (2004), 'Coherent representation of subjective risk aversion', in G. Szego (ed.), *Risk Measures for the 21st Century*, Wiley, Chichester, pp. 147–208.

Acerbi, C. and P. Simonetti (2002), 'Portfolio optimization with spectral measures of risk', working paper, Abaxbank, Milano.

Dokov, S., S. Stoyanov and S. Rachev (2008), 'Computing VaR and AVaR of skewed t distribution', *Journal of Applied Functional Analysis* **3**, 189–209.

Kim, Y., S. T. Rachev, M. L. Bianchi and F. J. Fabozzi (2009), 'Computing VaR and AVaR of infinitely divisible distributions', working paper.

Pflug, G. (2000), 'Some remarks on the value-at-risk and the conditional value-at-risk', in S. Uryasev (ed.), *Probabilistic Constrained Optimization: Methodology and Applications*. Kluwer Academic Publishers, Dordrecht, pp. 272–281.

Pflug, G. and W. Roemisch (2007), *Modeling, Measuring and Managing Risk*, World Scientific.

Rachev, S. T. and S. Mittnik (2000), *Stable Paretian Models in Finance*, Series in Financial Economics, John Wiley & Sons, New York.

Rockafellar, R. T. and S. Uryasev (2002), 'Conditional value-at-risk for general loss distributions', *Journal of Banking and Finance* **26 (7)**, 1443–1471.

Rudin, Walter (1970), *Real and Complex Analysis*, McGraw-Hill, New York.

Samorodnitsky, G. and M. S. Taqqu (1994), *Stable Non-Gaussian Random Processes*, Chapman & Hall, New York/London.

Sereda, E., E. Bronshtein, S. Rachev, F. Fabozzi, W. Sun, and S. Stoyanov (2009), 'Distortion risk measures in portfolio optimization', in J. Guerard (ed.), *The Handbook of Portfolio Construction: Contemporary Applications of Markowitz Techniques*, Springer, New York, pp. 649–674.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, Springer-Verlag, New York.

Stoyanov, S. (2005), *Optimal Financial Portfolios in Highly Volatile Markets*, PhD thesis, University of Karlsruhe.

Stoyanov, S., G. Samorodnitsky, S. Rachev and S. Ortobelli (2006), 'Computing the portfolio conditional value-at-risk in the α-stable case', *Probability and Mathematical Statistics* **26**, 1–22.

Stoyanov, S., S. Rachev and F. Fabozzi (2008), 'Probability metrics with applications in finance', *Journal of Statistical Theory and Practice* **2** (2), 253–277.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Zolotarev, V. M. (1986), *One-dimensional Stable Distributions* (Translation of Mathematical Monographs, Vol 65), American Mathematical Society, Washington, DC.

Chapter 7

Computing AVaR through Monte Carlo

The goals of this chapter are the following:

- To explore the accuracy of the Monte Carlo method in computing the *average value-at-risk* (AVaR) measure, taking advantage of the large-sample theory of the AVaR estimator.
- To consider the cases of both heavy-tailed return distribution models and light-tail return models and their implications for the approximation accuracy.
- To provide insight on the minimum sample size necessary to trust the asymptotic distribution as a model for the approximation error.
- To consider the method of tail truncation as a method for improving precision.

Notation introduced in this chapter:

Notation	*Description*
X^{tr}	A truncated version of the random variable X
$Z_{n,\epsilon}$	A random variable describing the approximation error of the Monte Carlo method
$Z^{tr}_{n,\epsilon}$	A random variable describing the approximation error of the Monte Carlo method for the truncated model

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

Important terms introduced in this chapter:

Term	*Concise explanation*
$L(x)$	A function which is slowly varying at infinity
asymptotic distribution	A probability distribution governing the properties of a statistical estimator as the sample size increases indefinitely.
Central limit theorem	A limit theorem describing the asymptotic behavior of sums of independent and identically distributed random variables under the condition of finite variance for the summands. The limit distribution is the normal distribution.
Generalized central limit theorem	The broadest generalization of the classical CLT allowing for heavy-tailed summands with infinite variance. The limit distribution is the class of stable distributions.

7.1 Introduction

We have already considered the question of how risk can be calculated. An important prerequisite is a measure of risk with suitable properties. We discussed standard deviation, value-at-risk (VaR), and the general family of coherent risk measures in Chapter 5 and AVaR in Chapter 6, which arises as an important member of the family of coherent risk measures.

The measure of risk by itself, however, is not sufficient for the problem of estimating the risk of a given portfolio. It has to be combined with a probabilistic model for the risk factor returns, encapsulating information about both their stand-alone and joint behaviors. A good measure of risk used together with a unrealistic probabilistic model can lead to bad decisions. We need the combination of a good measure of risk and a realistic multivariate model to compute portfolio risk properly.

An acceptable probabilistic model has to take into account the phenomena observed in empirical data, such as heavy-tailed behavior, clustering of volatility, and short- and long-range dependence. The dependence structure between risk drivers can be captured by means

of a copula function. Building a realistic, multidimensional model is a major topic which is beyond the scope of this chapter. We focus only on the one-dimensional case in which computing the risk of a single variable, a stock for example, is a much more tractable problem. While this setting may seem simplified, we can always view the task of computing total portfolio risk as a one-dimensional problem in which portfolio returns are described by means of a one-dimensional random variable.

In Chapter 6, we provided arguments indicating that AVaR is a good choice for a risk measure. Therefore, in this chapter, we explore how reliably AVaR can be calculated when it is combined with a given one-dimensional probability model. However, we do not discuss whether the model is adequate for a given time series or how its parameters should be calibrated. Our assumption is that the probability model has been already selected and calibrated.

It was discussed earlier that only very rarely does there exist a formula computing the AVaR of a given distribution. In Chapter 6, we provided examples for the Gaussian distribution, Student's t distribution, stable Paretian distributions, and a few other classes. If we are not aware of the exact distribution, or if it is hard to calculate explicitly, but we can sample from it relatively easily, then there is a general statistical numerical method we can take advantage of. This is the Monte Carlo method and its application to the problem of AVaR calculation is the main topic of this chapter.

In order to illustrate why the assumption that we may not know explicitly the distribution of a random variable and yet be able to sample from it makes sense in this context, consider the following example. Suppose that we can fit good one-dimensional models to the daily return time series of a given set of stocks which exhibit different fat-tailed behavior using, for instance, the Student's t distribution. Suppose also that the dependence model between them is nicely captured by a given copula function. Thus, a multivariate model can be constructed and sampled by combining the copula and the one-dimensional models. The returns of a long-only portfolio appear as a weighted average of the stock returns, but the

distribution of the portfolio returns is impossible to derive in closed-form. In fact, this is not too surprising because in a general real-world situation, analytic tractability may not be possible.[1] However, scenarios from this analytically intractable distribution are easy to obtain. We can first generate joint scenarios from the fitted multidimensional model and then compute the corresponding scenarios for portfolio returns by applying portfolio weights to the simulated stock returns. In this way, we obtain a sample from the portfolio return distribution without actually knowing it in closed-form.

By construction, AVaR computed through the Monte Carlo method depends on the underlying scenarios. If the scenarios are re-generated, the resulting AVaR will change. Therefore, from a theoretical viewpoint, the AVaR obtained in this way can be viewed as a random variable itself. It is reasonable to expect that the true AVaR should be close to, or even coincide with, the mean of this random variable and its variance should diminish as the sample size increases indefinitely. Non-strictly speaking, these properties are known as unbiasedness and consistency of the AVaR estimator.

Certain characteristics of the estimator indicate how reliable the method is. For example, we can compute a confidence interval and assess if the accuracy is good enough. This, however, is again analytically intractable because we do not know the distribution of the estimator in closed-form. As a result, one has to resort to approximations. There are two general methods:

- Parametric bootstrap: re-generate the sample a number of times and compute AVaR each time. In this way, we obtain a sample from the estimator.
- Asymptotic theory: take advantage of the central limit theorem (CLT), or a generalization of it, to compute the asymptotic distribution of the estimator. Use this distribution having re-adjusted it for the number of scenarios we use.

Both approaches can be viewed as providing approximations. In the bootstrap method, we rely on a sample drawn from the estimator. However, the empirical cumulative distribution function (c.d.f.)

is only an approximation of the true c.d.f. and, therefore, anything computed on the basis of the sample is only an approximation. In a similar vein, the asymptotic theory holds true for infinitely large samples and it is, therefore, an approximation in the case of a finite sample. The bootstrap method, however, relies on the assumption that we can re-generate the sample easily, which may not be the case when the portfolio contains complex derivatives, for example. Thus, practical considerations may force us to use the asymptotic theory rather than the computationally intensive bootstrap method. Numbers of scenarios such as 5,000 or 10,000 are typical choices. They are high enough to be considered large sample cases and applying the asymptotic theory is an acceptable approach.

In this chapter, we explore the asymptotic theory of the empirical AVaR estimator. We provide answers to the following two general questions:

- What is the asymptotic distribution of the empirical AVaR estimator?
- For what sample size can the asymptotic distribution be regarded as a good model for the approximation error of the Monte Carlo method?

Both questions remain valid if VaR is selected as a risk measure but they have much simpler answers in the case of VaR. We make comparisons to VaR where appropriate.

7.2 An Illustration of Monte Carlo Variability

In Tables 5.2 and 6.1 in Chapters 5 and 6, respectively, we provided the 95% confidence bounds of VaR and AVaR at the 1% tail probability, computed through the bootstrap method assuming the returns follow the standard normal distribution. Both tables show that the confidence interval becomes smaller and smaller as the number of simulations increases. This means that precision increases and if we can run the calculations with an infinitely large sample, we expect

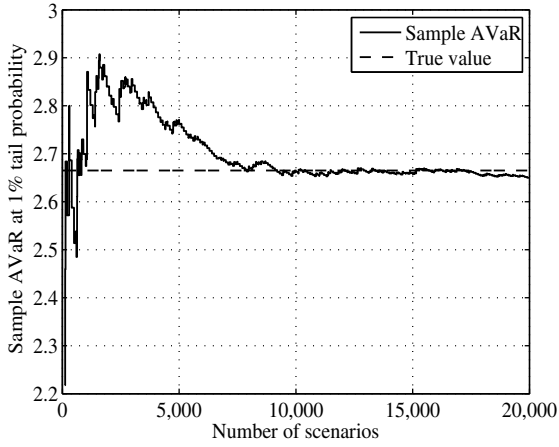

Figure 7.1: Convergence of sample $AVaR_{0.01}(X)$ where $X \in N(0, 1)$ to the true value of 2.66.

that the result will be exactly the VaR and AVaR at 1% tail probability of the standard normal distribution.

The first point that needs attention is if this is always true. That is, the tables illustrate this property for the standard normal distribution but what if we change the model to something different? This claim is always true for VaR due to the celebrated Glivenko–Cantelli theorem proving that the empirical c.d.f. converges almost surely to the c.d.f. of a given random variable as the sample size increases indefinitely. Therefore, the empirical quantile at any probability level converges almost surely to the corresponding quantile of the distribution.

For AVaR, however, the claim is not always true. There are distributions for which AVaR is infinite at any tail probability. For such distributions, while the empirical AVaR for any finite sample is a finite quantity, increasing the sample size does not lead to convergence since the quantity we are trying to approximate is unbounded by construction.

Figures 7.1 and 7.2 illustrate a case with convergence and a case with lack of convergence. We generated two large samples of 20,000 scenarios from the standard normal distribution and the standard

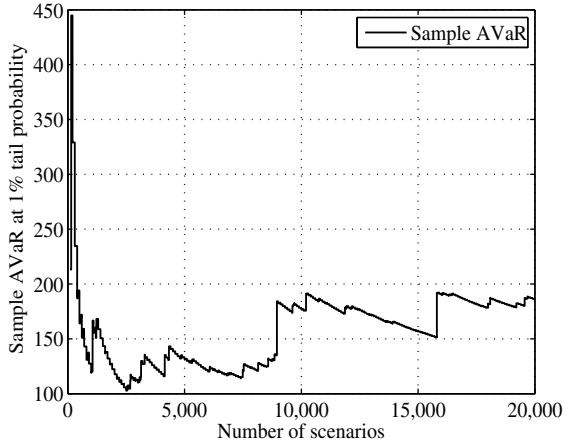

Figure 7.2: Lack of convergence of sample $AVaR_{0.01}(X)$ where X has the standardized Cauchy distribution.

Cauchy distribution. Then, we estimated the sample AVaR at 1% tail probability, starting from a small sub-sample and increasing the sample size until all the 20,000 scenarios are used. We took advantage of formula (6.3.1) from Chapter 6 to compute the sample AVaR. The figures show the two sample AVaRs as a function of the number of scenarios.

If X has a standard normal distribution, then $AVaR_{0.01}(X) = 2.66$. This is the value towards which the sample AVaR converges. In contrast, the AVaR of the Cauchy distribution is not bounded at any tail probability and the sample AVaR does not converge to any value. The AVaR is unbounded because the tails of the Cauchy distribution are so heavy that even the mathematical expectation of this distribution is infinite.

A condition which guarantees that $AVaR_\epsilon(X) < \infty$ arises from the inequality

$$AVaR_\epsilon(X) \leq -E\min(X, 0). \tag{7.2.1}$$

If X describes the return of a common stock, then the right-hand side can be interpreted as the average loss. Therefore, if the left tail of the

return distribution is such that the average loss is a finite number, then $AVaR_\epsilon(X) < \infty$ for any $\epsilon < 1$. This sufficient condition is only technical because, from a practical viewpoint, it makes no sense to use a theoretical model predicting an infinite average loss. Therefore, we can conclude that while in theory there may be distributions, the AVaR of which is unbounded at any tail probability, they are not appropriate for modeling returns of financial variables.

Even if purely technical, the condition in equation (7.2.1) is interesting as it emphasizes the intuitive expectation that it is the left tail of X that governs the properties of sample AVaR. We confirm this expectation in the following sections when discussing the asymptotic distribution of sample AVaR.

7.3 Asymptotic Distribution, Classical Conditions

The accuracy of the Monte Carlo method can be illustrated with the help of the bootstrap method. Such an example can be found in Figure 6.6 and the corresponding 95% confidence bounds are available in Table 6.1. The bootstrap method is a numerical technique allowing us to gain insight into the distribution of sample AVaR with the number of scenarios fixed. The distribution of sample AVaR is not known even for simple assumptions about X: for example, if X has a normal distribution. For this reason, we may resort to this numerical method to gain understanding about the accuracy of the Monte Carlo method.

Another standard method in statistics is to consider the asymptotic distribution of the estimator using CLT-type arguments. Before proceeding, let us summarize the properties discussed so far and introduce some notation.

The sample AVaR is defined through the expression

$$\widehat{AVaR}_\epsilon(X) = -\frac{1}{\epsilon}\int_0^\epsilon F_n^{-1}(p)dp, \tag{7.3.1}$$

where $F_n^{-1}(p)$ denotes the inverse of the sample c.d.f. $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\}$, in which $I\{A\}$ denotes the indicator function of the event A, and $X_1, \ldots, X_n$ is a sample of independent and identically distributed (i.i.d.) copies of a random variable X. If we assume that the random variable X describes the return distribution of a common stock, then the sample $X_1, \ldots, X_n$ represents the Monte Carlo simulations drawn from the distribution of X.

In the previous section, we noted that $-E(\min(X, 0)) < \infty$ is a sufficient condition. It can be demonstrated that it is actually necessary and sufficient for $AVaR_\epsilon(X) < \infty$. For this reason, by the Glivenko–Cantelli theorem, the same condition becomes necessary and sufficient for the following almost sure convergence:

$$\widehat{AVaR}_\epsilon(X) \xrightarrow{a.s.} AVaR_\epsilon(X) \quad \text{as} \quad n \to \infty. \tag{7.3.2}$$

In the introduction to this chapter, we noted that it is desirable that the AVaR estimator be unbiased. For sample AVaR, this property does not hold and the estimator is biased. It is possible to demonstrate that the bias is negative and is of order $O(n^{-1})$. Therefore, we can consider it negligible for the large-sample theory.[2]

In this section, we consider the asymptotic distribution of sample AVaR assuming the classical condition that second moment of X is finite, $EX^2 < \infty$. From a technical viewpoint, this condition implies that the tails of X do not decay too slowly. Under this assumption, we can establish the following CLT-type result,[3]

Theorem 7.3.1. Suppose that X is a random variable with finite second moment $EX^2 < \infty$. Furthermore, suppose that the c.d.f. of X is differentiable at $x = q_\epsilon$, where q_ϵ is the ϵ-quantile of X. Then, as $n \to \infty$,

$$\frac{\sqrt{n}}{\sigma_\epsilon}\left(\widehat{AVaR}_\epsilon(X) - AVaR_\epsilon(X)\right) \xrightarrow{w} N(0, 1) \tag{7.3.3}$$

where $\xrightarrow{w}$ denotes weak limit and

$$\sigma_\epsilon^2 = \frac{1}{\epsilon^2} D(\max(q_\epsilon - X, 0)). \tag{7.3.4}$$

The condition that the c.d.f. be differentiable at q_ϵ is only technical and is not restrictive. This condition is always satisfied for random variables which have density functions.

The result in the theorem is asymptotic in the sense that it becomes valid when the sample size increases indefinitely, $n \to \infty$. Therefore, for any finite n, the asymptotic distribution is only approximately true. Nevertheless, we can apply the result in the theorem in the following way. Suppose that we know the distribution of X which has a finite second moment, and we fix $n = 10{,}000$. We can calculate σ_ϵ as it depends only on the distribution of X and the choice of tail probability ϵ. The 95% confidence interval for the true AVaR can be calculated according to

$$\widehat{AVaR}_\epsilon(X) - \frac{1.96\sigma_\epsilon}{\sqrt{n}} \leq AVaR_\epsilon(X) \leq \widehat{AVaR}_\epsilon(X) + \frac{1.96\sigma_\epsilon}{\sqrt{n}},$$

in which 1.96 is approximately the 97.5% quantile of the standard normal distribution.

For some choices of X satisfying the conditions in the theorem, it may be hard to calculate σ_ϵ. In such cases, the standard approach is to estimate it using the available scenarios. If more convenient, the following equivalent representation in terms of conditional expectations can be used:

$$\sigma_\epsilon^2 = \frac{q_\epsilon^2}{\epsilon} - \frac{2q_\epsilon}{\epsilon}E(X|X \leq q_\epsilon) + \frac{1}{\epsilon}E(X^2|X \leq q_\epsilon) - (q_\epsilon - E(X|X \leq q_\epsilon))^2 \tag{7.3.5}$$

Thus, in order to estimate $E(X|X \leq q_\epsilon)$, we average the scenarios smaller than the corresponding sample quantile.

In showing how the theorem can be applied, we made one assumption which we may not be ready to accept without questioning. We assumed that $n = 10{,}000$ is sufficiently large in order to adopt the asymptotic distribution for the true distribution of sample AVaR. In the following sections, we demonstrate that for some popular distribution models in finance this is not true (i.e., we need more scenarios).

As far as the theory of probability is concerned, this question has received a lot of attention. It was at the heart of the research which led to results establishing the *rate of convergence* in the CLT. In our context, a rate-of-convergence theorem would indicate how large n should be in order to adopt the asymptotic distribution and make a minimal error. The error, in this case, is computed in terms of a probability metric between the distribution of the sample AVaR and the asymptotic model.

In the current chapter, we do not consider this much more involved area. We only demonstrate through a Monte Carlo study what choices of n are sufficiently high so that the asymptotic model is satisfactory for given choices for the distribution of X which are popular in finance. We also discuss a technique which improves the convergence rate and, thus, requires fewer scenarios for the same level of precision.

7.4 Rate of Convergence to the Normal Distribution

In this section, our goal is to investigate the effect of the tail behavior of X on the rate of convergence in (7.3.1). We are also interested in the question if the method of tail truncation improves convergence and by how much. Generally, the tail truncation method consists of "replacing" the tails of X with the tails of a thin-tailed distribution "far away" from the center of the distribution of X: for example, beyond the 0.1% and 99.9% quantiles. The tail truncation method has applications in finance for modeling the distribution of stock returns. A practical reason for adopting it is that a stock exchange may close if a severe market crash occurs, limiting in this way the loss that can be observed. This method also has application in derivatives pricing with a heavy-tailed distributional assumption for the return of the underlying.[4]

In the following sections, we start with the Student's t distribution and investigate the convergence rate in the limit relation (7.3.1) as degrees of freedom increase. The Student's t distribution has been

widely applied as a model for stock returns and while there is a closed-form expression for AVaR, we consider it in the context of the Monte Carlo method because it may serve as a guideline for the convergence rate of AVaR of other distributions with a similar tail behavior, closed-form expressions of AVaR for which do not exist.

We address the same questions with a truncated Student's t distribution in which the truncation is done in the simplest possible way – we set the values of the random variable which are beyond the 0.1% and 99.9% quantiles to be equal to the corresponding quantiles. As a result, small point masses appear at the 0.1% and 99.9% quantiles. We also focus on the class of stable distributions and truncated stable distributions, in which the same truncation technique is adopted as in the case of Student's t distribution.

7.4.1 The effect of tail thickness

The impact of the tail behavior on the rate of convergence in Theorem 7.3.1 is first studied when X has Student's t distribution, $X \in t(\nu)$, with $\nu \geq 3$. We impose the condition on the degrees of freedom in order for the random variable to have a finite variance. Taking advantage of the expression for the density

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \ x \in \mathbb{R},$$

it is possible to compute explicitly the variance in equation (7.3.4). In fact, for this purpose the expression in (7.3.5) is more appropriate. As a first step, we calculate the two conditional expectations:

$$E(X|X \leq q_\epsilon) = -\frac{1}{\epsilon} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{\sqrt{\nu}}{(\nu-1)\sqrt{\pi}} \left(1 + \frac{q_\epsilon^2}{\nu}\right)^{\frac{1-\nu}{2}}, \text{ if } \nu > 1. \tag{7.4.1}$$

$$E(X^2|X \leq q_\epsilon) = q_\epsilon E(X|X \leq q_\epsilon) + \frac{\nu}{\epsilon(\nu-2)} F_{\nu-2}\left(q_\epsilon \sqrt{\frac{\nu-2}{\nu}}\right), \text{ if } \nu > 2. \tag{7.4.2}$$

Table 7.1 The number of scenarios sufficient to accept the normal distribution as an approximate model for different values of ν and ϵ.

ν	$\epsilon = 0.01$	$\epsilon = 0.05$
3	70,000	17,000
4	60,000	9,000
5	50,000	7,000
6	23,000	4,500
7	14,000	4,200
8	13,000	4,100
9	12,000	4,000
10	12,000	3,900
15	11,000	3,850
25	10,000	3,800
50	10,000	3,750
∞	10,000	3,300

Plugging these expressions[5] in (7.3.5), we obtain the expression for the variance σ_ϵ^2.

Having obtained an expression for the variance allows us to use the test of Kolmogorov and address the question of how many simulations are needed in order to accept the hypothesis that the distribution of the random variable on the left-hand side of the limit relation (7.3.3),

$$Z_{n,\epsilon} = \frac{\sqrt{n}}{\sigma_\epsilon} \left(\widehat{AVaR}_\epsilon(X) - AVaR_\epsilon(X) \right), \tag{7.4.3}$$

is standard normal. If we accept the null hypothesis for a given value of n, then the standard normal distribution can be used as an approximate model and we can calculate not only confidence intervals but also other characteristics based on it.

Table 7.1 shows the values of n sufficient to accept the null hypothesis in the test of Kolmogorov for different degrees of freedom and tail probabilities. We chose $\epsilon = 0.01$ and $\epsilon = 0.05$ since these values are

frequently used in the financial industry for VaR estimation. These tail probabilities correspond to 99% and 95% VaR, respectively. The numbers in the table are calculated by generating independently 2,000 samples of the given size, and then from each sample we estimate (7.4.3). As a result, we obtain 2,000 scenarios from the distribution of the random variable $Z_{n,\epsilon}$. In this calculation, $AVaR_\epsilon(X)$ is computed taking advantage of the closed-form expression provided in Chapter 6.

In line with intuition, the numbers in Table 7.1 indicate that when the tail is heavier, we need a larger sample in order for the asymptotic law to be sufficiently close to the distribution of $Z_{n,\epsilon}$ in terms of the Kolmogorov metric. Another expected conclusion is that as the tail probability increases, a smaller sample turns out to be sufficient because a smaller tail probability means that a larger part of the sample is effectively used in the calculation.

In Table 7.2, we calculated the 95% confidence interval for AVaR when the sample size changes from 250 to 10,000 scenarios. We generated 2,000 independent samples and then computed (7.4.3) for each sample. Thus, the 95% confidence intervals are obtained from 2,000 scenarios of the random variable $Z_{n,\epsilon}$. As n increases, the two quantiles approach the corresponding quantiles of the standard normal distribution. Note that the largest $n = 10{,}000$ is generally below the sample sizes for $\epsilon = 0.01$ given in Table 7.1. Nevertheless, the relative discrepancies between the quantiles given in Table 7.2 and the corresponding standard normal distribution quantiles are less than 5% for $\nu \geq 6$.[6] The relative discrepancies between the quantiles given in Table 7.3 and the corresponding standard normal distribution quantiles for $n = 10{,}000$ have the same magnitude. However, in this case $n = 10{,}000$ is well above the sample sizes given in Table 7.1 for $\epsilon = 0.05$. As a result, we can conclude that even smaller samples than the ones given in Table 7.1 can lead to 95% confidence intervals obtained via resampling from the distribution of $Z_{n,\epsilon}$ being close to the corresponding 95% confidence interval obtained from the limit distribution, even though the Kolmogorov test fails for such samples. For instance, the relative deviation between the quantiles given

Table 7.2 The 95% confidence bounds generated from 2,000 simulations from the distribution of $Z_{n,\epsilon}$ in equation (7.4.3) with $\epsilon = 0.01$. The corresponding quantiles of $N(0, 1)$ are $q_{2.5\%} = -1.96$ and $q_{2.5\%} = 1.96$.

	$n = 250$		$n = 500$		$n = 1{,}000$		$n = 5{,}000$		$n = 10{,}000$	
ν	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$
3	−1.110	2.011	−1.257	2.173	−1.352	2.202	−1.633	2.037	−1.664	2.007
4	−1.337	2.144	−1.442	2.229	−1.543	2.082	−1.744	2.230	−1.756	2.176
5	−1.441	2.153	−1.529	2.224	−1.728	2.190	−1.843	2.060	−1.807	2.009
6	−1.522	2.134	−1.618	2.033	−1.701	2.115	−1.848	1.987	−1.955	1.982
7	−1.627	2.050	−1.668	1.975	−1.827	2.043	−1.841	2.048	−1.913	2.014
8	−1.655	2.028	−1.760	2.145	−1.836	2.032	−1.898	2.034	−1.866	1.939
9	−1.720	1.938	−1.753	2.146	−1.798	2.075	−1.866	2.005	−1.905	2.007
10	−1.747	1.925	−1.809	1.980	−1.762	2.078	−1.822	1.950	−1.962	2.000
15	−1.813	1.751	−1.848	1.896	−1.891	1.956	−1.969	1.941	−1.968	1.873
25	−1.848	1.760	−1.933	2.028	−1.897	1.950	−1.939	1.957	−1.899	1.923
50	−1.898	1.948	−1.962	1.900	−1.971	1.973	−1.961	1.914	−1.895	1.948
∞	−1.921	1.761	−1.976	1.920	−1.964	1.822	−1.869	1.907	−2.004	1.937

Table 7.3 The 95% confidence bounds generated from 2,000 simulations from the distribution of $Z_{n,\epsilon}$ in equation (7.4.3) with $\epsilon = 0.05$. The corresponding quantiles of $N(0, 1)$ are $q_{2.5\%} = -1.96$ and $q_{2.5\%} = 1.96$.

	$n = 250$		$n = 500$		$n = 1{,}000$		$n = 5{,}000$		$n = 10{,}000$	
ν	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$
3	−1.422	2.110	−1.543	2.016	−1.549	1.981	−1.725	1.947	−1.883	1.987
4	−1.647	2.169	−1.737	2.235	−1.787	2.171	−1.900	2.226	−1.849	2.115
5	−1.749	2.081	−1.811	2.096	−1.757	2.148	−1.868	2.015	−1.937	2.100
6	−1.810	2.071	−1.896	2.030	−1.921	1.941	−1.958	1.998	−1.886	2.032
7	−1.786	2.215	−1.824	1.990	−1.809	2.086	−1.986	2.030	−1.916	2.015
8	−1.932	2.131	−1.870	2.058	−1.755	2.090	−1.937	2.014	−1.915	1.952
9	−1.848	2.139	−1.884	2.081	−1.930	2.023	−1.995	1.964	−1.863	2.048
10	−1.906	2.103	−2.021	1.966	−1.839	2.087	−2.009	1.930	−1.989	1.995
15	−1.797	1.905	−1.929	2.056	−1.944	1.952	−1.924	1.973	−1.947	1.979
25	−1.958	1.950	−1.994	1.956	−1.939	1.968	−2.085	1.993	−1.894	1.944
50	−1.986	1.927	−1.980	1.823	−1.962	1.883	−1.911	1.969	−2.002	1.935
∞	−2.013	1.828	−1.953	1.869	−1.975	1.893	−2.034	1.958	−1.903	1.944

in Table 7.2 for $n = 5{,}000$ and the corresponding standard normal distribution quantiles are below 7% for $n \geq 6$, which is a small deviation for all practical purposes.

As a result of this analysis, we can conclude that for the purposes of building confidence intervals for $AVaR_\epsilon(X)$ when $X \in t(\nu)$, with $\nu \geq 6$ and $\epsilon = 0.01, 0.05$, we can safely employ the asymptotic law when the sample size we use for AVaR estimation contains more than 5,000 scenarios. If the Student's t distribution is fitted to daily stock returns time series, such values for ν are very common.

Figure 7.3 illustrates the differences in the convergence rate when X has a Student's t distribution with $\nu = 3$, which corresponds to heavier tails, and $\nu = 10$. Since high degrees of freedom imply more light tails, smaller samples are sufficient for the density of (7.4.3) to be closer to the standard normal density.

7.4.2 The effect of tail truncation

The stochastic stability of sample AVaR increases dramatically after tail truncation. In this section, we repeat the calculations from the previous section but when X has Student's t distribution with the left tail truncated at $q_{0.1\%}$ quantile.

We adopt the simplest possible truncation method. The random variable X^{tr} is said to be a truncated version of X at $q_{0.1\%}$ quantile if

$$X^{tr} = XI\{q_{0.1\%} \leq X\} + q_{0.1\%}I\{X < q_{0.1\%}\}$$

in which $I\{A\}$ denotes the indicator of the event A, and $q_{0.1\%}$ is the corresponding quantile of X. The tail truncation introduces small point masses at the two quantile levels.

The two conditional expectations in (7.3.5) can be related to the corresponding conditional expectations of X. In the following, we assume that the tail probability ϵ is larger from the tail probability of the left truncation point, $\epsilon > 0.001$. Under this assumption, the

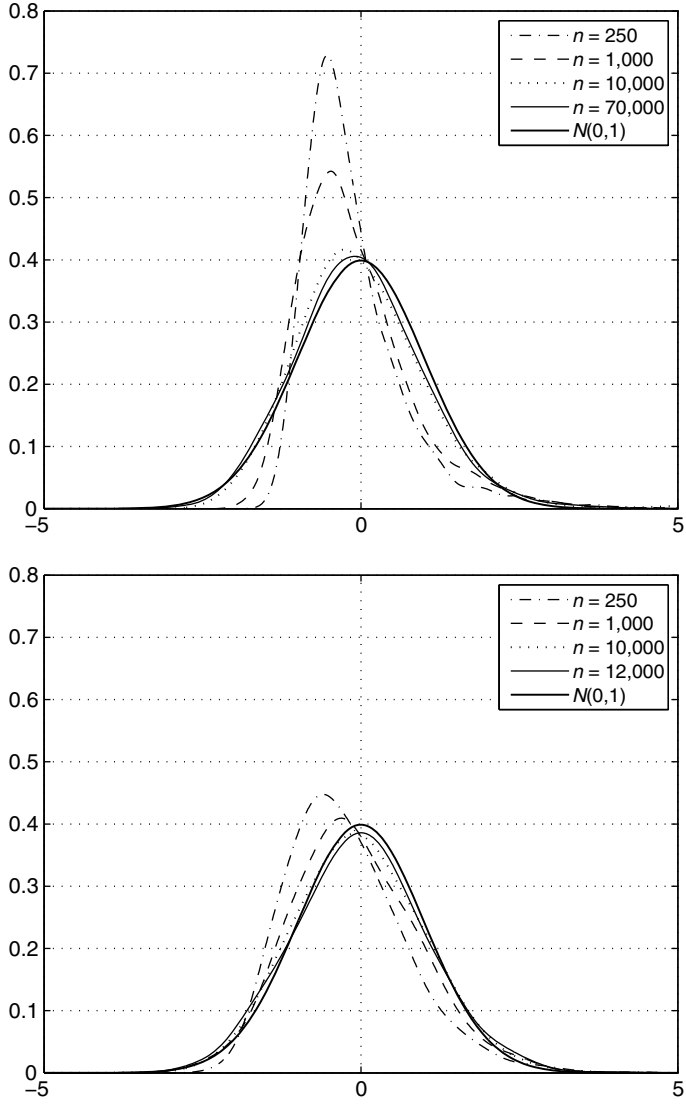

Figure 7.3: The density of (7.4.3) approaching the $N(0, 1)$ density as the sample size increases with $\nu = 3$ (top) and $\nu = 10$ (bottom).

ϵ-quantile of X is the same as the ϵ-quantile of X^{tr} and, therefore,

$$E(X^{tr}|X^{tr} \leq q_\epsilon) = E(X|X \leq q_\epsilon) - \frac{0.001}{\epsilon} E(X|X \leq q_{0.1\%}) + \frac{0.001 q_{0.1\%}}{\epsilon} \tag{7.4.4}$$

$$E((X^{tr})^2|X^{tr} \leq q_\epsilon) = E(X^2|X \leq q_\epsilon) - \frac{0.001}{\epsilon} E(X^2|X \leq q_{0.1\%}) + \frac{0.001 q_{0.1\%}^2}{\epsilon} \tag{7.4.5}$$

If we assume that X has Student's t distribution, $X \in t(\nu)$, then the conditional expectations of X in equations (7.4.4) and (7.4.5) can be computed according to formulae (7.4.1) and (7.4.2). Plugging the expressions for the conditional expectations of Y in the expression for σ_ϵ^2, we obtain the variance of the asymptotic distribution. Furthermore, the AVaR of the truncated Student's t distribution, $AVaR_\epsilon(X^{tr})$, can be calculated through equation (7.4.4), taking advantage of the closed-form expression for $AVaR_\epsilon(X)$.

In the following, we investigate the convergence rate of

$$Z_{n,\epsilon}^{tr} = \frac{\sqrt{n}}{\sigma_\epsilon} \left(\widehat{AVaR}_\epsilon(X^{tr}) - AVaR_\epsilon(X^{tr}) \right), \tag{7.4.6}$$

for different degrees of freedom to the standard normal distribution and we compare the results to the ones in the previous section.

Table 7.4 is the counterpart of Table 7.1 for the truncated distribution. It is impressive how the sample size sufficient to accept the null hypothesis in the Kolmogorov test decreases after tail truncation. The most dramatic change is in the case $\nu = 3$. Now we need only 12,000 scenarios compared to 70,000 in the non-truncated case.

Tables 7.5 and 7.6 are the counterparts of Tables 7.2 and 7.3. The relative deviation of the quantiles $q_{2.5\%}$ and $q_{97.5\%}$ of the random variable $Z_{n,\epsilon}^{tr}$ in (7.4.6) from those of the standard normal distribution are below 7% for all degrees of freedom and $n = 10{,}000$, and, with a few exceptions, for $n = 5{,}000$. Compare Figure 7.4 and the top plot in Figure 7.3 for an illustration of the improvement in the

Table 7.4 The number of scenarios sufficient to accept the normal distribution as an approximate model for different values of ν and ϵ.

ν	$\epsilon = 0.01$	$\epsilon = 0.05$
3	12,000	4,000
4	11,500	3,600
5	11,000	3,300
6	11,000	3,200
7	10,500	3,100
8	10,000	3,000
9	10,000	3,000
10	10,000	3,000
15	10,000	2,950
25	10,000	2,900
50	10,000	2,900
∞	10,000	2,900

convergence rate. These results indicate that the asymptotic distribution can be used to obtain a 95% confidence bound for the sample AVaR for all degrees of freedom if the sample size contains more than 5,000 scenarios.

The increase of convergence rate when using the tail truncation method comes at the cost of a bias which is introduced by truncating the left tail. Rearranging equation (7.4.4), we obtain

$$AVaR_\epsilon(X) - AVaR_\epsilon(X^{tr}) = \frac{0.001}{\epsilon} E(X|X \leq q_{0.1\%}) - \frac{0.001 q_{0.1\%}}{\epsilon} > 0.$$

The magnitude of the bias depends on how heavy the tails of X are. The heavier the tails, the larger the bias is and, therefore, the larger the improvement in the convergence rate is. The bias when $X \in t(\nu)$ computed as a percentage of $AVaR_\epsilon(X)$ is provided in Table 7.7.

7.4.3 Infinite variance distributions

A critical assumption behind the limit result in Theorem 7.3.1 is the finite variance of X. To be more precise, the condition of finite

Table 7.5 The 95% confidence bounds generated from 2,000 simulations from the distribution of $Z_{n,\epsilon}^{tr}$ in equation (7.4.6) with $\epsilon = 0.01$. The corresponding quantiles of $N(0, 1)$ are $q_{2.5\%} = -1.96$ and $q_{2.5\%} = 1.96$.

	$n = 250$		$n = 500$		$n = 1{,}000$		$n = 5{,}000$		$n = 10{,}000$	
ν	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$
3	−1.723	1.699	−1.847	1.932	−1.850	1.958	−1.966	1.921	−1.860	1.936
4	−1.759	1.694	−1.863	1.819	−1.903	1.860	−1.989	1.942	−1.964	1.886
5	−1.808	1.536	−1.884	1.871	−1.926	1.932	−1.961	1.964	−1.782	2.066
6	−1.947	1.565	−1.937	1.759	−2.002	1.734	−2.057	1.946	−1.981	1.958
7	−1.960	1.524	−1.960	1.666	−1.965	1.844	−2.101	1.932	−1.981	1.927
8	−2.002	1.567	−2.015	1.693	−1.903	1.802	−1.952	1.856	−1.917	1.928
9	−1.963	1.552	−2.030	1.748	−2.106	1.779	−1.965	2.026	−1.932	1.938
10	−2.003	1.596	−2.119	1.709	−2.034	1.850	−1.925	1.813	−1.990	1.952
15	−2.090	1.485	−2.159	1.650	−2.065	1.786	−1.983	1.847	−2.035	1.855
25	−2.183	1.502	−2.084	1.578	−2.093	1.747	−2.016	1.806	−1.954	1.877
50	−2.272	1.509	−2.089	1.632	−2.042	1.726	−1.938	1.914	−2.056	1.970

Table 7.6 The 95% confidence bounds generated from 2,000 simulations from the distribution of $Z_{n,\epsilon}^{tr}$ in equation (7.4.6) with $\epsilon = 0.05$. The corresponding quantiles of $N(0, 1)$ are $q_{2.5\%} = -1.96$ and $q_{2.5\%} = 1.96$.

	$n = 250$		$n = 500$		$n = 1{,}000$		$n = 5{,}000$		$n = 10{,}000$	
ν	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$	$q_{2.5\%}$	$q_{97.5\%}$
3	−1.815	2.116	−1.866	2.041	−1.939	2.018	−1.944	1.975	−2.045	1.874
4	−1.756	2.150	−1.811	2.073	−2.052	2.060	−1.923	1.973	−1.922	1.854
5	−1.820	1.954	−1.971	2.032	−1.916	2.036	−1.826	1.960	−1.941	1.883
6	−1.899	2.089	−1.981	2.036	−2.012	2.012	−1.955	1.933	−1.921	2.011
7	−2.001	2.032	−1.921	1.997	−1.949	1.980	−1.980	1.936	−2.016	1.915
8	−1.888	1.995	−1.922	2.050	−1.907	1.917	−1.942	1.911	−1.910	1.903
9	−2.017	2.003	−1.892	1.918	−1.899	2.017	−1.931	2.001	−2.009	1.967
10	−1.928	1.814	−1.992	1.960	−1.870	1.949	−1.845	2.076	−1.992	1.898
15	−2.059	1.983	−2.020	2.007	−1.961	1.922	−1.953	1.870	−1.936	1.874
25	−1.999	1.854	−2.038	1.945	−1.889	2.028	−2.031	1.916	−1.975	1.890
50	−1.960	1.898	−2.028	1.898	−1.947	1.906	−2.015	2.002	−1.959	1.911

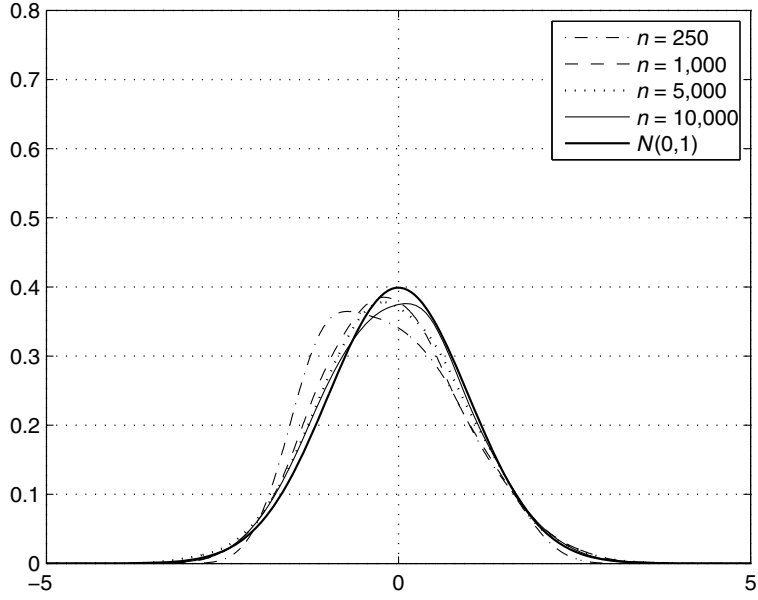

Figure 7.4: The density of (7.4.6) approaching the $N(0,1)$ density as the sample size increases with $\nu = 3$ and $\epsilon = 0.01$.

variance can be loosened to finite downside semi-variance,

$$D\max(-X, 0) < \infty,$$

because it is the behavior of the left tail which is important. As a consequence, the sample AVaR of distributions with infinite variance, but finite downside semi-variance, may still follow Theorem 7.3.1.

However, there are infinite variance distributions for which

$$D\max(-X, 0) = \infty$$

and, therefore, the limit result in Theorem 7.3.1 does not hold for them. Such is the class of stable distributions.[7]

Stable distributions are introduced by their characteristic functions and except for a couple of representatives, generally no closed-form expressions for their densities and c.d.f.s are known. If $\alpha < 2$, then X has infinite variance. If $1 < \alpha \leq 2$, then X has finite

Table 7.7 The magnitude of the bias introduced through the truncation method as a percentage of $AVaR_\epsilon(X)$, $X \in t(\nu)$.

ν	$\epsilon = 0.01$	$\epsilon = 0.05$
3	7.418%	2.682%
4	4.814%	1.569%
5	3.641%	1.122%
6	2.996%	0.891%
7	2.596%	0.754%
8	2.326%	0.665%
9	2.134%	0.602%
10	1.990%	0.556%
15	1.608%	0.437%
25	1.352%	0.359%
50	1.185%	0.311%
∞	1.110%	0.289%

mean and the AVaR of X can be calculated. In the calculations in this section, we use the semi-analytic formula provided in section 6.2.1 of Chapter 6.

Even though we know that Theorem 7.3.1 does not hold for a stable distribution with $\alpha < 2$, we simulate 2,000 draws from the random variable in equation (7.4.3), in which σ_ϵ is estimated from a generated sample by estimating the corresponding conditional moments. In theory, the second conditional moment explodes but for any finite sample its estimate is a finite number. Our goal is to see what happens when Theorem 7.3.1 does not hold. Figure 7.5 illustrates such a divergent case in which $\alpha = 1.5$ and $\epsilon = 0.05$. The lack of convergence is quite obvious.

Stable distributions with $\alpha < 2$ in combination with a tail truncation method have been proposed as a model for the returns of the underlying in derivatives pricing. It is interesting to see how much the simple truncation technique we applied in the previous section can change Figure 7.5. With its left tail truncated according

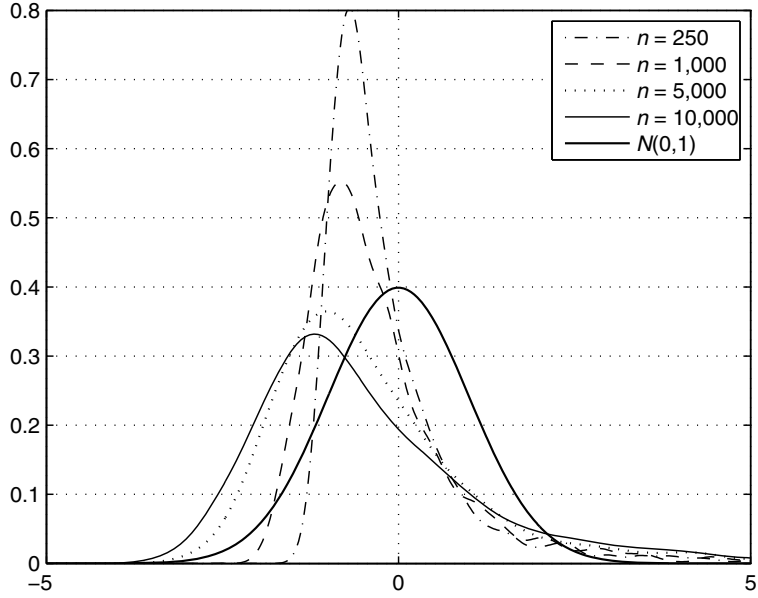

Figure 7.5: Lack of convergence, X has a stable distribution with $X \in S_{1.5}(1, 0, 0)$ and $\epsilon = 0.05$.

to our simple method, Theorem 7.3.1 holds. Figure 7.6 illustrates this change. We observe a quick convergence rate, similar to the one illustrated in Figure 7.4 for the Student's t distribution. The bias introduced by the tail truncation method in this case can be calculated in a way similar to the one we employed for the Student's t distribution. If $X \in S_{1.5}(1, 0, 0)$, then the bias is 17.08% of $AVaR_\epsilon(X)$.

As a result, we can conclude that when X has a standardized symmetric stable distribution with an index of stability of 1.5, then the asymptotic distribution of sample AVaR is not the normal distribution. Nevertheless, as we increase the number of scenarios, the Monte Carlo method provides estimates with an increasing accuracy, and, with an infinitely large sample, we can calculate precisely the AVaR of X at any tail probability. The normal distribution, however, does not describe the approximation error.

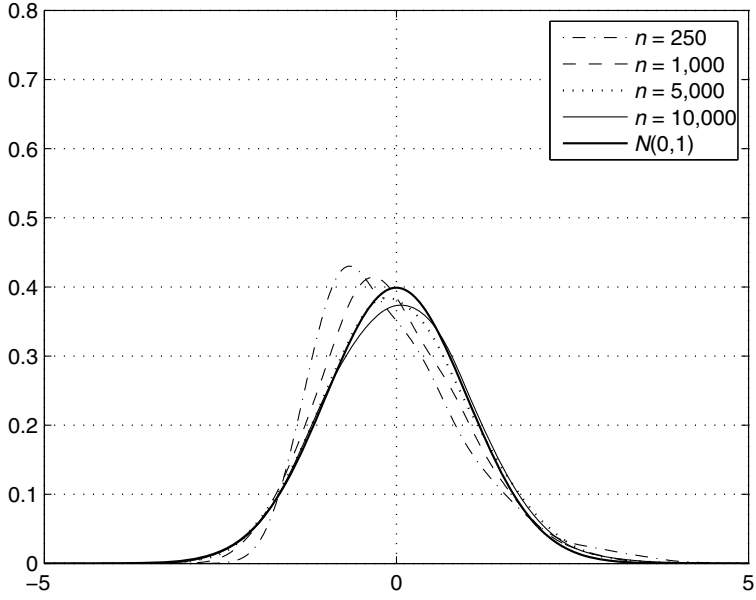

Figure 7.6: After tail truncation at $q_{0.1\%}$ and $q_{99.9\%}$, there is a fast convergence to $N(0,1)$, $\alpha = 1.5$ and $\epsilon = 0.05$.

7.5 Asymptotic Distribution, Heavy-tailed Returns

In section 7.4.3, we demonstrated that the regularity condition $EX^2 < \infty$ in the limit theorem 7.3.1 is quite significant. We did this by finding a random variable X for which the limit result in Theorem 7.3.1 does not hold even though $AVaR_\epsilon(X) < \infty$. Intuition suggests that for distributions with $EX^2 = \infty$ but $AVaR_\epsilon(X) < \infty$, there should be a probability law governing the quality of approximation in the Monte Carlo scheme which, apparently, cannot be the normal distribution. Efforts to find such a probability law should result in an extension of Theorem 7.3.1 in which the condition of finite second moment is relaxed and the probability law should contain the normal distribution as a special case.

Without discussing the details, we mention that infinite variance distributions have been proposed as plausible models for financial data. Especially when high frequency returns are concerned, intra-day stock returns for example, an infinite variance model may turn out to be realistic.[8]

Since computing AVaR from a sample of scenarios reduces to summation of a transformation of i.i.d. random variables, we can look at the generalized CLT (GCLT) theorem and apply it to the problem of finding a limit distribution of sample AVaR. GCLT concerns summation of i.i.d. random variables and from a theoretical viewpoint it provides the most general result possible for the limit distribution of such sums. It contains the CLT as a special case. The limit distributions arising in the GCLT are the stable distributions which we introduced in Chapter 6. An important consequence of the GCLT is that there are no distributions other than the family of stable laws which can govern the limit behavior of sums of i.i.d. random variables. This is also known as the *domains of attraction* property. Therefore, we can expect that resorting to the GCLT, it would be possible to find a generalized version of Theorem 7.3.1.

In order to state the more general limit result for sample AVaR, we need to characterize the domains of attraction of stable distributions as they are going to play an essential role. The characterization result relies on the concept of slowly varying functions. A positive function $L(x)$ is said to be slowly varying at infinity if the following limit relation is satisfied:

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1, \quad \forall t > 0. \tag{7.5.1}$$

We took advantage of this concept in the appendix to Chapter 6 to characterize tail behavior in order to study the problem under which conditions spectral risk measures are bounded. In this case, it characterizes the domains of attraction.[9]

Theorem 7.5.1. Let $X_1, \ldots, X_n$ be i.i.d. with c.d.f. $F(x)$. There exist $a_n > 0$, $b_n \in \mathbf{R}$, $n = 1, 2, \ldots$, such that the distribution of

$$a_n^{-1}[(X_1 + \ldots + X_n) - b_n]$$

converges as $n \to \infty$ to $S_\alpha(1, \beta, 0)$ if and only if both

(i) $x^\alpha[1 - F(x) + F(-x)] = L(x)$ is slowly varying at infinity.

(ii) $\dfrac{1 - F(x) - F(-x)}{1 - F(x) + F(-x)} \to \beta$ as $x \to \infty$.

The sequence a_n must satisfy

$$\lim_{n\to\infty} \frac{nL(a_n)}{a_n^\alpha} = \begin{cases} (\Gamma(1-\alpha)\cos(\pi\alpha/2))^{-1} & \text{if} \quad 0 < \alpha < 1, \\ 2/\pi & \text{if} \quad \alpha = 1, \\ \left(\frac{\Gamma(2-\alpha)}{\alpha-1}|\cos\frac{\pi\alpha}{2}|\right)^{-1} & \text{if} \quad 1 < \alpha < 2. \end{cases} \tag{7.5.2}$$

The sequence b_n may be chosen as follows:

$$b_n = \begin{cases} 0 & \text{if} \quad 0 < \alpha < 1, \\ na_n \displaystyle\int_{-\infty}^{\infty} \sin(x/a_n) dF(x) & \text{if} \quad \alpha = 1, \\ n \displaystyle\int_{-\infty}^{\infty} x dF(x) & \text{if} \quad 1 < \alpha < 2. \end{cases} \tag{7.5.3}$$

In all cases, the sequence a_n can be represented as $a_n = n^{1/\alpha} L_0(n)$ where $L_0(n)$ is a function slowly varying at infinity.

If the index α characterizing the tails of the c.d.f. $F(x)$ in condition (i) satisfies $\alpha \geq 2$, then the tail index of the limiting distribution equals $\alpha^* = 2$. Thus, the relationship between the tail index of the limiting distribution, which we denote by α^*, and the tail index in condition (i) can be generalized as $\alpha^* = \min(\alpha, 2)$. If $\alpha > 2$, then $EX_1^2 < \infty$ and we are in the setting of the classical CLT. The centering and normalization can be done by choosing $b_n = nEX_1$ and $a_n = n^{1/2}\sigma_{X_1}$, where σ_{X_1} denotes the standard deviation of X_1. The

case $\alpha = 2$ is more special because the variance of X_1 is infinite and a_n cannot be chosen in this fashion. Moreover, the proper normalization cannot be obtained by computing the limit $\alpha \to 2$ in equation (7.5.2). Under the simpler assumptions that the function $L(x)$ in condition (i) equals a constant, Zolotarev and Uchaikin (1999) provide the formula $a_n = (n \log n)^{1/2} A^{1/2}$.

The generalized result providing the limit distribution of sample AVaR is given below. A proof of it can be found in the appendix to this chapter.

Theorem 7.5.2. Suppose that X is a random variable with c.d.f. $F(x)$ which satisfies the following conditions:

(a) $x^\alpha F(-x) = L(x)$ is slowly varying at infinity.
(b) $\int_{-\infty}^{0} x dF(x) < \infty$.
(c) $F(x)$ is differentiable at $x = q_\epsilon$, where q_ϵ is the ϵ-quantile of X.

Then, there exist $c_n > 0$, $n = 1, 2 \ldots$, such that for any $0 < \epsilon < 1$,

$$c_n^{-1} \left(\widehat{AVaR}_\epsilon(X) - AVaR_\epsilon(X) \right) \xrightarrow{w} S_{\alpha^*}(1, 1, 0), \tag{7.5.4}$$

in which $\xrightarrow{w}$ denotes weak limit, $1 < \alpha^* = \min(\alpha, 2)$, and $c_n = n^{1/\alpha^* - 1} L_0(n)/\epsilon$ where L_0 is slowly varying at infinity. Furthermore, the c_n are representable as $c_n = a_n/n\epsilon$ where a_n stands for the normalizing sequence in Theorem 7.5.1 and must satisfy the condition in equation (7.5.2).

If $\alpha > 2$ in condition (a), then $\int_{-\infty}^{0} x^2 dF(x) < \infty$ and the limiting distribution is the standard normal distribution. Thus, under this condition we obtain the result in Theorem 7.3.1. In this case, the normalizing sequence c_n should be calculated using $d_\epsilon^2 = D(q_\epsilon - X)_+$,

$$c_n = n^{-1/2} d_\epsilon/\epsilon = n^{-1/2} \sigma_\epsilon,$$

where σ_ϵ is defined in equation (7.3.4).

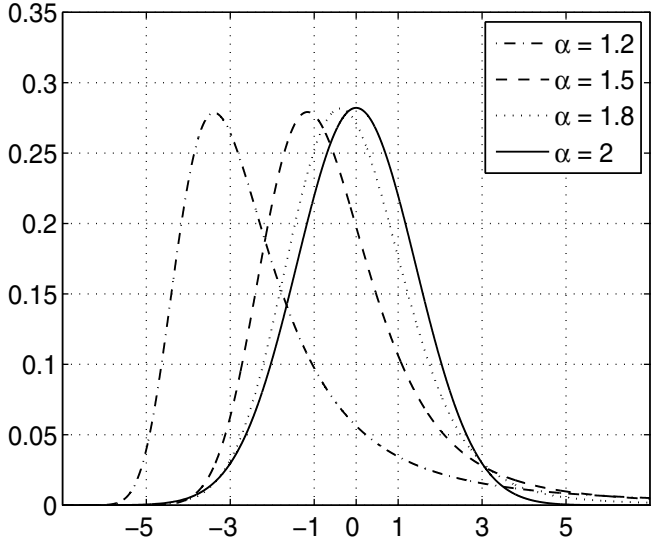

Figure 7.7: Densities of the limiting stable distribution corresponding to different tail behavior.

By definition, the AVaR is the negative of the average of the quantiles of X beyond a reference quantile q_ϵ. For this reason, as we noted already, it is only the behavior of the left tail of X which matters and both assumptions (a) and (b) in Theorem 7.5.2 concern the left tail only. Condition (c), just as in the more limited Theorem 7.3.1, is only technical. It is automatically satisfied if X has a density function.

The limiting stable distribution is totally skewed to the right, $\beta = 1$. However, the observed skewness in the shape of the distribution decreases as $\alpha \to 2$ (see Figure 7.7). At the limit, when $\alpha = 2$, the limiting distribution is Gaussian and is symmetric irrespective of the value of β. Therefore, the degree of the observed skewness in the limiting distribution is essentially determined by the tail behavior of X, or by the value of α, and is not influenced by any other characteristic.

When $\epsilon \to 1$, then AVaR approaches the mean of X (or the sample average if we consider the sample AVaR),

$$\lim_{\epsilon \to 1} AVaR_\epsilon(X) = EX.$$

Unfortunately, there is no such continuity in equation (7.5.4) unless X has a finite variance. That is, generally it is not true that the weak limit in equation (7.5.4) holds for the sample average letting $\epsilon \to 1$. The reason is that if $\epsilon = 1$, then both tails of the distribution of X matter and the limiting stable distribution can have any $\beta \in [-1, 1]$. The condition $DX < \infty$ is sufficient to guarantee that the limiting distribution is normal for any $\epsilon \in (0, 1]$ and in this case there is continuity in equation (7.5.4) as $\epsilon \to 1$.

As an illustration of the singularity at $\epsilon = 1$, consider the following example. Suppose that the right tail of X is heavier than the left tail and, as a consequence,

$$\int_{-\infty}^{q_\epsilon} x^2 dF(x) < \infty, \quad \text{for any} \quad \epsilon < 1,$$

but $EX^2 = \infty$. Under this assumption, the limiting distribution of the sample AVaR is normal for any $\epsilon < 1$. If $\epsilon = 1$, then the limiting distribution becomes stable non-Gaussian due to the heavier right tail. Thus, there is a change in the limiting distribution of the sample AVaR with $\epsilon < 1$ and the sample average.

The result in (7.5.4) is not as easy to apply as the simpler result in Theorem 7.3.1. Difficulties arise because it is harder to calculate the normalizing sequence c_n. If we assume that we know the distribution of X and we can calculate or estimate the slowly varying function $L_0(n)$ for a given choice of n (e.g. $n = 10{,}000$), and also the tail exponent α of X, then we can apply the result in (7.5.4) in the following way. The 95% confidence interval for AVaR can be computed according to,

$$\widehat{AVaR}_\epsilon(X) - q_{2.5\%} c_n \leq AVaR_\epsilon(X) \leq \widehat{AVaR}_\epsilon(X) + q_{97.5\%} c_n$$

where $c_n = n^{1/\alpha^* - 1} L_0(n)/\epsilon$ and $q_{2.5\%}$ and $q_{97.5\%}$ are the 2.5% and 97.5% quantiles from the distribution $S_{\alpha^*}(1, 0, 0)$ where $\alpha^* = \min(\alpha, 2)$. The two quantile levels can be computed by numerical inversion if the c.d.f. of the stable distribution can be approximated using either the fast Fourier transform method or the integral representations of Zolotarev. For details on numerical work with stable distributions, see, for example, Stoyanov and Racheva-Iotova (2004) and the references therein.

In the next section, we provide examples of how the normalizing sequence c_n can be computed for two particular choices for the distribution of X – stable Paretian and Student's t distributions.

7.6 Rate of Convergence, Heavy-tailed Returns

The result in Theorem 7.5.2 provides the limiting distribution but does not provide any insight on the rate of convergence. That is, it does not give an answer to the question of how many scenarios are needed in order for the distribution of the left-hand side in equation (7.5.4) to be sufficiently close to the distribution of the right-hand side in terms of a selected probability metric. In this section, we provide illustrations of the stable limit theorem and the rate of convergence assuming particular distributions of X.

7.6.1 Stable Paretian distributions

We remarked that stable Paretian distributions are stable distributions with tail index $\alpha < 2$. This distinction is made since their properties are very different from the properties of the normal distribution which appears as a stable distribution with $\alpha = 2$. For example, in contrast to the normal distribution, stable Paretian distributions have heavy tails exhibiting power decay. In the field of finance, stable Paretian distribution were proposed as a model for stock returns and other financial variables.[10]

Denote by X the random variable describing the return of a given stock. In this section, we assume that $X \in S_\alpha(\sigma, \beta, \mu)$ with $1 < \alpha < 2$,

$\beta \neq 1$, and our goal is to apply the result in Theorem 7.5.2 which provides a tool for computing the confidence interval of the sample AVaR of X on condition that the Monte Carlo method is used with a large number of scenarios. Since by assumption $\alpha > 1$, guaranteeing convergence of the sample AVaR to the theoretical AVaR in almost sure sense. In the case of stable distributions, the quantity $AVaR_{\epsilon}(X)$ can be calculated using a semi-analytic expression given in Stoyanov et al. (2006).

In order to apply the result in Theorem 7.5.2, first we have to check if the conditions are satisfied and then choose the scaling constants c_n. For this purpose, we use the following property of stable Paretian distributions.[11]

Property 7.6.1. Let $X \in S_{\alpha}(\sigma, \beta, \mu)$ $0 < \alpha < 2$. Then

$$\lim_{\lambda \to \infty} \lambda^{\alpha} P(X > \lambda) = C_{\alpha} \frac{1+\beta}{2} \sigma^{\alpha}$$

$$\lim_{\lambda \to \infty} \lambda^{\alpha} P(X < -\lambda) = C_{\alpha} \frac{1-\beta}{2} \sigma^{\alpha}$$

where

$$C_{\alpha} = \left(\int_0^{\infty} x^{-\alpha} \sin(x) dx \right)^{-1} = \begin{cases} \frac{1-\alpha}{\Gamma(2-\alpha)\cos(\pi\alpha/2)}, \alpha \neq 1 \\ 2/\pi, \qquad\qquad \alpha = 1 \end{cases}$$

This property provides the asymptotic behavior of the left tail of the distribution. We further assume that $\beta \neq 1$ since in this case the asymptotic behavior of the left tail is different.[12] Condition (b) is satisfied because of the assumption $1 < \alpha < 2$ and, finally, condition (c) is satisfied for any choice of $0 < \epsilon < 1$ since all stable distributions have densities. Therefore, all assumptions are satisfied and the result in Theorem 7.5.2 holds with $\alpha^* = \alpha$ and the scaling constants c_n should be chosen in the following way:

$$c_n = n^{1/\alpha - 1} \left(\frac{1-\beta}{2} \right)^{1/\alpha} \frac{\sigma}{\epsilon}.$$

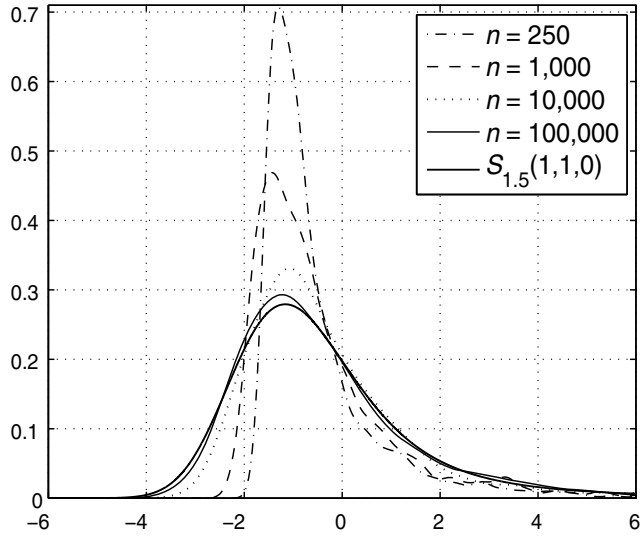

Figure 7.8: The density of the sample AVaR as n increases with $\beta = 0.7$ and $\epsilon = 0.01$.

Note that in this case, the skewness in the distribution of X translates into a different scaling of the normalizing constants. If X is negatively skewed ($\beta < 0$), the scaling factor is larger than if X is positively skewed ($\beta > 0$).

We carry out a Monte Carlo study assuming $X \in S_{1.5}(\beta, 1, 0)$ where $\beta = \pm 0.7$ and two choices of the tail probability $\epsilon = 0.01$ and $\epsilon = 0.05$. We generate 2,000 samples from the corresponding distribution, the size of which equals $n = 250; 1{,}000; 10{,}000;$ and 100,000.

Figure 7.8 illustrates the convergence rate for the case $\epsilon = 0.01$ as the number of scenarios increases. While from the plot it seems that $n = 100{,}000$ results in a density which is very close to that of the limiting distribution, the Kolmogorov test fails. The convergence rate is much slower in the heavy-tailed case than in the setting of the classical CLT. Apparently, many more scenarios are needed in this heavy-tailed case in comparison to the finite-variance case.

The plots in Figure 7.9 indicate that as the tail probability ϵ increases, the behavior of the sample AVaR distribution improves.

Furthermore, the behavior improves when X turns from being negatively to positively skewed.

7.6.2 Student's t distribution

We used Student's t distribution in section 7.4.1. It can be demonstrated[13] that for the Student's t distribution,

$$\lim_{x\to\infty} x^{\nu} F(-x) = \nu^{\nu/2-1} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\pi}}. \tag{7.6.1}$$

This distribution is interesting because by varying the degree of freedom parameter, we can explore what happens to the rate of convergence as we shift between the classical result in Theorem 7.3.1 and the more general result in Theorem 7.5.2.

The result in this proposition and Theorem 7.5.2 imply that for $\nu > 2$, the limiting distribution of the sample AVaR is the Gaussian distribution. If $1 < \nu \leq 2$, then the limiting distribution is stable with $\alpha^* = \nu$. If $\nu \leq 1$, then the AVaR of X diverges. The scaling constants c_n should be chosen in a different way depending on the value of ν,

$$c_n = \begin{cases} n^{-1/2}\sigma_\epsilon, & \text{if } \nu > 2 \\ n^{1/\nu-1} A_\nu/\epsilon, & \text{if } 1 < \nu < 2 \end{cases} \tag{7.6.2}$$

where σ_ϵ is given in equation (7.3.4) and

$$A_\nu^\nu = \nu^{\nu/2-1} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma(\nu/2)\sqrt{\pi}} \frac{\Gamma(2-\nu)}{\nu-1} |\cos(\pi\nu/2)|.$$

The value of the constant A_ν is obtained by taking into account the limit in (7.6.1) and the condition in equation (7.5.2). The case $\nu > 2$ is covered by our result in Theorem 7.3.1. This case is in the classical setting of the CLT as the variance of X is finite.

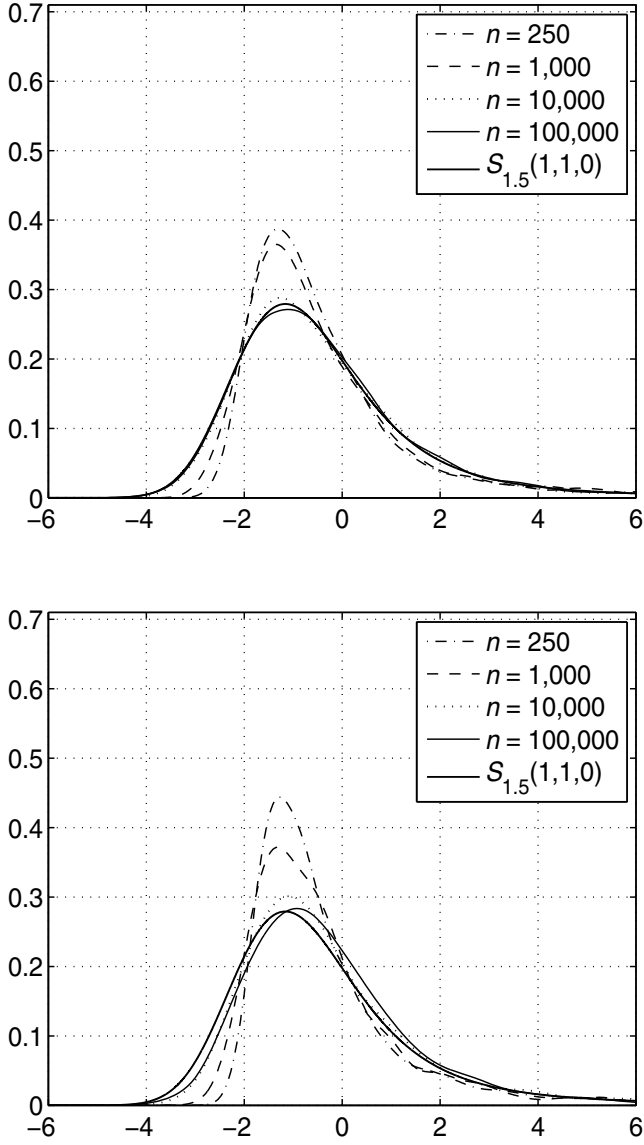

Figure 7.9: The density of the sample AVaR as n increases with $\beta = 0.7$ (top) and $\beta = -0.7$ (bottom) and $\epsilon = 0.05$.

We carry out a Monte Carlo experiment in order to study the convergence rate of the sample AVaR distribution to the limiting distribution. We fix the degrees of freedom, the number of simulations to 100,000, and $\epsilon = 0.05$. Next we generate 2,000 samples from which the sample AVaR is estimated. Thus we obtain 2,000 estimates of $AVaR_\epsilon(X)$, $X \in t(\nu)$. Finally, we calculate the Kolmogorov distance

$$\rho(G_\nu, G) = \sup_x |G_\nu(x) - G(x)|$$

where G_ν is the c.d.f. of the sample AVaR approximated by the sample c.d.f. obtained with the 2,000 estimates, and G is the c.d.f. of the limiting distribution $S_{\alpha^*}(1, 1, 0)$ where $\alpha^* = \min(\nu, 2)$.

Figure 7.10 shows the values of $\rho(G_\nu, G)$ as ν varies from 1.05 to 3. The horizontal line shows the critical value of the Kolmogorov statistic: if the calculated $\rho(G_\nu, G)$ is below the critical value, we accept the hypothesis that the sample AVaR distribution is the same as the limiting distribution, otherwise we reject it. Since we use a sample c.d.f. to approximate $G_\nu(x)$, the solid line fluctuates a little but we notice that for $\nu \leq 1.5$ and $\nu \geq 2.5$ it seems that 100,000 scenarios are enough in order to accept the limiting distribution as a model. For the middle values, larger samples are needed. This observation indicates that the rate of convergence of the sample AVaR distribution to the limiting distribution deteriorates as ν approaches 2 and is slowest for $\nu = 2$. This finding can be summarized in the following way by considering all possible cases for ν:

- $\nu > 2$. As ν decreases from larger values to 2, the tail thickness increases, which results in higher absolute moments becoming divergent, $E|X|^\delta = \infty$, $\delta \geq \nu$. The limiting distribution of sample AVaR is the Gaussian distribution but the tails becoming thicker results in a deterioration of the convergence rate to the Gaussian distribution.
- $\nu = 2$. The limiting distribution of sample AVaR is the Gaussian distribution even though the variance of X is infinite. This case is not covered by Theorem 7.3.1 but is contained in the more general Theorem 7.5.2.

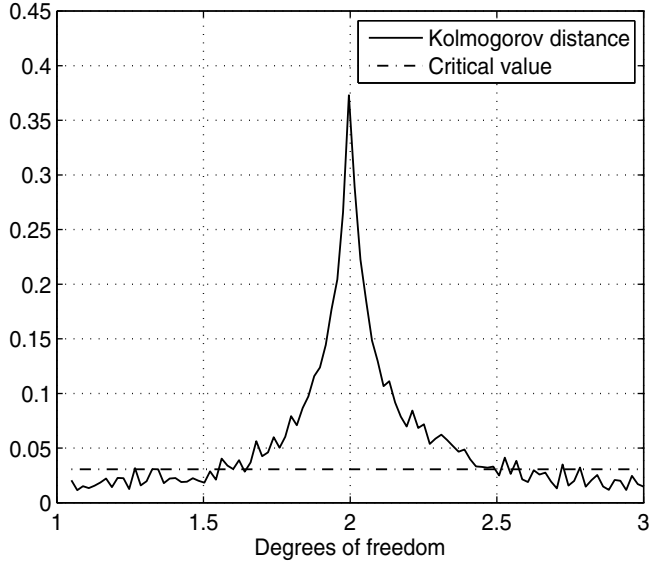

Figure 7.10: The Kolmogorov distance between the sample AVaR distribution of $X \in t(\nu)$ obtained with 100,000 simulations and the limiting distribution.

- $1 < \nu < 2$. We continue decreasing ν and the tails become so thick that they start influencing the limit distribution which is stable Paretian, $S_\nu(1, 1, 0)$, and already depends on ν. However, the convergence rate starts improving.
- $0 < \nu \leq 1$. The tails of X become so heavy that $AVaR_\epsilon(X) = \infty$. Consequentially, no asymptotic theory exists.

Therefore, from the standpoint of the limit distribution of sample AVaR, the case $\nu = 2$ represents a threshold at which a phase transition occurs. This behavior is nothing specific to the Student's t distribution but is generic to families of distributions including representatives with heavy and light tails. As a result, such a phase transition in the probability law governing the approximation accuracy of sample AVaR will be observed with other classes of distributions. This has to be taken into account when using such

families to model stock returns since, as theory implies, different tail behavior leads to different approximation quality of the Monte Carlo method when using samples of equal sizes.

7.7 On the Choice of a Distributional Model

In the discussion of the distribution of sample AVaR, we emphasized the importance of the assumed probabilistic model for the underlying return distribution. We demonstrated that its tail behavior determines the asymptotic theory of sample AVaR.

In fact, the assumption for the return distribution is important not only for the asymptotic theory, which is a question of secondary importance, but it determines to a large extent how appropriate the AVaR estimate itself is. This remark is valid with respect not only to AVaR, but also to any other downside tail risk measure. Therefore, as we noted in the introduction to this chapter, choosing a model for the return distribution is an important question which deserves special attention.

In this section, we do not aim at covering the topic in detail. Our goal is to provide insight into one aspect, which is based on a result from the theory of probability metrics known as a *pre-limit theorem*.

7.7.1 Tail behavior and return frequency

There is no fundamental theory in finance that can derive the probabilistic model of stock returns processes from basic principles. Therefore, the search for a probabilistic model is an empirical question. A good model should be able to account for phenomena which have been observed in the data through empirical studies. Such phenomena include:

- volatility clustering (ARCH-effects)
- temporal dependence of the tail behavior
- short- and long-range dependence
- non-Gaussian, heavy-tailed and skewed distributions for the "building blocks" of the time-series model (e.g., innovations).[14]

To what degree these properties are pronounced in the data depends largely on the return frequency but also on whether we consider normal or distressed markets. In normal markets, lower frequency returns (e.g., monthly) tend to be less heavy tailed and with less volatility clustering. The fact that the thin-tailed normal distribution can be very misleading for risk estimation has been pointed out in numerous papers and books.[15] It was also acknowledged by the Financial Services Authority[16] in a discussion paper about the reasons for the banking crisis of 2008: "Short-term observation periods plus assumption of normal distribution can lead to large underestimation of probability of extreme loss events."[17]

Consequentially, if we would like the probabilistic model to have a consistent behavior across different frequencies, it has to be heavy tailed at higher frequencies (e.g., daily) and if we aggregate the returns to a lower frequency through the model (e.g., monthly), they have to become less heavy tailed (e.g., possibly close to the normal distribution for monthly returns). The return aggregation can be illustrated as a process in which we draw sample paths of 22 daily returns (i.e., typical number of trading days in a month) and sum them to compute one monthly return.[18]

In case we use a time series model, one component of the daily returns sum is the sum of the daily residuals, which are assumed to be i.i.d. random variables. Therefore, one determinant of the distribution of the monthly return is the behavior of this sum. In the following discussion, we consider only this component as it can be studied generically without referring to the particular form of the time series model.

Denote the sum of the residuals by

$$S_n = \sum_{i=1}^{n} \epsilon_i \tag{7.7.1}$$

where ϵ_i are i.i.d. copies of ϵ_1. Delving further into the question of what distribution for ϵ_1 is appropriate, we confront a contradiction. If we assume, as supported by empirical studies,[19] a heavy-tailed

distribution with an infinite variance for ϵ_1, then the behavior of S_n is also heavy tailed with the same tail behavior as ϵ_1. On the other hand, a thin-tailed distribution for S_n, supported by empirical evidence, can be obtained only if we assume that ϵ_1 is not heavy tailed.

There is a way to resolve the puzzle by noticing that empirical data in normal markets suggest a decrease in tail thickness as the number of summands in (7.7.1) increases.[20] Therefore, we can resolve the contradiction by assuming a model which is in the normal domain of attraction: that is, for large n, S_n is close to the normal distribution, but such that for smaller n, S_n has a behavior *similar* to that of a heavy-tailed model.

An example of such a model is the truncated stable distribution discussed in section 7.4.3 but if both tails are truncated, for example at 0.1% and 99.9% quantiles. The truncated stable distribution has a bounded support and, therefore, has finite moments of all orders. In effect, the sum in (7.7.1) is approximately normal when n is large. However, if n is small (e.g., $n = 5$), the sum is well described by a stable distribution because the truncation points are deep into the tails and the overall shape of the distribution is close to the shape of the corresponding stable distribution.

This is illustrated in Figure 7.11 through a Monte Carlo experiment. We generated 150 samples of 10,000 scenarios from a truncated standardized symmetric stable distribution with $\alpha = 1.7$. Using the generated samples, we computed 10,000 scenarios of S_5 and S_{150} as defined in (7.7.1) and then we fitted a stable and a normal distribution. The tail index of S_5 was estimated to be $\hat{\alpha} = 1.75$ by the method of maximum likelihood. It is higher than 1.7 but significantly smaller than 2, which corresponds to the normal distribution. The top plot of Figure 7.11 clearly shows that the fitted stable density describes S_5 much better than the fitted normal density. Actually, the fitted stable density can hardly be distinguished from the non-parametric kernel estimate of the scenarios density. The outcome is quite different for S_{150}. From the bottom plot of Figure 7.11, we can see that the normal distribution describes S_{150} perfectly well.

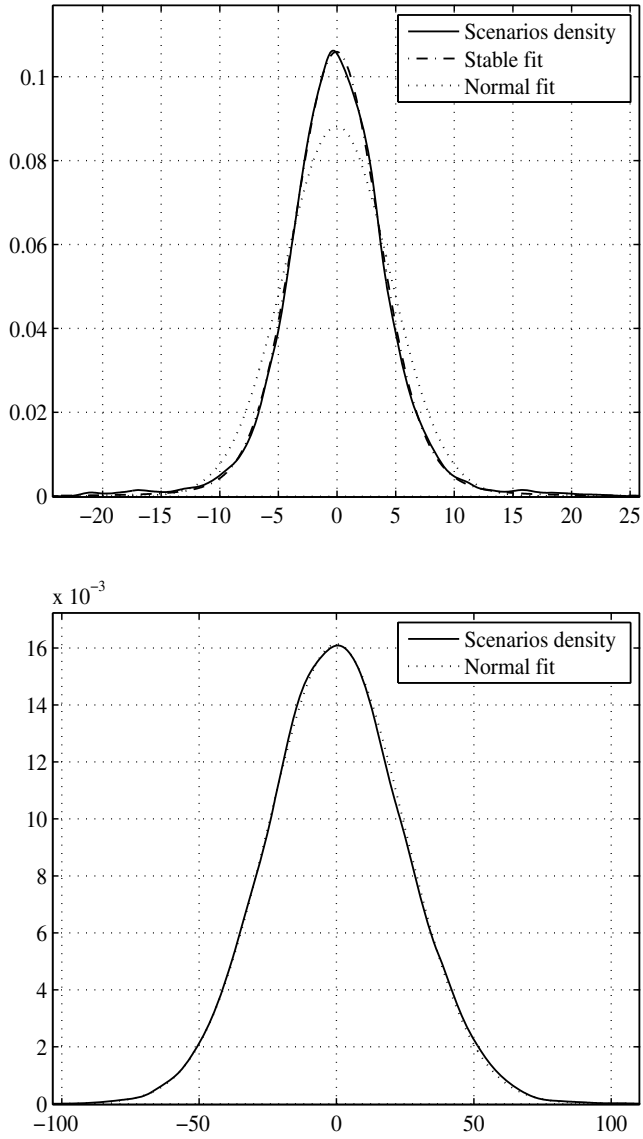

Figure 7.11: Approximating S_5 (top) and S_{150} (bottom) as defined in (7.7.1) of a truncated symmetric stable distribution with $\alpha = 1.7$ via Monte Carlo. The maximum likelihood estimate of the tail index of the stable fit on the top plot is $\hat{\alpha} = 1.75$.

The fact that a stable distribution can be accepted as a model for S_n even though ϵ_1 is in the domain of attraction of the normal distribution is known as a pre-limit theorem. The distance between a properly normalized sum of i.i.d. random variables and a stable distribution in this context is studied by means of the following probability metric:

$$k_h(F, G) := \sup_{x \in \mathbb{R}} |F * h(x) - G * h(x)| \tag{7.7.2}$$

where $*$ stands for convolution and the "smoothing" function $h(x)$ stands for a c.d.f. with a bounded continuous density function, $\sup_x |h'(x)| \leq c(h) < \infty$. The convolution $F * h(x)$ can be viewed as a c.d.f. smoothed with a kernel function equal to $h'(x)$. In this way, we do not compare directly the c.d.f.s F and G but kernel-based smooth approximations of them.[21]

In the pre-limit theorem, just as in other limit theorems, we use appropriately scaled sums. With a slight abuse of the classical notation, we define

$$S_{n,\alpha} = n^{-1/\alpha} \sum_{i=1}^{n} \epsilon_i. \tag{7.7.3}$$

The distance between $S_{n,\alpha}$ and an α-stable random variable is bounded by means of the following semidistance between the random variables X and Y:

$$d_{c,\gamma}(X, Y) = \sup_{|t|>c} \frac{|\varphi_X(t) - \varphi_Y(t)|}{|t|^\gamma} \tag{7.7.4}$$

where c and γ are two positive constants, and $\varphi_X(t) = Ee^{itX}$ denotes the characteristic function (ch.f.) of X. The semidistance $d_{c,\gamma}(X, Y)$ does not take into account the differences in the tails between X and Y since for any $c > 0$, the origin is excluded in the calculation of the supremum in (7.7.4) and it is a well-known fact from the theory of probability that the behavior of the ch.f. at $t = 0$ determines the tail behavior of the random variable.[22]

Furthermore, if $d_{0,\gamma}(S_{n,\alpha}, Y) < \infty$ for some $\gamma > \alpha$, where Y is strictly α-stable random variable, then $S_{n,\alpha}$ is in the domain of attraction of

Y. From our discussion so far, it is evident that we are interested in the case in which $S_{n,\alpha}$ is in the domain of attraction of the normal distribution and, therefore, $d_{0,\gamma}(S_{n,\alpha}, Y) = \infty$. As a consequence, we consider $d_{c,\gamma}(X, Y)$ with $c > 0$ which is bounded.

The result known as the *central pre-limit theorem*, which we adapt from Klebanov et al. (2009), is stated below.

Theorem 7.7.1. (central pre-limit theorem) Let ϵ, ϵ_j, $j > 1$ be i.i.d. random variables and $S_{n,\alpha} = n^{-1/\alpha} \sum_{j=1}^{n} \epsilon_j$. Suppose that Y is a strictly α-stable random variable. Let $\gamma > \alpha$ and $\Delta > \delta$ be arbitrary given positive constants and let $n < (\Delta/\delta)^{\alpha}$ be an arbitrary positive integer. Then

$$k_h(S_{n,\alpha}, Y) \leq \inf_{a>0} \left(\sqrt{2\pi} \frac{d_{\delta,\gamma}(\epsilon, Y)(2a)^{\gamma}}{n^{\frac{\gamma}{\alpha}-1}\gamma} + 2\frac{c(h)}{a} + 2\Delta a \right) \tag{7.7.5}$$

As consequence of the theorem, the c.d.f. of normalized sum of i.i.d. random variables is close to the c.d.f. of the corresponding α-stable random variable for "mid-size values" of n. We also see that for these values of n, the closeness of $S_{n,\alpha}$ to a strictly α-stable random variable depends on the body of the distribution of ϵ. Finally, the metric $k_h(S_{n,\alpha}, Y)$ emphasizes the closeness of the "middle parts" of the distributions of $S_{n,\alpha}$ and Y, which coincides with our conclusions based on Figure 7.11.

In order for the analysis to be complete, note that if ϵ happens to be in the domain of attraction of an α-stable distribution, then Theorem 7.7.1 turns into a rate of convergence theorem for GCLT. In this case, as we noted, the distance $d_{\delta,\gamma}(\epsilon, Y)$ is finite for $\delta = 0$ and, therefore, the theorem holds for arbitrarily large n.

7.7.2 Practical implications

From a practical viewpoint, we can think of the bound on the right-hand side of equation (7.7.5) in the following way. Suppose that the random variable ϵ is such that for some $\gamma > \alpha$, the distance $d_{\delta,\gamma}(\epsilon, Y)$ remains reasonably small even for small values of $\delta > 0$. Then, we can choose Δ and compute a value for a such that the bound on

the right-hand side of (7.7.5) remains small even for relatively large values of n. Therefore, for a given random variable ϵ, there exists a constant N_ϵ such that for $n < N_\epsilon$ the distance $k_h(S_{n,\alpha}, Y)$ is small and we can accept the strictly α-stable random variable as a model for the central part of the distribution of $S_{n,\alpha}$.

The constant N_ϵ can be thought of as the natural scale of ϵ. While there is no sudden change in regime, on an intuitive level we can think that below N_ϵ, a strictly α-stable distribution determines the properties of the body of ϵ and above it, this is done by the normal distribution. In case we would like to build a model which is to be used for aggregation of returns across frequencies, the magnitude of the natural scale can be one criterion for model selection. The natural scales of different random variables are computed in Grabchak and Samorodnitsky (2009).

Theorem 7.7.1 also has important implications for the application of stable distributions as a model for stock returns. In a very broad sense, it justifies the use of stable laws as an approximate model for sums of i.i.d. random variables. However, under the assumption that in usual market conditions the stock returns at lower frequencies possess thinner tails, there is a caveat. The distance $k_h(S_{n,\alpha}, Y)$ emphasizes the differences in the body part of the distributions, which means that the α-stable law can be accepted as a good model for the central region of $S_{n,\alpha}$. The tails of $S_{n,\alpha}$, however, converge faster than the tails of Y by construction. As a consequence, the AVaR of Y at a small tail probability (e.g., below 1%) is larger than the AVaR of $S_{n,\alpha}$. Therefore, under the current assumptions, the AVaR computed using information only from the extreme tail of the approximating α-stable distribution is conservative and may overestimate the corresponding true AVaR.

Nevertheless, in times of severe market crashes, we may have reasons to believe that the aggregation property we assumed in normal markets does not hold for the period of the market crash. Thus, we may accept the hypothesis that the tail behavior of lower-frequency returns for that period does not differ much from the tail behavior of the higher-frequency returns. If we assume that ϵ is in the domain of attraction of an α-stable distribution, then, as we noted,

Theorem 7.7.1 turns into a rate of convergence theorem for GCLT. This implies that the approximation Y can be a good model not only for the body of the distribution but also for the tail behavior and, as a result, the AVaR of Y can be a reasonable approximation for the true AVaR.

Turning back to the question of choosing a return distribution when markets are normal, Theorem 7.7.1 suggests that we adopt a model in which we preserve the shape of stable distributions for the body part but change the tails to be more quickly decaying than those of stable laws. We considered a very crude method for tail truncation in section 7.4.3 in which we directly remove the tail. A more refined approach concerns the so-called *tempered stable distributions* which are used as a model for the residuals in a time series process and are successfully applied in option pricing theory.[23]

7.8 Summary

In this chapter, we explored the application of the Monte Carlo method for estimating AVaR. The Monte Carlo method is a very common numerical technique which is applied when the probabilistic model is involved and analytic calculations are not feasible, or even when the probabilistic model is not known in an analytic form but is relatively easy to draw scenarios from.

A generic issue with the Monte Carlo method is the variability of statistical estimators due to their dependence on the generated scenarios. This variability introduces an error which can be studied and estimated through the asymptotic distribution of the estimator.

In this chapter, we provided a classical limit result valid under the assumption that the random variable describing the return of a common stock has a finite second moment. In this case, the limit distribution is the normal distribution. We also provided a generalized result the limit distribution in which is a totally skewed stable distribution. This result applies when the random variable modeling the return distribution is heavy tailed and has infinite variance.

The limit distributions in these theorems describe the approximation accuracy of the Monte Carlo method. They are more precise

when the generated sample is larger. The number of scenarios needed in order to trust the limit distribution for assessing the accuracy of the Monte Carlo method can be inferred from the rate of convergence to the limit distribution. Generally speaking, the most important factor for the convergence rate is how heavy the tails of the probabilistic model are. We studied this issue with Student's t distribution, which contains both finite and infinite variance representatives. We found out that the lighter the tails are, the fewer scenarios we need to estimate AVaR at the same precision level. As the tail of the distribution becomes heavier, a phase transition occurs at some point and the limit distribution becomes dependent on the tail behavior. The convergence rate, however, improves.

Concerning the question of improving the rate of convergence, we suggested the method of tail truncation, which introduces a bias but reduces dramatically the sample size necessary to accept the limit distribution as a model for the approximation error of the Monte Carlo method.

7.9 Technical Appendix

In the technical appendix to this chapter, we provide a proof of the stable limit result in Theorem 7.5.2. We do not provide a separate proof of Theorem 7.3.1 since it arises as a special case. Nevertheless, a separate proof based on the influence functions approach can be found in Stoyanov and Rachev (2008*a*). A more detailed discussion of the proof in this technical appendix is available in Stoyanov and Rachev (2008*b*).

7.9.1 Proof of the stable limit result

In order to develop the limit theorem, we need a few additional facts related to building a linear approximation to AVaR and estimating the rate of improvement of the linear approximation. They are collected in the following proposition.

Proposition 7.9.1. Suppose X is a random variable with c.d.f. F which satisfies the condition $E\max(-X, 0) < \infty$ and F is differentiable at the

ϵ-quantile of X. Denote by F_n the sample c.d.f. of $X_1, \ldots, X_n$ which is a sample of i.i.d. copies of X. There exists a linear functional Δ defined on the difference $G - F$ where the functions G and F are c.d.f.s, such that

$$\left|\phi(F_n) - \phi(F) - \Delta(F_n - F)\right| = o(\rho(F_n, F)) \tag{7.9.1}$$

where $\rho(F_n, F) = \sup_x |F_n(x) - F(x)|$ stands for the Kolmogorov metric and

$$\phi(G) = -\frac{1}{\epsilon}\int_0^{\epsilon} G^{-1}(p)dp$$

in which G^{-1} is the inverse of the c.d.f. G. The linear functional Δ has the form

$$\Delta(F_n - F) = \frac{1}{\epsilon}\int_{-\infty}^{q_\epsilon} (q_\epsilon - x)d(F_n(x) - F(x)). \tag{7.9.2}$$

where q_ϵ is the ϵ-quantile of X.

Proof. For a detailed proof, see Stoyanov and Rachev (2008*b*). □

Corollary 7.9.1. Under the assumptions in the proposition,

$$\left|\phi(F_n) - \phi(F) - \Delta(F_n - F)\right| = o(n^{-1/2}). \tag{7.9.3}$$

Proof. By the Kolmogorov theorem, the metric $\rho(F_n, F)$ approaches zero at a rate equal to $n^{-1/2}$ which indicates the rate of improvement of the linear approximation $\Delta(F_n - F)$. □

The main result is given in Theorem 7.5.2. The idea is to use the linear approximation $\Delta(F_n - F)$ of the AVaR functional in order to obtain an asymptotic distribution as $n \to \infty$. We reproduce below the proof given in Stoyanov and Rachev (2008*b*).

Proof. By the result in Theorem 7.9.1,

$$\phi(F_n) - \phi(F) = \Delta(F_n - F) + o(n^{-1/2}) \tag{7.9.4}$$

where ϕ is the AVaR functional and $\Delta(F_n - F)$ is given in (7.9.1). Simplifying the expression for $\Delta(F_n - F)$, we obtain

$$\phi(F_n) - \phi(F) = \frac{1}{n\epsilon} \sum_{i=1}^{n} \left[(q_\epsilon - X_i)_+ - E(q_\epsilon - X_i)_+ \right] + o(n^{-1/2}) \quad (7.9.5)$$

It remains to apply the domains of attraction characterization in Theorem 7.5.1 to the right-hand side of equation (7.9.5). For this purpose, consider the expression

$$\sum_{i=1}^{n} Y_i - nEY_1 \quad (7.9.6)$$

where $Y_i = (q_\epsilon - X_i)_+$ are i.i.d. random variables. Denote by $F_Y(x)$ the c.d.f. of Y. The left-tail behavior of X assumed in (a) implies $x^\alpha(1 - F_Y(x)) = L(x)$ as $x \to \infty$ where $L(x)$ is the slowly varying function assumed in (a). This is demonstrated by

$$\begin{aligned} x^\alpha(1 - F_Y(x)) &= x^\alpha P(\max(q_\epsilon - X, 0) > x) \\ &= x^\alpha P(X < q_\epsilon - x) \\ &\sim x^\alpha P(X < -x) \end{aligned} \quad (7.9.7)$$

Furthermore, the asymptotic behavior of the left tail of Y is $F_Y(-x) = 0$ which holds for any $x \geq -q_\epsilon$. As a result, condition (i) from Theorem 7.5.1 holds.

Condition (b) implies that the tail exponent α in (a) must satisfy the inequality $\alpha > 1$. Therefore, subtracting nEY_1 in (7.9.6) is a proper centering of the sum as suggested in (7.5.3) in Theorem 7.5.1. Note that if $\alpha \geq 2$, then Y is in the domain of attraction of the normal distribution and the same choice of centering is appropriate. Thus, the tail index of the limiting distribution satisfies $1 < \alpha^* = \min(\alpha, 2)$.

Finally, computing condition (ii) in Theorem 7.5.1 from the tail behavior of Y yields $\beta = 1$. Essentially, this follows because $F_Y(-x) = 0$ if $x \geq -q_\epsilon$.

Therefore, all conditions in Theorem 7.5.1 are satisfied and, as a result, there exists a sequence of normalizing constants a_n satisfying

(7.5.2), such that

$$a_n^{-1}\left(\sum_{i=1}^{n} Y_i - nEY_1\right) \xrightarrow{w} S_{\alpha^*}(1,1,0). \tag{7.9.8}$$

as $n \to \infty$. In order to apply this result to sample AVaR, we need (7.9.8) reformulated for the average rather than the sum of Y_i. Thus, a more suitable form is

$$n\epsilon a_n^{-1}\left(\frac{1}{n\epsilon}\sum_{i=1}^{n}(Y_i - EY_i)\right) \xrightarrow{w} S_{\alpha^*}(1,1,0). \tag{7.9.9}$$

as $n \to \infty$.

As a final step, we apply the limit result in (7.9.9) to equation (7.9.5). Multiplying both sides of (7.9.5) by $n\epsilon a_n^{-1}$ yields the limit

$$n\epsilon a_n^{-1}(\phi(F_n) - \phi(F)) \xrightarrow{w} S_{\alpha^*}(1,1,0) \tag{7.9.10}$$

as $n \to \infty$. It remains only to verify if the normalization does not lead to explosion of the residual. Indeed,

$$n\epsilon a_n^{-1} o(n^{-1/2}) = \frac{n^{1/2}}{a_n} o(1) = o(1)$$

because the factor $n^{1/2}/a_n$ approaches zero by the asymptotic behavior of a_n given in the domains of attraction characterization in Theorem 7.5.1. □

Notes

1. Analytic tractability very often hinges on simplified assumptions about the nature of the objects being studied. It is a very desirable modeling feature since it allows conceiving model properties in full detail. Requiring more realistic assumptions, however, usually detracts from mathematical elegance and we are forced to rely on numerical techniques instead.
2. For further details about the bias of sample AVaR, see Trindade et al. (2007).
3. For a proof, see Stoyanov and Rachev (2008*a*).

4. See Rachev et al. (2005) and the references therein.
5. For more details about how these expressions are derived, see Stoyanov and Rachev (2008*a*).
6. If we generate a sample of 2,000 scenarios from the standard normal distribution, a relative deviation below 6% between the estimated quantile $q_{2.5\%}$ and the corresponding standard normal quantile happens with about 95% probability, and below 7.7% with about 99% probability.
7. See Chapter 6 for more details on stable distributions.
8. For more information, see Rachev and Mittnik (2000).
9. For further information about stable distributions and their properties, see Samorodnitsky and Taqqu (1994).
10. See Rachev and Mittnik (2000).
11. See Samorodnitsky and Taqqu (1994).
12. See Samorodnitsky and Taqqu (1994).
13. For a proof, see Stoyanov and Rachev (2008*b*).
14. See Rachev et al. (2007) for more details on time series models.
15. See, for example, Rachev and Mittnik (2000).
16. Financial Services Authority is an independent body that regulates the financial services industry in the UK.
17. See Financial Services Authority (2009), p. 42.
18. We assume that the returns are logarithmic. This is the usual assumption when modeling stock returns.
19. See Rachev and Mittnik (2000).
20. The discussion is based on Klebanov et al. (2009) and Grabchak and Samorodnitsky (2009).
21. For more information about kernel methods, see section 6.4.5 of Chapter 6 and the references therein.
22. For additional information, see Klebanov et al. (2009) and Rachev (1991).
23. See Kim, Rachev, Bianchi and Fabozzi (2008) and Kim, Rachev, Chung and Bianchi (2008).

References

Financial Services Authority (2009), 'A regulatory response to the global banking crisis', *Discussion Paper* 2, March.

Grabchak, M. and G. Samorodnitsky (2009), 'Do financial returns have finite or infinite variance? A paradox and an explanation', working paper.

Kim, Y., S. T. Rachev, D. Chung and M. L. Bianchi (2008), 'A modified tempered stable distribution with volatility clustering', in J. O. Soares, J. P. Pina and M. C. Lopes (eds), *New Developments in Financial Modelling*, Cambridge Scholars Publishing, pp. 344–365.

Klebanov, L. B., S. T. Rachev and F. J. Fabozzi (2009), *Ill-Posed Problems in Probability and Stability of Random Sums*, Nova, New York.

Rachev, S., F. Fabozzi and C. Menn (2005), *Fat-tails and Skewed Asset Return Distributions*, Wiley, New York.

Rachev, S., S. Mittnik, F. Fabozzi, S. Foccardi and Teo Jasic (2007), *Financial Econometrics: From Basics to Advanced Modeling Techniques*, Wiley, New York.

Rachev, S. T. (1991), *Probability Metrics and the Stability of Stochastic Models*, Wiley, New York.

Rachev, S. T. and S. Mittnik (2000), *Stable Paretian Models in Finance*, Series in Financial Economics, John Wiley & Sons, New York.

Samorodnitsky, G. and M. S. Taqqu (1994), *Stable Non-Gaussian Random Processes*, Chapman & Hall, New York, London.

Stoyanov, S. and S. Rachev (2008a), 'Asymptotic distribution of the sample average value-at-risk', *Journal of Computational Analysis and Applications* **10**, 465–483.

Stoyanov, S. and S. Rachev (2008b), 'Asymptotic distribution of the sample average value-at-risk in the case of heavy-tailed returns', *Journal of Applied Functional Analysis* **3**, 443–461.

Stoyanov, S. and B. Racheva-Iotova (2004), 'Univariate stable laws in the field of finance – approximations of density and distribution functions', *Journal of Concrete and Applicable Mathematics* **2 (1)**, 38–57.

Stoyanov, S., G. Samorodnitsky, S. Rachev and S. Ortobelli (2006), 'Computing the portfolio conditional value-at-risk in the α-stable case', *Probability and Mathematical Statistics* **26**, 1–22.

Trindade, A. A., S. Uryasev, A. Shapiro and G. Zrazhevsky (2007), 'Financial prediction with constrained tail risk', forthcoming in *Journal of Banking and Finance*.

Zolotarev, V. M. and V. V. Uchaikin (1999), *Chance and Stability, Stable Distributions and their Applications*, Brill Academic Publishers, Boston.

Chapter 8

Stochastic Dominance Revisited

The goals of this chapter are the following:

- To explore the relationship between preference relations and quasi-semidistances.
- To introduce a universal description of probability quasi-semidistances in terms of a Hausdorff structure.
- To provide examples with first-, second-, and higher-order stochastic dominance and to introduce primary, simple, and compound stochastic orders.
- To explore new stochastic dominance rules based on a popular risk measure.
- To provide a utility-type representation of probability quasi-semidistances and to describe the degree of violation utilized in almost stochastic orders in terms of quasi-semidistances.

Notation introduced in this chapter:

Notation	*Description*
$\preceq$	A binary relation or a preference relation
$\preceq_\tau$	The specialization pre-order of a given topology τ

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

Notation	*Description*
$\preceq_d$	A preference relation induced by a quasi-semidistance d
$d^{-1}(x, y)$	The dual of a quasi-semidistance $d(x, y)$
$\preceq_{d^{-1}}$	A preference relation induced by the dual of a quasi-semidistance d
$r(A, B)$	The Hausdorff semimetric between two sets A and B
$h_{\lambda,\phi,\mathfrak{B}}$	The Hausdorff structure of a probability quasi-semidistance
$L^*_\lambda(X, Y)$	The Lévy quasi-semidistance between X and Y
$\mathbb{L}^*_{\lambda,n}(X, Y)$	A quasi-semidistance metrizing n-th order stochastic dominance
$\zeta^*_n(X, Y)$	A quasi-semidistance based on the Zolotarev probability metric metrizing the Rothschild–Stiglitz stochastic order
$\pi^*_\lambda(X, Y)$	A quasi-semidistance based on the Hausdorff probability metric
$\mathbb{AV}_{\lambda,\phi,\mathfrak{B}}$	A probability quasi-semidistance metrizing a stochastic order based on the risk measure average value-at-risk
$\zeta^*_{\mathcal{U}}(X, Y)$	A quasi-semidistance based on the utility functions in the class $\mathcal{U}$
$\widetilde{r}_\lambda(f, g)$	The Hausdorff semimetric between two functions f and g
$\boldsymbol{\kappa}^*(X, Y)$	The Kantorovich quasi-semidistance
$v_\mu(X, Y)$	A ratio measuring the degree of violation of the stochastic order $\preceq_\mu$ generated by the quasi-semidistance μ

Important terms introduced in this chapter:

Term	*Concise explanation*
quasi-semidistance	A functional satisfying the identity and the triangle inequality axioms
preference relation	A relation which is reflexive and transitive; a pre-order

Term	*Concise explanation*
metrizable stochastic order	A stochastic order which can be described by means of a quasi-semidistance
dual stochastic order	The stochastic order induced by the dual of a given quasi-semidistance
topology	A family of sets satisfying certain properties that are used to define a topological space
specialization pre-order	A pre-order generated by a topology describing the natural order of the points of the corresponding topological space
almost stochastic dominance	A stochastic dominance corresponding to all preferences in a sub-category of given category (e.g. almost first-order stochastic dominance)

8.1 Introduction

From a historical perspective, stochastic dominance (SD) rules were first introduced in relation to the normative expected utility theory describing choice under uncertainty. The notions of first-order stochastic dominance (FSD) and second-order stochastic dominance (SSD) were used to prescribe the behavior of unsatiable investors and unsatiable, risk-averse investors, respectively. Since its introduction, the significance of SD analysis has increased enormously. In portfolio theory, for example, new families of risk measures have been introduced but consistency with FSD and SSD is always sought. In areas other than finance, SD finds application in diverse fields such as economics, insurance, agriculture, and medicine. We discussed SD relations in more detail in Chapter 3. For additional information about the applications in different areas, see Levy (2006).

In this chapter, we discuss a new concept describing SD relations which is based on the notion of a *quasi-semidistance*. We consider probability quasi-semidistances, which represent an extension of the notion of probability semidistances.[1] In the context of SD relation,

quasi-semidistances allow measuring by how much a given prospect X dominates another prospect Y or, in case they are incomparable, they allow measuring the degree of violation of the SD rule. The notion of degree of violation is closely connected with the notion of *almost stochastic dominance* developed to explain observed deviations from the rational behavior prescribed by expected utility theory. For additional details about almost stochastic dominance, see Levy (2006).

The new concept distinguishes between a few types of stochastic orders nested in each other, such that a stochastic order from a given category cannot imply a stochastic order from categories in which it is nested. This is, essentially, a corollary of the subdivision of probability distances into a primary, simple, and compound type which we discussed in Chapter 4. Here is an example concerning SD relations. The smallest category includes stochastic orders based on certain characteristics of the underlying prospects. For example, the mean-variance order belongs to this category as it is based on inequalities between the means and the variances of the corresponding prospects. The second smallest category includes stochastic orders based on inequalities between certain transformations of the cumulative distribution functions (c.d.f.s). Both FSD and SSD belong to this category. As a consequence, the mean-variance order can imply neither FSD nor SSD. The same holds for any mean-risk order, where risk is measured by an arbitrary risk measure.

Comparing the mean-variance, or more generally the mean-risk, approach and the SD approach, we can conclude that the former leads to optimization problems that are practical. Even though the SD approach is more general, it does not provide a method for construction of a portfolio from several individual securities (see, for example, Levy (2006)). We believe that the framework discussed in this chapter is a step towards resolving this shortcoming.

The goal of this chapter is to discuss two types of representations of probability quasi-semidistances – a Hausdorff representation and a utility-type representation. We provide examples for n-th order stochastic dominance, stochastic orders based on average value-at-risk (AVaR), and stochastic orders arising from classes of investors.

The appendix to this chapter includes a more technical discussion about the utility-type representation and also a more general discussion about the links between preference relations and topology. Finally, we briefly describe the structural classification of probability distances which is based on Rachev (1991).

8.2 Metrization of Preference Relations

In this section, we introduce notation and discuss an approach towards describing a preference relation through a quasi-semidistance.

Let S denote the space of all combinations of goods, services, and assets, which we also call *baskets*. A preference relation on S, denoted by $\preceq$, is introduced by a binary relation such that $x \preceq y$, if y is at least as preferable as x. There are a few assumptions that are usually made:

1. The binary relation is assumed *reflexive*, i.e. $x \preceq x$, for all $x \in S$.
2. The binary relation is assumed *transitive*, i.e. if $x \preceq y$ and $y \preceq z$, then $x \preceq z$ for any $x, y, z \in S$.

If $x \preceq y$ and $y \preceq x$, then we say that x and y are indistinguishable or equivalent from the standpoint of the preference order.

We do not discuss the adequacy of the reflexivity and the transitivity assumptions. We assume that they characterize every preference relation and, as a consequence, the preference relation $\preceq$ represents a pre-order defined on S. For a detailed discussion of these axioms, see Anand (1995).

The most direct way to describe a preference relation defined in this way is through the corresponding binary relation. However, this is not practical because we have to make a list of all pairs (x, y) such that $x \preceq y$. A generic and more practical approach to describe a preference relation is by means of a quasi-semidistance. In section 2.3 of Chapter 2, we provided a definition of metric and semidistance spaces which are closely related notions. In this section, we define the notion of a quasi-semidistance.

A *quasi-semidistance* is a function $d(x, y) : S \times S \to [0, \infty]$ satisfying the properties:

i. The *identity property*: if $x = y$, then $d(x, y) = 0$.

ii. The *triangle inequality*: $d(x, y) \leq \mathbb{K}(d(x, z) + d(z, y))$ for any $x, y, z \in S$ in which $K \geq 1$.

If $\mathbb{K} = 1$, then the quasi-semidistance turns into a quasi-semimetric. In fact, comparing the definitions of a quasi-semidistance and a semidistance, we can generalize that semidistances are symmetric quasi-semidistances.[2]

Every quasi-semidistance defines a pre-order and, therefore, a preference relation in the following way. The basket y is at least as preferable as another basket x, $x \preceq_d y$, if $d(x, y) = 0$. It is straightforward to verify that the transitivity property of $\preceq_d$ is a consequence of the triangle inequality for $d(x, y)$ and reflexivity follows from the identity property.

Since every quasi-semidistance defines a preference relation, we can ask the converse question, which is explored in detail in section 8.8.1. In this section, we only mention that the answer is in the negative. Therefore, the set of all preference relations can be divided into two parts: (1) those that arise from quasi-semidistances and (2) those that do not arise in this fashion. The preference relations that do arise from quasi-semidistances we call *quasi-metrizable* or simply *metrizable*.

Finally, note that from a given quasi-semidistance, we can always construct a semidistance using the dual. One approach to do that is to calculate the maximum between the quasi-semidistance and its dual,

$$\rho(x, y) = \max(d(x, y), d(y, x)). \tag{8.2.1}$$

It is straightforward to verify that in addition to the identity property and the triangle inequality, μ satisfies the symmetry property $\mu(x, y) = \mu(y, x)$. The representation in (8.2.1) implies that the quasi-semidistance $d(x, y)$ is consistent with any convergence in $\rho(x, y)$. That is, if $x_1, x_2, \ldots$ is a sequence converging to x in $\rho(x, y)$, $\lim_{n\to\infty} \rho(x_n, x) \to 0$, then necessarily $\lim_{n\to\infty} d(x_n, x) \to 0$.

We exploit this property in the classification of stochastic orders, which we link to the corresponding classification of the metric $\rho(x, y)$ which is obtained through the symmetrization transform in (8.2.1).

8.3 The Hausdorff Metric Structure

In section 8.2, we do not specify the nature of the points in the space S. Suppose that S is the space of one-dimensional random variables defined on a probability space $(\Omega, \mathfrak{A}, \Pr)$ taking values in $(\mathbb{R}, \mathfrak{B}_1)$, where $\mathfrak{B}_1$ is the σ-field of all Borel subsets of $\mathbb{R}$.[3] In this setting, a quasi-metrizable preference order on S turns into a stochastic order with a quasi-semidistance $d(X, Y)$ defined on the space of all joint distributions S^2 generated by the pairs of random variables (X, Y), which we denote with capital letters. We proceed with the definition of the universal Hausdorff representation of quasi-semidistances on S^2 which we also call *probability quasi-semidistances* as they metrize preference relations between random quantities. The terms and the notation are consistent with the discussion in section 2.4 of Chapter 2.

Consider the Hausdorff metric $r(A, B)$ defined on the space of all subsets of $\mathbb{R}$. Let $\mathfrak{B} \subseteq \mathfrak{B}_1$ and define a function $\phi : S^2 \times \mathfrak{B}^2 \to [0, \infty]$ satisfying the following relations:

I. If $P(X = Y) = 1$, then $\phi(X, Y; A, B) = 0$ for all $A = B \in \mathfrak{B}$.

II. There exists a constant $K_\phi \geq 1$ such that for all $A, B, C \in \mathfrak{B}$ and random variables X, Y, Z

$$\phi(X, Y; A, B) \leq K_\phi(\phi(X, Z; A, C) + \phi(Z, Y; C, B))$$

Let $d(X, Y)$ be a probability quasi-semidistance. The representation of $d(X, Y)$ in the following form:

$$\begin{aligned} d(X, Y) &= h_{\lambda,\phi,\mathfrak{B}}(X, Y) \\ &:= \sup_{A\in\mathfrak{B}} \inf_{B\in\mathfrak{B}} \max \left\{ \frac{1}{\lambda} r(A, B), \phi(X, Y; A, B) \right\} \end{aligned} \tag{8.3.1}$$

is called the Hausdorff structure of $d(X, Y)$. In this representation, $r(A, B)$ is the Hausdorff metric in the set $\mathfrak{B}$, λ is a positive number, and the function ϕ satisfies relations I and II above.

It can be demonstrated that the function $h_{\lambda,\phi,\mathfrak{B}}(X, Y)$ defined above is indeed a quasi-semidistance.

Theorem 8.3.1. The function $h_{\lambda,\phi,\mathfrak{B}}(X, Y)$ defined in equation (8.3.1) is a quasi-semidistance.

Proof. The identity property and the triangle inequlity are essentially metric properties. The proof follows from the arguments in Rachev (1991) proving that the Hausdorff representation is a probability metric. □

Applying the symmetrization in equation (8.2.1) to the representation in (8.3.1), we obtain the Hausdorff representation of probability metrics. For more information, see Chapter 4 in Rachev (1991).

The following example illustrates the significance of the Hausdorff representation. It turns out that every probability quasi-semidistance is representable in the form in (8.3.1). Consider an arbitrary probability quasi-semidistance $\mu(X, Y)$. It has the trivial form $h_{\lambda,\phi,\mathfrak{B}}(X, Y) = \mu(X, Y)$ where the set $\mathfrak{B}$ is a singleton: for example, $\mathfrak{B} = \{A_0\}$, and $\phi(X, Y; A_0, A_0) = \mu(X, Y)$.

In the limit cases $\lambda \to 0$ and $\lambda \to \infty$, the Hausdorff structure turns into a structure of a uniform type. The following limit relations hold.

Theorem 8.3.2. Let $d(X, Y)$ have the representation in (8.3.1). Then, as $\lambda \to 0$, $d(X, Y)$ has a limit equal to

$$h_{0,\phi,\mathfrak{B}}(X, Y) = \sup_{A \in \mathfrak{B}} \phi(X, Y; A, A) \tag{8.3.2}$$

As $\lambda \to \infty$, the limit $\lim_{\lambda\to\infty} \lambda h_{\lambda,\phi,\mathfrak{B}}(X, Y) = h_{\infty,\phi,\mathfrak{B}}(X, Y)$ exists and equals

$$h_{\infty,\phi,\mathfrak{B}}(X, Y) = \sup_{A \in \mathfrak{B}} \inf_{B \in \mathfrak{B}, \phi(X,Y;A,B)=0} r(A, B) \tag{8.3.3}$$

Proof. For a proof, see Stoyanov et al. (2009*b*). □

The main building block of the Hausdorff representation, the function $\phi(X, Y; A, B)$, can be interpreted in the following way. It calculates the performance of X relative to Y over two events A and B. If $\phi(X, Y; A, B) = 0$ for some A and B, then, according to the preference order definition, Y performs at least as X with respect to the two events. As we demonstrate in the next section, in some cases there is a straightforward interpretation in the sense that ϕ calculates the deviation of the probability of X belonging to A relative to the probability of Y belonging to B. In other cases, the relationship of X to A and Y to B is not so direct.

Besides the function ϕ, the definition in (8.3.1) includes also the Hausdorff metric $r(A, B)$ in order to take into account the degree of dissimilarity between the events A and B. If we want to calculate the degree of deviation between X and Y on one and the same events, i.e. $A = B$, then we can use the limit case given in (8.3.2). In this case, if Y outperforms X with respect to all events $A = B$, i.e. $\phi(X, Y; A, A) = 0$ for all A, then Y is at least as preferable as X.

The Hausdorff representation of a quasi-semidistance in (8.3.1) can be translated into a different form which is more open to interpretation.

Theorem 8.3.3. Suppose that a probability quasi-semidistance admits the Hausdorff representation $h_{\lambda,\phi,\mathfrak{B}}$ given in (8.3.1). Then, the probability quasi-semidistance also enjoys the following representation:

$$h_{\lambda,\phi,\mathfrak{B}}(X, Y) = \inf\{\epsilon > 0 : v(X, Y; \lambda\epsilon) < \epsilon\} \tag{8.3.4}$$

where

$$v(X, Y; t) = \sup_{A\in\mathfrak{B}} \inf_{B\in A(t)} \phi(X, Y; A, B) \tag{8.3.5}$$

in which $A(t)$ is the collection of all elements B of $\mathfrak{B}$ such that the Hausdorff metric $r(A, B)$ is not greater than t.

Proof. The proof is constructed in the same way as the proof of Theorem 4.2.1 in Rachev (1991). □

We can interpret equation (8.3.5) in the following way. Fix an event A and a tolerance level $t > 0$. Using the Hausdorff metric, take all events that do not deviate from A more than as implied by the tolerance level: that is, build the set $A(t) = \{B \in \mathfrak{B} : r(A, B) < t\}$. With A fixed, compute the minimum performance deviation between X and Y running through all events for Y, which are within the tolerance level. As a next step, compute the maximum of those minimal deviations by varying A.

By varying the tolerance level t, we control the size of the admissible sets relative to A. The larger t is, the more the admissible events may deviate from the event A and, therefore, the larger potential there is for deviation in the performance of Y relative to X. At the other extreme, when $t = 0$, the deviation in performance is estimated over one and the same event.

Finally, in equation (8.3.4) we calculate the smallest tolerance level such that the largest of those minimal performance deviations is smaller than it. Note that, depending on the nature of the random variables X and Y and the choice of ϕ, this smallest tolerance level may actually be infinite: that is, the quasi-semidistance may be unbounded.

The parameter λ in both (8.3.1) and (8.3.4) allows calculating limit quasi-semidistances arising naturally from the general case. This is demonstrated in Theorem 8.3.2. If we view $h_{\lambda,\phi,\mathfrak{B}}(X, Y)$ defined in (8.3.1) as a function of the parameter λ, it appears that it is a monotonic, non-increasing function.

Theorem 8.3.4. The quasi-semidistance $h_{\lambda,\phi,\mathfrak{B}}(X, Y)$ defined in (8.3.1) is a non-increasing function of $\lambda > 0$.

Proof. For any fixed $A \in \mathfrak{B}$ and $0 < \lambda_1 < \lambda_2$,

$$\max\left\{\frac{1}{\lambda_2}r(A, B), \phi(X, Y; A, B)\right\} \leq \max\left\{\frac{1}{\lambda_1}r(A, B), \phi(X, Y; A, B)\right\}$$

for all $B \in \mathfrak{B}$. Therefore, the same inequality is preserved after computing sequentially the infimum with respect to B and

then the supremum with respect to A. In effect, $h_{\lambda_2,\phi,\mathfrak{B}}(X, Y) < h_{\lambda_1,\phi,\mathfrak{B}}(X, Y)$. □

This result implies that the limit computed in (8.3.2) is an upper bound of $h_{\lambda,\phi,\mathfrak{B}}(X, Y)$, i.e.

$$h_{\lambda,\phi,\mathfrak{B}}(X, Y) \leq h_{0,\phi,\mathfrak{B}}(X, Y).$$

Thus, if $h_{0,\phi,\mathfrak{B}}(X, Y)$ is finite, it means that $h_{\lambda,\phi,\mathfrak{B}}(X, Y)$ is finite as well.

8.4 Examples

In this section, we provide examples of quasi-semidistances with a Hausdorff structure. We also provide examples of quasi-semidistances metrizing different stochastic dominance orders.[4]

The structure of the quasi-semidistance determines whether the induced stochastic order is based essentially on inequalities between certain characteristics such as mean, volatility, etc., inequalities based on c.d.f.s, or inequalities directly between functions of the corresponding random variables. In line with the theory of probability metrics, the first order type we call *primary*; the second, *simple*; and the third, *compound*.[5] The formal definition is as follows.

Definition 8.4.1. A metrizable stochastic order $\preceq_d$ is called primary, simple, or compound if the probability semidistance arising from a symmetrization transform, such as the one given in (8.2.1), is primary, simple, or compound, respectively.

From the point of view of finance, the stochastic order behind the mean-variance framework is primary. In contrast, FSD and SSD are simple orders, as we demonstrate below. A theoretical advantage of this categorization is the inclusion

$$\text{primary orders} \subset \text{simple orders} \subset \text{compound orders}$$

which implies that a primary order cannot induce a simple order, which, in turn, cannot induce a compound order. This is a consequence of the corresponding relations between primary, simple, and compound probability metrics.

8.4.1 The Lévy quasi-semidistance and first-order stochastic dominance

Consider the choice $\phi(X, Y; (-\infty, x], (-\infty, y]) = (F_X(x) - F_Y(y))_+$, where $(x)_+ = \max(x, 0)$. In this case, the sets A and B are of the form $(-\infty, a]$, $a \in \mathbb{R}$. The representation in (8.3.1) becomes

$$L^*_\lambda(X, Y) = \sup_{x\in\mathbb{R}} \inf_{y\in\mathbb{R}} \max \left\{ \frac{1}{\lambda}|x - y|, (F_X(x) - F_Y(y))_+ \right\} \tag{8.4.1}$$

The quasi-semidistance defined above also equals

$$L^*_\lambda(X, Y) = \inf\{\epsilon > 0 : (F_X(x) - F_Y(x + \lambda\epsilon))_+ < \epsilon, \forall x \in \mathbb{R}\}$$

which can be demonstrated by applying the result in Theorem 8.3.3. Applying the symmetrization transform in (8.2.1) leads to the parametric version of the celebrated Lévy metric

$$\begin{aligned} L_\lambda(X, Y) = \inf\{\epsilon > 0 : F_X(x - \lambda\epsilon) - \epsilon \le F_Y(x) \le F_X(x + \lambda\epsilon) \\ + \epsilon, \forall x \in \mathbb{R}\} \end{aligned}$$

and for this reason we call $L^*_\lambda(X, Y)$ the *Lévy quasi-semidistance*. The two limit cases in Theorem 8.3.2 can be calculated explicitly and they equal

$$L^*_0(X, Y) = \sup_{x\in\mathbb{R}} (F_X(x) - F_Y(x))_+$$

$$L^*_\infty(X, Y) = \sup_{t\in[0,1]} (F^{-1}_Y(t) - F^{-1}_X(t))_+$$

where $F^{-1}_X(t) = \sup\{x : F_X(x) < t\}$ is the inverse c.d.f. of X.

Figure 8.1 illustrates the Lévy quasi-semidistance with $\lambda = 1$ and its dual.[6] Like the Lévy metric discussed in Chapter 2, $L^*_1(X, Y)$ can be interpreted in terms of a distance between the graphs of F_X and

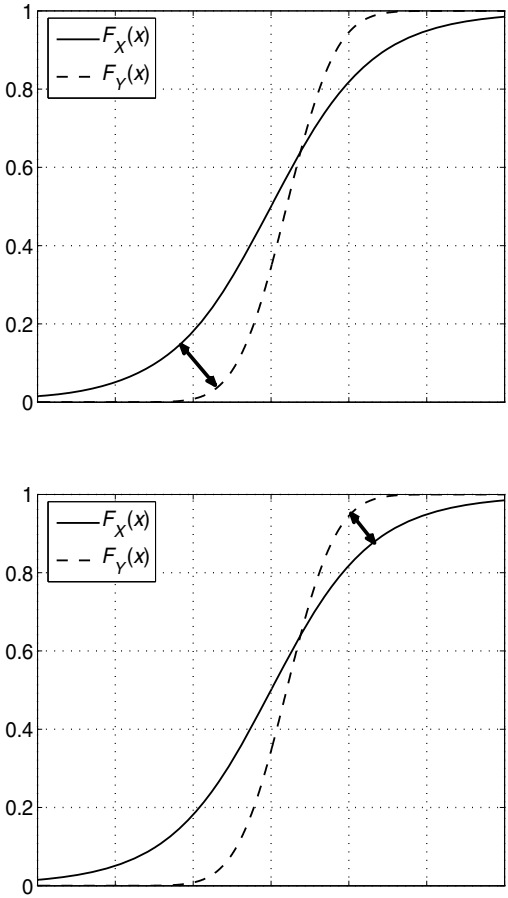

Figure 8.1: Illustration of the Lévy quasi-semidistance with $\lambda = 1$ and its dual. $L_1^*(X, Y)\sqrt{2}$ is the maximum distance between F_X and F_Y computed where $F_X \geq F_Y$ along a 45 degree direction (top plot). In a similar way, $L_1^*(Y, X)\sqrt{2}$ is the maximum distance between F_X and F_Y computed where $F_Y \geq F_X$ along a 45 degree direction (bottom plot).

F_Y computed along a 45 degree direction. In contrast to the classical Lévy metric, however, in the case of $L_1^*(X, Y)$ we compute the maximal distance where $F_X \geq F_Y$. The arrow on the top plot of Figure 8.1 shows where the maximum is attained.

In a similar vein, the dual $L_1^*(Y, X)$ can be interpreted in terms of the maximal distance between the graphs of F_X and F_Y computed along a 45 degree direction but where the converse inequality holds, $F_X \leq F_Y$. This is illustrated on the bottom plot of Figure 8.1.

The Lévy quasi-semidistance is an important example because it can be used to metrize FSD. In fact, the definition in (8.4.1) induces the dual order[7] and is, therefore, monotonic with respect to FSD in the sense of Proposition 8.8.2 which is provided in the appendix to this chapter. Recall that FSD can be introduced by an inequality between the corresponding c.d.f.s,[8]

$$X \preceq_{FSD} Y \quad \Longleftrightarrow \quad F_Y(x) \leq F_X(x), \forall x \in \mathbb{R}. \tag{8.4.2}$$

The dual order, $\preceq_{FSD^{-1}}$, can be expressed in a similar way:

$$X \preceq_{FSD^{-1}} Y \quad \Longleftrightarrow \quad F_X(x) \leq F_Y(x), \forall x \in \mathbb{R}. \tag{8.4.3}$$

Theorem 8.4.1. The functional $L_\lambda^*(X, Y)$ defined in (8.4.1) metrizes $\preceq_{FSD^{-1}}$.

Proof. For a proof, see Stoyanov et al. (2009*b*). □

As a corollary to this result, it follows that the dual quasi-semidistance $L_\lambda^*(Y, X)$ metrizes FSD. Also, note that the reasoning does not depend on the choice of λ and, therefore, the result is valid for any $\lambda > 0$. In effect, there are many quasi-semidistances inducing FSD. In section 8.6, we provide yet another example which does not belong to the class $L_\lambda^*(Y, X)$.

Since symmetrization of L_λ^* leads to the Lévy metric, which is a simple probability metric, it follows that FSD is a simple order.

8.4.2 Higher-order stochastic dominance

The reasoning in section 8.4.1 can be applied to the more general case of higher-order stochastic dominance. Stochastic dominance of order n, $\preceq_n$, can be introduced by means of an inequality involving

the corresponding c.d.f.s,[9]

$$X \succeq_n Y \quad \Longleftrightarrow \quad F_X^{(n)}(x) \leq F_Y^{(n)}(x), \forall x \in \mathbb{R}. \tag{8.4.4}$$

where $F_X^{(n)}(x)$ stands for the n-th integral of the c.d.f. of X which can be defined recursively as

$$F_X^{(n)}(x) = \int_{-\infty}^{x} F_X^{(n-1)}(t)dt. \tag{8.4.5}$$

with the initial condition $F_X^{(1)}(x) = F_X(x)$.

Repeating the arguments in Theorem 8.4.3, it can be demonstrated that the quasi-semidistance

$$\mathbb{L}_{\lambda,n}^*(X, Y) = \sup_{x\in\mathbb{R}} \inf_{y\in\mathbb{R}} \max \left\{ \frac{1}{\lambda}|x - y|, \left(F_X^{(n)}(x) - F_Y^{(n)}(y)\right)_+ \right\} \tag{8.4.6}$$

metrizes the dual of the n-th order stochastic dominance and, as a result, $\mathbb{L}_{\lambda,n}^*(Y, X)$ metrizes $\preceq_n$. In this more general case, however, it is not clear a priori if $\mathbb{L}_{\lambda,n}^*(X, Y) < \infty$. For the Lévy quasi-semidistance this question is redundant because the limit $L_0^*(X, Y)$ is always finite and Theorem 8.3.4 guarantees boundedness of the Lévy quasi-semidistance. The limit of (8.4.6) as $\lambda \to 0$ equals

$$\mathbb{L}_{0,n}^*(X, Y) = \sup_{x\in\mathbb{R}} \left(F_X^{(n)}(x) - F_Y^{(n)}(x)\right)_+$$

and by Theorem 8.3.4 we can conclude that $\mathbb{L}_{\lambda,n}^*(X, Y) < \infty$ if $\mathbb{L}_{0,n}^*(X, Y) < \infty$.

We develop a set of sufficient conditions involving another probability quasi-semidistance.

Theorem 8.4.2. The following inequality holds true, provided that $EX^k = EY^k,\ k = 1, 2, \ldots, n-2$, $E|X|^{n-1} < \infty$ and $E|Y|^{n-1} < \infty$,

$$\mathbb{L}_{0,n}^*(X, Y) \leq \int_{\mathbb{R}} \left(\int_{-\infty}^{x} \frac{(x-t)^{n-2}}{(n-2)!} d(F_X(t) - F_Y(t)) \right)_+ dx < \infty \tag{8.4.7}$$

Proof. For a proof, see Stoyanov et al. (2009*b*). □

The inequality in (8.4.7) is an inequality between two quasi-semidistances. The upper bound is the *Zolotarev quasi-semidistance*

$$\zeta_n^*(X, Y) = \int_{\mathbb{R}} \left(\int_{-\infty}^{x} \frac{(x-t)^{n-1}}{(n-1)!} d(F_X(t) - F_Y(t)) \right)_+ dx \tag{8.4.8}$$

which can also be represented as

$$\zeta_n^*(X, Y) = \frac{1}{(n-1)!} \int_{\mathbb{R}} \left(E(x-X)_+^{n-1} - E(x-Y)_+^{n-1} \right)_+ dx$$

The Zolotarev quasi-semidistance itself can be used to metrize the n-order stochastic dominance under the assumed moment conditions. However, $\mathbb{L}_{\lambda,n}^*(X, Y)$ is strictly weaker as there is no lower bound of it that can be expressed in terms of $\zeta_{n-1}^*(X, Y)$. Therefore, this can be viewed as an illustration of how one and the same stochastic order can arise from two different quasi-semidistances.

The conclusion that FSD is a simple order can be extended to the n-th order stochastic dominance by noticing that symmetrizing ζ_{n-1}^* leads to a simple probability semidistance.

The approach discussed in this section can be applied without modification to the fractional and the inverse orders discussed in Ortobelli et al. (2009). From a theoretical viewpoint, they belong to the class of simple stochastic orders as the probability semidistances arising from applying the symmetrization transform are simple.

The inequality (8.4.7) has an interesting interpretation in terms of the relationship between SSD and the Rothschild–Stiglitz stochastic dominance order (RSD) which is provided in section 3.3.3 of Chapter 3. The interpretation can also be viewed as an application of Proposition 8.8.3. Recall that the Rothschild–Stiglitz stochastic order, $\preceq_{RSD}$, is introduced in the following way:

$$X \succeq_{RSD} Y \iff \begin{cases} EX = EY, \\ \int_{-\infty}^{x} F_X(t)dt \leq \int_{-\infty}^{x} F_Y(t)dt, \ \forall\, x \in \mathbb{R} \end{cases}$$

and it implies SSD: that is, if $X \succeq_{RSD} Y$, then $X \succeq_{SSD} Y$. In section 3.7.2 of Chapter 3, we discussed the relationship $X \succeq_{SSD} Y \Rightarrow X \succeq_3 Y$ and, therefore, if $X \succeq_{RSD} Y$, then $X \succeq_3 Y$.

We can demonstrate that a quasi-semidistance metrizing RSD is $\zeta_2^*(Y, X)$ defined in equation (8.4.8). Conditions that guarantee $\zeta_2^*(X, Y) < \infty$ are $EX = EY$ and existence of second moments, $EX^2 < \infty$ and $EY^2 < \infty$. As a consequence, $\zeta_2^*(Y, X) = 0$ implies $X \preceq_{RSD} Y$ because both conditions defining RSD are satisfied. The converse follows from the definition of ζ_2^* in (8.4.8). As a result, ζ_2^* metrizes RSD.

The inequality $\mathbb{L}_{0,3}^*(Y, X) \leq \zeta_2^*(Y, X)$ proved in Theorem 8.4.7 indicates that $\mathbb{L}_{0,3}^*(Y, X) = 0$ if $X \preceq_{RSD} Y$. Therefore, RSD implies the order metrized by $\mathbb{L}_{0,3}$ which is in fact third-order stochastic dominance: that is, $X \succeq_{RSD} Y \Rightarrow X \succeq_3 Y$. Even though the relationship between RSD and third-order stochastic dominance can be directly illustrated through the inequality in Theorem 8.4.2, there does not seem to be an inequality illustrating the relationship between RSD and SSD. Nevertheless, the relationship between them can be derived taking advantage of different arguments.

Finally, the stochastic order induced by the quasi-semidistance ζ_n^* can be viewed as a generalization of RSD. The reason is that X and Y need to satisfy the following conditions in order to guarantee $\zeta_n^*(X, Y) < \infty$:

$$EX^k = EX^k,\ k = 1, \ldots, n-1 \quad \text{and} \quad E|X|^n < \infty,\ E|Y|^n < \infty.$$

8.4.3 The H-quasi-semidistance

In the formal construction of the Hausdorff representation given in (8.3.1), we do not impose symmetry with respect to X and Y because our goal is to describe quasi-semidistances. Nevertheless, in some cases the resulting functional can be symmetric because, for example, the assumed set $\mathfrak{B}$ is very general. This should not be regarded as a weird consequence of the generality of the construct. Rather, it is an expected outcome because symmetric probability quasi-semidistances are probability semidistances.

Consider the choice $\phi(X, Y; A, B) = (P(X \in A) - P(Y \in B))_+$, where $A, B \in \mathfrak{B}$ are arbitrary events and $\mathfrak{B}$ is the Borel σ-field. The Hausdorff representation equals

$$\pi_\lambda^*(X, Y) = \sup_{A \in \mathfrak{B}} \inf_{B \in \mathfrak{B}} \max \left\{ \frac{1}{\lambda} r(A, B), (P(X \in A) - P(Y \in B))_+ \right\}.$$

The function ϕ can be also expressed in terms of the complements of the events A and B, $\phi(X, Y; A, B) = (P(Y \in B^c) - P(X \in A^c))_+$, in which A^c denotes the complement of A and $P(Y \in B^c) = 1 - P(Y \in B)$. Since the complement of a given event also belongs to $\mathfrak{B}$, the functional above equals also

$$\pi_\lambda^*(X, Y) = \sup_{A \in \mathfrak{B}} \inf_{B \in \mathfrak{B}} \max \left\{ \frac{1}{\lambda} r(A, B), \left| P(X \in A) - P(Y \in B) \right| \right\}. \tag{8.4.9}$$

Applying the symmetrization transform in (8.2.1) to $\pi_\lambda^*(X, Y)$, we obtain the $\pi H_\lambda(X, Y)$ probability semidistance which is between the Prokhorov metric and the total variation metric from a topological viewpoint: that is, the topology generated by πH_λ is finer than the topology of the Prokhorov metric and the coarser than the topology of the total variation metric. This metric is also known as the *Hausdorff probability metric*, or the *H-metric*. For additional details, see section 4.1 in Rachev (1991).

The two limit cases from Theorem 8.3.2 can be derived using the arguments in Lemma 4.1.5 in Rachev (1991). They equal

$$\pi_0^*(X, Y) = \sup_{A \in \mathfrak{B}} |P(X \in A) - P(Y \in A)|$$

and

$$\pi_\infty^*(X, Y) = \inf \left\{ \epsilon > 0 : \inf_{B \in A(\epsilon)} |P(X \in A) - P(Y \in B)| = 0, \ \forall A \in \mathfrak{B} \right\}$$

in which $A(\epsilon) = \{C \in \mathfrak{B} : r(A, C) < \epsilon\}$. Notice that $\pi_0^*(X, Y)$ is in fact the total variation metric which, being a metric, satisfies the symmetry property $\pi_0^*(X, Y) = \pi_0^*(Y, X)$. It is curious that $\pi_0^*(X, Y)$ is

symmetric even though it arises as a limit from $\pi^*_\lambda(X, Y)$, which is a non-symmetric quasi-semidistance.

Repeating the arguments in the proof of Theorem 8.4.1, we can conclude that the stochastic order generated by (8.4.9) is actually an equivalence relation. Indeed, $\pi^*_\lambda(X, Y) = 0$ if and only if $|P(X \in A) - P(Y \in A)| = 0$ for all $A \in \mathfrak{B}$. Since $\mathfrak{B}$ denotes the entire Borel σ-field, this implies that the probability laws associated with X and Y are identical.

8.4.4 AVaR generated stochastic orders

In this section, we provide an example of a probability quasi-metric generated from a coherent risk measure that admits the representation given in (8.3.2). The coherent risk measure is the *average value-at-risk* (AVaR), also known as *conditional value-at-risk*, which is defined as

$$AVaR_\epsilon(X) = -\frac{1}{\epsilon}\int_0^\epsilon F_X^{-1}(t)dt \tag{8.4.10}$$

where $0 < \epsilon < 1$ is called *tail probability* and X is a random variable describing the return distribution of an investment. AVaR is interpreted as the average loss, provided that the loss is larger than the ϵ-quantile. A detailed discussion of this risk measure was provided in Chapters 6 and 7.

Another representation of (8.4.10), which is essentially a consequence of the general representation of coherent risk measures given in Artzner et al. (1998), equals

$$AVaR_\epsilon(X) = \sup_{A\in\mathfrak{A}_\epsilon} -\int_0^1 F_X^{-1}(t)d\nu_A = -\inf_{A\in\mathfrak{A}_\epsilon}\int_0^1 F_X^{-1}(t)d\nu_A \tag{8.4.11}$$

where $\mathfrak{A}_\epsilon = \{A \subset [0, 1] : \lambda(A) = \epsilon\}$, in which $\lambda(A)$ is the Lebesgue measure of A and ν_A is a uniform probability measure on the set A. The family of sets A_ϵ can be interpreted as the collection of all sets A such that $F_X^{-1}(A)$ is an ϵ-probability event, $P(X \in F_X^{-1}(A)) = \epsilon$. The interval $[0, \epsilon] \in \mathfrak{A}_\epsilon$ yields the AVaR at tail probability ϵ.

Consider the following choice for the building block ϕ of the Hausdorff representation in (8.3.1):

$$\phi(X, Y; A, B) = \left(\int_0^1 F_X^{-1}(t) d\nu_B - \int_0^1 F_Y^{-1}(t) d\nu_A \right)_+ \tag{8.4.12}$$

in which $A, B \in \mathfrak{B}$, where $\mathfrak{B} = [0, \epsilon] \cup \mathfrak{B}_1 \subseteq \mathfrak{A}_\epsilon$ because the interval $[0, \epsilon]$ needs to be in $\mathfrak{B}$. It is easy to verify that the axiomatic properties hold and this is a valid choice for ϕ in the Hausdorff representation. The resulting quasi-semidistance

$$\begin{aligned} \mathbb{AV}_{\lambda,\epsilon,\mathfrak{B}}(X, Y) = \sup_{A \in \mathfrak{B}} \inf_{B \in \mathfrak{B}} \max \Bigg\{ & \frac{1}{\lambda} r(A, B), \\ & \left(\int_0^1 F_X^{-1}(t) d\nu_B - \int_0^1 F_Y^{-1}(t) d\nu_A \right)_+ \Bigg\} \end{aligned} \tag{8.4.13}$$

is an AVaR generated quasi-semidistance. In the special case when $\mathfrak{B} = \{[0, \epsilon]\}$, then

$$\mathbb{AV}_{\lambda,\epsilon,\{[0,\epsilon]\}}(X, Y) = (AVaR_\epsilon(Y) - AVaR_\epsilon(X))_+.$$

The stochastic order $\preceq_{\mathbb{AV}_\mathfrak{B}}$ induced by the quasi-semidistance $\mathbb{AV}_{\lambda,\epsilon,\mathfrak{B}}$ can be interpreted in the following way. Suppose that X and Y are two random variables describing the returns of two stocks. If $X \preceq_{\mathbb{AV}_\mathfrak{B}} Y$, then the average loss of X in events occurring with probability ϵ is always not smaller than the corresponding average loss of Y. The events that we consider in this comparison depend on the choice of $\mathfrak{B}$ but the most extreme ones, $F_X^{-1}([0, \epsilon])$ and $F_Y^{-1}([0, \epsilon])$, are always included.

A couple of properties are collected in the next theorem.

Theorem 8.4.3. The following relations hold true.

(a) If $X \preceq_{\mathbb{AV}_\mathfrak{B}} Y$, then $AVaR_\epsilon(Y) \leq AVaR_\epsilon(X)$ for any admissible choice of $\mathfrak{B}$.

(b) The limit of $\mathbb{AV}_{\lambda,\epsilon,\mathfrak{B}}(X, Y)$ as $\lambda \to 0$ equals

$$\mathbb{AV}_{0,\epsilon,\mathfrak{B}}(X, Y) = \sup_{A\in\mathfrak{B}} \left(\int_0^1 F_X^{-1}(t) d\nu_A - \int_0^1 F_Y^{-1}(t) d\nu_A \right)_+ \tag{8.4.14}$$

(c) If $X = EY$ is a constant, then $\mathbb{AV}_{\lambda,\epsilon,\mathfrak{B}}(EY, Y) = AVaR_\epsilon(Y - EY)$ and, thus, equals the deviation measure behind the AVaR risk measure.

(d) Suppose that $\mathfrak{B}_1 \subseteq \mathfrak{B}_2$. Then, the stochastic order $\preceq_{\mathbb{AV}_{\mathfrak{B}_2}}$ implies the stochastic order $\preceq_{\mathbb{AV}_{\mathfrak{B}_1}}$.

(e) If $X \preceq_{FSD} Y$, then $X \preceq_{\mathbb{AV}_{\mathfrak{B}}} Y$ for any admissible choice of $\mathfrak{B}$. The converse is not true.

Proof. For a proof, see Stoyanov et al. (2009*b*). □

An expected corollary from the results above is that AVaR is consistent with FSD. Assuming that the random variables X and Y describe asset returns, it is the structure of the admissible family $\mathfrak{B}$ which determines whether only events including losses are considered in $\preceq_{\mathbb{AV}_{\mathfrak{B}}}$, i.e. negative returns, or both profits and losses, i.e. positive and negative returns.

8.4.5 Compound quasi-semidistances

The examples in the previous sections share a common feature. If X and Y are two random variables such that $F_X(x) = F_Y(x), \forall x \in \mathbb{R}$, then the corresponding quasi-semidistances equal zero. In this section, we consider compound quasi-semidistances in the form in (8.3.4) which are essentially characterized by the following feature: if $X = Y$ in almost sure sense, then they turn into zero.

Consider a function $d(x, y)$ defined on $\mathbb{R} \times \mathbb{R}$, which is a quasi-semidistance. Define the function v in the representation in (8.3.4) to be

$$v(X, Y; t) = P(d(X, Y) > t).$$

Then, the functional

$$\mu_\lambda(X, Y) = \inf\{\epsilon > 0 : P(d(X, Y) > \lambda\epsilon) < \epsilon\}$$

is a compound quasi-semidistance.

The stochastic order generated from $\mu_\lambda(X, Y)$ is of a compound type. Suppose that $d(X, Y) = (X - Y)_+$. Under this assumption, $\mu_\lambda(X, Y) = 0$ if and only if $P((X - Y)_+ > \epsilon) = 0, \forall\epsilon > 0$ which means that $X \leq Y$ in almost sure sense.

There are also other ways to construct compound quasi-semidistances which do not enjoy a non-trivial Hausdorff representation. For additional information, see Stoyanov et al. (2008).

8.5 Utility-type Representations

From the perspective of economic theories that describe choice under uncertainty, some stochastic orders arise from the preferences of a given class of economic agents. For example, according to classical expected utility theory, FSD arises from the class of *non-satiable* investors who have non-decreasing utility functions. Thus, if all non-satiable investors do not prefer Y to X, then $X \preceq_{FSD} Y$. Likewise, second-order stochastic dominance arises from the non-satiable, risk-averse investors who have non-decreasing, concave utility functions. In the same manner, n-th order stochastic dominance can be introduced through the preference relations of a class of investors the utility functions of whom are characterized by certain properties involving derivatives of higher order. For a more detailed discussion, see Chapter 3.

Consider the preference relation of an investor with a utility function $u(x), x \in \mathbb{R}$. The preference relation is characterized by the expected utility: that is, $X \preceq_u Y$ if and only if $Eu(X) \leq Eu(Y)$. As a result, one natural quasi-semidistance metrizing the preference relation is

$$\zeta_u^*(X, Y) = (Eu(X) - Eu(Y))_+.$$

Indeed, it can be directly verified that $X \preceq_{\zeta_u^*} Y \Leftrightarrow X \preceq_u Y$.

This approach can be generalized to a given class of investors $\mathcal{U}$. The arising stochastic order $\preceq_{\mathcal{U}}$ is introduced in the following way: $X \preceq_{\mathcal{U}} Y$ if and only if $X \preceq_u Y, \forall u \in \mathcal{U}$. In this case, one natural quasi-semimetric metrizing $\preceq_{\mathcal{U}}$ has the form

$$\zeta^*_{\mathcal{U}}(X, Y) = \sup_{u \in \mathcal{U}} \left(\int_{\mathbb{R}} u(x) d(F_X(x) - F_Y(x)) \right)_+ \tag{8.5.1}$$

which equals $\zeta^*_{\mathcal{U}}(X, Y) = \sup_{u \in \mathcal{U}} (Eu(X) - Eu(Y))_+$ if the corresponding expected utilities are finite. Thus, the condition $\zeta^*_{\mathcal{U}}(X, Y) = 0$ guarantees that $X \preceq_u Y, \forall u \in \mathcal{U}$, and therefore the stochastic order generated by the quasi-semidistance in (8.5.1) coincides with the stochastic order of the class $\mathcal{U}$. Since the representation in (8.5.1) is directly liked to the class $\mathcal{U}$, we call it a *utility-type representation*.

Some properties of (8.5.1) are collected in the following theorem.

Theorem 8.5.1. Suppose that the functional defined in (8.5.1) is finite. Under this assumption, it is a probability quasi-semidistance which metrizes the stochastic order $\preceq_{\mathcal{U}}$.

Proof. The identity property is obvious, if $F_X(x) = F_Y(x), \forall x \in \mathbb{R}$, then $\zeta^*_{\mathcal{U}}(X, Y) = 0$. The triangle inequality follows from the properties of the $(y)_+$ function.

From the definition in (8.5.1), it follows that if $\zeta^*_{\mathcal{U}}(X, Y) = 0$, then $X \preceq_u Y, \forall u \in \mathcal{U}$. Therefore, $\preceq_{\zeta^*_{\mathcal{U}}} \Rightarrow \preceq_{\mathcal{U}}$. The converse relationship follows by construction, if $Eu(X) \leq Eu(Y), \forall u \in \mathcal{U}$, then $\sup_{u \in \mathcal{U}} (Eu(X) - Eu(Y))_+ = 0$. As a result, $\preceq_{\zeta^*_{\mathcal{U}}} \Leftrightarrow \preceq_U$. The assumption of boundedness of $\zeta^*_{\mathcal{U}}(X, Y)$ is technical and is required to make sure the order $\preceq_{\zeta^*_{\mathcal{U}}}$ is well-defined over all pairs (X, Y). □

Additional properties for the functions in $\mathcal{U}$ have to be specified in order to guarantee that $\zeta^*_{\mathcal{U}}(X, Y)$ is finite. Usually this is done by imposing certain growth conditions. For additional details, see Rachev (1991).

Stoyanov et al. (2009*a*) consider a functional similar to (8.5.1) which is constructed to be consistent with cumulative prospect theory.[10] They demonstrate that the class of investors with balanced views,

introduced in Stoyanov et al. (2009*a*), is sufficient to metrize FSD. In order to be consistent with the definition in (8.5.1), we illustrate this with a subclass. Consider all investors with bounded, non-decreasing Lipschitz utility functions, $u(x) : |u(x) - u(y)| \leq K|x - y|, \forall x, y \in \mathbb{R}$, where $0 < K \leq 1$. Denote this class of utility functions with $\mathcal{U}_L$. Under these assumptions, the quasi-semidistance $\zeta^*_{\mathcal{U}_L}(X, Y)$ is bounded,

$$\zeta^*_{\mathcal{U}_L}(X, Y) \leq \int_{\mathbb{R}} (F_Y(x) - F_X(x))_+ dx,$$

and metrizes FSD. See sections 8.8.3 and 8.8.4 for further details.

Note that both $\zeta^*_{\mathcal{U}_L}$ from this example and the Lévy quasi-semidistance in (8.4.1) metrize FSD. This does not necessarily mean that there is an inequality between $\zeta^*_{\mathcal{U}_L}$ and L^*_λ. From a topological viewpoint, the topologies generated by the two quasi-semimetrics may be completely different and yet their specialization orders can be the same. The link with topology is considered in more detail in section 8.8.1 in the appendix to this chapter.

The quasi-semidistance in (8.5.1) is not a universal representation like the Hausdorff construction in (8.3.1). Therefore, even though any metrizable stochastic order is generated by a quasi-semidistance, there may not exist a quasi-semidistance with a utility-type representation metrizing it. An example of a stochastic order which implies SSD but for which no representation in terms of a class of investors is known was provided in section 3.7.4 of Chapter 3.

In the theory of probability distances, there is a representation similar to the utility-type representation. It is called the *zeta structure* and, unlike the Hausdorff construction for probability distances, it is not a universal representation. Additional details concerning the structural classification of probability distances are provided in section 8.8.5 in the appendix to this chapter.

Finally, whether a utility-type order is primary or simple depends on how rich the family $\mathcal{U}$ is. As an extreme example, if $\mathcal{U}$ contains only one utility function (i.e., there is only one investor), $\zeta^*_{\mathcal{U}}$ generates a primary order.

8.6 Almost Stochastic Orders and Degree of Violation

In section 3.5 of Chapter 3, we discussed that the classical expected utility theory is prescriptive: that is, it determines what the rational behavior of economic agents should be. Empirical work in the field of behavioral finance has identified examples in which people behave differently from the rational prescription of expected utility theory. The following paradox,[11] among many others, illustrates a discrepancy of this kind.

Suppose that an investor having a utility function of the type

$$u_0(x) = \begin{cases} x, & x \le x_0 \\ x_0, & x > x_0 \end{cases}$$

faces the following two alternatives:

$$\text{Alternative A:} \begin{cases} \$1, & p_1 = 1/10 \\ \$10 \text{ mln}, & p_2 = 9/10 \end{cases}$$

$$\text{Alternative B:} \begin{cases} \$2, & p_1 = 1/10 \\ \$3, & p_2 = 9/10. \end{cases}$$

It is easy to verify that neither of the c.d.f.s of A and B dominate the other with respect to FSD because the two c.d.f.s cross. In practice, many investors, if not all of them, would prefer A to B. Nevertheless, an investor with a utility function $u_0(x)$ and $x_0 = 2$ prefers B to A because alternative B has a higher expected utility.[12]

The reason for this paradoxical result is that the FSD criterion is based on the set of all investors with non-decreasing utility functions and $u_0(x)$ is a utility function of this type. This set may include preferences that can be regarded as extreme, pathological, or simply unrealistic. Yet, since from a mathematical viewpoint these preferences describe the behavior of some non-satiable investors, they cannot be excluded and, as a result, neither of the two alternatives dominates the other in terms of FSD.

A way to address some of the paradoxes arising from expected utility theory is discussed in Leshno and Levy (2002) and Bali et al. (2009). They suggest considering a subset of the corresponding investors set because, as we discussed, paradoxes arise from non-realistic choices of utility functions. The stochastic order arising from this smaller set of investors is called *almost stochastic order*.

The general idea is to develop conditions that the utility functions in a given set need to satisfy which depend on the degree of violation of the stochastic order arising from the larger investors set. For instance, consider $s_1 = \{x : F_X(x) - F_Y(x) < 0\}$ and $s_2 = \{x : F_Y(x) - F_X(x) < 0\}$. The degree of violation of $X \preceq_{FSD} Y$ is defined as the ratio

$$\epsilon = \frac{\int_{s_1}(F_Y(x) - F_X(x))dx}{\int_{\mathbb{R}} |F_X(x) - F_Y(x)|dx}$$

and the corresponding condition on the non-decreasing utility functions is derived to be $u'(x) \leq \inf_x u'(x)(1/\epsilon - 1)$.

The degree of violation of FSD can be expressed in terms of a quasi-semidistance metrizing FSD. Consider the Kantorovich quasi-semidistance

$$\boldsymbol{\kappa}^*(X, Y) = \int_{\mathbb{R}} (F_Y(x) - F_X(x))_+ dx. \tag{8.6.1}$$

It can be demonstrated that it metrizes FSD by repeating the arguments in Theorem 8.4.1. For additional information, see also Stoyanov et al. (2009*a*).

The degree of violation ϵ can be related to $\boldsymbol{\kappa}^*(X, Y)$ in the following way:

$$\epsilon/(1 - \epsilon) = \boldsymbol{\kappa}^*(X, Y)/\boldsymbol{\kappa}^*(Y, X).$$

As a result, the corresponding condition becomes

$$u'(x) \leq \inf_x u'(x) \frac{\boldsymbol{\kappa}^*(Y, X)}{\boldsymbol{\kappa}^*(X, Y)}.$$

This example is interesting as it illustrates a generic property. If X and Y are two prospects such that their c.d.f.s do not coincide

completely, then the ratio

$$v_\mu(X, Y) = \frac{\mu(X, Y)}{\mu(Y, X)}, \tag{8.6.2}$$

in which $\mu(X, Y)$ is some quasi-semidistance m measures the degree of violation of the stochastic order $X \preceq_\mu Y$ metrized by μ.

Figure 8.2 illustrates the Kantorovich quasi-semidistance defined in (8.6.1). The shaded area on the top plot equals $\boldsymbol{\kappa}^*(X, Y)$ and the shaded area on the bottom plot, equals $\boldsymbol{\kappa}^*(Y, X)$. As a result, the degree of violation coefficient $v_{\boldsymbol{\kappa}^*}(X, Y)$ corresponding to $\boldsymbol{\kappa}^*$ represents the ratio of the two shaded areas. If the $v_{\boldsymbol{\kappa}^*}$ is very large, then the area on the top plot is very small compared to the area on the bottom plot, implying that Y almost dominates X with respect to FSD. Conversely, if $v_{\boldsymbol{\kappa}^*}$ is very small, then the area on the bottom plot is very small compared to the area on the top plot, implying that X almost dominates Y with respect to FSD.

The same reasoning can be applied to $v_\mu(X, Y)$ with an arbitrary quasi-semidistance μ. In this case, however, the conclusion concerns almost stochastic dominance with respect to the induced order $\preceq_\mu$.

8.7 Summary

In this chapter, we considered a general systematic approach towards describing stochastic dominance rules by means of quasi-semidistances. We provided a universal representation of quasi-semidistances, which we call the Hausdorff representation in line with a similar universal representation in the theory of probability metrics. The theoretical framework allows for a categorization of stochastic orders to a primary, simple, and compound type.

A number of examples supporting the theoretical construct were discussed pertaining to FSD and the n-th order stochastic dominance in general. We introduced a stochastic order based on the average value-at-risk measure which illustrates how the quasi-semidistances approach can be used to generate new stochastic orders.

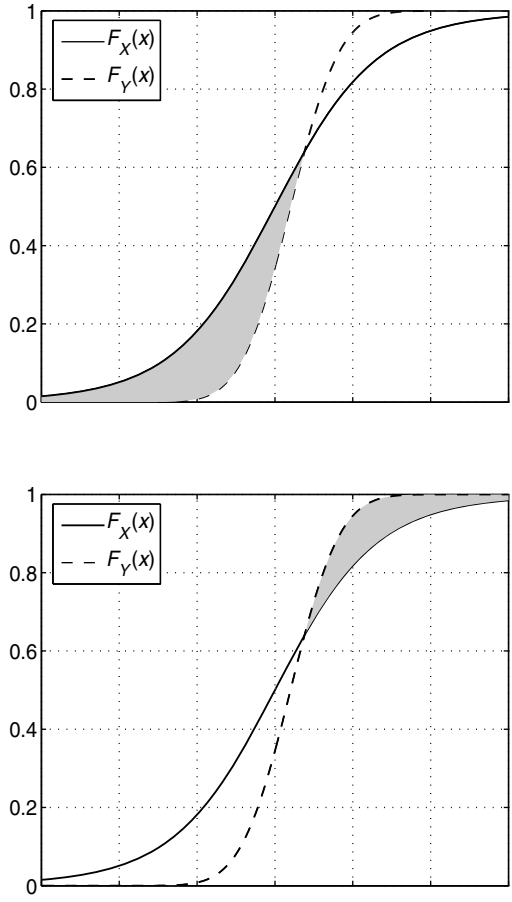

Figure 8.2: Illustration of the Kantorovich quasi-semidistance and its dual. The shaded area on the top plot equals $\boldsymbol{\kappa}^*(X, Y)$ and the shaded area on the bottom plot equals $\boldsymbol{\kappa}^*(Y, X)$.

We also considered stochastic orders arising from classes of investors and a utility-type quasi-semidistance metrizing them. An expected outcome from the theoretical framework is that not all metrizable stochastic orders have a utility-type representation.

Finally, we discussed a way to measure the degree of violation of a stochastic order and how it is related to the notion of almost stochastic dominance.

8.8 Technical Appendix

In this appendix, we start with a more general discussion of the relationship between preference relations, topology, and the notion of metrization. Next, we focus on utility-type representations of quasi-semidistances and cumulative prospect theory. Finally, we briefly describe the structural classification of probability distances which provides intuition for the Hausdorff structure and utility-type representation of probability quasi-semidistances.

8.8.1 Preference relations and topology

In section 8.2, we concluded that every quasi-semidistance defines a preference relation. Discussing the converse question, whether a given preference relation is representable through a quasi-semimetric, requires the more advanced topic of a topological space.

A topological space is a set X together with a collection of subsets τ which satisfies the following axioms:[13]

(a) The empty set and X belong to τ.
(b) The union of any collection of sets in τ is also in τ.
(c) The intersection of any finite collection of sets from τ also belongs to τ.

The collection τ is called a topology on X. In our case, $X = S$. In every topology, there is a natural pre-order which is also known as the *specialization pre-order*. It is denoted by $\preceq_\tau$ and is defined in the following way: for any $x, y \in S$, $x \preceq_\tau y$, if and only if y is contained in all elements of τ that contain x.[14] Intuitively, the element y is more special than the element x because it is contained in more open sets, hence the name of the pre-order.

For example, consider the two-point set $X = \{0, 1\}$ with the topology $\tau = \{\emptyset, \{1\}, \{0, 1\}\}$. It can be directly verified that τ satisfies the

three conditions given above. The specialization order in this case coincides with the natural order of the numbers 0 and 1. Indeed, since the element {1} is contained in all open sets that contain the element {0}, it follows that $0 \preceq_\tau 1$. The cases $0 \preceq_\tau 0$ and $1 \preceq_\tau 1$ are trivial.

In effect, we can conclude that any topology τ on the space of baskets S induces a preference order through the specialization preorder. In this setting, the notion of equivalence between baskets can be related to the notion of topological indistinguishability. That is, if $x \preceq_\tau y$ and $y \preceq_\tau x$, then the open sets which contain x contain also y and vice versa. Therefore, the open neighborhoods of x and y are identical and, therefore, x and y can be considered indistiguishable from a topological viewpoint.

If a given topology on S induces a preference order, then is it true that any preference order arises as the specialization order of some topology? In this more general setting, the answer is in the affirmative. In fact, for a given preference order there are many topologies generating it as a specialization order and the finest of them is also known as the *Alexandrov topology*. There is a representation result according to which there exists a one-to-one correspondence between the set of Alexandrov topologies on S and the set of all preference relations on S. For more details, see Steiner (1966). We can link this result to the discussion of quasi-semimetrics if we can link the notion of topology to the notion of quasi-semimetric. Every quasi-semimetric generates a topology through the system of open balls $B_{x,\epsilon} = \{y : d(x, y) < \epsilon\}$. A topology, however, may not arise from a quasi-semimetric in this manner. If the converse is true, we say that the topology is *quasi-metrizable*. Therefore, even though there is an Alexandrov topology behind any preference order on S, there may not be a quasi-semimetric generating it: that is, it may not be quasi-metrizable.

As a result of this discussion, a given preference order arises from a quasi-semidistance if and only if we can find a quasi-metrizable topology that generates it as a specialization order. Preference orders arising in this fashion we call *quasi-metrizable* or simply *metrizable*. A review of different sets of necessary and sufficient conditions for

quasi-metrizability can be found in Andrikopoulos (2007). Therefore, in case a number of topologies generating a given preference order appear to be quasi-metrizable, then there will be a number of quasi-semidistances generating one and the same preference order. By the same token, if none of the topologies generating a given preference order are quasi-metrizable, then there exist no quasi-semidistances generating it. It is also possible that different quasi-semidistances generate one and the same topology: that is, they are topologically equivalent. In this case, one and the same specialization order is generated by all topologically equivalent quasi-semidistances.

8.8.2 Quasi-semidistances and preference relations

In this section, we discuss in more detail the connection between quasi-semidistances and preference relations. The discussion is generic with no assumptions about the nature of the space S.

We noted that the preference order $\preceq_d$ defined through a quasi-semidistance is a pre-order and, therefore, a preference relation. Every preference relation induces a dual one by considering the converse relation. The dual of $\preceq_d$, denoted by $\preceq_{d^{-1}}$, is introduced in the following way: $x \preceq_{d^{-1}} y$ if and only if $y \preceq_d x$. It turns out that if a given preference relation is generated by a quasi-semidistance, then its dual is also generated by a quasi-semidistance.

Proposition 8.8.1. Suppose that $\preceq_d$ is generated by the quasi-semidistance $d(x, y)$. Then, the dual preference relation $\preceq_{d^{-1}}$ is generated by $d^{-1}(x, y) = d(y, x)$ which is also a quasi-semidistance.

Proof. For a proof, see Stoyanov et al. (2009*b*). □

The quasi-semidistance generating the dual order is monotonic with respect to primary order $\preceq_d$ in the following sense.

Proposition 8.8.2. Suppose that $x \preceq_d y \preceq_d z$, where $x, y, z \in S$. Then, $d^{-1}(x, y) \leq d^{-1}(x, z)$ and also $d^{-1}(y, z) \leq d^{-1}(x, z)$ in which $d^{-1}(x, y) = d(y, x)$.

Proof. For a proof, see Stoyanov et al. (2009*b*). □

This result shows how the quasi-semidistances concept can be used to construct monotonic functionals relative to a given metrizable preference relation, which can be exploited in approximation problems.

Another advantage of the theoretical framework is that it provides a way of comparing preference relations if there exists an inequality between the corresponding quasi-semidistances.

Proposition 8.8.3. Suppose that $d_1(x, y)$ and $d_2(x, y)$ are two quasi-semidistances and that $d_1(x, y) \leq d_2(x, y)$. Under these assumptions, if $x \preceq_{d_2} y$, then $x \preceq_{d_1} y$.

Proof. For a proof, see Stoyanov et al. (2009*b*). □

Note that this result does not imply the converse, i.e. if $x \preceq_{d_2} y \Rightarrow x \preceq_{d_1} y$ for all $x, y \in S$, then there exists an inequality between the corresponding quasi-semidistances.

8.8.3 Construction of quasi-semidistances on classes of investors

The discussion in this section is based on Stoyanov et al. (2009*a*). The approach is consistent with cumulative prospect theory (CPT) and the corresponding notation described in section 3.5 of Chapter 3.

We begin by introducing some notation. The class of bounded S-shaped value functions we denote by $\mathcal{S}$. The elements of $\mathcal{S}$ are bounded real-valued functions $v(x) : \mathbb{R} \to \mathbb{R}$ with the following property,

$$v(x) = \begin{cases} v^-(x), & x < 0 \\ 0, & x = 0 \\ v^+(x), & x > 0 \end{cases}$$

where $v^-(x) < 0$ is a monotonically increasing convex function and $v^+(x) > 0$ is a monotonically increasing concave function.

According to CPT, individuals make a choice between two risky prospects X and Y by computing the subjective expected values according to

$$V(X) = \int_{-\infty}^{0} v(x)d[w^{-}(F_X(x))] + \int_{0}^{\infty} v(x)d[-w^{+}(1 - F_X(x))]$$

where

$$w^{-}(0) = w^{+}(0) = 0 \text{ and } w^{-}(1) = w^{+}(1) = 1.$$

and then compare $V(X)$ and $V(Y)$. If $V(X) \geq V(Y)$, then Y is not preferred to X. If $V(X) = V(Y)$, then the individual is indifferent. Note that if the individuals do not weight the cumulative probabilities, then $V(X)$ reduces to $V(X) = Ev(X)$.

Suppose that all investors which we consider are indifferent between X and Y, $V_j(X) = V_j(Y)$, for all $j \in \mathcal{J}$. Note that $\mathcal{J}$ is a general set, not necessarily countable. In order to study the implications of this assumption on the distribution functions of X and Y, we consider the functional

$$\zeta_{\mathcal{J}}(X, Y) = \sup_{j \in \mathcal{J}} \left| V_j(X) - V_j(Y) \right|, \tag{8.8.1}$$

where

$$\begin{aligned} V_j(X) - V_j(Y) = &\int_{-\infty}^{0} v_j(x)d[w_j^{-}(F_X(x)) - w_j^{-}(F_Y(x))] \\ &+ \int_{0}^{\infty} v_j(x)d[w_j^{+}(1 - F_Y(x)) - w_j^{+}(1 - F_X(x))]. \end{aligned}$$

Note that in the case of no subjective weighting, this expression reduces to

$$V_j(X) - V_j(Y) = \int_{-\infty}^{\infty} v_j(x)d(F_X(x) - F_Y(x)).$$

The functional $\zeta_{\mathcal{J}}(X, Y)$ is the largest difference between the values assigned by the investors to X and Y running through all investors. If the functional in (8.8.1) equals zero, then this means that all investors that we consider are indifferent between X and Y. In fact, $\zeta_{\mathcal{J}}(X, Y)$ has

metric properties. In particular, if $X \stackrel{d}{=} Y$, then $\zeta_{\mathcal{J}}(X, Y) = 0$ although the converse may not hold. Therefore, the fact that $\zeta_{\mathcal{J}}(X, Y) = 0$ does not necessarily imply equality in distribution between X and Y as it depends on how rich the set $\mathcal{J}$ is. The next theorem establishes a sufficient condition for the converse relationship.

Theorem 8.8.1. Suppose that the set $\mathcal{V}_{\mathcal{J}} = \{v_j, j \in \mathcal{J}\} \subseteq \mathcal{S}$ contains the functions

$$v^{-}_{x_0,n}(x) = \begin{cases} -1/n, & x < x_0 \\ x_0 - x - 1/n, & x \in [x_0, x_0 + 1/n) \\ 0, & x \geq x_0 + 1/n \end{cases} \tag{8.8.2}$$

where $n = 1, 2, ...$ and $x_0 + 1/n \leq 0$ and

$$v^{+}_{x_0,n}(x) = \begin{cases} 0, & x < x_0 - 1/n \\ x - x_0 + 1/n, & x \in [x_0 - 1/n, x_0) \\ 1/n, & x \geq x_0 \end{cases} \tag{8.8.3}$$

where $n = 1, 2, ...$ and $x_0 - 1/n \geq 0$. Suppose also that the weighting functions w^{-} and w^{+} are continuous. Then, $\zeta_{\mathcal{J}}(X, Y)$ is a simple probability metric which means that $\zeta_{\mathcal{J}}(X, Y) = 0 \Longleftrightarrow X \stackrel{d}{=} Y$.

Proof. The proof is provided in Stoyanov et al. (2009*a*). □

This result implies that if the set of investors is so large that it contains the value functions defined in (8.8.2) and (8.8.3), then $\zeta_{\mathcal{J}}(X, Y) = 0$ indicates that the c.d.f.s of X and Y coincide. Note that the particular form of the weighting functions is immaterial. The only properties needed are that they are non-decreasing and continuous.

The reasoning outlined above can be used to construct a functional consistent with the FSD order. Consider

$$\zeta^{*}_{\mathcal{J}}(X, Y) = \sup_{j \in \mathcal{J}} \left(V_j(X) - V_j(Y) \right)_{+}, \tag{8.8.4}$$

where $(x)_{+} = \max(x, 0)$. The interpretation of (8.8.4) is as follows. The distance between X and Y equals the largest difference $V_j(X) - V_j(Y)$

running through all investors who do not prefer Y to X. In this case, the condition $\zeta^*_{\mathcal{J}}(X, Y) = 0$ implies that all investors prefer Y to X because in this case $V_j(X) \leq V_j(Y),\ \forall j \in \mathcal{J}$.

Theorem 8.8.2. $\zeta^*_{\mathcal{J}}(X, Y)$ is a probability quasi-semimetric and if $F_Y(x) \leq F_X(x), \forall x \in \mathbb{R}$, then $\zeta^*_{\mathcal{J}}(X, Y) = 0$. Furthermore, if the set of value functions contains the set defined in (8.8.2) and (8.8.3), and the weighting functions are continuous, then $\zeta^*_{\mathcal{J}}(X, Y) = 0$ implies $F_Y(x) \leq F_X(x), \forall x \in \mathbb{R}$.

Proof. The proof is provided in Stoyanov et al. (2009*a*). □

Similarly to $\zeta_{\mathcal{J}}(X, Y)$, if the class $\mathcal{J}$ is not rich enough, then the condition $\zeta^*_{\mathcal{J}}(X, Y) = 0$ does not imply inequality between the distribution functions but only between certain characteristics of X and Y.

8.8.4 Investors with balanced views

This section extends the discussion in section 8.8.3 introducing the notion of investors with balanced views, which is consistent with CPT. This section is based on Stoyanov et al. (2009*a*).

In section 8.8.3, we introduce the quasi-semidistance $\zeta^*_{\mathcal{J}}(X, Y)$ which is consistent with FSD provided that the set $\mathcal{J}$ contains a set of value functions. One last condition we need to check is whether we can choose a class of investors which is sufficiently large and at the same time (8.8.1) and (8.8.4) are bounded. Otherwise, if (8.8.1) and (8.8.4) take only two values – zero and infinity, the construct is meaningless. In this section, we provide upper bounds on $\zeta_{\mathcal{J}}(X, Y)$ and $\zeta^*_{\mathcal{J}}(X, Y)$ introducing additional assumptions which concern the rate of change of $v_j(x)$ and the weighting functions. From a mathematical viewpoint, they can be regarded as smoothness assumptions but because of the particular relationship between $v_j(x)$ and $w(x)$, we call the set $\mathcal{J}$ investors with balanced views. The main result is provided below.

Theorem 8.8.3. Consider the set $\mathcal{V}_{\mathcal{J}}$ of value functions $v_j \in \mathcal{S}$ satisfying the Lipschitz condition $|v_j(x) - v_j(y)| \leq K_{v_j}|x - y|$

and the weighting functions satisfy the Lipschitz conditions $|w_j^-(x) - w_j^-(y)| \leq K_{w_j}|x-y|$ and $|w_j^+(x) - w_j^+(y)| \leq K_{w_j}|x-y|$ where $0 < K_{v_j}K_{w_j} \leq 1$. The following inequalities hold;

$$\zeta_{\mathcal{J}}(X,Y) \leq \int_{\mathbb{R}} |F_X(x) - F_Y(x)|dx \tag{8.8.5}$$

$$\zeta^*_{\mathcal{J}}(X,Y) \leq \int_{\mathbb{R}} (F_Y(x) - F_X(x))_+ dx \tag{8.8.6}$$

Proof. The proof is provided in Stoyanov et al. (2009*a*). □

The Lipschitz conditions imply that the value function and the weighting functions do not change too quickly. For example, if we compare two outcomes x and $x+h,\ h>0$, then the Lipschitz condition suggests that $v_j(x+h) - v(x) \leq K_{v_j}h$ which means that the difference between the assigned values by v_j of the j-th investor is bounded by $K_{v_j}h$. Likewise, we can interpret the Lipschitz condition for the weighting function.

The condition in the theorem, $0 < K_{w_j}K_{v_j} \leq 1$, means that if the value function of a given investor is changing too quickly (K_{v_j} is high), then the weighting functions of the corresponding investor should have a constant K_{w_j} bounded from above by $1/K_{v_j}$. In effect, the combined condition in the theorem means that the individuals that we consider are balanced in their views. A steeper value function should be compensated by a more flat weighting function and vice versa. If the value function and the weighting functions are differentiable, then the Lipschitz conditions translate into bounds on their first derivatives, $|dv_j(x)/dx| \leq K_{v_j}$ and $|dw^{-/+}(x)/dx| \leq K_{w_j}$.

The class of Lipschitz value functions includes (8.8.3) and (8.8.2) with a constant $K_v = 1/n \leq 1$. Thus, investors with balanced views are a sufficiently large class with suitable properties. On the basis of this class using (8.8.1) and (8.8.4), we can draw conclusions about the relation between X and Y.

8.8.5 Structural classification of probability distances

In this section, we discuss briefly a structural classification of probability metrics. This classification is the basis of the Hausdorff

structure and the utility-type representations of probability quasi-semidistances discussed in this chapter.

Chapter 4 was devoted to a classification of probability semidistances $\mu(P)$ ($P \in \mathcal{P}_2$) with respect to various partitionings of the set $\mathcal{P}_2$ into classes $\mathcal{PC}$ such that $\mu(P)$ takes a constant value on each $\mathcal{PC}$. For instance, if $\mathcal{PC} := \mathcal{PC}(P_1, P_2) := \{P \in \mathcal{P}_2 : T_1P = P_1, T_2P = P_2\}$, $P_1, P_2 \in \mathcal{P}_1$ and $\mu(P') = \mu(P'')$ for each P', $P'' \in \mathcal{PC}$ then μ was said to be a simple semidistance. Analogously, if

$$\mathcal{PC} := \mathcal{PC}(\bar{a}_1, \bar{a}_2) := \{P \in \mathcal{P}_2 : h(T_1P) = \bar{a}_1, h(T_2P) = \bar{a}_2\}$$

and $\mu(P') = \mu(P'')$ as P', $P'' \in \mathcal{PC}(\bar{a}_1, \bar{a}_2)$ then μ was said to be a primary distance.

In the present section, we classify the probability semidistances on the basis of their metric structure. There are three basic types of representations – the Hausdorff structure, the lambda structure, and the zeta structure. The material in this section is based on Rachev (1991).

The Hausdorff structure

The intuition behind the Hausdorff structure is based on the Hausdorff semimetric in the space of all subsets in a given metric space and representations of the Lévy metric similar to it.

The definition of Hausdorff probability semidistance structure (briefly, h-structure) is based on the notion of *Hausdorff semimetric* in the space of all subsets of a given metric space (S, ρ):

$$\begin{aligned} r(A, B) &= \inf\{\varepsilon > 0 : A^\varepsilon \supseteq B, B^\varepsilon \supseteq A\} \\ &= \max\{\inf\{\varepsilon > 0 : A^\varepsilon \supseteq B\}, \inf\{\varepsilon > 0 : B^\varepsilon \supset A\}\} \end{aligned} \tag{8.8.7}$$

where A^ε is the open ε-neighborhood of A.

From the definition in (8.8.7), another representation of the Hausdorff follows immediately:

$$r(A, B) := \max(r', r'') \tag{8.8.8}$$

where

$$r' := \sup_{x\in A} \inf_{y\in B} \rho(x, y)$$

and

$$r'' := \sup_{y\in B} \inf_{x\in A} \rho(x, y).$$

As an example of a probability metric with a representation close to that of equality (8.8.8), let us consider the following parametric version of the Lévy metric for $\lambda > 0$, $X, Y \in \mathfrak{X}(\mathbb{R})$

$$\begin{aligned}\mathbf{L}_\lambda(X, Y) :=& \mathbf{L}_\lambda(F_X, F_Y) := \inf\{\varepsilon > 0 : F_X(x - \lambda\varepsilon) - \varepsilon \le F_Y(x) \\ &\le F_X(x + \lambda\varepsilon) + \varepsilon \quad \forall x \in \mathbb{R}\}. \end{aligned} \tag{8.8.9}$$

Obviously, $\mathbf{L}_\lambda$ is a simple metric in $\mathfrak{X}(\mathbb{R})$ for any $\lambda > 0$, and $\mathbf{L} := \mathbf{L}_1$ is usual Lévy metric. Moreover, it is not difficult to check that $\mathbf{L}_\lambda(F, G)$ is a metric in the space $\mathcal{F}$ of all d.f.s. Considering $\mathbf{L}_\lambda$ as a function of λ, we see that $\mathbf{L}_\lambda$ is non-increasing on $(0, \infty)$ and the following limit relations hold

$$\lim_{\lambda\to 0} \mathbf{L}_\lambda(F, G) = \boldsymbol{\rho}(F, G) \quad F, G \in \mathcal{F} \tag{8.8.10}$$

and

$$\lim_{\lambda\to 0} \lambda\mathbf{L}_\lambda(F, G) = \mathbf{W}(F, G). \tag{8.8.11}$$

In equality (8.8.10), $\boldsymbol{\rho}$ is the Kolmogorov metric in $\mathcal{F}$:

$$\boldsymbol{\rho}(F, G) := \sup_{x\in\mathbb{R}} |F(x) - G(x)|. \tag{8.8.12}$$

In equality (8.8.11), $\mathbf{W}(F, G)$ is the *uniform metric between the inverce functions* F^{-1}, G^{-1}

$$\mathbf{W}(F, G) := \sup_{0<t<1} |F^{-1}(t) - G^{-1}(t)| \tag{8.8.13}$$

where F^{-1} is the generalized inverse of F, $F^{-1}(t) := \sup\{x : F(x) < t\}$. Equality (8.8.10) follows from (8.8.9). Likewise, (8.8.11) is handled

by the equalities

$$\lim_{\lambda\to\infty} \lambda \mathbf{L}_\lambda(F,G) = \inf\{\delta > 0 : F(x) \le G(x+\delta), G(x) \le F(x+\delta) \quad \forall x \in \mathbb{R}\}$$
$$= \mathbf{W}(F,G).$$

Let us define the Hausdorff metric between two bounded functions on the real line $\mathbb{R}$. Let dm_λ ($\lambda > 0$) be the Minkovski metric on the plane $\mathbb{R}^2$, i.e. for each $A = (x_1, y_1)$ and $B = (x_2, y_2)$ we have $dm_\lambda(A,B) := \max\{(1/\lambda)|x_1 - x_2|, |y_1 - y_2|\}$. The Hausdorff metric r_λ ($\lambda > 0$) in the set $\mathcal{C}(\mathbb{R}^2)$ (of all closed non-empty sets $G \subset \mathbb{R}^2$) is defined as follows: for $G_1 \subseteq \mathbb{R}^2$ and $G_2 \subseteq \mathbb{R}^2$

$$r_\lambda(G_1,G_2) := \max\left\{\sup_{A\in G_1}\inf_{B\in G_2} dm_\lambda(A,B), \sup_{B\in G_2}\inf_{A\in G_1} dm_\lambda(A,B)\right\}. \tag{8.8.14}$$

We say that r_λ is generated by the metric dm_λ as in equality (8.8.8) the Hausdorff distance r was generated by ρ. Let $f \in D(\mathbb{R})$, the set of all bounded right-continuous functions on $\mathbb{R}$ having limits $f(x-)$ from the left. The set

$$\bar{f} = \{(x,y) : x \in \mathbb{R} \text{ and either } f(x-) \le y \le f(x) \text{ or } f(x) \le y \le f(x-)\}$$

is called the *completed graph* of the function f.

Definition 8.8.1. The metric

$$r_\lambda(f,g) := r_\lambda(\bar{f},\bar{g}) \quad f,g \in D(\mathbb{R})$$

is said to be the *Hausdorff metric in* $D(\mathbb{R})$.

It turns out that the Lévy metric admits a representation in terms of the Hausdorff semimetric r_λ defined in (8.8.14).

Theorem 8.8.4. For all $F, G \in \mathcal{F}$ and $\lambda > 0$

$$\mathbf{L}_\lambda(F,G) = r_\lambda(F,G).$$

Proof. For a proof, see Rachev (1991). □

In order to cover other probability metrics by means of the Hausdorff metric structure, the following generalization of the notion of Hausdorff metric r is needed. Let $\mathcal{FS}$ be the space of all real-valued functions $F_A : A \to \mathbb{R}$, where A is a subset of the metric space (S, ρ).

Definition 8.8.2. Let $f = f_A$ and $g = g_B$ be elements of $\mathcal{FS}$. The quantity

$$\widetilde{r}_\lambda(f, g) := \max(\widetilde{r}'_\lambda(f, g), \widetilde{r}'_\lambda(g, f)) \tag{8.8.15}$$

where

$$\widetilde{r}'_\lambda(f, g) := \sup_{x \in A} \inf_{y \in B} \max \left\{ \frac{1}{\lambda} \rho(x, y), f(x) - g(y) \right\}$$

is called the *Hausdorff semimetric* between the functions f_A, g_B.

Obviously, if $f(x) = g(y) =$ constant for all $x \in A$, $y \in B$ then $\widetilde{r}_\lambda(f, g) = r(A, B)$. Note that $\widetilde{r}_\lambda$ is a metric in the space of all upper semi-continuous functions with closed domains.

The next two theorems are simple consequences of a more general theorem, Theorem 8.8.7.

Theorem 8.8.5. The Lévy metric $\mathbf{L}_\lambda$ (8.8.9) admits the following representation in terms of metric $\widetilde{r}$ given in (8.8.15):

$$\mathbf{L}_\lambda(X, Y) = \widetilde{r}_\lambda(f_A, g_B)$$

where $f_A = F_X$, $g_B = F_Y$, $A \equiv B \equiv \mathbb{R}$, $\rho(x, y) = |x - y|$.

The Lévy metric $\mathbf{L}_\lambda$, thus, has two representations in terms of r_λ and in terms of $\widetilde{r}_\lambda$. Concerning the Prokhorov metric $\boldsymbol{\pi}_\lambda$ (4.7.34) only a representation in terms of $\widetilde{r}_\lambda$ is known. Namely, let $\mathcal{S} = \mathcal{C}((U, d))$ be the space of all closed non-empty subsets of a metric space (U, d) and let r be the Hausdorff distance (8.8.7) in $\mathcal{S}$. Any law $P \in \mathcal{P}_1(U)$ can be considered as a function on the metric space $(\mathcal{S}, r)$ because P is determined uniquely on $\mathcal{S}$, namely:

$$P(A) := \sup\{P(C) : C \in \mathcal{S}, C \subseteq A\} \text{ for any } A \in \mathcal{B}_1.$$

Define a metric $\widetilde{r}_\lambda(P_1, P_2)$ ($P_1, P_2 \in \mathcal{P}_1(U)$) by putting $A = B = \mathcal{S}$ and $\rho = r$ in Equality (8.8.15).

Theorem 8.8.6. For any $\lambda > 0$, the Prokhorov metric $\boldsymbol{\pi}_\lambda$ takes the form

$$\boldsymbol{\pi}_\lambda(P_1, P_2) = \widetilde{r}_\lambda(P_1, P_2) \qquad (P_1, P_2 \in \mathcal{P}_1(U))$$

where $U = (U, d)$ is assumed to be arbitrary metric space.

Next, taking into account Definition 8.8.2, we shall define the Hausdorff structure of probability semidistances.

Without loss of generality (see section 2.6.2), we assume that any probability semidistance $\mu(P)$, $P \in \mathcal{P}_2(U)$ has a representation in terms of pairs of U-valued random variables $X, Y \in \mathfrak{X} := \mathfrak{X}(U)$

$$\mu(P) = \mu(\mathrm{Pr}_{X,Y}) = \mu(X, Y).$$

Let $\mathcal{B}_0 \subseteq \mathcal{B}(U)$ and let the function $\phi : \mathfrak{X}^2 \times \mathcal{B}_0^2 \to [0, \infty]$ satisfy the relations

(a) if $\Pr(X = Y) = 1$ then $\phi(X, Y; A, A) = 0$ for all $A \in \mathcal{B}_0$;
(b) there exists a constant $K_\phi \geq 1$ such that for all $A, B, C \in \mathcal{B}_0$ and r.v. X, Y, Z

$$\phi(X, Z; A, B) \leq K_\phi[\phi(X, Y; A, C) + \phi(Y, Z, C, B)].$$

Definition 8.8.3. Let μ be probability semidistance. The representation of μ in the following form:

$$\mu(X, Y) = h_{\lambda,\phi,\mathcal{B}_0}(X, Y) := \max\{h'_{\lambda,\phi,\mathcal{B}_0}(X, Y), h'_{\lambda,\phi,\mathcal{B}_0}(Y, X)\} \quad (8.8.16)$$

where

$$h'_{\lambda,\phi,\mathcal{B}_0}(X, Y) = \sup_{A \in \mathcal{B}_0} \inf_{B \in \mathcal{B}_0} \max\left\{\frac{1}{\lambda} r(A, B), \phi(X, Y; A, B)\right\} \qquad (8.8.17)$$

is called the *Hausdorff structure* of μ, or simply *h-structure.*

In (8.8.17), $r(A, B)$ is the Hausdorff semimetric in the Borel σ-algebra $\mathcal{B}((U, d))$ (see (8.8.7) with $\rho \equiv d$), λ is a positive number. $B_0 \subseteq \mathcal{B}(U)$ and ϕ satisfies conditions (a) and (b).

Using the properties (a) and (b) we easily obtain the following lemma.

Lemma 8.8.1. Each μ in the form (8.8.16) is a probability semidistance in $\mathfrak{X}$ with a parameter $\mathbb{K}_\mu = K_\phi$.

In the limit cases $\lambda \to 0$, $\lambda \to \infty$ the Hausdorff structure turns into a uniform structure. More precisely, the following limit relations hold.

Lemma 8.8.2. Let μ have Hausdorff structure (8.8.16), then, as $\lambda \to 0$, $\mu(X, Y) = h_{\lambda,\phi,\mathcal{B}_0}(X, Y)$ has a limit which is defined to be be

$$h_{0,\phi,\mathcal{B}_0}(X, Y) = \max\left\{\sup_{A\in\mathcal{B}_0} \phi(X, Y; A, A), \sup_{A\in\mathcal{B}_0} \phi(Y, X; A, A)\right\}.$$

As $\lambda \to \infty$ the limit

$$\lim_{\lambda\to\infty} \lambda h_{\lambda,\phi,\mathcal{B}_0}(X, Y) = h_{\infty,\phi,\mathcal{B}_0}(X, Y) \tag{8.8.18}$$

exists and is defined to be

$$\max\left\{\sup_{A\in\mathcal{B}_0} \inf_{B\in\mathcal{B}_0,\phi(X,Y;A,B)=0} r(A, B), \sup_{A\in\mathcal{B}_0} \inf_{B\in\mathcal{B}_0,\phi(Y,X;A,B)=0} r(A, B)\right\}.$$

Proof. For a proof, see Rachev (1991). □

Example 8.8.1. (*Universal Hausdorff representation*). Each probability semidistance μ has the trivial form $h_{\lambda,\phi,\mathcal{B}_0} = \mu$ where the set $\mathcal{B}_0$ is a singleton, say $\mathcal{B}_0 \equiv \{A_0\}$, and $\phi(X, Y; A_0, A_0) = \mu(X, Y)$.

Example 8.8.2. (*Hausdorff structure of the Prokhorov metric $\boldsymbol{\pi}_\lambda$*). The Prokhorov metric (4.7.34) admits a Hausdorff structure representation $h_{\lambda,\phi,\mathcal{B}_0} = \mu$ where $\mathcal{B}_0$ is either the class $\mathcal{C}$ of all non-empty closed

subsets of U or $\mathcal{B}_0 \equiv \mathcal{B}(U)$ and $\phi(X, Y; A, B) = \Pr(X \in A) - \Pr(Y \in B)$, $A, B \in \mathcal{B}(U)$. As $\lambda \to 0$ and $\lambda \to \infty$ we obtain the limits

$$h_{0,\phi,\mathcal{B}_0} = \boldsymbol{\sigma} \quad \text{(distance in variation)}$$

and

$$h_{\infty,\phi,\mathcal{B}_0} = \ell_\infty.$$

Example 8.8.3. (*Lévy metric* $\mathbf{L}_\lambda(\lambda > 0)$ *in the space* $\mathcal{P}(\mathbb{R}^n)$). Let $\mathcal{F}(\mathbb{R}^n)$ be the space of all right continuous d.f.s F on $\mathbb{R}^n$. We extend the definition of the Lévy metric $(\mathbf{L}_\lambda, \lambda > 0)$ in $\mathcal{F}(\mathbb{R}^1)$ (see Definition (8.8.9)) considering the multivariate case $\mathbf{L}_\lambda$ in $\mathcal{F}(\mathbb{R}^n)$

$$\begin{aligned}\mathbf{L}_\lambda(P_1, P_2) :=& \mathbf{L}_\lambda(F_1, F_2) := \inf\{\varepsilon > 0 : F_1(x - \lambda\varepsilon\mathbf{e}) - \varepsilon \le F_2(x) \\ &\le F_1(x + \lambda\varepsilon\mathbf{e}) + \varepsilon \quad \forall x \in \mathbb{R}^n\} \end{aligned} \tag{8.8.19}$$

where F_i is the d.f. of P_i $(i = 1, 2)$ and $\mathbf{e} = (1, 1, \ldots, 1)$ is the unit vector in $\mathbb{R}^n$.

The Hausdorff representation of $\mathbf{L}_\lambda$ is handled by Representation (8.8.16) where $\mathcal{B}_0$ is the set of all multivariate intervals $(-\infty, x]$ $(x \in \mathbb{R}^n)$ and

$$\phi(X, Y; (-\infty, x], (-\infty, y]) := F_1(x) - F_2(y)$$

i.e., for r.v.s X and Y with d.f.s F_1 and F_2, respectively,

$$\begin{aligned}\mathbf{L}_\lambda(X, Y) = \mathbf{L}_\lambda(F_1, F_2) := \max\Bigg\{ &\sup_{x \in \mathbb{R}^n} \inf_{y \in \mathbb{R}^n} \max\left[\frac{1}{\lambda}\|x - y\|_\infty, F_1(x) - F_2(y)\right], \\ &\sup_{y \in \mathbb{R}^n} \inf_{x \in \mathbb{R}^n} \max\left[\frac{1}{\lambda}\|x - y\|_\infty, F_2(y) - F_1(x)\right]\Bigg\} \end{aligned} \tag{8.8.20}$$

for all $F_1, F_2 \in \mathcal{F}(\mathbb{R}^n)$ where $\|\cdot\|$ stands for the Minkovski norm in $\mathbb{R}^n$, $\|(x_1, \ldots, x_n)\|_\infty := \max_{1 \le i \le n} |x_i|$. Letting $\lambda \to 0$ in Definition (8.8.20) we get the *Kolmogorov distance* in $\mathcal{F}(\mathbb{R}^n)$:

$$\lim_{\lambda \to 0} \mathbf{L}_\lambda(F_1, F_2) = \boldsymbol{\rho}(F_1, F_2) := \sup_{x \in \mathbb{R}^n} |F_1(x) - F_2(x)|$$

The limit of $\lambda \mathbf{L}_\lambda$ as $\lambda \to \infty$ is given by (8.8.18), i.e.

$$\begin{aligned}
&\lim_{\lambda\to\infty} \lambda_\lambda(F_1, F_2) \\
&\quad = \inf\{\varepsilon > 0 : \inf[F_1(x) - F_2(y) : y \in \mathbb{R}^n, \|x - y\|_\infty \le \varepsilon] = 0, \\
&\qquad \inf[F_2(x) - F_1(y) : x \in \mathbb{R}^n, \|x - y\|_\infty \le \varepsilon] = 0 \quad \forall x \in \mathbb{R}^n\} \\
&\quad = \mathbf{W}(F_1, F_2) := \inf\{\varepsilon > 0 : F_1(x) \le F_2(x + \varepsilon \mathbf{e}), F_2(x) \le F_1(x + \varepsilon \mathbf{e}) \\
&\qquad \forall x \in \mathbb{R}^n\}. \qquad (8.8.21)
\end{aligned}$$

Example 8.8.4. (*Lévy probability distance* $\mathbf{L}_{\lambda,H}, \lambda > 0, H \in \mathcal{H}$). The Lévy metric $\mathbf{L}_\lambda$ (8.8.19) can be rewritten in the form

$$\begin{aligned}
\mathbf{L}_\lambda(F_1, F_2) :=& \inf\{\varepsilon > 0 : (F_1(x) - F_2(x - \lambda\varepsilon \mathbf{e}))_+ < \varepsilon, \\
& (F_2(x) - F_1(x - \lambda\varepsilon \mathbf{e}))_+ < \varepsilon \quad \forall x \in \mathbb{R}^n\} \quad (\cdot)_+ := \max(\cdot, 0)
\end{aligned}$$

which can be viewed as a special case ($H(t) = t$) of the *Lévy probability distance* $\mathbf{L}_{\lambda,H}(\lambda > 0, H \in \mathcal{H})$ defined as follows

$$\begin{aligned}
\mathbf{L}_{\lambda,H}(F_1, F_2) :=& \inf\{\varepsilon > 0 : \widetilde{H}(F_1(x) - F_2(x + \lambda\varepsilon \mathbf{e})) < \varepsilon, \\
& \widetilde{H}(F_2(x) - F_1(x + \lambda\varepsilon \mathbf{e})) < \varepsilon \quad \forall x \in \mathbb{R}^n\} \qquad (8.8.22)
\end{aligned}$$

where

$$\widetilde{H}(t) := \begin{cases} H(t) t \ge 0 \\ 0 \quad t \le 0. \end{cases}$$

$\mathbf{L}_{\lambda,H}$ admits a Hausdorff representation of the following type:

$$\begin{aligned}
\mathbf{L}_{\lambda,H}(F_1, F_2) = \max \Bigg\{ & \sup_{x\in\mathbb{R}^n} \inf_{y\in\mathbb{R}^n} \max \left[\frac{1}{\lambda} \|x - y\|, \widetilde{H}(F_1(x) - F_2(y)) \right], \\
& \sup_{y\in\mathbb{R}^n} \inf_{x\in\mathbb{R}^n} \max \left[\frac{1}{\lambda} \|x - y\|, \widetilde{H}(F_2(y) - F_1(x)) \right] \Bigg\}. \qquad (8.8.23)
\end{aligned}$$

The last representation of $\mathbf{L}_{\lambda,H}$ shows that $\mathbf{L}_{\lambda,H}$ is a simple distance with parameter $\mathbb{K}_{\mathbf{L}_{\lambda,H}} := K_H$ (see 2.3.3). Also, from (8.8.23) as $\lambda \to 0$, we get the *Kolmogorov probability distance*

$$\lim_{\lambda\to 0} \mathbf{L}_{\lambda,H}(F_1, F_2) = H(\rho(F_1, F_2)) = \boldsymbol{\rho}_H(F_1, F_2) := \sup_{x\in\mathbb{R}^n} H(|F_1(x) - F_2(x)|).$$

Analogously, letting $\lambda \to \infty$ in (8.8.23), we have

$$\lim_{\lambda \to \infty} \lambda \mathbf{L}_{\lambda,H}(F_1, F_2) = \mathbf{W}(F_1, F_2). \tag{8.8.24}$$

We prove equality (8.8.24) by arguments provided in the limit relation (8.8.21).

For additional examples and further information on the Hausdorff construction, see Rachev (1991).

The Lambda structure
The probability semidistance structure Λ in $\mathcal{X} = \mathcal{X}(U)$ is defined by means of a non-negative function ν on $\mathcal{X} \times \mathcal{X} \times [0, \infty)$ that satisfies the relationships: for all $X, Y, Z \in \mathfrak{X}$,

(a) If $\Pr(X = Y) = 1$ then $\nu(X, Y; t) = 0 \; \forall t \geq 0$
(b) $\nu(X, Y; t) = \nu(Y, X; t)$
(c) If $t' < t''$ then $\nu(X, Y; t') \geq \nu(X, Y; t'')$
(d) For some $K_\nu > 1$, $\nu(X, Z; t' + t'') \leq K_\nu[\nu(X, Y; t') + \nu(Y, Z, t'')]$.

If $\nu(X, Y; t)$ is completely determined by the marginals $P_1 = \Pr_X$, $P_2 = \Pr_Y$, we shall use the notation $\nu(P_1, P_2; t)$ instead of $\nu(X, Y; t)$. For the case $K_\nu = 1$, we provide the following definition.

Definition 8.8.4. The probability semidistance μ has a Λ-structure if it admits a Λ-representation: that is,

$$\mu(X, Y) = \Lambda_{\lambda,\nu}(X, Y) := \inf\{\varepsilon > 0 : \nu(X, Y; \lambda\varepsilon) < \varepsilon\} \tag{8.8.25}$$

for some $\lambda > 0$ and ν satisfying (a) to (d).

Obviously, if μ has a Λ-representation (8.8.25), then μ is a probability semidistance with $\mathbb{K}_\mu = K_\nu$. In Example 8.8.1 it was shown that each probability semidistance has a Hausdorff representation $h_{\lambda,\phi,\mathcal{B}_0}$. In the next theorem we shall prove that each probability semidistance μ with Hausdorff structure (see Definition 8.8.3) also has a Λ-representation. Hence, in particular, each probability semidistance has a Λ-structure as well as a Hausdorff structure.

Theorem 8.8.7. Suppose a probability semidistance μ admits the Hausdorff representation $\mu = h_{\lambda,\phi,\mathcal{B}_0}$ (8.8.16). Then μ enjoys also a Λ-representation

$$h_{\lambda,\phi,\mathcal{B}_0}(X, Y) = \Lambda_{\lambda,\nu}(X, Y) \tag{8.8.26}$$

where

$$\nu(X, Y; t) := \max\left\{\sup_{A\in\mathcal{B}_0}\inf_{B\in A(t)}\phi(X, Y; A, B), \sup_{A\in\mathcal{B}_0}\inf_{B\in A(t)}\phi(Y, X; A, B)\right\}$$

and $A(t)$ is the collection of all elements B of $\mathcal{B}_0$ such that the Hausdorff semimetric $r(A, B)$ is not greater than t.

Proof. For a proof, see Rachev (1991). □

Example 8.8.5. (Λ-*structure of the Lévy metric and the Lévy distance*). Recall the definition of the Lévy metric in $\mathcal{P}(\mathbb{R}^n)$ (see (8.8.19)):

$$\mathbf{L}_\lambda(P_1, P_2) := \inf\left\{\varepsilon > 0 : \sup_{x\in\mathbb{R}^n}(F_1(x) - F_2(x + \lambda\varepsilon\mathbf{e})) \le \varepsilon \right.$$
$$\left. \text{and } \sup_{x\in\mathbb{R}^n}(F_2(x) - F_1(x + \lambda\varepsilon\mathbf{e})) \le \varepsilon\right\}$$

where obviously F_i is the d.f. of P_i. By Definition 8.8.4, $\mathbf{L}_\lambda$ has a Λ-representation

$$\mathbf{L}_\lambda(P_1, P_2) = \Lambda_{\lambda,\nu}(P_1, P_2) \quad \lambda > 0$$

where

$$\nu(P_1, P_2; t) := \sup_{x\in\mathbb{R}^n}\max\{(F_1(x) - F_2(x + \lambda t\mathbf{e})), (F_2(x) - F_1(x + \lambda t\mathbf{e}))\}$$

and F_i is the d.f. of P_i. With an appeal to Theorem 8.8.7, for any $F_1,\ F_2 \in \mathcal{F}(\mathbb{R}^n)$, we conclude that the metric h defined below

admits a Λ-representation:

$$\begin{aligned} h(F_1, F_2) :&= \max\Big\{ \sup_{x\in\mathbb{R}^n} \inf_{y\in\mathbb{R}^n} \max\Big\{ \frac{1}{\lambda}\|x-y\|_\infty, F_1(x) - F_2(y)\Big\}, \\ &\qquad\qquad \sup_{x\in\mathbb{R}^n} \inf_{y\in\mathbb{R}^n} \max\Big\{ \frac{1}{\lambda}\|x-y\|_\infty, F_2(x) - F_1(y)\Big\}\Big\} \\ &= \Lambda_{\lambda,\nu}(P_1, P_2) \end{aligned}$$

where

$$\begin{aligned} \nu(P_1, P_2; t) = \max\Big\{ &\sup_{x\in\mathbb{R}^n} \inf_{y:\|x-y\|_\infty\le t} (F_1(x) - F_2(y)), \\ &\sup_{x\in\mathbb{R}^n} \inf_{y:\|x-y\|_\infty\le t} (F_2(x) - F_1(y))\Big\}. \end{aligned}$$

By virtue of the Λ-representation of the $\mathbf{L}_\lambda$ we conclude that $h(F_1, F_2) = \mathbf{L}_\lambda(F_1, F_2)$ which proves (8.8.20) and Theorem 8.8.5.

Analogously, consider the Lévy distance $\mathbf{L}_{\lambda,H}$ (8.8.22) and apply Theorem 8.8.7 with

$$\begin{aligned} \nu(X, Y; \lambda t) &= \nu(P_1, P_2; \lambda t) \\ &:= H\Big(\sup_{x\in\mathbb{R}^n} \max\{F_1(x) - F_2(x + \lambda t\mathbf{e}), \{F_2(x) - F_1(x + \lambda t\mathbf{e})\Big) \end{aligned}$$

to prove the Hausdorff representation of $\mathbf{L}_{\lambda,H}$ (8.8.23).

Example 8.8.6. (Λ-*structure of the Prokhorov metric* $\boldsymbol{\pi}_\lambda$). Let

$$\begin{aligned} \nu(P_1, P_2; \varepsilon) &:= \sup_{A\in\mathcal{B}(U)} \max\{P_1(A) - P_2(A^\varepsilon), P_2(A) - P_1(A^\varepsilon)\} \\ &= \sup_{A\in\mathcal{B}(U)} \{P_1(A) - P_2(A^\varepsilon)\} \end{aligned}$$

Then $\Lambda_{\lambda,\nu}$ is the Λ-representation of the Prokhorov metric $\boldsymbol{\pi}_\lambda(P_1, P_2)$. In this way, Theorem 8.8.6 is a corollary of Theorem 8.8.7.

Example 8.8.7. (Λ-*structure of the Ky Fan metric and Ky Fan distance*). The Λ-structure of the Ky Fan metric $\mathbf{K}\Lambda$ (see 4.7.57) and the Ky Fan distance $\mathbf{KF}_H$ (see 4.7.56) is handled by assuming

that in (8.8.25), $\nu(X, Y; \lambda t) := \Pr(d(X, Y) > \lambda t)$ and $\nu(X, Y; t) := \Pr(H(d(X, Y)) > t)$, respectively.

The zeta structure
Let $C^b(U)$ be the set of all bounded continuous functions on U. Then, for each subset $\mathfrak{F}$ of $C^b(U)$, the functional

$$\zeta_{\mathfrak{F}}(P_1, P_2) := \zeta(P_1, P_2; \mathfrak{F}) := \sup_{f \in \mathfrak{F}} \left| \int_U f \mathrm{d}(P_1 - P_2) \right| \tag{8.8.27}$$

on $\mathcal{P}_1 \times \mathcal{P}_1$ defines a simple probability semimetric in $\mathcal{P}_1$. The metric $\zeta_{\mathfrak{F}}$ was introduced by Zolotarev (1976) and it is called the *Zolotarev* $\zeta_{\mathfrak{F}}$*-metric* (or briefly $\zeta_{\mathfrak{F}}$*-metric*).

Definition 8.8.5. A simple semimetric μ having the $\zeta_{\mathfrak{F}}$*-representation*

$$\mu(P_1, P_2) = \zeta_{\mathfrak{F}}(P_1, P_2) \tag{8.8.28}$$

for some $\mathcal{F} \subseteq C^b(U)$, is called semimetric with ζ*-structure*.

Example 8.8.8. (*Engineer metric*). Let $U = \mathbb{R}$ and $\mathfrak{X}^{(1)}$ be the set of all real valued r.v.s X with finite first absolute moment, i.e. $E|X| < \infty$. In the set $\mathfrak{X}^{(1)}$ the engineer metric $\mathbf{EN}(X, Y) := |EX - EY|$ admits the ζ-representation, where $\mathcal{F}$ is a collection of functions:

$$f_N(x) = \begin{cases} -Nx < N \\ x \quad |x| \le N \\ N \quad x > N, N = 1, 2, \ldots \end{cases}$$

Example 8.8.9. (*Kolmogorov metric and* $\boldsymbol{\theta}_p$*-metric in the distribution function space*). Let $\mathcal{F} = \mathcal{F}(R)$ be the space of all d.f.s on $\mathbb{R}$. The Kolmogorov metric $\boldsymbol{\rho}(F_1, F_2) := \sup_{x \in \mathbb{R}} |F_1(x) - F_2(x)|$ in $\mathcal{F}$ has $\zeta_{\mathfrak{F}}$-structure. In fact

$$\boldsymbol{\rho}(F_1, F_2) = \|f_1 - f_2\|_\infty = \sup \left\{ \left| \int_{-\infty}^{\infty} u(x)(F_1(x) - F_2(x)) \mathrm{d}x \right| : \|u\|_1 \le 1 \right\} \tag{8.8.29}$$

Here and subsequently $\|\cdot\|_p$ $(1 \le p < \infty)$ stands for the $\mathcal{L}^p$-norm

$$\|u\|_p := \left\{ \int_{-\infty}^{\infty} |u(x)|^p \mathrm{d}x \right\}^{1/p} \qquad 1 \le p < \infty$$
$$\|u\|_\infty := \operatorname*{ess\,sup}_{x \in \mathbb{R}} |u(x)|.$$

Further, let us denote, by $\mathfrak{F}(p)$, the space of all (Lebesgue) a.e. differentiable functions f such that the derivative f' has $\mathcal{L}^p$-norm $\|f'\|_p \le 1$, hence, integrating by parts the right-hand side of (8.8.29) we obtain a ζ-representation of the uniform metric $\boldsymbol{\rho}$

$$\boldsymbol{\rho}(F_1, F_2) := \sup_{f \in \mathfrak{F}(1)} \left| \int_{-\infty}^{\infty} f(x) \mathrm{d}(F_1(x) - F_2(x)) \right| = \zeta(F_1, F_2; \mathfrak{F}(1)).$$

Analogously, we have a $\zeta_{\mathfrak{F}(q)}$-representation for $\boldsymbol{\theta}_p$-metric $(p \ge 1)$ (see 4.7.40):

$$\begin{aligned} \boldsymbol{\theta}_p(F_1, F_2) &:= \|F_1 - F_2\|_p \\ &= \sup \left\{ \left| \int_{-\infty}^{\infty} u(x)(F_1(x) - F_2(x)) \mathrm{d}x \right| : \|u\|_q \le 1 \right\} \\ &= \zeta(F_1, F_2; \mathcal{F}(q)). \end{aligned}$$

It turns out, however, that not all metrics admit a zeta representation. The following lemma shows that $\ell_p = \widehat{\mathcal{L}}_p$, $(p > 1)$ has no ζ-representation.

Lemma 8.8.3. If a s.m.s. (U, d) has more than one point and the minimal metric $\widehat{\mathcal{L}}_p$, $(p > 1)$ has a ζ-representation (8.8.27), then $p = 1$.

Proof. For a proof, see Rachev (1991). □

By Lemma 8.8.3 it follows, in particular, that there exist simple metrics that have no $\zeta_{\mathbb{F}}$-representation. In the case of $\widehat{\mathcal{L}}_p$-metric, however, we can find a $\zeta_{\mathbb{F}}$-metric which is topologically equivalent to $\widehat{\mathcal{L}}_p$. Nevertheless, there exist metrics for which this cannot be done: that is, it is impossible to find a topologically equivalent $\zeta_{\mathbb{F}}$-metric which

implies that the ζ-structure is not universal. For additional details and examples, see Rachev (1991).

An generalization of the notion of the ζ-structure, which represents a universal structure, is provided below.

Definition 8.8.6. We say that a probability semidistance μ admits a *ζ-structure* if μ can be written in the following way:

$$\mu(X, Y) = \bar{\zeta}(X, Y; \overline{\mathbb{F}}(X, Y)) = \sup_{f \in \overline{\mathbb{F}}(X,Y)} Ef \tag{8.8.30}$$

where $\mathbb{F}(X, Y)$ is a class of integrable functions $f : \Omega \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given on a probability space $(\Omega, \mathcal{A}, \Pr)$.

In general $\bar{\zeta}$ is not a probability semidistance, but each probability semidistance has a $\bar{\zeta}$-representation. Actually, for each probability semidistance μ, the equality (8.8.30) is valid where $\overline{\mathbb{F}}(X, Y)$ contains only a constant function $\mu(X, Y)$. For additional details and examples, see Rachev (1991).

We completed the investigation of the three universal metric structures (h, Λ, and $\bar{\zeta}$). The reason we call them universal is that each probability semidistance μ has h-, Λ- and $\bar{\zeta}$-representation simultaneously. Thus, depending on the specific problem under consideration, one can use one or another probability semidistance representation.

Notes

1. Probability semidistances are introduced in Chapter 2. In the appendix to Chapter 2, we briefly discuss probability quasi-semidistances and how they differ from probability semidistances.
2. In section 2.4 of Chapter 2, we defined probability quasi-semidistances and make a similar parallel with probability semidistances.
3. The discussion will not change in a fundamental way if we consider general random elements taking values in a general functional space. We consider one-dimensional random variables for the sake of simplicity.

4. Different types of stochastic dominance relations are discussed in detail in Chapter 3.
5. A detailed discussion of primary, simple, and compound probability metrics is provided in Chapter 4.
6. The dual quasi-semidistance $d^{-1}(x, y)$ corresponding to a given quasi-semidistance $d(x, y)$ is defined in the following way, $d^{-1}(x, y) = d(y, x)$. A more detailed discussion is provided in section 8.8.2 in the appendix to this chapter.
7. The dual order $\preceq_{d^{-1}}$ corresponding to a given order $\preceq_d$ is defined in the following way, $x \preceq_{d^{-1}} y$ if and only if $y \preceq_d x$. For a more detailed discussion, see section 8.8.2 in the appendix to this chapter.
8. For a detailed discussion, see section 3.3 of Chapter 3.
9. For a detailed discussion, see section 3.7.2 of Chapter 3.
10. More details are provided in sections 8.8.3 and 8.8.4 in the appendix to this chapter.
11. This paradox is discussed in Levy (2006).
12. Similar examples can be constructed for SSD and higher-order stochastic dominance as well. See Levy (2006) for additional examples.
13. See, for example, Kelly (1975).
14. See, for example, Steiner (1966).

References

Anand, P. (1995), *Foundations of Rational Choice under Risk*, Oxford University Press, Oxford.

Andrikopoulos, Athanasios (2007), 'The quasimetrization problem in the (bi)topological spaces', *International Journal of Mathematics and Mathematical Sciences*, Article ID 76904, doi:10.1155/2007/76904.

Artzner, P., F. Delbaen, J.-M. Eber and D. Heath (1998), 'Coherent measures of risk', *Mathematical Finance* **6**, 203–228.

Bali, T., O. Demirtas, H. Levy and A. Wolf (2009), 'Bonds versus stocks: Investors' age and risk taking', *Journal of Monetary Economics*, Forthcoming. Available at SSRN: http://ssrn.com/abstract=936648.

Kelly, John (1975), *General Topology*, Springer, New York.

Leshno, M. and H. Levy (2002), 'Preferred by "all" and preferred by "most" decision makers: Almost stochastic dominance', *Management Science* **48**, 1074–1085.

Levy, Haim (2006), *Stochastic Dominance: Investment Decision Making Under Uncertainty*, Springer, New York.

Ortobelli, L. S., S. Rachev, H. Shalit and F. Fabozzi (2009), 'Orderings and probability functionals consistent with preferences', *Applied Mathematical Finance* **16**, 81–102.

Rachev, S. T. (1991), *Probability Metrics and the Stability of Stochastic Models*, Wiley, New York.

Steiner, A. K. (1966), 'The lattice of topologies: structure and complementation', *Trans. Amer. Math. Soc.* **122 (2)**, 379–398.

Stoyanov, S., S. Rachev and F. Fabozzi (2008), 'Probability metrics with applications in finance', *Journal of Statistical Theory and Practice* **2** (2), 253–277.

Stoyanov, S., S. Rachev and F. Fabozzi (2009*a*), 'Construction of probability metrics on classes of investors', *Economics Letters* **103**, 45–48.

Stoyanov, S., S. Rachev and F. Fabozzi (2009*b*), 'Metrization of stochastic dominance rules', research report, Institut für Statistik und Mathematische Wirtschaftstheorie, Karlsruhe University.

Index

Note: an 'n' after a page reference refers to a note on that page.

absolute moments 18, 73, 74, 87, 89, 100, 112–13, 246
  central 155–6
  divergent 288
  finite first 121, 351
absolute risk-aversion
  Arrow-Pratt measure of 80n
  coefficient of 50, 51, 56
  constant 51
  decreasing 50, 57
Acerbi, C. 221, 249n
additive shifts 156, 186
additivity 34, 96
  see also sub-additivity axiom
Akilov, G. P. 124
Alexandrov topology 333
Allias's paradox 67
almost everywhere identity 100
almost stochastic dominance 306, 307, 330, 332
almost stochastic orders 328–30
  degree of violation utilized in 304
analytic tractability 301n
Anand, P. 308
Andrikopoulos, A. 334
approximate model 1, 101, 296
  appealing candidate for 220
  standard normal distribution can be used as 264
approximation 239, 242, 243, 255–6, 257, 283, 288, 335
  implications for accuracy 252
  kernel 212, 244, 294
  linear 298, 299
  normal 174
  probability law governing 277, 289–90
  reasonable 220, 297

A Probability Metrics Approach to Financial Risk Measures by Svetlozar T. Rachev, Stoyan V. Stoyanov and Frank J. Fabozzi

approximation (*Continued*)
smooth 211, 212, 294
step 212
approximation error 252, 256, 276, 298
Archimedean axiom 71
Arrow, Kenneth 43, 80n
Artzner, P. 175, 322
asset allocation puzzle 67
asset returns
empirically-confirmed phenomena about 4
joint behavior of 3
see also return distributions
asymptotic distribution 5, 253, 255, 256, 270, 271, 276, 297, 299
classical conditions 259–62
heavy-tailed returns 277–83, 284
autocorrelations 168, 172
autoregressive behavior 4
auxiliary variables 205, 206
AVaR (average value-at-risk) 160, 179, 180, 191–251, 322–4
advantages and disadvantages 5
consistent with SSD order 182
definition of 181
Monte Carlo-based estimation of 5, 209–11, 239, 252–303
average metrics 131, 132
axiomatic constructions 5, 9, 189n
asymmetric probability distances 5
dispersion measures and deviation measures 150
moment functions 101
probability metrics 3
risk measures 150
axiomatic description 155–6, 182
axioms of choice 47, 70, 71–2

back-testing
AVaR 218–20, 227
VaR 172–5
Bali, T. 329
bandwidth matrix 213–14, 215–16, 243, 249n
optimal choice for 218
banking crisis (2008) 291
Basel Committee on Banking Supervision 160, 188n
Baucells, M. 68, 69
behavioral finance 67, 98, 328
Benartzi, S. 67
benchmark tracking problems 149, 224
benchmarks 16–17, 96, 101
deterministic 12–13
multiple 149
Berkes and Phillips's Lemma 38
Bernoulli, Daniel 42, 44, 45–6, 48
Bernoulli, Nicolas 44
Bianchi, M. L. 302n
bias 260, 276, 298, 301n
magnitude of 271, 275
bias status quo 67
Billingsley, P. 29, 30, 94, 95
binary relations 308
binomial distribution 174
Birnbaum, Z. W. 23
Birnbaum-Orlicz distance 8, 23, 123, 126
compound 134
uniform 84, 127, 135
Birnbaum-Orlicz metric 107, 135
Blackwell, D. 33
bootstrap methods 171, 255–6, 259

Borel algebras 7, 21, 25, 28, 30, 31, 123, 127, 310, 321, 322, 345
  finite unions of rectangles 95–6
  isomorphism 32, 33, 34
  non-empty sets 36, 37
  probability measures 29, 37, 38, 87
bounded functions 21, 122, 342

Cantor sets 34
capital reserves 160, 163, 173
cash invariance *see* translation invariance
Cauchy distribution 257–8
c.d.f. (cumulative distribution function) 14, 46–7, 51–4, 64, 72, 73, 168, 193–4, 245, 248n, 249n, 274, 279, 300, 317, 318, 328
  coincident 329–30, 337
  continuous 198, 234, 239
  differentiable 260–1, 298–9
  empirical 168, 212, 213, 214, 255–6, 257
  inequality in terms of 63, 181–2, 314
  inverse 196, 197, 226, 229, 231, 234–5, 237–8, 260, 299, 315
  kernel estimate of 213, 242
  limiting distribution 288
  lottery described by 46
  random variables with 10, 65, 74–5, 280, 295, 298–9
  smooth 242, 294
  stable distribution 283
  theoretical 212, 242
  transformations of 69, 307
  true 250n, 256
characteristic functions 249n, 274, 294
Chebyshev's inequality 153
choice
  axioms of 47, 70, 71–2
  inappropriate for risk measures 149
  nonrealistic, utility functions 329
  optimal, bandwidth matrix 218
  portfolio problem, efficient sets and 58–9
  risky prospects 336
  under uncertainty 3, 5, 6, 40–82, 98, 306, 325
Chung, D. 302n
closed-form expressions 169, 172, 191, 193, 199, 200, 202, 205, 207, 209, 255, 263, 265, 270, 274
CLT (central limit theorem) 1, 5, 13, 255, 259, 260
  classical 253, 279, 285, 286, 287
  problems related to 106
  rate of convergence in 262
  see also GCLT
coefficients 72, 248n
  absolute risk-aversion 50, 51, 56
coherent risk measures 147, 175–8, 193, 196, 220, 221, 226–7, 231, 240
  average of 232, 233
  generating a family of 228
  probability quasi-metric generated from 322
  strictly expectation-bounded 180, 181
Cohn, D. L. 29, 30, 32

coin-tossing
  fair 44–5
  unfair 174
colog measure 156
co-minimal distance 97, 99, 119–20, 127
co-minimal metrics 96–7, 118, 141
common stocks 46, 165, 174
  FSD order of 62
  log-return of 61
  payoffs of 60, 62, 76
  price of 60, 74–5, 107
  random values of 107
  see also return distributions
completeness axiom 71
compound distances 4, 139, 141
  and moment functions 99–105
  examples of 131–5
compound metrics 19, 25, 100, 136–7
  average 17, 104–5, 107, 132, 133–4
compound quasi-semidistances 324–5
conditional expectations 198, 229, 239, 243, 261, 268, 270
conditional loss distribution 198–9, 228–31
conditional probability 77
confidence intervals 170–1, 173, 174, 210, 255, 256, 261, 264, 265, 268, 282, 284
confidence levels 161, 164, 165, 168, 169
consistency 24, 25, 62, 70, 80, 255, 306
  risk measure 59, 146, 150, 160, 181–2, 183
continuity conditions 71, 94–5, 120
  technical 47, 72
continuous functions 198, 224
  bounded 92, 119, 128–9, 342, 351
  non-decreasing 23, 122, 131
  non-negative 131, 137
  space of 30
  strictly increasing 21
convergence 21, 246
  absolute 247
  see also rate of convergence; weak convergence
convex risk measures 183–4
convexity property 157, 177, 183–4, 243, 244
copula function 169, 253–4
covariance matrix 166, 169, 207, 218
CPT (cumulative prospect theory) 43, 66–9, 70, 98, 335, 336, 338
Cramer, Gabriel 44, 46
credit risk 149, 188n

Debreu, Gérard 43, 80n
decision-making 42, 46, 53, 223
  risk an essential factor in 147
DeGiorgi, E. 182
degree of violation 304, 328–30
degrees of freedom 199, 248n, 262, 263, 264, 268, 270–1, 288
  varying 286
densities 202, 218, 239, 263, 274, 285
  absolute difference between 112
  normal 212–13, 214, 242–3, 268, 292
  random variable 108–9
  stable 201, 284
density functions 249n, 281

analytic 243
bounded continuous 294
kernel density estimator 213
multivariate 244
random variables 212, 261
dependence 172, 182
captured by means of copula function 253–4
presumed 169
short- and long-range 4, 253, 290
stochastic 3
unrealistic model 96
varying structure 89, 93
derivatives 73, 76, 172, 211, 325, 352
complex 256
partial 214, 242
pricing 262, 275
see also first derivatives; second derivatives
deviation measures 4–5, 8, 147, 156, 179, 182, 189n
axiomatic construction of 150
downside 158
lower-range-dominated 180, 181
probability metrics and 184–7
probability quasi-metrics and 183, 187–8
symmetric 157, 185, 186
see also MAD; standard deviation
differentiability 243
lack of 211, 215
dispersion measures 4–5, 147, 150–8
axiomatic description of 182
convex 150
probability metrics and 158–9
risk measures and 179–81
upside/downside 155
very natural generic way of defining 187
distance spaces 19–23
distribution functions 5, 13, 21, 52, 53, 54–5, 60, 75, 93, 113, 127, 188n
coincidence of 19, 90, 91–2, 99, 100
compound probability semimetric 85
conditions for stochastic dominance involving 63
distances defined on the space of 4, 90
equal 94
generalized inverse of 84, 104
inverse of 108, 161, 162, 174, 225
investment opportunities compared directly through 3
L_p-metrics between 8, 15–16, 98, 108
marginal 104
probability 48, 65
space of 22, 23, 90
uniform metric between 12
see also c.d.f.
distributional hypotheses 5, 167
accepting or rejecting 175
distributional models 290–7
diversification effect 94, 163, 164, 175, 196
convexity property and 177, 183
Dokov, S. 204
domains of attraction 220, 278, 294, 295, 296, 300, 301
dual stochastic order 70, 306

Dudley, R. M. 16, 29, 30, 32, 36, 95, 125
Dunford, N. 23

Ellsberg paradox 67
EN (engineer's metric) 8, 10, 13, 18, 21, 22, 27, 84, 87, 89, 117, 351
equity premium puzzle 67
equivalence 22, 28, 118
ETL (expected tail loss) 237–42
Euclidean space 117
exceedances 173–4, 219
expected payoffs 49, 52, 53, 54, 79
  inequality a necessary condition for SSD 54
  infinite 45
expected returns 3, 59, 87, 158, 165, 166–7, 193, 194, 207
  EN computes distance between 10
  infinite 196
  non-random return equal to 64
  portfolio risk always greater than negative of 180
  random variables with 12, 89
expected utility 40, 42, 43, 44–51, 52, 53, 54, 55, 58, 59, 66, 72, 74, 78, 98
  alternative to 68
  finite 326
  paradoxes arising from 329
  preference relation characterized by 325
  rational prescription of 67, 307, 328
  stochastic dominance rules first introduced in relation to 306
  uniqueness of representation 70, 75
  see also CPT
expected values 98
  subjective 69
exponential distribution 78, 192
exponential smoothing algorithm 168

Fabozzi, F. J. 302n
failure rate function 78
fair value 44–5
fat-tailed behavior 4–5
financial economics 3, 7, 9, 17, 18, 19, 27, 28
Financial Services Authority (UK) 291, 302n
first derivatives 339
  non-negative 49
  non-positive 61
first difference pseudomoments 124
Föllmer, H. 184, 189n
Fourier transform 31, 249n, 283
fractional integrals 74
framing effect 67
frequency returns
  higher 4, 278, 296
  lower 291, 296
FSD (first-order stochastic dominance) 40, 44, 52–3, 54–5, 59, 62–3, 66, 70, 306, 307, 314, 319, 324, 325, 328, 330
  consistency with 181–2, 337, 338
  criteria developed for 3
  degree of violation of 329
  investors with balanced views sufficient to metrize 326–7
  Lévy quasi-semidistance and 315–17

log-return distributions and random payoffs 60
relationship between SSD, TSD and 74
functional analysis 23, 27
functionals 9, 68, 69, 87, 89, 92, 94, 119, 134, 139, 140, 142–3, 156, 185, 187–8, 317, 321, 325, 326, 336, 337, 351
AVaR 299–300
co-minimal 83, 86, 90, 97, 114
continuous 224
distortion 226–7, 249n
linear 299
minimal 114
monotonic 335
risk-quantifying 4
symmetric 320
uncertainty-quantifying 4–5
see also coherent risk measures; dispersion measures; ideal probability metrics; minimal norms; moment functions

gambling and betting puzzles 67
Gaussian distribution 254, 281, 286, 288
GCLT (generalized CLT theorem) 1, 5, 253, 278
rate of convergence for 112, 295, 297
Glivenko-Cantelli theorem 250n, 257, 260
Grabchak, M. 296, 302n

Hadar, J. 80n
Hanoch, G. 80n
Hausdorff metric structure 21, 31, 304, 305, 307, 310–25, 340–9
hazard rate function 78
heavy-tailed behavior 4, 168, 169, 172, 211, 222, 230, 252, 253, 258, 262, 271, 291–2, 297, 298
asymptotic distribution 277–83
rate of convergence 283–90
stable Paretian distributions 200
Hennequin, P. L. 13
Heukamp, F. 68, 69
Hewitt, E. 37
Heyde, C. 189n
Hill estimator 220
historical method 167, 208, 211
deficiency of 168–9
Holton, Glyn A. 188n
homogeneity property 106–7, 110
positive 156, 157, 158, 176–7, 178, 180, 183–8, 203, 243–4
Hwang, S. 68
hybrid method 168–9, 208–9

ideal probability metrics 105–6
conditions for boundedness of 112–14
interpretation and examples of 107–12
ideal semimetrics 85
identity property 20, 21, 25, 85, 101, 106, 309, 311
obvious 326
see also almost everywhere identity
i.i.d. (independent and identically distributed) observations 167–8, 202, 220, 260, 278–9, 291, 294, 295, 296, 299, 300
inconsistency *see* paradoxes
independence axiom 71

index of stability 276
see also tail exponent
index sets 117, 119
indicator-type events 174
infimum 103, 104, 313–14
infinite variance distributions 271–4, 278, 292, 297, 298
integers 112, 129, 204, 209, 239, 240
arbitrary positive 295
Internet 203
invariance property 177–9, 180, 187
see also translation invariance
investment opportunities 3
possible outcomes 52

joint distributions 25, 94, 100
space of 4, 7, 26, 28, 310
JP Morgan 160

Kahneman, Daniel 43, 67, 68, 69
Kalashnikov, V. 79, 80n
Kantorovich, L. V. 124
Kantorovich distance 121, 124
Kantorovich metric 8, 14, 27, 90, 99, 103, 105, 124–5
dual representation 15–16, 98, 121–2
finite 66
weighted 225
Kantorovich quasi-semidistance 329–30
Kaufman, R. 31
Kelly, John 354
Kemperman, J. H. B. 119
kernel methods 211–18, 242–5, 294
Kim, Y. 204, 302n
Klebanov, L. B. 295, 302n
Knight, F. H. 188n
Kolmogorov distance 10–11, 12–13, 65, 288–9, 346–7
Kolmogorov metric 15, 27, 64–5, 80n, 89, 90, 93, 97, 105, 108, 127, 265, 299, 341, 351–2
Kantorovich interpreted along the lines of 14
relationship between Lévy and 13
Kolmogorov test 175, 264, 270, 285
Kolmogorov-Rachev metric 111, 112, 113
Kolmogorov-Smirnov test 175
Kruglov distance 8, 23, 26
Kuratowski, K. 31
kurtosis 230
Ky Fan distance 132, 133, 135, 350–1
Ky Fan metrics 8, 16–17, 22, 27, 132–3, 135, 350–1

Lamantia, F. 167
lambda structure 340, 348–51
Lebesgue measure 30, 246, 322, 352
Leshno, M. 329
Levin, V. L. 119
Lévy, H. 80n, 307, 329, 354n
Lévy distance 347, 349, 350
Lévy metric 8, 27, 90, 93, 315–16, 317, 340–1, 342, 343, 346, 347, 349
relationship between Kolmogorov and 13
Lévy quasi-semidistance 305, 327
and first-order stochastic dominance 315–17
guaranteed boundedness of 318

Lévy stable distributions 193, 220
limit theorems 106, 298–301
  see also CLT
limiting distributions 279, 280, 281, 283, 285, 286, 288, 300
  stable non-Gaussian 282
Lipschitz conditions/functions 98, 121–2, 327, 338–9
Loeve, M. 36
loss-aversion effect 67, 68
loss distribution 149, 220, 248n
  conditional 198–9, 228–31
lotteries 42–8, 71, 74
L_p-metrics 17–18, 19, 22, 27, 89, 100–1, 102, 107, 117, 132, 135, 139
  between distribution functions 8, 15–16, 98, 108
Lukacs, E. 16, 18
Lusin, N. 30

MAD (mean absolute deviation) 17, 146, 153–4, 156, 158, 182
marginal distributions 3, 91, 102–3, 104
market crashes 262, 296
market risk 188n
  computing exposure to 160
  standard variables 149
Markov kernels 95
Markowitz, H. M. 159
mathematical expectation 151, 166, 180, 200, 205, 207, 215, 228
  conditional loss distribution 198–9
  infinite 199, 203, 245, 248n, 258
  probability laws 87
  random variables 48, 195–6
MATLAB (software package) 203
maximal distance/metrics 103–4, 136, 137
max-stable distribution 220
Mazukiewicz, S. 30
measure theory 30
minimal distances 93–4, 96, 118, 122, 123, 127, 131, 134, 137
  maximal and 103–4
  primary 89–90, 115, 135, 141–2
  relationship between co-minimal and 97
minimal metrics 96, 114, 135
minimal norms 83, 86, 90, 97–9, 114, 127–8, 130, 142
minimization formula 198, 205, 228
  geometric interpretation for AVaR 234–7
minimum performance deviation 313
Minkovski metric/norm 342, 346
Mittnik, S. 204, 302n
moment-based conditions 245–8
moment functions 83, 85
  compound distances and 99–105
  examples of 135–44
moments
  deviation between 117
  finite second 66, 153, 297
  implied coincidence of all characteristics 19
  integer 112
  lower partial 111
  marginal 104, 105, 139, 140
  see also absolute moments; tail moments

monotonic functions 247, 317, 334
  convex/concave 335
  non-increasing 313
monotonicity property 175–6, 179, 180, 181, 233
Monte Carlo method 169–70, 221–2
  computing AVaR through 5, 209–11, 239, 252–303
  true merits of 171–2
Morgenstern, Oskar *see* Von Neumann-Morgenstern
MTL (median tail loss) 192, 234
multivariate intervals 346
multivariate statistical models 169, 209, 254

Neumann, John von 37
  see also Von Neumann-Morgenstern
non-Gaussian distributions 290
  stable 282
non-parametric methods 167
non-satiable investors 43, 48–9, 52–3, 72, 182, 328
  risk-averse 50–1, 54, 55–6, 58, 59, 61–2, 63, 73, 111, 306, 325
normal distributions 12, 65, 152, 154, 155, 166, 201, 218, 248n, 253, 259, 283, 296, 297
  AVaR of 193, 200
  closed-form expressions for 193, 199
  density of 212, 214
  domain of attraction of 294, 295, 300
  multivariate 165, 169, 207, 212
  rate of convergence to 262–77
  symmetric around the mean 167
  thin-tailed 262, 291, 292
  see also standard normal distribution
normalization 127, 202, 230, 279–80, 282, 283, 285, 294, 295, 300–1
null hypothesis 264, 270
numerical methods 202–3, 254, 259

operational risk 149, 188n
optimization problems 205–6, 307
option pricing theory 297
Orlicz's condition 23
  see also Birnbaum-Orlicz
Ortobelli, S. 74, 319
outliers 189n, 200

paradoxes 328, 329, 354
  see also Allias; Ellsberg; St Petersburg
parametric bootstrap 171, 255–6
Pareto distribution 200, 247, 254, 283–6, 289
payoff distributions 60, 61, 62, 74, 150, 164, 172, 239
payoffs 48, 57, 70, 77, 80n, 165, 177
  mean 56
  random 46–7, 60, 62, 162, 164, 175, 176, 177, 182
  return versus 59–63, 68, 74–6
  target 58
  see also expected payoffs
Pflug, G. 189n, 227, 248n, 249n
Polish space 29, 30, 32
  nonempty closed subsets 125

portfolio returns 13, 17, 169, 178, 205, 242, 243
AVaR of 204, 207, 209
daily 167, 175, 219
distribution functions of 60
kernel approximation/estimator of 212, 213, 214
observed 167, 204, 208
described/interpreted as random variables 12, 64, 66, 106–7, 177, 183, 249n, 254
realized 218
standard deviation of 167
uncertainty of 158, 163
variance of 166
see also expected returns; return distributions
positive linear transform 72, 75, 80n
positivity axiom 157
power function 247
Pratt, John W. 80n
preference order 72, 111
preference relations 47, 72, 184, 193, 305
characterized by expected utility 325
defined on probability distributions of random variables 42
metrization of 308–10
quasi-semidistances and 304, 334–5
risk-averse investors 156
topology and 308, 332–4
preferences
characterizations of 42–4, 47, 49, 51, 54, 66, 111
natural 67
numerical representation of 41–2, 46, 47, 48
pre-limit theorems 4, 290, 294
central 295
primary distances 4, 86–90, 114–17
primary metrics 19, 91, 99–100, 116, 117
discrete 84
primary distances and 86–90
probabilistic models 172, 175, 291, 297, 298
acceptable 253
assumed 290
estimating the stochastic stability of 5
non-realistic/unrealistic 149, 253
probability distances/metrics 1–2, 7–39
asymmetric 5
axiomatic construction of 3, 5
classification of 4, 83–145, 339–53
definitions of 9, 19, 24–8
deviation measures and 184–7
direct application in finance 4
dispersion measures and 158–9
risk measures and 223–6
stochastic dominance and 63–6
see also TPM
probability distributions 46, 47, 48, 51, 70, 71, 74, 148, 192, 223, 253
assumed 163
c.d.f. of 72
possibility to uniquely define 249n
regarded as objective 42

probability laws 96, 97–8, 322
  characteristics of 88
  governing approximation 277, 289–90
  space of 87
probability quasi-distances 5, 28
probability quasi-metrics 187–8
probability quasi-semidistances 5, 305, 313, 326
  arbitrary 311
  bounded 327
  compound 324–5
  construction on classes of investors 335–8
  Hausdorff 307, 310, 312, 314, 320–2
  non-symmetric 322
  preference relations and 304, 334–5
  symmetric 28, 309, 320
  utility-type 307, 331
probability quasi-semimetrics 28, 309
probability semidistances 8, 25, 26, 28, 98, 103, 131, 320, 321, 353
  asymmetric 5
  axioms of 27, 93
  classification of 118–19, 340
  co-minimal functional induces 97
  compound 314
  definition of Hausdorff structure 340, 344–5, 348–9
  extension of the notion 306
  obtaining lower and upper bounds of 104
  primary 314
  simple 93, 314, 319
probability (semi-)metrics 9, 18, 22, 26, 27, 184, 185, 225
  compound 19, 85, 86
  ideal 85
  minimal 85, 89
  primary 85, 86, 114
  probability quasi-semimetrics satisfy all properties of 189n
  simple 85, 86, 95, 130, 351
probability space 27, 35, 37, 85, 88, 353
  non-atomic 28, 36, 38
  random variables defined on 7, 26, 28, 310
  "rich enough" 94
Prokhorov compactness criteria 95
Prokhorov metric/distance 124–5, 126, 134–5, 321, 343, 344
  Hausdorff structure of 345–6
  $\wedge$-representation of 350
prospect selection 5
pseudometric space 21
psychological effects 67

Rabin, M. 67
Rachev, S. T. 2, 79, 80n, 112, 114, 119, 138, 139, 141, 156, 204, 298, 299, 301n, 302n, 308, 311, 312, 321, 326, 340, 342, 345, 348, 349, 352, 353
Rachev metric 110–11, 112
  see also Kolmogorov-Rachev
Racheva-Iotova, B. 283
random quantities 176
  measuring distances between 2, 9, 64, 85, 86, 158, 159, 223
  metrized preference relations between 310

random variables 14, 25, 27, 36, 37, 38, 46, 73, 89, 91, 92, 98, 149, 151, 152, 157, 161–2, 181, 184, 186, 193, 207, 223–4, 244, 248, 257, 268, 270, 275, 277, 283, 313, 314, 322, 323, 324
assuming they describe random profits 78
AVaR of/viewed as 191, 255
bivariate 96–7
c.d.f. of 10, 65, 74–5, 238, 280, 295, 298–9
coherent properties depend on interpretation of 150
coincident 19, 90, 100, 158
convergence in distribution 113, 144n
corresponding absolute moments of 18
defined on probability space 7, 26, 28, 310
densities of 108–9
density functions of 212, 261
dependent 24, 102
described/interpreted as portfolio returns 12, 64, 66, 106–7, 110, 182, 183, 249n, 254
discrete 48, 239
dispersion measures of 146, 147, 156
distribution of 264, 265
distributional assumption for 222
elementary outcomes of 41, 48
finite mean 212, 233
functional which measures "closeness" between 106
i.i.d. 202, 220, 260, 278, 291, 294, 295, 296, 300
independent and identically distributed 253
indistinguishable 86
interpreted as random payoff 175
inverse c.d.f. of 226, 229, 237–8
lotteries interpreted as 42
mathematical expectation of 48, 195–6
mean absolute deviation of 146
n-dimensional 212
non-negative 180
one-dimensional 7, 9, 19–20, 254, 310, 353
positive 78, 79, 156
present instant 148
probability metrics defined on pairs of 24
real-valued 79, 195; *see also* r.v.s.
regarded as random payoff of common stock 60
return of common stock described by 178
scaled 106, 108–9
second lower partial moment of 58
stable distribution 200, 202, 203, 247, 263
symmetric 111, 155, 188n
tail behavior of 77, 80, 192, 230, 231, 234, 245, 294, 297
uncertainty of 146, 147, 154
unknown 153
U-valued 8, 28, 94, 344
varying the dependence structure between 93
zero-mean 101

random vectors 7, 24
bivariate 93
n-dimensional 117, 213
two-dimensional 89
rate of convergence 112, 295, 297
exact estimates of 106
heavy-tailed returns 283–90
normal distribution 262–77
rational behavior 40, 43, 66–7, 307, 328
real numbers 42, 47, 74, 75, 149, 249n
space of 21–2, 24
reliability theory 77, 80n
residuals 4, 297, 301
daily 291
return distributions 10, 12, 14, 59–64, 75, 80, 89, 92, 93, 96, 98, 111, 161, 162, 165, 175, 180, 207, 225, 260, 262
arbitrary 172
assumed 222, 290
AVaR of 182, 198, 228, 230, 237
benchmark 16, 224
choosing when markets are normal 290
daily 101, 170, 209, 219
differences between 15
heavy-tailed 252
impossible to derive in closed-form 255
mean 11
random variable describing 322
realistic hypothesis for 4
risk measure for 150, 245
risk profiles 228
statistical models for 182, 209
tail of 219–20, 222, 226, 252, 258–9, 297
see also frequency returns
return versus payoff 59–63, 68
stochastic dominance and 74–6
reward
computing the measure of 3
expected 41
risk and 159
risk-aversion 43, 64, 150, 181, 306
non-satiable 50–1, 54, 55–6, 58, 59, 61–2, 63, 73, 111, 306, 325
preference relations 156
see also absolute risk-aversion
risk-aversion function 221–3, 226, 228, 233, 245, 248
importance of proper selection 227
inverse of 247
risk-driving factors 172, 174
risk-free assets 158, 178
risk measures 4, 158, 159–79
arbitrary 307
axiomatic construction of 150
convex 183–4, 243
dispersion measures and 179–81
distortion 226–8, 249n
inappropriate choice for 149
infinite 222–3
probability metrics and 223–6
stochastic orders and 146, 181–2
see also coherent risk measures; spectral risk measures
RiskMetrics Group 160, 165, 167, 170
risky prospects 43–4, 69, 336
assets 48, 178
investment 175, 176
stock 158
robust estimator 192, 230–1

Rockafellar, R. T. 156–7, 189n, 248n, 249n, 250n
Roemisch, W. 189n, 227, 249n
RSD (Rothschild-Stiglitz stochastic dominance) 55–6, 66, 79, 80n, 305, 319–20
Rüschendorf, L. 112
Russel, W. R. 80n
r.v.s (space of real-valued random variables) 10, 16, 351
  dependent *U*-valued 138

Samorodnitsky, G. 200, 296, 302n
Satchell, S. 68
Savage, Leonard 42–3
scaling constants 218, 284–5, 286
Schied, A. 184, 189n
Schwartz, J. 23
second derivatives
  arbitrary 61
  negative 50, 57
  zero 57
semi-analytic expressions 203, 220, 228, 275, 284
semi-continuous functions 343
semidistance space 22, 308
semimetric space 9, 20
semi-standard deviation 156, 157, 158, 159
  downside 180, 182
  positive/negative 146, 154–5
  upside 182
Sereda, E. 249n
Siegel, J. 67
Simonetti, P. 221
Simonoff, J. S. 249n
simple distances/metrics 4, 19, 90–9, 118–31, 134–5, 341, 352
  see also Kantorovich; Kolmogorov
skewness 89, 113, 200, 201, 203, 230, 281, 290, 297
  negative 57, 58, 73, 155, 285, 286
  positive 57, 58, 73, 111, 155, 285, 286
Skorokhod metric 8, 21
smooth functions 213, 215, 240, 241, 294
smoothing 111, 112, 215–18
  exponential 168
  see also Kolmogorov-Rachev
smoothness assumptions 338
s.m.s. (separable metric space) 26, 27, 28, 30, 32, 33, 36, 37, 38, 87, 99, 102, 115, 123, 125, 127, 131, 132, 352
  arbitrary 25
  Polish 29
  see also u.m.s.m.s.
Souslin sets 30
specialization pre-order 304, 306, 333–4
spectral risk measures 5, 179, 182, 191, 193, 220–3, 245–8
SSD (second-order stochastic dominance) 40, 44, 53–5, 56, 58, 59, 62, 63–4, 70, 78–9, 181, 306, 307, 314, 327, 354n
  consistency with 80, 182, 183
  criteria developed for 3
  log-return distributions and random payoffs 60
  relationship between FSD, TSD and 74
  RSD and 319–20
St Petersburg Paradox 42, 44–6, 48
stability property 202

stable distributions 167, 247, 249n, 253, 274, 294, 295, 302n
AVaR for 200–4, 228
c.d.f. approximated 283
domains of attraction 220, 278, 296
Lévy 193, 220
limiting 281, 282
Paretian 200, 254, 283–6, 289
standardized symmetric 276, 292
tempered 204, 297
totally skewed 297
truncated 263, 275, 292
stable laws 202, 278, 296, 297
standard deviation 3, 101, 148, 151–3, 165, 166–7, 182, 188n, 193, 195, 199, 224, 248n, 253, 279
deficiencies as a risk measure 159
infinite 223
measured in dollars/percentage points 149
properly scaled 200, 207
proxy for risk 18, 159
tail 229, 230
see also semi-standard deviation
standard normal distribution 111, 167, 170, 171, 174, 199, 256, 258, 261, 264, 265, 268, 270, 271, 280, 302
c.d.f. of 213, 214, 242, 243
density of 214, 242–3
independent observations on 240, 241
Monte Carlo example for 221
VaR and AVaR of 257
state-preference approach 43
statistical models 239
multivariate 169, 209, 254
risk-aversion function and 223
status quo 41, 67
Steiner, A. K. 354
step functions 213, 241
Stiglitz, J. E. *see* RSD
stochastic dominance 41, 51–63, 68, 304–55
probability metrics and 63–6
return versus payoff and 74–6
stochastic dominance relations 72–4, 76–80
application of probability metrics theory in 70
classification of 5
stochastic dominance rules 5, 44, 306, 330
stochastic orders 78, 80, 325
AVaR-generated 322–4
classification of 310
compound 304, 314–15
generated by quasi-semidistance 326, 327
induced 314, 320, 323
metrizable 306, 314, 326, 327
modifying 79
more refined 71
nested in each other 307
primary 304, 314–15
risk measures and 146, 181–2
simple 304, 314–15, 319
see also FSD; RSD; SSD; TSD
Stoyanov, S. V. 159, 187, 202, 245, 249n, 283, 284, 298, 299, 301n, 302n, 311, 317, 319, 324, 325, 326–7, 329, 334–5, 337, 338, 339
Stromberg, K. 37

Student's *t* distribution 167, 193, 199, 247, 248n, 254, 268, 276, 283, 286–90, 298
  AVaR of 200, 202, 204
  truncated 263, 270
  widely applied as model for stock returns 262–3
sub-additivity axiom 128, 147, 157, 176, 177, 179, 180, 183, 187, 192, 243
subjectivity 42–3, 147
sufficient conditions 54, 66, 99, 128, 246, 247, 259, 260, 318, 337
  necessary and 333–4
supremum 90, 98, 103, 104, 294, 314
  essential 246
symmetrization transform 310, 311, 314, 315, 317, 319, 321
symmetry axiom/property 5, 20, 22, 25, 27–8, 106, 128, 130, 188n, 189n, 200, 309, 320, 321–2

tail behavior 5, 15, 78, 174, 195, 196–7, 204, 205, 207, 221, 228, 281
  asymptotic 284, 300, 301
  random variables 77, 80, 192, 230, 231, 234, 245, 294, 297
  return distributions 219–20, 222, 226, 252, 258–9, 297
  return frequency 290–5
  thickness effect 263–8
  truncation effect 262, 268–71, 275–6
  see also ETL; heavy-tailed behavior; MTL; tail exponent; tail moments; thin-tailed distribution
tail exponent 201, 203, 282, 300
tail moments 191, 192, 230
  higher-order 229
Taqqu, M. S. 200, 302n
Thaler, R. 67
thin-tailed distribution 262, 291, 292
time-series model 4–5, 290
tolerance level 313
topology 22, 29–30, 31, 33, 304, 306, 321, 327, 352–3
  preference relations and 308, 332–4
  quasi-metrizable 333
Tortrat, A. 13
total variation metrics 105, 108, 109–10, 111–12, 321
TPM (Theory of Probability Metrics) 20, 86, 99, 119, 135
  basic problem in 118
  stochastic dominance relations and 44, 70
  u.m.s.m.s. plays an important role in 29
tracking error 18, 96–7, 105, 224
  bounds for 101
transformations 115
  c.d.f. 69, 307
transitivity 71, 309
translation invariance 147, 157, 158, 177, 180, 184–8
triangle inequality 20, 22, 25–6, 106, 115, 309, 311, 326
  best possible improvement 102–3, 138, 140
  functional satisfying 305
  refinement for maximal metrics 137
Trindade, A. A. 301n

TSD (third-order stochastic dominance) 40, 44, 56–8, 62, 73
  relationship between FSD, SSD and 74
  relationship between RSD and 320
Tversky, Amos 43, 67, 68, 69

Uchaikin, V. V. 280
u.m. (universally measurable metric space) 8, 28, 29, 30
u.m.s.m.s. (universally measurable separable metric spaces) 8, 29–35, 38, 94, 95, 120, 122, 135, 136
uncertainty
  choice under 3, 5, 6, 40–82, 98, 306, 325
  risk and 4, 41, 44, 146–90
uniform metrics 12, 21, 105, 108, 352
  see also Kantorovic; Kolmogorov
unrealistic models
  dependence 96
  probabilistic 253
Uryasev, S. 248n, 249n, 250n
Urysohn's Metrization Theorem 22–3
utility functions 41, 48, 56, 58–60, 68, 72–7, 78, 80
  concave 49–50, 54, 55, 59, 325
  exponential 50–1, 61
  linear 50
  logarithmic 45, 50, 53, 57
  non-decreasing 52, 54, 59, 325, 327, 328, 329
  nonrealistic choices of 329
  power 51, 53
  quadratic 50
  quasi-semidistance based on 305
utility-type representations 304, 307–8, 325–7, 331, 332, 340

value functions 40–1, 43, 69, 98
  assumed radially asymmetric 68
  bounded S-shaped 335, 337, 338, 339
Van der Vaart, A. W. 249n
VaR (value-at-risk) 146, 147, 160–72
  advantages and disadvantages 5
  back-testing of 172–5
  conditional 3, 194
  numerical calculation of 204
  see also AVaR
variance 148, 151, 166
volatility 12, 13, 112, 314
  clustering 4, 168, 169, 172, 253, 290, 291
Von Neumann-Morgenstern theory 41–2, 46–8, 70
  basic result of 72
  lotteries in 74
  superior alternative to 43

weak convergence 21, 84, 94
  Lévy metric metrizes 11–12, 13
weighting functions 41, 43, 221, 226
  continuous 337, 338
  Lipschitz condition for 339
  non-decreasing 69
Whitmore, G. A. 80n

yield curves 2, 9, 20, 172

zeta structure 340, 351–3
Zolotarev, V. M. 94, 124, 200, 280, 283
Zolotarev ideal metric 66, 110, 111, 112
Zolotarev quasi-semidistance 319
Zolotarev semimetric 128–9, 351