Portfolio Optimization with Quantile-based Risk Measures

by

Gerardo José Lemus Rodriguez

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

March 1999

Author .....
Department of Electrical Engineering and Computer Science
March 11, 1999

Certified by .....
Roy E. Welsch
Professor of Statistics and Management Science
Thesis Supervisor

Certified by .....
Alexander Samarov
Principal Research Associate
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Portfolio Optimization with Quantile-based Risk Measures

by

Gerardo José Lemus Rodriguez

Submitted to the Department of Electrical Engineering and Computer Science
on March 11, 1999, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

Abstract

In this thesis we analyze Portfolio Optimization risk-reward theory, a generalization of the mean-variance theory, in the cases where the risk measures are quantile-based (such as the Value at Risk (VaR) and the shortfall). We show, using multicriteria theory arguments, that if the measure of risk is convex and the measure of reward concave with respect to the allocation vector, then the expected utility function is only a special case of the risk-reward framework.

We introduce the concept of pseudo-coherency of risk measures, and analyze the mathematics of the Static Portfolio Optimization when the risk and reward measures of a portfolio satisfy the concepts of homogeneity and pseudo-coherency. We also implement and analyze a sub-optimal dynamic strategy using the concept of consistency which we introduce here, and achieve a better mean-VaR than with a traditional static strategy.

We derive a formula to calculate the gradient of quantiles of linear combinations of random variables with respect to an allocation vector, and we propose the use of a non-parametric statistical technique (local polynomial regression - LPR) for the estimation of the gradient. This gradient has interesting financial applications where quantile-based risk measures like the VaR and the shortfall are used: it can be used to calculate a portfolio sensitivity or to numerically optimize a portfolio. In this analysis we compare our results with those produced by current methods.

Using our newly developed numerical techniques, we create a series of examples showing the properties of efficient portfolios for pseudo-coherent risk measures. Based on these examples, we point out the danger for an investor of selecting the wrong risk measure and we show the weaknesses of the Expected Utility Theory.

Thesis Supervisor: Roy E. Welsch
Title: Professor of Statistics and Management Science

Thesis Supervisor: Alexander Samarov
Title: Principal Research Associate

Acknowledgments

The idea behind this thesis was proposed to me by Alexander Samarov a nice Fall day of 1996, (September 10 to be exact!). From that day on, after countless discussions, lots of encouragement and help from Alex and Roy Welsch, multiple days in the library, the computer room, encountering dead ends, discovering breakthroughs, writing and editing in Boston, Meéxico, Paris, New York and London, this thesis is finally finished. Without Alex and Roy, this work would have never taken place; I wholeheartedly thank them.

I also would like to thank the readers, Dr. Amar Gupta, professors Sanjoy Mitter and John Tsitsiklis (who also taught me two graduate classes, and was at my Oral Qualifying Exam!), for their multiple comments and suggestions. Then, the spotlight turns to my friends, (new and old), but particularly to Ante, who started with me at MIT, and who is helping me to end this stage of my life, and to my host-mum, Kate Baty (who taught me all I know about ice-hockey). I thank my family (my parents, Enrique and Esperanza, and my brothers Enrique and Alexandra) for their encouragement and understanding.

And last, but not least, to Cati, who has been the main force motivating me to actually finish this thesis, and making sure I am ready to walk down both aisles in June.

London, March 7, 1999.[1]

[1]The research was partially funded with a Scholarship from the DGAPA (UNAM), the NSF grant DNS-9626348, the Sloan School of Management, the course 6 department, and personal funds.

Contents

List of Figures

List of Tables

List of Symbols

Vectors and matrices will be denoted with boldface letters (e.g. $\mathbf{V}$). Sets will be denoted with calligraphic letters (e.g. $\mathcal{V}$). Random variables will be denoted by a letter accentuated with a $\tilde{}$ (e.g. $\tilde{r}$).

$\mathcal{N}$	the set of all positive integers.
$\mathcal{N}_0$	the set of all nonnegative integers.
$\mathcal{Z}$	the set of all integers.
$\Re$	the real line.
$\Re^m$	the m-dimensional Euclidean space.
$\hat{a}$	estimator of the variable a.
$\mathbf{x}'$	the transpose of the vector $\mathbf{x} \in \Re^m$.
$\mathbf{x}'\mathbf{y}$	inner product in $\Re^m$.
$\mathbf{x} > 0$	all components of vector $\mathbf{x}$ are positive.
$\mathbf{x} \geq 0$	all components of vector $\mathbf{x}$ are nonnegative.
$\text{diag}(c_1, \cdots, c_m)$	the diagonal matrix with diagonal elements $(c_1, \cdots, c_m)$.
$\mathbf{e}_i$	the i-th unit vector (i-th column of the identity matrix).
$\mathbf{0}$	the zero vector of size i (all components are equal to zero)
$\mathbf{1}$	the one vector of size i (all components are equal to 1)
u_s	the unit step function.
$q_\alpha(\mathbf{x})$	the quantile function at the α percentile.

Financial terms:

$\tilde{b}_{j,i}, \tilde{b}_i$	simple gross return of asset i at the j-th period.
b_f	simple gross return of the risk-free asset at the j-th period.
c_f	risk-free constant for risk measures.
e_α	shortfall at level α.
$P_{j,i}, P_i$	price of the i-th financial asset at the j-th period.
$\tilde{r}_{j,i}, \tilde{r}_i$	simple return of asset i at the j-th period.
r_f	simple return of the risk-free asset.
$\rho(\mathbf{x})$	pseudo-coherent risk measure.
$S_{j,i}, S_i$	price of i-th stock at the j-th period.
VaR, VaR_α	Value-at-risk at level α.
W_t	wealth at end of period t.
W	wealth at end of the period (for the single period case).
$\mathbf{x}$	allocation vector (cash units).
$\mathcal{X}$	set of constraints for $\mathbf{x}$.
$\mathbf{y}_j, \mathbf{y}$	percentage allocation vector at the j-th period.
$\mathcal{Y}$	set of constraints for $\mathbf{Y}$.

Chapter 1

Introduction

Financial portfolio optimization is a mature field which grew out of the Markowitz's mean-variance theory, and the theory of expected utility. Both theories rely on the numerical representation of the preference relation investors have for assets with random outcomes. It is also assumed that investors are averse to the variability of random outcomes (or risk). Once a numerical representation of the investors' behavior is obtained, it is possible, in practice, to use different optimization methods to compute the optimal allocation of assets for a particular investor.

When Markowitz developed his original theory, he did not use the variance as the only measure of risk; he proposed the semivariance as one of the other measures. However, for both theoretical and computational reasons, the use of the variance is the most accepted since it allows, not only a very detailed theoretical analysis of the properties of optimal portfolios (such as the efficient frontier), but also the use of the quadratic optimization methods.

The Mean-variance theory has some limitations, when the random outcome of assets follows a non-normal distribution. Although in those cases the expected utility function could be used to optimize portfolios, practitioners have had the tendency to keep the concepts of "reward" and "risk" of a portfolio separated, assigning a numerical quantity to each concept. In particular, financial practitioners have developed new risk measures which are quantile-based, such as Value-At-Risk (VaR) and the shortfall.

This triggered our decision to analyze, for some of those new risk measures, the *risk-*

reward theory of portfolio optimization, which is in fact the generalization of the mean-variance theory. As with the mean-variance theory, there are efficient portfolios and efficient frontiers, but their characteristics depend on the definition of risk being used. We examine in detail the mathematics of efficient portfolios and efficient frontiers for risk measures which are homogeneous functions of portfolio weights (such as the VaR and the shortfall).

Once we have established a framework to compare random assets, we extended the static case to the dynamic one, by simply stipulating that the optimal dynamic trading strategy has the best risk-reward measures. Most of the analysis of dynamic strategies relies on the use of utility functions and their maximizations; in contrast, we analyze a simple example in which both risk and reward are optimized.

1.1 Contributions of the thesis

- We introduce the analysis of the risk-reward theory from the multicriteria optimization theory, claiming that the risk-reward theory applies to a broader set of case than the expected utility theory.

- We analyze the properties of optimal portfolios for pseudo-coherent risk measures (i.e., risk measures which are homogeneous functions of portfolio weights and have a risk-free condition); in particular we analyze the efficient frontier for cases when a risk-free asset is present, and when shortsales are allowed.

- We derive a formula for the gradient of a quantile with respect to the linear weights of random assets.

- We propose the use of the local polynomial regression (lpr) for the estimation of the gradient of a quantile, and illustrate this technique on applications which compute the gradient of quantile-based measures of risk.

- We implement and compare gradient and non-gradient based optimization methods for quantile-based risk measures.

- We implement and analyze a simple example of dynamic optimization using the risk-reward theory.

1.2 Outline.

The thesis is composed of 7 chapters, including this introduction, and three appendices. In Chapter 2 the notation for the single period asset allocation case is introduced, and some assets that challenge the classical mean-variance theory are reviewed. The need for a more general asset allocation theory is highlighted by those assets.

In Chapter 3 we review definitions of classic and modern risk-reward measures of financial portfolios (such as coherent and pseudo-coherent risk measures, standard deviation, VaR, shortfall). Using the modern risk measures, we generalize the classic mean-variance theory, and call it the *risk-reward* theory. The use of different risk measures solves the problem of non-normality of assets from Chapter 2. However, since the expected utility theory is entrenched in the field of Economics, we propose an alternative method to study the relationship between the expected utility theory and the risk-reward methodology, based on the multicriteria optimization theory. Once the risk-reward theory is established, we use it to analyze the properties of optimal portfolios for the special case where the risk measures are pseudo-coherent. In Chapter 3 we also include several examples of optimal portfolios and efficient frontiers.

The risk gradient is fully explored in Chapter 4, since it is a very useful analysis tool for trades. We derive a new formula for the gradient of a quantile of linear combinations of random variables; this formula has direct application to the gradient of quantile-based risk measures. In practice, the gradients have to be efficiently estimated, and we review and propose an estimation method for the gradient formula which uses local linear regression.

While in Chapter 3 we analyze the theory behind the risk-reward portfolio optimization theory, in Chapter 5 we overview the optimization of functions involving quantiles, which can be directly used for the portfolio optimization using quantile-based risk measures. While non-gradient based methods for the optimization of portfolios are already available (and we review some of them), we propose the use of a gradient-based nonlinear method for the

optimization of general quantile-based risk measures (including the shortfall); we do so using the estimation techniques developed in Chapter 4.

Chapter 6 goes beyond all the previous chapters which are dedicated to the single period case, and studies the dynamic case of portfolio optimization, introducing the notation commonly used to describe multiperiod asset allocation. That chapter also reviews some of the previous attempts to solve the dynamic case, such as expected utility maximization in the dynamic case, continuous-time analysis, and the dynamic option replication. We set the theoretical foundations of dynamic asset allocation in the risk-reward framework; we analyze and implement a simple example which optimizes the VaR and the expected return in the dynamic case.

Finally, Chapter 7 contains our conclusions and suggestions of future research to be done in this field.

The content of the appendices is the following:

Appendix A. Here we include the mathematical notation used in the thesis, definitions of quantiles (important to define quantile based risk measures), as well as a brief overview of the nonlinear optimization method used in the portfolio optimization algorithm. A section on local linear regression is included for completeness.

Appendix B. We review the preference relation of the financial assets theory, the multicriteria optimization theory in the risk-reward framework analysis, and different theories allowing the ranking of assets with random outcomes (the mean-variance, the utility theory, and the stochastic dominance).

Appendix C. The characteristics of the data used for several examples and experiments in this thesis are detailed here.

Chapter 2

Finance background

In this chapter we briefly review all the finance nomenclature and definitions required for the static portfolio optimization, following closely classic books such as [35, 32] and [14].

In section 2.1 the basic static asset allocation problem is posed, and the notation is defined. The characteristics of the allocation problem depend upon the underlying assets available, and section 2.2 reviews the probability distributions associated with financial instruments, in particular with non-normal distributions (section 2.2.1 and 2.2.2). The financial literature that contradicts the normality assumption for random outcomes of financial portfolios is a strong argument against the use of the mean-variance framework, and is the reason why new measures of risk have been introduced.

2.1 Asset allocation in the static case

The objective of the static portfolio optimization theory (also known as the single period optimization) consists in the selection of an optimal allocation of an investor's wealth in different investment alternatives, such that the investor obtains the "best" possible outcome at the end of one investment period. In general, techniques heavily depend upon the preferences of each individual investor.

The basic asset allocation problem for a single period can be defined as follows: Let us assume there are two trading periods during which the investor is allowed to perform

transactions; the initial trading period 0 and the final trading period T. Let W_0 be the initial amount of wealth available to invest across m random assets, and if it is available, one risk-free asset. Each one of the assets has an initial price $P_{0,i}$ (for the asset i at period 0), and a final price $\tilde{P}_{T,i}$ (for the same asset i at the end of period T). The prices $\tilde{P}_{T,i}$ are non-negative random variables whose values become known to the investors at period T. The risk-free asset will have an initial price P_f and a certain final price $b_f P_f$, where b_f will be a constant known as the *simple gross risk-free return*; while the constant $r_f = b_f - 1$ will be known as the *simple risk-free return.* The random vector $\tilde{\mathbf{b}} = [\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_m]'$ is composed of the *simple gross returns* $\tilde{b}_i = \tilde{P}_{T,i}/P_{0,i}$ and has a multivariate joint distribution F. The *simple return* $\tilde{r}_i$ is defined as $\tilde{b}_i - 1$, and the simple return vector is $\tilde{\mathbf{r}} = [\tilde{r}_1, \tilde{r}_2, \ldots, \tilde{r}_m]'$. The analysis is almost identical; in the static case simple returns are usually used, while in the dynamic case it is easier to analyze final payoffs, by using gross returns. The possible values the random variables $\tilde{P}_{T,i}$, $\tilde{\mathbf{b}}$ and $\tilde{\mathbf{r}}$ may have, are denoted as $P_{T,i}$, $\mathbf{b}$ and $\mathbf{r}$ respectively; and are known at the end of the trading period. An investor is assumed to be *non-satiable*, i.e., to always prefer more money than less; in the expected utility case this implies monotonically increasing utility functions.

Investments can be characterized by an $m \times 1$ vector $\mathbf{x}$ of commitments to the various random assets; x_i is the commitment to asset i and is proportional to the amount invested in the i^{th} asset; $\mathbf{x}$ is also be known as the *allocation vector* or *decision vector*, or vector of portfolio weights. The $m \times 1$ vector $\mathbf{y}$ is the *percentage allocation vector* where each y_i represents the percentage of the initial wealth W_0 invested in the i-th asset; the percentage vector is related to the vector of commitments as $\mathbf{x} = W_0\mathbf{y}$. If there is no risk-free asset available, the vector $\mathbf{x}$ can be constrained to be an element of the set $\mathcal{X} = \{\mathbf{x}|\mathbf{x}'\mathbf{1} = W_0\}$, also known as the *budget constraint.* Other constraints can be added, such as the *no shortselling restriction* $\mathbf{x} > \mathbf{0}$; if $x_i < 0$, then the asset i has been *sold short*; similar constraints can be set for the percentage vector. In the case when there is a risk free asset available, the budget constraint can be enforced implicitly by investing the allocation vector $\mathbf{x}$ in the m risky assets, and $W_0 - \mathbf{x}'\mathbf{1}$ in the risk-free asset.

When no risk-free asset is available, the final wealth $\tilde{W}(\mathbf{x})$ as a function of the decision

vector is

$$\tilde{W}(\mathbf{x}) = \mathbf{x}'\tilde{\mathbf{b}} = \mathbf{x}'(\mathbf{1} + \tilde{\mathbf{r}}) = W_0\mathbf{y}'\tilde{\mathbf{b}}, \tag{2.1}$$

assuming the budget constraint previously mentioned is enforced. The quantity $\mathbf{x}'\tilde{\mathbf{r}} = \tilde{W}(\mathbf{x}) - W_0$ is known as the *net worth* [6]. If a risk-free asset is available, then the final wealth can be expressed as

$$\tilde{W}(\mathbf{x}) = \mathbf{x}'\tilde{\mathbf{b}} + (W_0 - \mathbf{x}'\mathbf{1})b_f = W_0(\mathbf{x}'\tilde{\mathbf{b}} + b_f - \mathbf{x}'\mathbf{1}b_f), \tag{2.2}$$

where the net worth is now defined as $\mathbf{x}'\tilde{\mathbf{r}} + (W_0 - \mathbf{x}'\mathbf{1})r_f$. At period 0, the final wealth $\tilde{W}$ defined in equation (2.2) is a function of the random variable $\tilde{\mathbf{b}}$, and has a set of possible outcomes; W denotes a possible value that $\tilde{W}$ can take, which is known at the end of the trading period. The financial assets are assumed to give no dividends. Asset prices are always assumed positive. This is assured if we also assume *limited liability*; i.e., an asset has limited liability if there is no possibility that it require any additional payments after its purchase. An *arbitrage portfolio* $\mathbf{x}_a$ is defined as a decision vector summing to zero; $\mathbf{x}_a'\mathbf{1} = 0$. An *arbitrage opportunity* arises if there is an arbitrage portfolio $\mathbf{x}_a$ such that $\mathbf{x}_a'\mathbf{b} \geq 0$ for all possible realizations $\mathbf{b}$ of $\tilde{\mathbf{b}}$, and $E[\mathbf{x}_a'\tilde{\mathbf{b}}] > 0$. An arbitrage opportunity is a riskless way of making money; if such a situation were to exist, the underlying economic model would not be in equilibrium [50].

It is useful at this point to also define the *compound return* $\tilde{rc}_{T,i} = \ln(\tilde{P}_{T,i}/P_{0,i})$. Sometimes it is assumed that the compound returns $\tilde{rc}_T$ follow a multivariate normal distribution, and the prices follow a Geometric Brownian Motion. In that case, simple returns follow a log-normal distribution. Also, for very small trading periods the approximation

$$\ln(P_{T,i}/P_{0,i}) \simeq (P_{T,i} - P_{0,i})/P_{0,i}$$

can be made, which means that in some cases simple returns can be approximated as random variables with a normal multivariate joint distribution; it must be remembered that normal returns contradict the limited liability assumption. Also, the natural logarithm function cannot be applied to the case where the final price of an asset is 0 (a common case for some

financial assets such as options); hence, we stick to price ratios or simple returns in our analysis.

Optimal Asset Allocation

Once the preference relation of an investor is established, it is possible in some cases to determine either the optimal asset allocation that will satisfy an investor, or at least an efficient portfolio. All of Chapter 5 is be devoted to different portfolio optimization techniques.

The goal of an optimal asset allocation is to select the optimal vector $\mathbf{x}^*$ that gives the "best" final wealth $\tilde{W}$ with distribution function $F_{\tilde{W}}(\cdot)$. Approaches to solving this problem depend on definitions of preference relations which allow us to rank the possible final wealths; in appendix B.1 we review some common representations of preference relations.

2.2 Probability distributions of financial returns

In practice, the total amount of financial assets an investor can select is extraordinarily large. For that reason, the m assets usually selected are only a small subset of the available universe of financial assets. The returns of the m assets selected will be assumed to have a joint multivariate distribution F. Different assets selected will have different joint distributions. We will assume that there are no arbitrage opportunities with the m assets selected; i.e., there is no x_a such that $\mathbf{x}_a'\mathbf{b} \geq 0$ for all possible outcomes $\mathbf{b}$ of the random variable $\tilde{\mathbf{b}}$ at the end of the trading period, and $E[\mathbf{x}_a'\tilde{\mathbf{b}}] > 0$. If the multivariate distribution F would allow arbitrage opportunities, one could increase wealth without making an initial investment, which goes against current economic theories. To rule out arbitrage opportunities, the concept of *risk neutral* probability measures is used (see [50]). A very important result establishes that there are no arbitrage opportunities, if and only if, a risk neutral probability measure Q on Ω exists (a finite sample space with $K < \infty$ elements, each element being a possible state of the world). This result should be remembered to select synthetic distributions and examples to test portfolio optimization methods, as well as risk measures.

Because wealth $\tilde{W}$ is the result of a linear combination of the m random variables (the random vector $\tilde{\mathbf{b}}$), the unidimensional distribution function $F_{\tilde{W}}$ will depend on both the

linear weights of the portfolio ($\mathbf{x}$), and on the distribution function of the m assets (F). The class of distribution functions for the final wealth that can be generated from a linear combination of assets with random returns is:

$$\mathcal{F} = \{F_{\tilde{W}} | \tilde{W} = \mathbf{x}'\tilde{\mathbf{b}}, \mathbf{x} \in \mathcal{X}, \tilde{\mathbf{b}} \sim F\}. \tag{2.3}$$

The characteristics of the set will therefore change depending on the kind of financial assets being used. For example, if the returns of the financial assets follow a Gaussian multivariate distribution, the return of the wealth will also be Gaussian.

2.2.1 Non-normality of financial assets

The normality assumption for the continuously compounded return is widely used to model the dynamics of common stock prices (as described in appendix C.1). When the inter-trading period observed is small, the normality assumption is a good approximation even if the simple return is log-normally distributed.

Previous research has shows that U.S common stock returns are distributed with more returns in the extreme tails [2, 24]. The distribution of Japanese security returns and other assets such as precious metals also exhibit significant kurtosis [1, 4]. It has been pointed out [56, 57] that investors' preferences for higher moments are important for portfolio selection, and that skewness and kurtosis cannot be diversified by increasing the size of portfolio [5]. Research exploring the deviations from the normality assumption abound, such as [3, 16, 23]. The classical linear market model consistent with the Capital Asset Pricing Model (CAPM) is:

$$\tilde{r}_j = \alpha_j + \beta_j \tilde{r}_k + \epsilon_j \qquad j = 1, \cdots, m$$

where the random variable $\tilde{r}_j$ represents the return of the j^{th} asset, $\tilde{r}_k$ is the market return, $\beta_j = \text{Cov}(\tilde{r}_j, \tilde{r}_k)/\text{var}(\tilde{r}_k)$, is the systematic variance of asset j, and ϵ_j is a zero mean random error. The CAPM holds if the market is efficient, stable, and if all investors have concave utility functions (such as quadratic utility functions). Some researchers [26] have proposed

a higher moment market model, such as the cubic market model:

$$\tilde{r}_j = \alpha_j + \beta_j \tilde{r}_k - \gamma_j \tilde{r}_k^2 + \delta_j \tilde{r}_k^3 + \epsilon_j \qquad j = 1, \cdots, m \tag{2.4}$$

where $\gamma_j = \text{Cov}(\tilde{r}_j, \tilde{r}_k^2)/E[(\tilde{r}_k - E[\tilde{r}_k])^3]$ is the systematic skewness of asset j, and $\delta_j = \text{Cov}(\tilde{r}_j, \tilde{r}_k^3)/E[(\tilde{r}_k - E[\tilde{r}_k])^4]$ is the systematic kurtosis of asset j. The higher moments introduce nonlinearities in the dependence of the individual return of asset j with respect to the market return $\tilde{r}_k$. This result indicates that the relation between the return of a stock relative to the market will not be linear as the classical $CAPM$ model suggests, but that the sensitivity of the return of a single stock depends on the level of the market return.

Assuming that the market is in equilibrium and that all investors have concave utility functions, [35], it can be shown that the market portfolio is an efficient portfolio, and this is how the CAPM theory is derived. However, in a general risk-reward framework (which will be introduced in the following chapter) we will not be able to assume that utility functions are concave. Because some quantile-based measures of risk (such as VaR) are non-convex, the uniqueness of the optimal solution will depend on the the joint distribution of returns of the underlying assets. The only general restriction for the distribution of returns is the no-arbitrage condition; hence, when investors behave in a risk-reward framework, the CAPM formula will be a special case of the no-arbitrage theory.

2.2.2 Non-normality introduced by options

For some financial assets, such as common stock, the normality assumption is considered as a very good approximation [58]. However, financial assets such as options (see appendix C.2 for the definition) can introduce nonlinearities and asymmetries to the portfolios [12, 45, 33]), (see appendix C.3 for a brief description of some strategies). The use of options in portfolios was precisely what led practitioners [38] to define new measures of risk able to determine the exposure to downside losses.

Example 2.2.1 In figure 2-1 we can notice the significant asymmetry of the distributions of portfolios which include options. The data was generated using 10000 samples, for three

different option-based strategies ("write-call" and "long-put" at 50 %, as well as a "write-call" at 100 %; in appendix C.4.2 we explain the exact procedure used to generate the returns). In future chapters we will refer to these data returns as the "Option-based strategies data". This data was generated to emphasize the non-normality of some financial assets. The Long Put (L.P.) strategy at 50 % is an asymmetric distribution with heavy left tail (greater downside risk). The Write Call (W.C.). strategy at 100 % has a very heavy left tail, as can be seen in the histogram and the empirical cumulative distribution function. The W.C. strategy at 50 % is also asymmetric *and* multimodal. Because of the non-normality, symmetric measures of risk as the standard deviation cannot be applied; they do not distinguish between heavy left tails and heavy right tails.

In figure 2-2 we can see the distribution function of two options (one put and one call, the "Put-Call" data, explained in appendix C.4.4), and its underlying asset (a common stock with Gaussian continuous returns). These data are used to show some of the weaknesses of the VaR risk measure. Both the Put and the Call have a very asymmetric distribution function. The options are evaluated using the Black and Scholes equation which assumes there are no-arbitrage opportunities. The joint distribution function F linking the put, call and underlying asset cannot be modeled using normality assumptions.

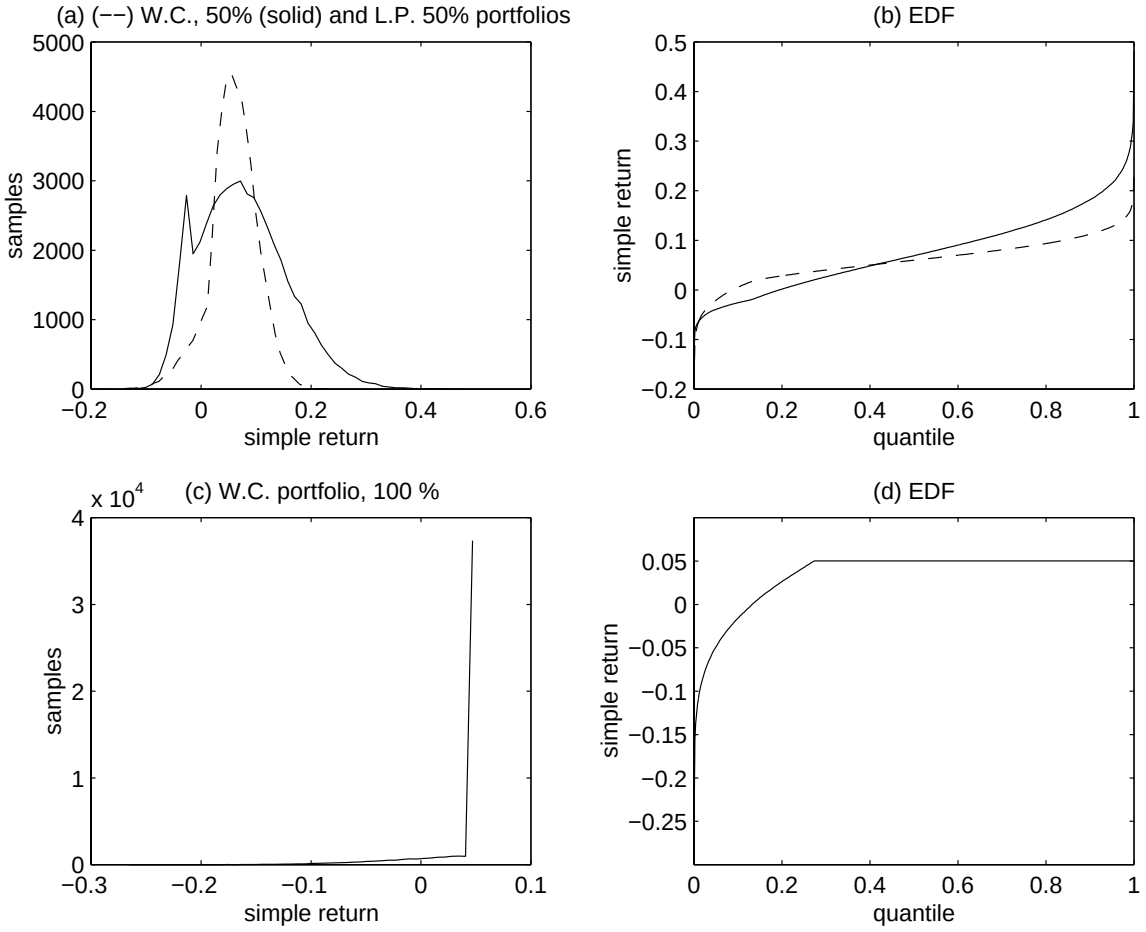

Figure 2-1: Examples of non-normal (asymmetric) distributions. Option-based strategies data.

EDF: empirical cumulative distribution function.
W.C.: "write-call" strategy. L.P.: "long-put" strategy.

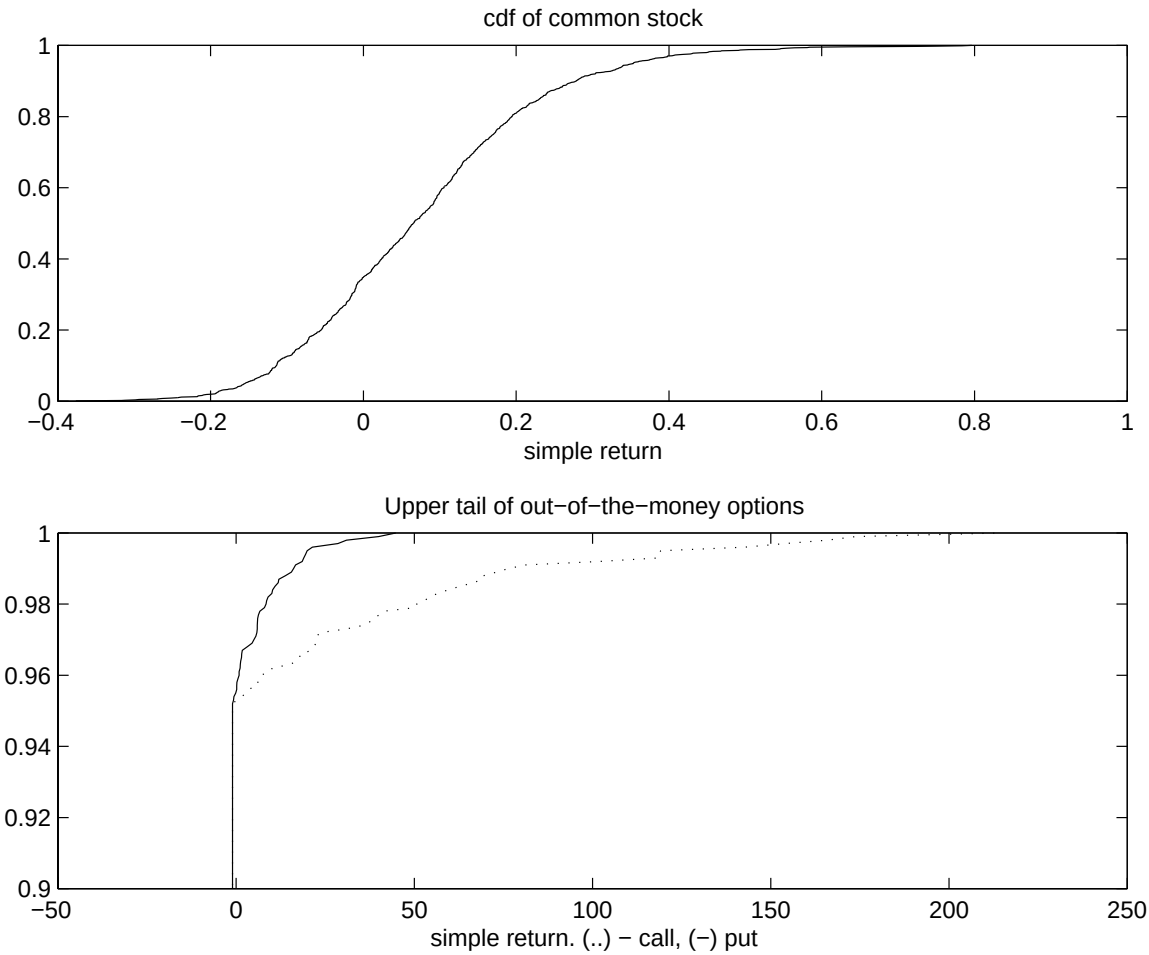

Figure 2-2: Far out of the money options. Put-Call data.

cdf: cumulative distribution function.

The top figure is the cdf of a normal return. In the bottom figure we see examples of non-normal distributions of the financial options returns. The options only have "right" tails, as can be seen from the cumulative distribution functions.

Chapter 3

Risk-reward framework

The portfolio optimization problem is a very subjective matter; it depends greatly on the ordering of the probability distributions of the returns of the assets considered. The basic problem of comparing different financial assets with random returns has been very widely researched using the expected utility theory, the stochastic dominance, and the multicriteria methodology (reviewed in appendix B). So far, the multicriteria methodology has been limited to the mean-variance case; in this chapter we extend it to the risk-reward cases. In section 3.1 we briefly compare the different preference relations defined in the financial literature.

While the expected utility theory assigns a scalar number to the random wealth, the risk-reward methodology assigns a performance vector of size 2 to each random wealth. For that reason, we use concepts available in the multicriteria optimization theory to analyze the generalized risk-reward framework; this analysis yields similar results as a previous research which links the risk-reward and the expected utility theories, but can be generalized to include almost any kind of risk and reward measures. The preference relations obtained using a risk-reward framework will be introduced in section 3.2. The reward and risk definitions for a financial asset are presented in sections 3.2.1 and 3.2.2; the Value-at-Risk (VaR), a very important measure recently defined is reviewed in detail in section 3.2.2.

The relationship of the risk-reward framework and the expected utility theory is analyzed using a multicriteria point of view in section 3.3. In section 3.3.1 we introduce a new

interpretation of the quadratic and the semivariance utility functions, from a risk-reward point of view which uses the concept of the value function. The same concept is used to give a new interpretation of the relationship between the quantile-based measures of risk (such as the VaR and the shortfall) and the utility theory (section 3.3.2). We intend to show that the utility theory can be treated as a special case of the risk-reward framework.

The risk-reward framework allows us to compare different assets or combinations of assets so that we can select the one our preference relation considers the "best". When there are m different financial assets available, we will be able to alter the "performance" (the combination of the risk-reward measures) of an investor's portfolio, by changing the linear weights. The advantage of the risk-reward framework is that classic methods such as the mean-variance and the utility theory optimization are special cases of the general risk-reward framework; in section 3.4 the basic optimization cases are posed from the risk-reward perspective; in section 3.4.1 we pose and solve the special case a risk-free asset is available, and the risk measure is pseudo-coherent (see section 3.2.2). For the latter, we derive an elegant solution (section 3.5). The case where no risk-free asset is available is analyzed in section 3.4.2, and the effect of noise in the optimization is reviewed in section 3.6.

The basic optimization problem, in the risk-reward framework, serves as basis for the computation of the efficient frontier (section 3.5); the pseudo-coherent risk offers again some elegant solutions.

3.1 Preference relations for risk averse investors

We still have not defined any particular function $\mathbf{h}(\tilde{W})$, or a value function $v(\tilde{W})$; this is a task more suited to Economists. Most agree that investors can be characterized by their *nonsatiability*, (i.e., investors always prefer more money to less, see section 2.1), and their risk behavior; investors can be risk-averse, risk-neutral or even risk-seekers [32, 35]. Risk is not universally defined, and each investor may approach decision-making under uncertainty with different risk definitions. However, certain restrictions on the available set of financial instruments used lead to the same behavior no matter which risk definition we use; for example, if assets follow a multivariate normal distribution, the mean-variance methodology

suffices to describe the investors' behavior.

In particular, some methods used to establish a preference relation are:

The expected utility theory: By far the most accepted, it assumes the investor's preference relation is complete, so that it can be expressed via a scalar value function $v(\tilde{W}) = E\{U(\tilde{W})\}$ where $U(\cdot)$ is the utility function. The concept of expected utility dominance can also be defined for a broad class of utility functions. This is equivalent to establishing an infinite vector $\mathbf{p}$ in which each element is an expected utility member of the broad class. The risk and return behaviors are implicitly encoded depending on the utility function used.

The stochastic Dominance: Using the concept of efficiency in an infinite performance space, finds preference relations between classes of utility functions. Because it uses the complete cumulative distribution function (CDF) of random variables to define preferences, it can also be seen as a generalization of the risk-reward framework.

The risk-reward: Given certain explicit definitions of risk and reward behavior, the two dimensional efficient frontier is computed so that each investor can chose her own E-point. The mean-variance is the most famous approach here, although the mean-semivariance and the safety-first techniques also belong to this category. If the risk is convex, and the reward concave over the decision vector $\mathbf{x}$, there is a value function representation for the performance space.

The value function: It assumes scalar function $v(\tilde{W})$ exist. Some research has been done using value functions that represent the risk-reward approach with additive models. However, some multiplicative models also exist.

Moments of distribution: The several moments of a distribution (such as skewness and kurtosis) to give a preference relation of some financial assets has been proposed in [35, 26]. In that case, the performance space will have a dimension greater than 2.

In some cases, particularly when the financial assets have a joint multivariate elliptical distribution, theoretical links between the different approaches have been established.

3.2 Risk-reward criteria

Because of the arbitrary nature of utility functions, there have been attempts to depart from the utility framework altogether and to use criteria based on more objective concepts. The risk-reward criteria represent the preference relation of an investor using the Pareto preference. A two dimensional vector

$$\mathbf{h}(\tilde{W})' = [reward(\tilde{W}), -risk(\tilde{W})]$$

(or $\mathbf{h}(\mathbf{x})$), called *performance vectors* and composed of reward and risk measures of the random return $\tilde{W}$, can be used to compare and rank random returns, and an efficient frontier can be computed using theorems B.1.1 to B.1.3. The negative sign assigned to the risk value is due to the fact that most investors want to minimize risk. The use of a performance vector of size two, in which each component is specifically designed to measure the risk-reward performance of a portfolio, was first proposed by Markowitz [44]. The use of a risk-reward performance vector was proposed by Encarnación, [39], who uses a lexicographic rule to rank the returns. Other risk-reward frameworks [21] were introduced, such as the multiplicative risk-reward models,

$$v(\tilde{W}) = risk(\tilde{W})reward(\tilde{W}),$$

or even some more general forms,

$$v(\tilde{W}) = g(risk(\tilde{W}), reward(\tilde{W})),$$

where $v(\cdot)$ and $g(\cdot, \cdot)$ are scalar functions.

We will focus on the Pareto optimal dimensional parameter space. As usual, the most difficult task is to select the adequate risk and reward measures of a portfolio, to approximate

the investors' behavior accurately.

Most of the risk-reward research was done analyzing returns instead of absolute wealth, but assuming simple returns are used, then the linear relation $\tilde{W} = W_0(1 + \tilde{r}_T)$ applies.

3.2.1 Reward measures

We loosely define *reward* as a function of the desirability of a financial asset described by a random variable ($\tilde{A}$, for example); $reward(\tilde{A}) > reward(\tilde{B})$ implies that $\tilde{A}$ is preferred to $\tilde{B}$ if the investor is indifferent to risk.

Some common measures of reward include the *mean*, or expected wealth, $E\tilde{[}W]$. However, using our definition, an expected utility function can be used as a reward measure, e.g., the linear utility function which computes the mean, or the log-normal utility function. We can also use quantiles (as the median, or others) to measure the reward of a financial asset, which would be a non-parametric function of return. This is a new concept, and can be related to the notion of stochastic dominance. Quantiles, (including the median) are homogeneous measures of reward, which theoretically may offer some advantages. However, in practice, the use of the median increases the unreliability of optimization algorithms.

3.2.2 Risk measures

A risk function will be a scalar function $risk(\tilde{W})$ associated with the random outcome of a financial asset ($\tilde{W}$). Risk will be assumed to be an undesirable characteristic of the random outcome $\tilde{W}$, related to the possibility of losing wealth. The characteristics of a risk function have been proposed in [6], defining the coherent risk measures. We can assume that we have two different wealth random outcomes, $\tilde{A}$ and $\tilde{B}$, which are the random outcomes of two portfolios, $\mathbf{x}_A$ and $\mathbf{x}_B$, such that $A = \mathbf{x}'_A\mathbf{1}$ and $B = \mathbf{x}'_B\mathbf{1}$. A *coherent* risk measure has the following properties,

i. Sub-additivity: $risk(\tilde{A} + \tilde{B}) \leq risk(\tilde{A}) + risk(\tilde{B})$.

ii. Homogeneity: $risk(\gamma\tilde{A}) = \gamma risk(\tilde{A})$, for any $\gamma > 0$.

iii. Risk-free condition: $risk(\tilde{A}+\gamma b_f) = risk(\tilde{A})+\gamma c_f$, for any real γ; c_f will be a risk-free constant that will depend on the definition of $risk(\cdot)$.

Coherency can also be written in terms of the portfolios $\mathbf{x}_A$ and $\mathbf{x}_B$, if desired.

The sub-additivity and the homogeneity imply the convexity of the coherent risk measure. A risk measure is *pseudo-coherent* (denoted by ρ) if it has the homogeneity and the risk-free condition properties, but not the sub-additivity property; the VaR is one example of a pseudo-coherent risk measure. Pseudo-coherency could also be a characteristic of reward measures; pseudo-coherent measures of reward are denoted as ϱ (e.g., the mean and the median of the wealth will be pseudo-coherent).

In appendix B.3.1 a risk averse individual is defined within the expected utility theory. In the risk-reward framework, we will define *risk aversion* as follows:

Given two assets with random payoffs $\tilde{A}$ and $\tilde{B}$, where $reward(\tilde{A}) = reward(\tilde{B})$, a *risk averse* person will select $\tilde{A}$ if $risk(\tilde{A}) \leq risk(\tilde{B})$. We will prefer definitions of risk which are also compatible with the expected utility theory. Some well known risk measures are the following:

The standard deviation

One of the oldest risk functions, it assumes the risk is proportional to the standard deviation of a random variable $\tilde{W}$:

$$risk(\tilde{W}) = \sigma_{\tilde{W}} = \sqrt{E\left\{(\tilde{W} - E\{\tilde{W}\})^2\right\}} \tag{3.1}$$

The standard deviation is a coherent risk measure with c_f constant equal to 0.

Lower partial moments

Other attempts to define risk include Harlow's research [30]. In his work he introduces *lower partial moments* (LPMs) in order to use only the left-hand tail of the return distribution. He defines an LPM for the probability distribution of a portfolio outcome $\tilde{W}(\mathbf{x})$ with a *target rate* τ as:

$$risk(\tilde{W}) = LPM_n = E[(\tau - \tilde{W})^n u_s(\tau - \tilde{W})], \tag{3.2}$$

(where u_s is the unit step function, definition A.1.2). Harlow uses $n = 2$ and $\tau = 0$ (also known as *semivariance*); he also recommends the use of $n = 1$, but is opposed to the use of $n = 0$ since it does not measure the dispersion of a loss once it falls below the target rate. The case when $n = 0$ is a type of the safety-first criteria.

In the case where the assets follow a normal joint distribution, the optimization result should give the same reward as the mean-variance optimal portfolio. However when financial assets do not follow a joint normal distribution (and have asymmetric distributions) asymmetric functions of risk like LPMs yield different optimal portfolios, which give a better protection against the risk defined by semivariance. When $n = 1$ or 2, the function $(\tau - W)^n u_s(\tau - W)$ is convex, and assuming that its expectation is finite, the risk measure LPM will be a convex risk measure (although in general it is not homogeneous, and has no risk-free condition property). The definition of LPM can be modified such that homogeneity is obtained.

The Value at risk (VaR)

Among all the possible definitions of disaster, one of the most often used by practitioners is the so called Value-at-Risk [38, 58]. The definition of VaR for a portfolio is the financial loss, relative to the mean,

$$VaR_{\text{mean}} = VaR_\alpha = E[\tilde{W}] - q_\alpha. \tag{3.3}$$

where q_α is the quantile function ($\Pr[\tilde{W} \leq q_\alpha] = \alpha$, see the definition A.2.1).

Sometimes the VaR is defined as the absolute financial loss, that is, relative to zero or without reference to the expected value,

$$VaR_{\text{absolute}} = VaR_{a,\alpha} = -q_\alpha. \tag{3.4}$$

We have developed functions of risk derived from the VaR, but that also measure the dispersion of returns given that we fall below the VaR.

The VaR is homogeneous, and has a risk-free constant c_f equal to 0 for the mean-centered

VaR, and $c_f = -b_f$ for the absolute VaR. However, it is not convex in general.

The shortfall

Another important objective function is

$$e_\alpha(\tilde{W}) = q_\alpha - E\{\tilde{W}|\tilde{W} \leq q_\alpha\}. \tag{3.5}$$

The function $e_\alpha(\mathbf{x})$ measures the expected loss below the disaster level q_α; and thus measures the risk beyond Q_p.

The function

$$\check{s}_\alpha(W) = s(q_\alpha) = \int_{-\infty}^{q_\alpha} F_{\tilde{W}}(w)dw = \alpha e_\alpha(\tilde{W})$$

appears in the definition of the second order stochastic dominance described in section B.2. However, the second order stochastic dominance also requires that the inequality $s(y) = \int_{-\infty}^{y}(F_{\tilde{A}}(t) - F_{\tilde{B}}(t))dt \leq 0$ is valid for all $y \in \Re$; if the shortfall is optimized, the stochastic dominance is not necessary obtained.

The shortfall can be related to the concept of stochastic dominance:

$$\tilde{s}_\alpha(\tilde{W}) \;\; = \;\; \alpha e_\alpha(\tilde{W}) - I_{ref}(q_\alpha) \tag{3.6}$$

where $I_{ref}(q_\alpha) \equiv \int_{-\infty}^{q_\alpha} F_{\tilde{W}_{ref}}(t)dt$. In this case, $\tilde{W}_{ref}$ is a reference random wealth with a distribution function $F_{\tilde{W}_{ref}}(\cdot)$, related to benchmarking, as will be explained in section 3.4 (see [36]). The shortfall is a coherent risk measure, assuming there are finitely many states of the nature (as described in [6]).

Absolute shortfall

Similarly to the VaR, we can also define the *absolute* shortfall risk as follows:

$$e_{a,\alpha}(\tilde{W}) = -E\{\tilde{W}|\tilde{W} \leq q_\alpha\}. \tag{3.7}$$

The shortfall is homogeneous with a risk-free factor $c_f = 0$; the absolute shortfall is also

homogeneous but with a risk-free factor $c_f = -b_f$.

Example 3.2.1 *Non-convexity of the VaR.* An example of the non-convexity of the VaR is shown in figure 3-1(a). Assume it is possible to have the portfolio $\mathbf{y}_c' = [0, -1]$ which represents selling one normalized call, and $\mathbf{y}_p' = [-1, 0]$ which represents selling one normalized put. We can form a portfolio of only two assets, two far out-of-the-money options, a put and call, with a linear combination of $\mathbf{y}_c$ and $\mathbf{y}_p$, so that $\mathbf{y}_{lc}(\lambda) = \lambda \mathbf{y}_c + (1-\lambda)\mathbf{y}_p$. As a function of λ, we plotted the absolute $VaR_{0.05}$ of the portfolio $\mathbf{y}_{lc}(\lambda)$, which is the non-convex graph depicted in 3-1(a).

In this example, the underlying asset was assumed to have a continuously compounded return of 15%, and a volatility of 20%. The risk-free asset return is 5%, the time to expiration of the options is 1/2 year, and an *absolute* $VaR_{0.05}$ was computed; the options are described in more detail in appendix C.4.4. Setting the initial price of the underlying asset as 1, we generated 1000 samples of prices for the expiration date, using a log-normal distribution (as in equation (C.3)) for the underlying asset. From the price distribution, we computed what would be the final price distributions for the two out-of-the-money options (using the definitions in appendix C.2). We obtained 1000 samples of the joint price distributions. For each portfolio, we used the technique described in section A.2.4 to compute a quantile estimator at 5% (which is the negative of the $VaR_{0.05}$).

Although the absolute $VaR_{0.05}$ is non-convex, the set of financial assets was limited to the two options (no investment possible in the underlying assets), and the portfolios did not comply with the budget constraint $\mathbf{y}'\mathbf{1} = 1$; the example assumed that pure shortselling portfolios as $\mathbf{y}_c$ was allowed.

In figure 3-1(b) we show the equivalent absolute shortfall risk measure (using the estimator described in equation (A.15)) for the same linear combination of portfolios as in 3-1(a). For this example, the absolute shortfall is clearly convex.

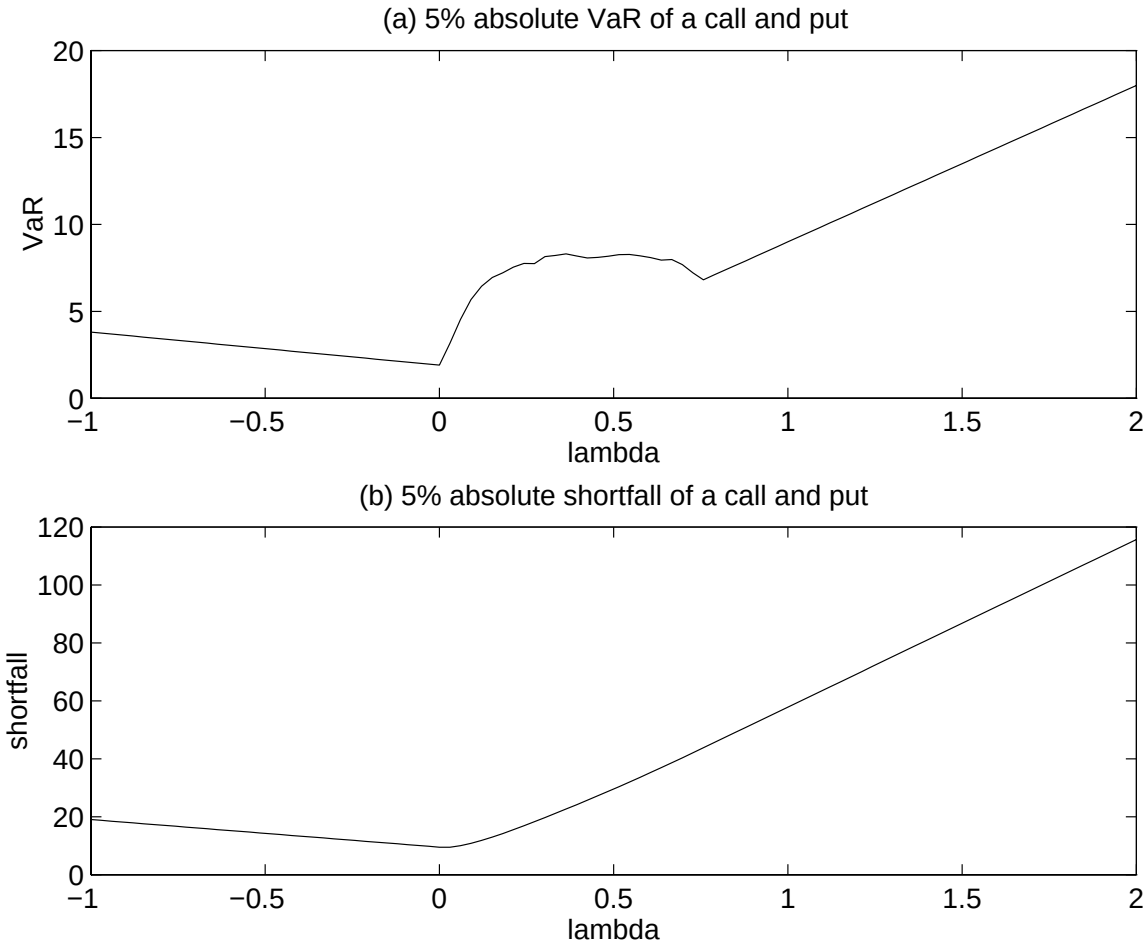

Figure 3-1: Non-convexity of VaR. (a) Absolute VaR. (b) Absolute shortfall.

3.3 Relationship between the risk-reward and the utility theories

Given the elegance of utility theory, many researchers have searched for equivalences between the risk-reward approach and the utility approach. From theorem B.1.5, if the reward measure and the risk measure are both functions of $\mathbf{x}$ and are also concave over a convex set $\mathcal{X}$, then the efficient frontier will also be convex, and can be calculated by selecting an "appropriate" vector $\lambda' = [\lambda_1, \lambda_2]'$. In this case, the risk-reward approach will have a scalar value function, formed by the weighting function

$$v(\mathbf{x}) = \lambda_1 reward(\mathbf{x}) - \lambda_2 risk(\mathbf{x}). \tag{3.8}$$

If that is the case, this weighting function is concave, and shares many of the nice properties of the utility theory (uniqueness of optimal solutions, market equilibrium, [32, 35]), although it is more general. From the definition of the asset allocation problem, we can see that the set $\mathcal{X}$ is indeed convex. The risk and reward functions that are concave over $\mathcal{X}$ are:

i. The mean.

ii. The concave utility functions.

iii. The negative of the variance.

iv. The negative of LPM_n of order 1 and 2.

v. The shortfall.

Selecting an appropriate vector λ, combinations of i - iv can be represented via expected utility functions (as shown in the next sections). Still, it is assumed that each investor will have a different λ that better fits her risk appetite.

For particular classes of the joint distribution F, the other risk functions can also be concave, and the corresponding risk-reward preference will have a value function representation. The quantile function is not concave for arbitrary distributions, as found by [6]; identification

of the class of distributions that allow its convexity is an interesting problem which needs to be solved. The shortfall is convex assuming there are finitely many states of the nature [6]. If the performance set $\mathcal{P}$ is not concave, then the Pareto relation cannot be expressed as a sum (and even worse, there is no value function that can represent the preference relation, see appendix B).

3.3.1 The mean-variance and the mean-LPM vs. the utility theory

The mean-variance framework uses, as the name indicates, a $reward(\mathbf{x}) = \mathbf{x}'E[\tilde{\mathbf{b}}]$ and a $risk(\mathbf{x}) = E[(\mathbf{x}'\tilde{\mathbf{b}} - \mathbf{x}'E[\tilde{\mathbf{b}}])^2]$; the weighting function (3.8) can be expressed in terms of the quadratic utility function:

$$E[U(\tilde{W}(\mathbf{x}))] = E[\lambda_1 \mathbf{x}'\tilde{\mathbf{b}} - \lambda_2(\mathbf{x}'\tilde{\mathbf{b}} - \mathbf{x}'E[\tilde{\mathbf{b}}])^2].$$

The value function required will be the expected value of this quadratic equation, $E[U(\cdot)]$. If the rates of return are multivariate elliptic (i.e. an affine transformation of a spherically symmetric distribution, which means it includes the multivariate normal joint distribution), a Taylor series expansion of an arbitrary expected utility function $E[U(\tilde{W})]$ is:

$$E[U(\tilde{W})] = U(E[\tilde{W}]) + \frac{1}{2!}u''(E[\tilde{W}])\sigma^2(\tilde{W}) + E[\text{H.O.M.}],$$

where $E[\text{H.O.M.}]$ is a term than includes moments of order higher than 2.

Hence, in the mean-variance framework, optimal portfolios will be confined to lay along the "efficient frontier" in a mean-variance space. However, optimal portfolios for arbitrary distributions and preferences cannot be represented within the efficient frontier.

Lower Partial Moments

Lower partial moments can use the expected mean as a reward measure;

$$reward(\mathbf{x}) = \mathbf{x}'E[\tilde{\mathbf{b}}]$$

(or another concave utility function), and the LPM_n as a risk measure:

$$risk(\mathbf{x}) = E[(\tau - \mathbf{x}'\tilde{\mathbf{b}})^n u_s(\tau - \mathbf{x}'\tilde{\mathbf{b}})].$$

The weighting function (3.8) adapted to the lower partial moments case can be expressed using the following utility function

$$E[U(\tilde{W}(\mathbf{x}))] = E[\lambda_1 \mathbf{x}'\tilde{\mathbf{b}} - \lambda_2(\tau - \mathbf{x}'\tilde{\mathbf{b}})^n u_s(\tau - \mathbf{x}'\tilde{\mathbf{b}})]. \tag{3.9}$$

The value function required will be the expected value of this piecewise utility function $E[U(\cdot)]$, which will be concave if $n = 1$ or 2, but not if $n = 0$.

This utility has both advantages and disadvantages: although it seems to better describe the investors' behavior, the wealth elasticity (see appendix B.3.1) is negative for possible outcomes $\mathbf{x}'\mathbf{b} \leq \tau$ of the random variable $\mathbf{x}'\tilde{\mathbf{b}}$, and zero for $\mathbf{x}'\mathbf{b} > \tau$. In the dynamic case, this means that the investor is non-consistent, or that her investing behavior changes based on the quantity of wealth she has. The dynamic behavior is further described in section 6.4.1.

3.3.2 The shortfall and the VaR vs. the utility theory

Everyone wants to find out if risk functions represented by quantile functions have an expected utility function. In cases where the quantile functions are concave with respect to $\mathbf{x} \in \mathcal{X}$ (even if they are not concave for the $\mathbf{x} \in \Re^m$), theorems B.1.3 and B.1.5 apply. Therefore, there will be an additive value function formed by the weighted sum of the *risk* and *reward* measures. The weighting functions of the shortfall can be represented as

$$v(\mathbf{x}) = \lambda_1 reward(\mathbf{x}) - \lambda_2 e_\alpha(\mathbf{x}'\tilde{\mathbf{b}}),$$

and the VaR can be represented as

$$v(\mathbf{x}) = \lambda_1 reward(\mathbf{x}) - \lambda_2 VaR_\alpha(\mathbf{x})$$

Table 3.1: The Allais paradox in risk-reward scenario

Lottery	mean	$q_{0.01}$	$VaR_{0.01}$
p1	1	1	0
p2	1.39	0	1.39
p3	0.5	0	0.5
p4	0.11	0	0.11

for appropriate nonnegative λ_1, λ_2; at least one of them is nonzero. There will be an expected utility representation of those weighting functions as long as the quantile or shortfall functions can be represented as a function of moments. For some special distribution functions such as elliptic distributions, there is a formula that involves the first two moments; arbitrary distributions might use more moments. Risk functions based on quantiles might be more suited to deal with arbitrary distributions, than methods using more than two moments. For arbitrary distributions, the exact value function can only be approximated with an expected utility function in certain ranges; in some cases the semivariance utility function seems to work quite well.

3.3.3 The Allais paradox in the risk-reward framework

We want to use the Allais paradox to show that in some cases the expected utility representation of a preference relation may not exist, whereas risk-reward representation might.

The Allais paradox (described in appendix B.3.2) can be used in the risk-reward framework, specifically in a mean-$VaR(0.01\%)$ (absolute) context, as shown in table 3.1.

Choosing $p1$ over $p2$ and $p3$ over $p4$ is consistent with the Pareto optimal preference relationship that uses a mean-quantile performance vector; $p3$ is certainly better than $p4$, and $p1$ is indifferent to $p2$, which is not a contradiction. The VaR and the shortfall risk measures are unable to rank the assets. However, if the mean-quantile had an expected utility representation, it would not be able to rank $p3$ over $p4$; hence the risk-reward methodology might not always have an expected utility representation.

3.4 The risk-reward approach to portfolio optimization

Once we know which preference relation to use, the numerical optimization method is one of the following two cases:

i. The maximization of a scalar value function $v(\tilde{W}(\mathbf{x}))$, constraining the decision vector $\mathbf{x}$ to the set $\mathcal{X}$, which includes the budget constraint defined in section 2.1 and other requirements (as non-negativity of the weights, for example). The expected utility approach corresponds to this case; in the expected utility framework, the value function corresponds to the expectation of a utility function $v(\tilde{W}) = E[U(\tilde{W})]$. The optimization problem is:

$$\max v(\tilde{W}(\mathbf{x})) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{X}. \tag{3.10}$$

ii. the maximization of an arbitrary component of the performance vector, $h_i(\tilde{W}(\mathbf{x}))$, constraining both the decision vector to be an element of the set $\in \mathcal{X}$, and the remaining components of the performance vector to predefined values; $h_k(\tilde{W}(\mathbf{x})) \geq \rho_k$, $k \neq i$, $k = 1, \ldots, Q\}$. In the two dimensional risk-reward case, $\mathbf{h}(\mathbf{x}) = [reward(\mathbf{x}), -risk(\mathbf{x})]'$; hence, we will have either (for a predefined risk level L_p):

$$\max reward(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{X},\ risk(\mathbf{x}) \leq L_p, \tag{3.11}$$

or, for a predefined reward level R_p;

$$\min risk(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{X},\ reward(\mathbf{x}) \geq R_p, \tag{3.12}$$

Efficient frontier methods like the mean variance are computed following this procedure, although the special nature of the mean-variance problem allows the computation of only two optimal portfolios. The remaining ones can be obtained as linear combinations of two optimal portfolios, a phenomenon known as the *two-mutual fund separation*,

[35, 32]. Homogeneous and convex risk measures also offer the mutual fund separation [60]. Depending on the chosen particular combination of risk and reward measures, it might be easier to find the solution constraining one particular component of the performance vector (e.g., for the mean-variance case the reward is constrained). This method results in a set of efficient portfolios, since it satisfies the theorem B.1.1 as well as the definition of the e-portfolio.

The budget constraint is represented as $\mathcal{X} = \{\mathbf{x}|\mathbf{x}'\mathbf{1} = 1\}$, although additional constraints such the as the non-negativity of the decision vector ($\mathbf{x} > \mathbf{0}$) can be added. The optimization of a value function is considerably simpler, and has already been very well studied. However, investors tend to select optimal portfolios computed using optimization problems of the second category.

Usually the computation of the complete frontier is perceived as a "naive" method; the Academic literature presents several alternatives which compute directly an optimal decision vector (goal optimization, penalty functions, etc). Unfortunately, since the selection of the optimal depends heavily on the investors' behavior, it is not possible to select those alternatives. Mutual Fund separation theorems might be useful in some cases; but if the non-negativity constraint is enforced, they are useless [32].

Nonlinear programming methods are used to solve both cases. Gradient-based optimization algorithms require the computation of the gradient of the performance vector $\nabla_{\mathbf{x}}\mathbf{h}(\mathbf{x})$; therefore either an explicit form of the gradient or an estimate must be available. In other cases, finite difference approximations of the gradient are sufficient for the algorithm to converge to an optimal solution. For some particular cases the optimization only requires a quadratic programming algorithm (i.e. mean-variance), or involves the maximization of a concave function (i.e. expected utility of risk averse individuals). There are a couple of interesting applications that can be derived from the static optimization problem, and which are already in practice.

Index tracking Also known as benchmarking [36]; let's assume we have a reference portfolio with a random outcome $\tilde{W}_{ref}$ and a CDF F_{ref}. If we define the *index tracking error*

$\tilde{err}(\mathbf{x})$ as

$$\tilde{err}(\mathbf{x}) = \tilde{W}(\mathbf{x}) - \tilde{W}_{ref}, \tag{3.13}$$

then we can define the index tracking problem as

$$\min risk(\tilde{err}(\mathbf{x})) \quad \text{s.t.} \quad reward(\tilde{err}(\mathbf{x})) = 0, \mathbf{x} \in \mathcal{X}. \tag{3.14}$$

The function $\tilde{e}(\mathbf{x})$ is known as the residual error.

Similarly, the *index enhancing* problem becomes:

$$\max reward(\tilde{err}(\mathbf{x})) \quad \text{s.t.} \quad risk(\tilde{err}(\mathbf{x})) = 0, \mathbf{x} \in \mathcal{X}. \tag{3.15}$$

However, this instance may turn out to be infeasible, depending on the chosen $\tilde{W}_{ref}$.

3.4.1 Optimization with risk-free asset

What follows is a new derivation of the properties of optimal portfolios when m risky assets and one risk-free asset are available, and shortsales of the asset are allowed.

Let us assume we use a pseudo-coherent risk measure (which we denote as ρ) with a risk-free constant c_f, and as reward measure we select the mean return of the portfolio. Then, the optimization can be analyzed as:

$$\tilde{W}(\mathbf{x}) = \mathbf{x}'\tilde{\mathbf{b}} + (W_0 - \mathbf{x}'\mathbf{1})b_f, \tag{3.16}$$

(similar to equation (2.1) from section 2.1). Note that the decision vector $\mathbf{x}$ represents the vector of cash commitments. If we decide to optimize the problem following the format of equation (3.12), constraining the expected return of the optimal portfolio to be equal to a predefined wealth level W_p, (where the gross return is $b_p = W_p/W_0$), we have to solve the following problem:

$$\min \rho(\tilde{W}(\mathbf{x})) \quad \text{s.t.} \quad E[\tilde{W}(\mathbf{x})] = W_p.$$

Equation (3.12) holds for an inequality constraint; however, for the analysis, we assume that we should enforce an equality constraint to compute the optimal solution. This assumes that the risk and reward measures are selected such that there must be a tradeoff, otherwise it would be possible to define some risk and reward measures that lead to non diversified portfolios. The Lagrangian is

$$L(\mathbf{x}, \lambda) = \rho(\mathbf{x}'\tilde{\mathbf{b}} + (W_0 - \mathbf{x}'\mathbf{1})b_f) - \lambda(\mathbf{x}'E[\tilde{\mathbf{b}}] + (W_0 - \mathbf{x}'\mathbf{1})b_f - W_p);$$

using the risk-free condition (assuming for simplicity that $c_f = 0$), the Lagrangian can also be expressed as

$$L(\mathbf{x}, \lambda) = \rho(\mathbf{x}) - \lambda(\mathbf{x}'E[\tilde{\mathbf{b}}] + (W_0 - \mathbf{x}'\mathbf{1})b_f - W_p).$$

For optimality, the gradient of the Lagrangian should satisfy the Kuhn-Tucker condition:

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*) = \nabla_{\mathbf{x}} \rho(\mathbf{x}^*) - \lambda^*(E[\tilde{\mathbf{b}}] - b_f\mathbf{1}) = \mathbf{0}, \tag{3.17}$$

where $\mathbf{x}^*$ and λ^* are respectively the optimal decision vector and the Lagrange multiplier. Pre-multiplying equation (3.17) by $\mathbf{x}^{*\prime}$, using the homogeneity properties of the coherent risk and the return constraint $E[W(\mathbf{x})] = W_p$ expressed as $\mathbf{x}'E[\tilde{\mathbf{b}}] - \mathbf{x}'\mathbf{1}b_f = W_p - W_0 b_f$, and solving for the Lagrange multiplier λ^*, we obtain:

$$\lambda^* = \frac{\rho(\mathbf{x}^*)}{W_p - W_0 b_f}. \tag{3.18}$$

Substituting the optimal Lagrange multiplier λ^* in equation (3.17), we obtain

$$E[\tilde{\mathbf{b}}] - b_f\mathbf{1} = \frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}^*)}{\rho(\mathbf{x}^*)}(W_p - W_0 b_f). \tag{3.19}$$

If we define the *generalized* $\breve{\beta}_j(\mathbf{x})$ as[1]

$$\breve{\beta}_j(\mathbf{x}) = \frac{\frac{\partial \rho(\mathbf{x})}{\partial x_j}}{\rho(\mathbf{x})} \qquad \text{for} \qquad j = 1, \ldots, m \tag{3.20}$$

[1]In the case when the risk-free constant $c_f \neq 0$ (e.g., for the absolute VaR and the shortfall risk measures),

(which for elliptic distributions turns out to be similar to the classic mean-variance definition of β, see section 2.2.1), then

$$E[\tilde{b}_j] - b_f = \breve{\beta}_j(\mathbf{x}^*)(W_p - W_0 b_f);$$

if we set $W_0 = 1$ and use simple returns, for the jth asset the equation (3.19) becomes:

$$E[\tilde{r}_j] - r_f = \breve{\beta}_j(\mathbf{x}^*)(r_p - r_f), \qquad (where \qquad r_p = W_p/W_0 - 1),$$

which is easily recognized as the prototype of the CAPM, and has been widely studied for distributions represented with a finite number of moments (for example, up to 4 moments were analyzed by [26]). Of course, to derive a CAPM model from this formula we would have to make further assumptions about the investors' behavior, which is not a straightforward procedure for arbitrary distributions and non-convex risk measures. It is probable that the distributions restricted by the no-arbitrage condition yield a unique solution.

Optimization with pseudo-coherent risk measures. This same procedure could be applied to derive similar formulas when we have pseudo-coherent reward measures, i.e., $reward(\mathbf{x}) = \varrho(\mathbf{x})$, where ϱ represents pseudo-coherent reward measures, such as the median. If the constraint $\varrho(\mathbf{x}) = W_p$ holds, and the pseudo-coherent reward measure has a risk-free constant d_f, the condition for optimality is:

$$\nabla_{\mathbf{x}}\varrho(\mathbf{x}) - d_f\mathbf{1} = \frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}^*) - c_f\mathbf{1}}{\rho(\mathbf{x}^*) - c_f\mathbf{1}'\mathbf{x}^*}(W_p - W_0 d_f). \tag{3.21}$$

The median is a good example of a pseudo-coherent risk measure, with $d_f = b_f$. However, a practical and reliable method of optimization for the median as reward measure is not available at the moment.

---

the full formula for the generalized $\breve{\beta}_j$ is:

$$\breve{\beta}_j(\mathbf{x}) = \frac{\frac{\partial\rho(\mathbf{x})}{\partial x_j} - c_f}{\rho(\mathbf{x}) - c_f\mathbf{1}'\mathbf{x}}, \qquad \text{for} \qquad j = 1, \ldots, m.$$

3.4.2 Optimization without risk-free asset

If no risk-free asset is available, then the wealth equation becomes

$$\tilde{W}(\mathbf{x}) = \mathbf{x}'\mathbf{b}. \tag{3.22}$$

and we now have to enforce explicitly the budget constraint $\mathbf{x}'\mathbf{1} = W_0$ in the optimization problem:

$$\min \rho(W(\mathbf{x})) \qquad \text{s.t.} \qquad \mathbf{x}'E[\tilde{\mathbf{b}}] = W_p, \qquad \mathbf{x}'\mathbf{1} = W_0.$$

We are assuming the optimization must be equally constrained. The Lagrangian is

$$L(\mathbf{x}, \lambda) = \rho(\mathbf{x}) - \lambda_1(\mathbf{x}'E[\tilde{\mathbf{b}}] - W_p) - \lambda_2(\mathbf{x}'\mathbf{1} - W_0),$$

where $\lambda' = [\lambda_1, \lambda_2]$. For optimality, the gradient of the Lagrangian should satisfy the Kuhn-Tucker condition:

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*) = \nabla_{\mathbf{x}} \rho(\mathbf{x}^*) - \lambda_1^* E[\tilde{\mathbf{b}}] - \lambda_2^* \mathbf{1} = \mathbf{0},$$

where $\mathbf{x}^*$ and the vector λ^* are respectively the optimal decision vector and the Lagrange multiplier vector. Pre-multiplying equation (3.17) by $\mathbf{x}^{*\prime}$, using the return constraint, and the homogeneity properties of the coherent risk, we obtain the equation:

$$\rho(\mathbf{x}^*) - \lambda_1^* W_p - \lambda_2^* = 0.$$

Without further assumptions, it is not possible to advance much beyond this result; if we assume the returns have a joint elliptic distribution, we recover the same results already obtained for the mean-variance case.

3.4.3 Numerical algorithm

For the implementation of a numerical algorithm, we assume that n samples of the random vector $\tilde{\mathbf{b}}$, b_1, b_2, $\cdots$, b_n are available. The samples could either result from a Monte-Carlo

simulation, or from historical data; but they must be independent and identically distributed (i.i.d.) samples. In Chapter 5 we describe several algorithms that could be used to optimize portfolios with quantile-based risk measures. For the following examples we used the gradient-based algorithm described in section 5.3.

Example 3.4.1 In the first example, depicted in figure 3-2, we assumed that three financial stocks with returns following a joint Gaussian multivariate distribution (described in appendix C.4.1) were available. With a set of samples, we computed the optimal weights as a function of the desired expected annual return using four different methods: the mean-variance, the mean-semivariance, the mean-shortfall and the mean-VaR. The optimization was done with no risk-free asset, but shortsales were allowed. The figure 3-2 only shows the weights of the optimal portfolios for the first two assets.

Since the joint multivariate distribution is Gaussian, the mean-variance numerical method yield the most accurate optimal weights; hence, deviations from the mean-variance optimal portfolio correspond to errors introduced by the algorithms used to compute the other mean-risk optimal portfolios. In this example, the mean-semivariance weights completely coincide with those of the mean-variance method. From figure 3-2, we can appreciate the large magnitude of errors introduced by the mean-VaR method when compared with other optimization methods. The mean-shortfall behaves relatively better than the mean-VaR method. Here, our optimization numerical algorithm is another source of errors. However, even if the optimal weights do not correspond exactly to the mean-variance optimal weights, when the comparison is made in the performance space (see example 3.5.1), the mean-VaR and the mean-shortfall methods perform similarly to the mean-variance optimal portfolios.

Example 3.4.2 For the second example, depicted in figure 3-3, we use financial assets with a significant asymmetric distribution function, the "Option-based strategies data", described in appendix C.4.2 and shown in figure 2-1. We only show the optimal weights for two of the assets. The "write-call" strategy at 50% is designed to have a heavy left tail (downside), while the "long-put" strategy at 50% has a lighter left tail. Intuitively, the "write-call" strategy seems *riskier* than the "long-put"; hence, we would expect optimal portfolios to

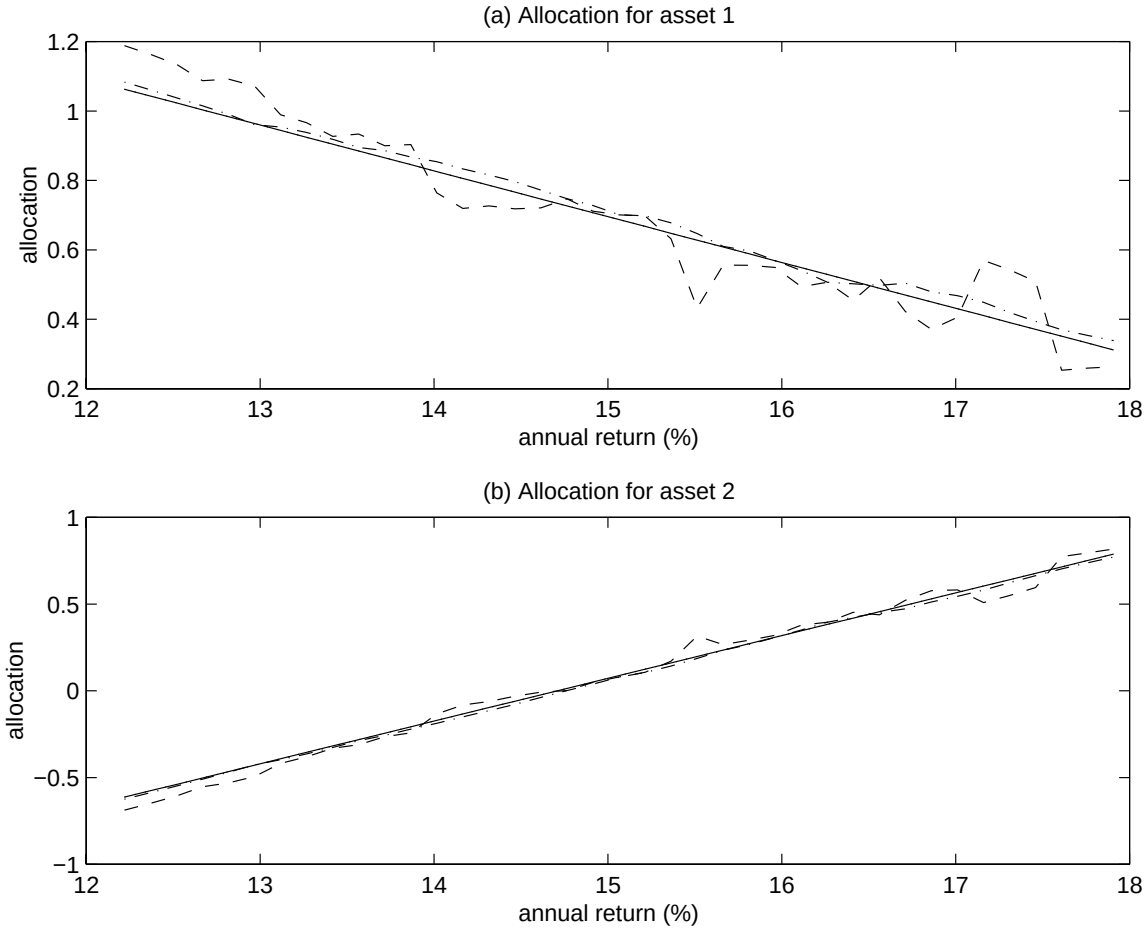

Figure 3-2: Weights of an optimal portfolio. Gaussian data. (no risk-free asset, shortsales allowed)

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

The dotted line coincides with the solid line. Deviations from the solid line represent errors introduced by the numerical.

allocate a greater percentage to the "long-put" strategy, and at the same time avoid the "write-call" strategy.

In this example, if we focus on the optimal weights for a certain return (no risk-free asset, and shortsales allowed), we clearly see that the optimal mean-variance gives a different optimal weight than the other methods, the mean-semivariance, the mean-VaR and the mean-shortfall. We will see in section 3.5 that when we compare the efficient frontiers of the different methods, each method yields the "optimal" portfolio according to its own performance measure.

It is very difficult to select the "best" method, since each one will be the "optimal" according to its own definition; still, in this example some interesting features of the optimal weight deserve further analysis. The mean-variance and the mean-semivariance methods increase the amount of asset 1 (figure 3-3) as the desired expected return is increased, while the mean-VaR and the mean-shortfall methods *decrease* the allocation of this asset as a function of the desired expected return. At this point the definition of risk is crucial; if the "write-call" strategy is more risky (i.e., the probability of having a loss is higher), then a risk averse investor would choose to have less of this asset.

In the case of the "long-put" strategy, although all methods increase the optimal weight as the desired expected return increases, the slope is much steeper for the mean-VaR and the mean-shortfall methods, which reflects again how each risk definition generates different attitudes towards risk. At a simple return of 6%, the optimal weights of the "long-put" strategy produced by the mean-VaR and the mean-shortfall methods are significantly larger than those of the mean-variance and the mean-semivariance methods, which indicates that the VaR and the shortfall risk measures consider asset 2 less risky.

This is one example in which the mean-VaR and the mean-shortfall methods yield "better" optimal portfolios, since they select optimal weights which better fit a risk-averse investor's behavior.

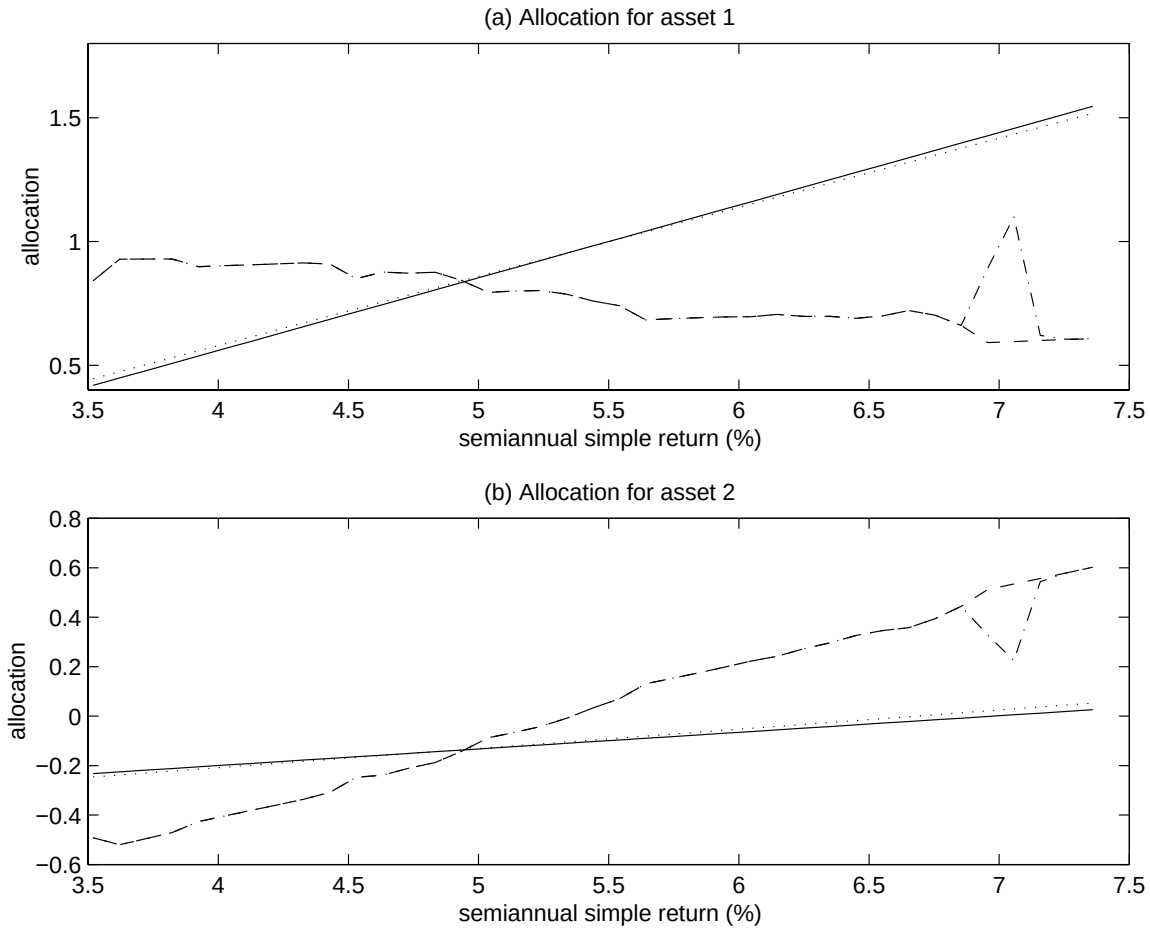

Figure 3-3: Weights of an optimal portfolio. Option-based strategies data. (no risk-free asset, shortsales allowed)

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 1: "write-call" strategy at 50 %. asset 2: "long-put" strategy at 50 %.

Different algorithms give completely different "optimal" weights when the distributions are asymmetric. The mean-shortfall and the mean-VaR coincide almost everywhere, except in a point where the mean-shortfall method fails to converge to a solution.

3.5 Risk-reward efficient frontiers

In the previous sections we described our attempts to identify the optimal portfolio. The efficient frontier technique is based on the premise that an investor selects only efficient portfolios, but that each investor may select a different portfolio. Hence, what the best mathematical programming methods can do is to identify the efficient frontier, and let the investor chose he desired performance vector.

The efficient frontier analysis has been widely developed for the mean-variance case [43, 44, 46, 35, 41, 27, 15, 32, 13], the mean-LPM_N (or semivariance) [42, 30, 45], and the safety-first [18, 52]. We take ideas from the efficient frontier analyses in the mean-variance case, and extend it to the risk-reward scenario.

When the risk measure is convex and the reward measure is concave, the efficient frontier is convex. It is possible to develop mathematical programming algorithms to find the risk-reward efficient frontier, using the mean-variance framework.

Example 3.5.1 *Convex efficient frontiers (i).* In figure 3-4 we see the familiar mean-variance efficient frontier, as well as three other efficient frontiers: the mean-semivariance, the mean-shortfall and the mean-VaR. The optimization is done with no risk-free asset, and allowing shortsales. With Gaussian case, regardless of the optimization method used, the frontiers are convex, and almost identical. Example 3.4.1 gives the weights of the efficient portfolios (figure 3-2, note that some numerical errors were introduced due to sampling) and we notice that in the performance space (figure 3-4) the graphs look almost identical. Still, in the mean-VaR space the efficient frontier is very noisy (due to numerical errors), and even looks as if it were not convex.

Example 3.5.2 *Convex efficient frontiers (ii).* Even when the financial assets have asymmetric returns, the efficient frontier can be convex, as figure 3-5 shows.

Here, shortsales are allowed, but no risk-free asset is available. The data being used comes from the "Option-based strategies data". This example clearly shows how if we compute the optimal weights using one mean-risk method, they can become inefficient when represented in a different mean-risk space. In figure 3-5(b), the mean-semivariance optimal weights

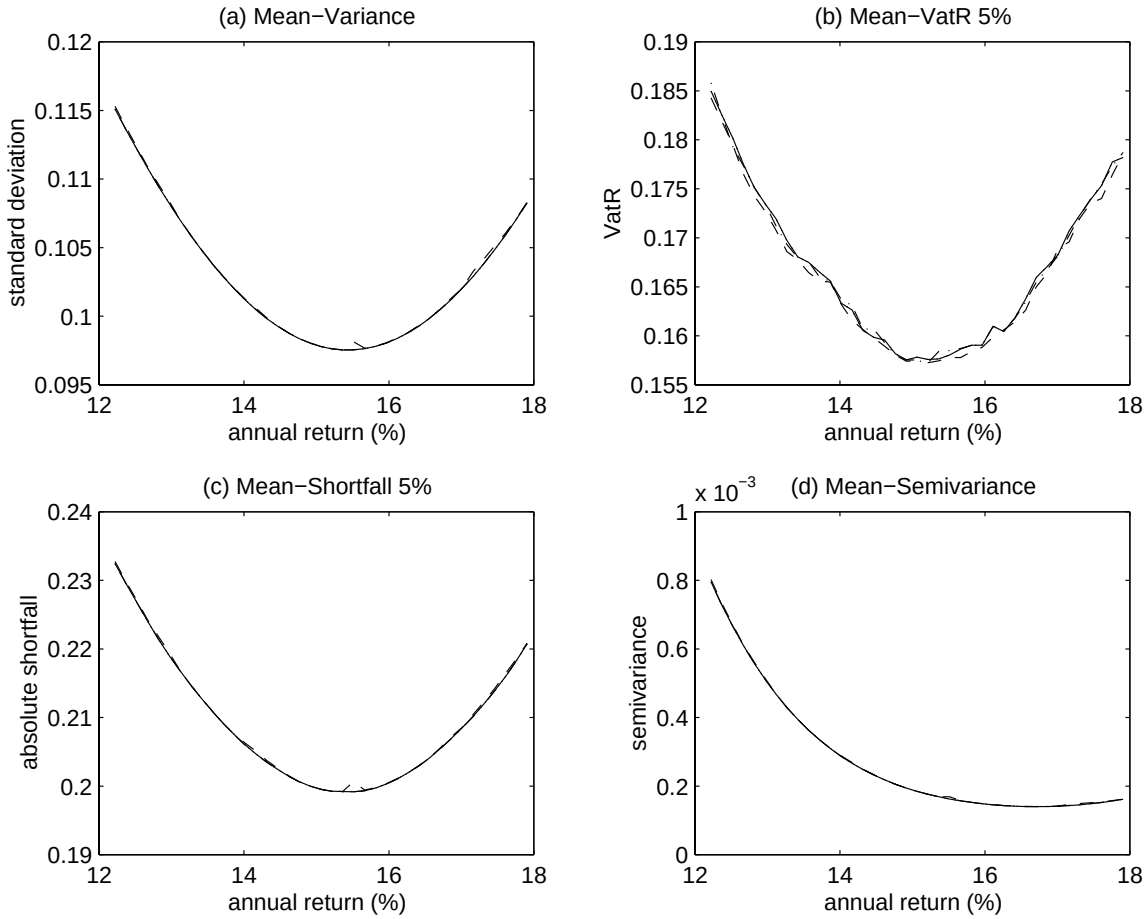

Figure 3-4: Efficient frontiers. Gaussian data. (no risk-free asset, shortsales allowed)

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

Examples of convex efficient frontiers. The frontiers coincide due to the use of simulated Gaussian data.

plotted in a mean-VaR space trace a curve with a greater VaR risk. Obviously, in the mean VaR space, the mean-VaR optimal weights outperform all other methods. The mean-VaR efficient frontier is not smooth, and appears to be non-convex, due to numerical errors. The same analysis could be done for the other graphs. In cases where only the performance measures are available, it is difficult to establish which method is superior.

In the cases where the efficient frontiers are convex it is possible to use a weighting function as the value function representing the preference relation.

It must be noted that the financial data used here is simulated with the Black and Scholes formula, which assumes there are no arbitrage opportunities.

Example 3.5.3 *Non-convex efficient frontiers.* We emphasize the fact that convex performance measures produce convex efficient frontiers. When the risk measure is not convex, the efficient frontier is not always convex. In the second example (figure 3-6) we use financial assets that with a significant asymmetric distribution function ("Put-Call data", described in appendix C.4.4); shortsales are allowed, and no risk-free asset is available. In figure 3-6(b), we see that the optimal frontier generated by $VaR_{0.05}$ is *not* convex. Not only it is not-convex, but the risk for $VaR_{0.05}$ seems to *decrease* as the expected return increases! This case outlines the danger of blindly following any particular risk measure.

3.5.1 Pseudo-coherent risk with risk-free asset

The pseudo-coherent risk case with a risk-free constant c_f and no shortsales allowed is fully analyzed in this section, using the results previously. Assuming that holding all the initial wealth W_0 on the risk-free asset ($\mathbf{x}_f = \mathbf{0}$) is an efficient portfolio, and that an efficient decision vector $\mathbf{x}^*$ with expected return r_p is available, a linear combination of them,

$$(1-\gamma)\mathbf{x}_f + \gamma\mathbf{x}^*,$$

can be substituted in equation (3.16), and results in the random final wealth variable

$$\tilde{W}(\gamma, \mathbf{x}) = \gamma\mathbf{x}^{*\prime}\tilde{\mathbf{b}} + (W_0 - \gamma\mathbf{x}^{*\prime}\mathbf{1})b_f.$$

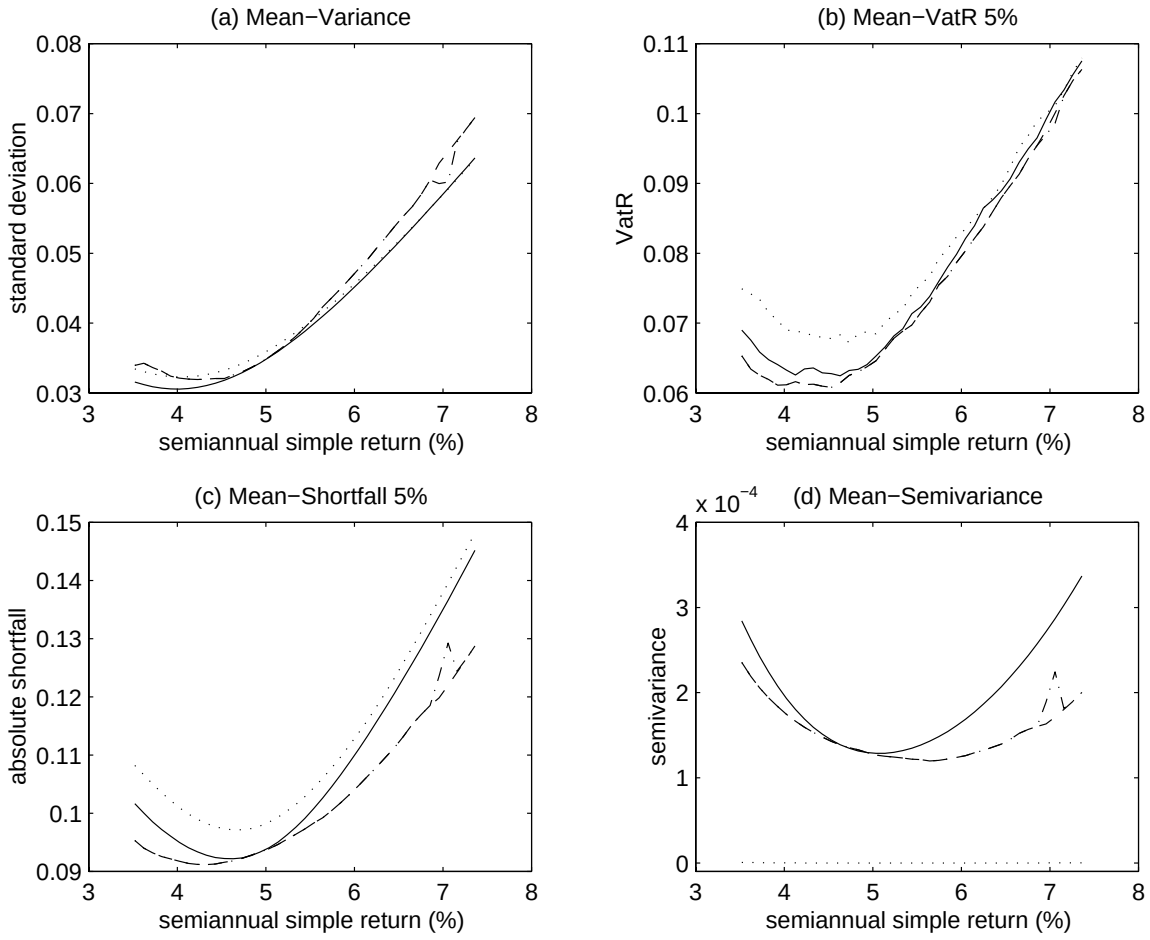

Figure 3-5: Efficient frontiers. Option-based strategies data. (no risk-free asset, shortsales allowed)

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

Example of convex efficient frontiers. Plotting the performance measures of the mean-variance optimal weights in a mean-shortfall space shows that they do not belong to the efficient frontier.

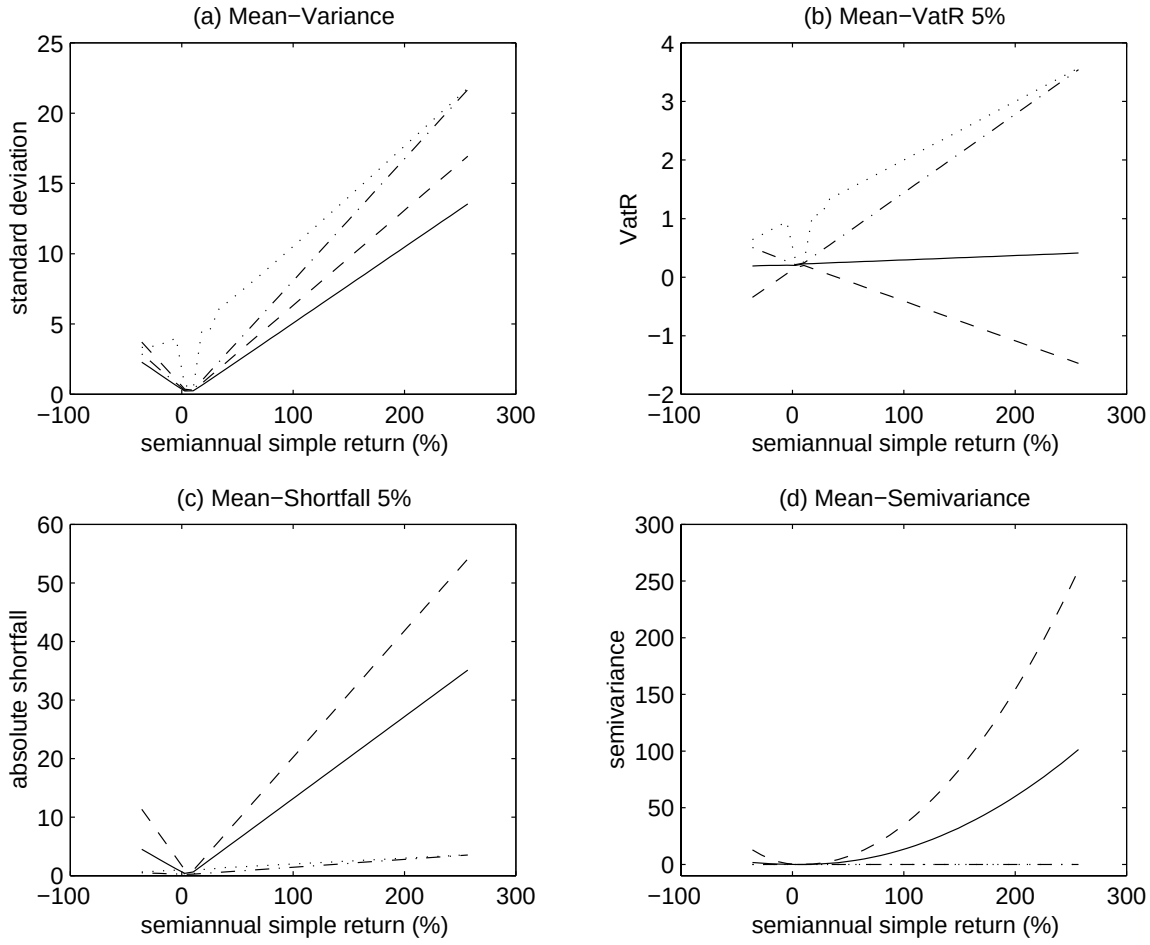

Figure 3-6: Efficient frontiers. Put-Call data (no risk-free asset, shortsales allowed).

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

(b) is an example of a non-convex efficient frontier; notice the mean-VaR efficient frontier.

If we substitute $\gamma\mathbf{x}^*$ in the equation (3.19), using the properties of homogeneous functions (see A.1.3) and the risk-free asset property of risk functions, we can derive

$$\frac{\nabla_{\mathbf{x}}\rho(\gamma\mathbf{x}^*) - c_f\mathbf{1}}{\rho(\gamma\mathbf{x}^*) - c_f\gamma\mathbf{1}'\mathbf{x}^*}(E[\tilde{W}(\gamma\mathbf{x}^*)] - b_f) = \frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}^*) - c_f\mathbf{1}}{\rho(\mathbf{x}^*) - c_f\mathbf{1}'\mathbf{x}}(E[\tilde{W}(\mathbf{x}^*)] - b_f).$$

We can see that the linear combination of the risk-free asset portfolio with any efficient portfolio satisfies the optimality condition derived in the previous section, which means that homogeneous risk measures have a mutual fund separation. With a risk-free asset, the efficient frontier is also restricted to a certain curve; taking any component of equation (3.19) we can write the risk $\rho(\mathbf{x}^*)$ as a function of the desired level of wealth, W_p:

$$\rho(\mathbf{x}^*) = \left(\frac{\frac{\partial\rho(\mathbf{x}^*)}{\partial x_j} - c_f}{E[\tilde{b}_j] - b_f}\right)(W_p - W_0 b_f) + c_f\mathbf{1}'\mathbf{x}^*. \tag{3.23}$$

In the specific cases where c_f is equal to 0, the efficient frontier is a line (e.g., for the standard deviation, the VaR_α and the shortfall), and the optimal assets have the mutual fund separation.

3.5.2 Examples of efficient frontiers

Example 3.5.4 *Linear efficient frontiers.* Following the procedures already explained for the previous examples, we generate figures 3-7 and 3-8 (with the "Option-based strategies data" and the "Put-Call data" respectively, allowing shortsales, and introducing the risk-free asset). In these figures we clearly see that the efficient frontiers are *linear* for pseudo-coherent risk measures with a risk-free constant (the standard deviation, the VaR and the shortfall). Of particular importance is the fact that the efficient frontier is a line even in the case of the "Put-Call data" and the mean-VaR optimization method (figure 3-8). This case is very important, since it shows that when there is a risk-free asset, the risk measure does not need to be convex with respect to $\mathbf{x}$ for the efficient frontier to be a line.

Example 3.5.5 *Mutual Fund separation.* Examples of mutual fund separation are shown in figures 3-9 and 3-10 where we plotted the graphs of the optimal weights with respect to

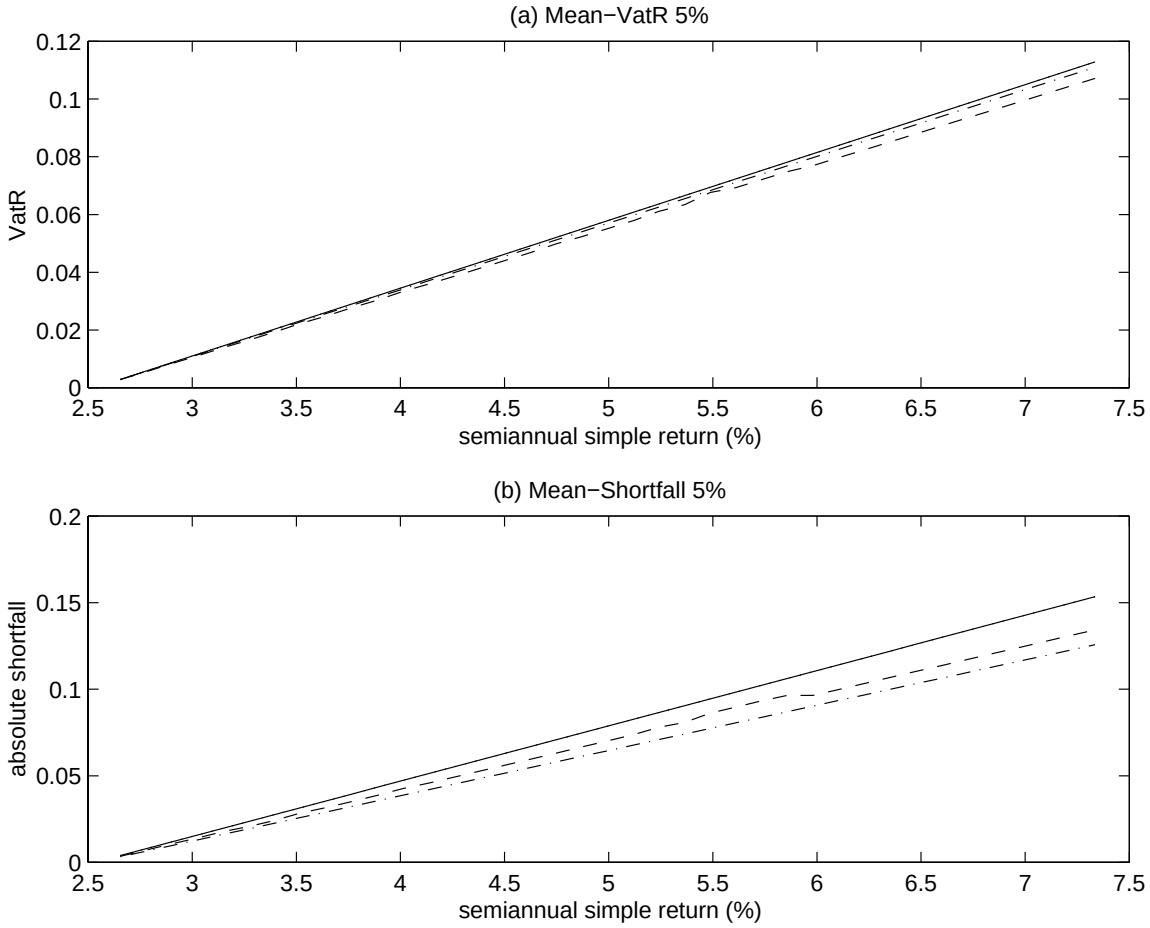

Figure 3-7: Efficient frontiers. Option-based strategies data (with risk-free asset, shortsales allowed).

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

Example of a linear efficient frontier when a risk-free asset is included. The data used is not Gaussian.

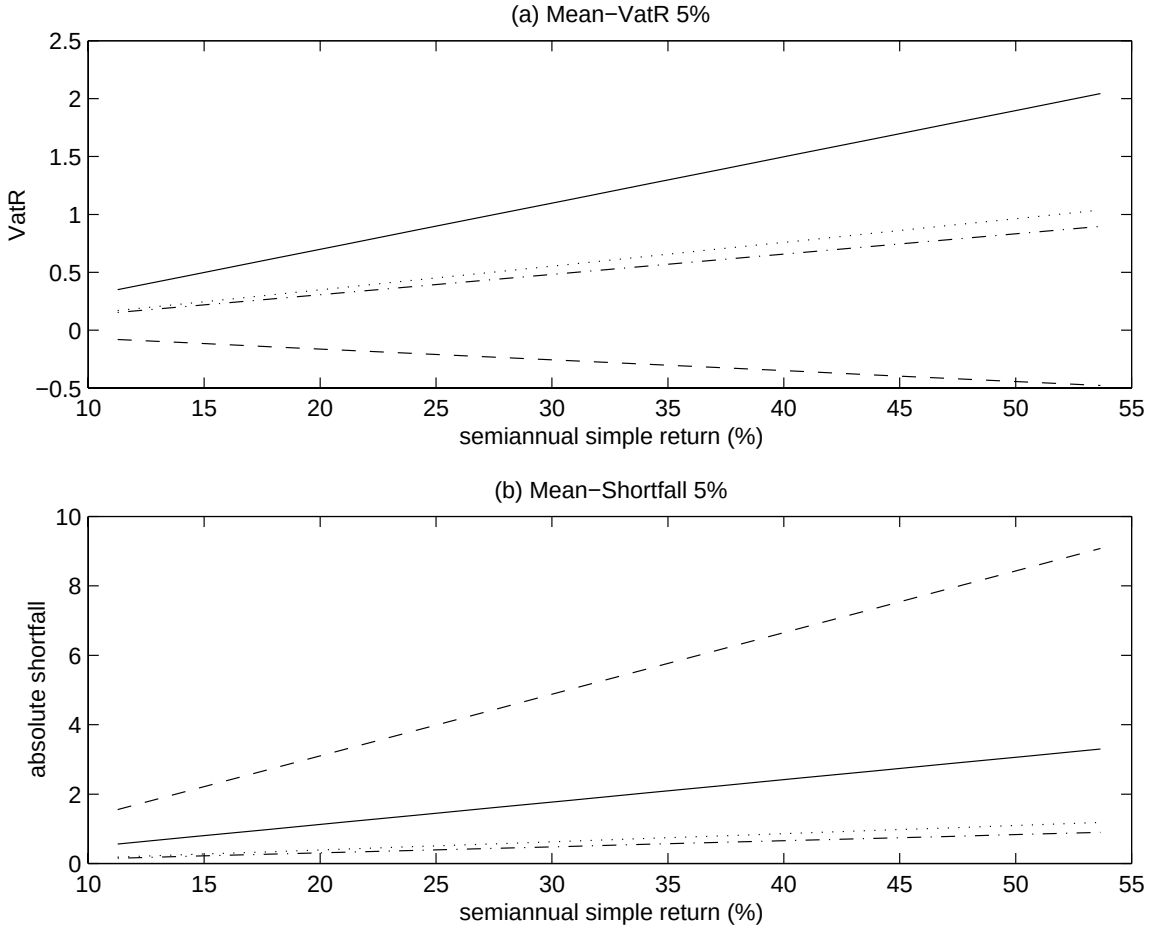

Figure 3-8: Efficient frontiers. Put-Call data. (with risk-free asset, shortsales allowed).

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

Example of linear efficient frontiers when a risk-free asset is included, even when the risk measure (VaR) is *not* convex.

different desired returns; shortsales are allowed, and a risk-free asset is available. Figure 3-10 clearly shows the mutual fund separation, even in the case of the mean-VaR optimization with the "Put-Call" data. This is important, since we already know that the "Put-Call" data usually magnify the shortcomings of the mean-VaR method. Figure 3-9 shows the mutual fund separation for the mean-variance, mean-semivariance and mean-shortfall methods when non-Gaussian data generated by the "Option-based strategies data" is used. The mean-VaR data is not a line, but the deviations are due to as errors introduced by the numerical algorithm. In any case, this example shows how the mean-VaR method can be very sensitive to sampling errors, while the mean-shortfall method is more robust. The mutual fund separation is a very important tool, since it allows the computation of two optimal portfolios which can be linearly combined to produce all other possible optimal portfolios. With mean-VaR optimal portfolios, the mutual fund separation also suggests an alternative way to produce a better estimate of the optimal portfolio; we can generate optimal portfolios at different desired level of returns, and then do a linear regression to compute a better estimate of the optimal portfolio at the desired level of return.

Example 3.5.6 *No risk-free asset.* In the previous example we confirmed that if a risk-free asset is available, the mutual fund separation exists. However, when the no risk-free asset is available, and the risk measures are not convex (as VaR), the mutual fund separation may not hold. In figure 3-11 we present one example: the optimal weights do not follow a line for all possible desired expected returns. In 3-11(a) and (b) we see that the mean-variance weights are always linear (as expected from the mean-variance theory), but the optimal weights for the mean-semivariance, the mean-shortfall and the mean-VaR are not linear, even if shortsales are allowed. This example uses the "Put-Call" data to show one case where the mutual fund separation does not exist.

Experiment 3.5.1 *Real data.* We use real stock return data to determine the optimal portfolio for the case where shortsales are allowed, and no risk-free asset is available (see appendix C.4.3 for details). Results are shown in figures 3-12 and 3-13. Figure 3-12 shows the efficient frontier, and as in the case of the Gaussian data, the performance of the optimal portfolios

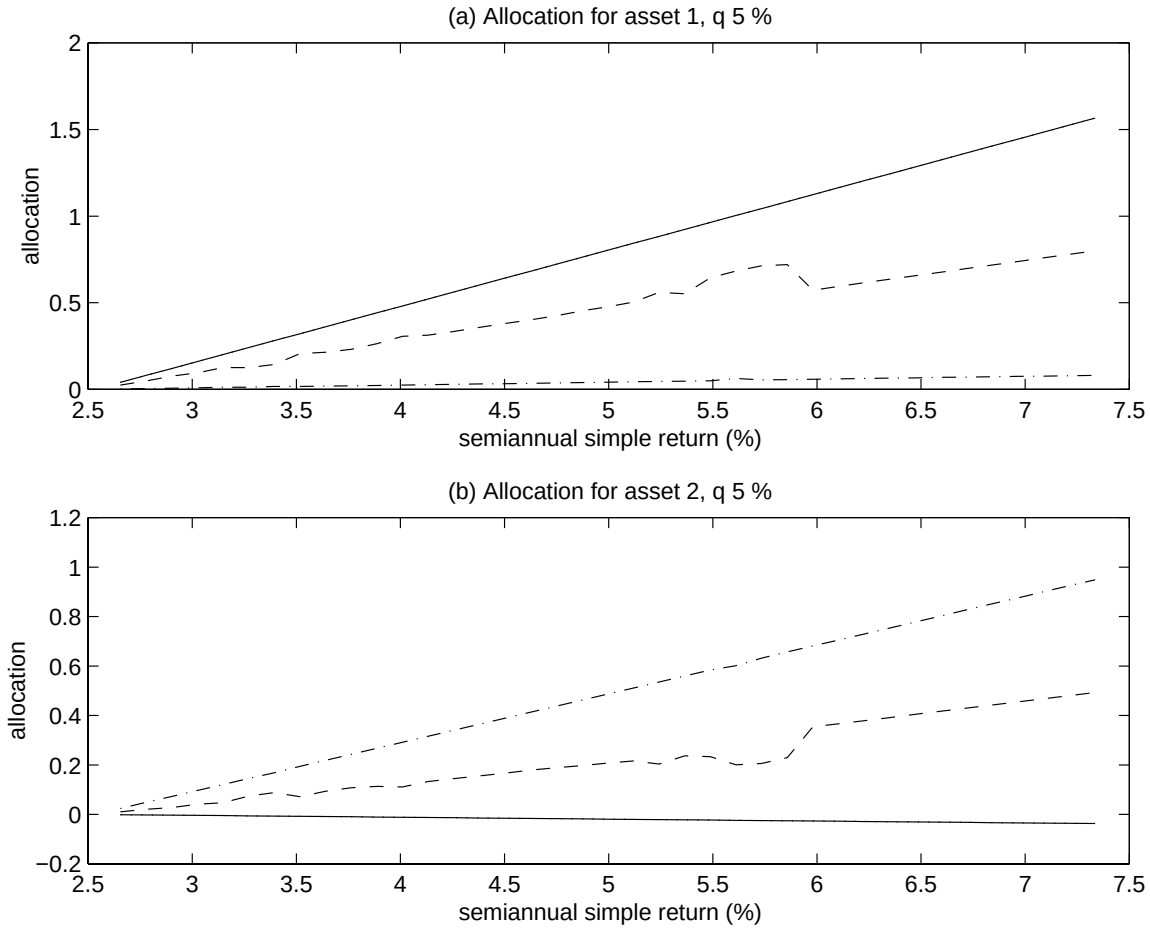

Figure 3-9: Weights of an optimal portfolio. Option-based strategies data (with risk-free asset, shortsales allowed).

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 1: "write-call" at 50 %. asset 2: "long-put" at 50 %.

Example of the mutual fund separation. The optimal weights can be obtained as linear combinations of two efficient portfolios. The deviations from the line are due to numerical errors.

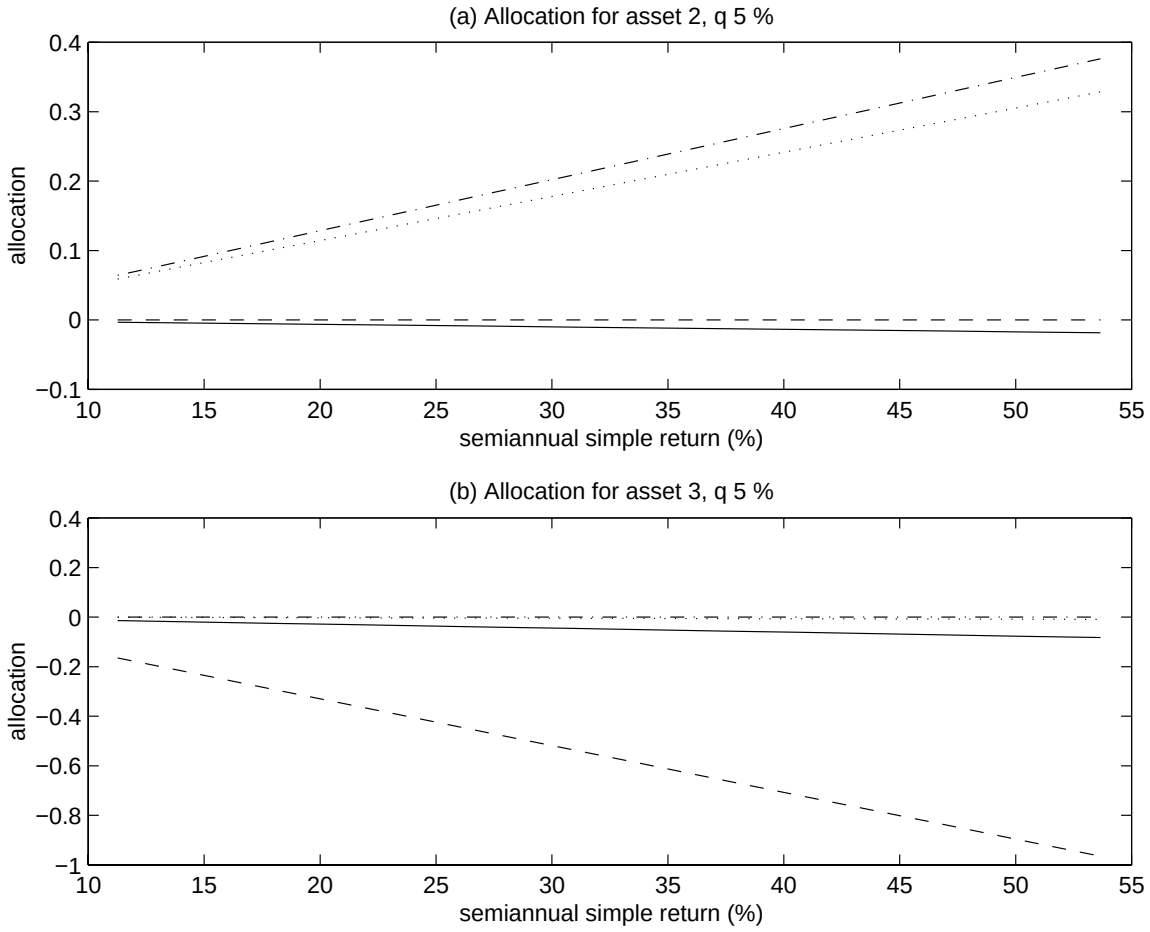

Figure 3-10: Weights of an optimal portfolio. Put-Call data. (with risk-free asset, shortsales allowed).

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 2: Put. asset 3: Call.

The mutual fund separation (when a risk-free asset is included) exists even if the risk measure is not convex.

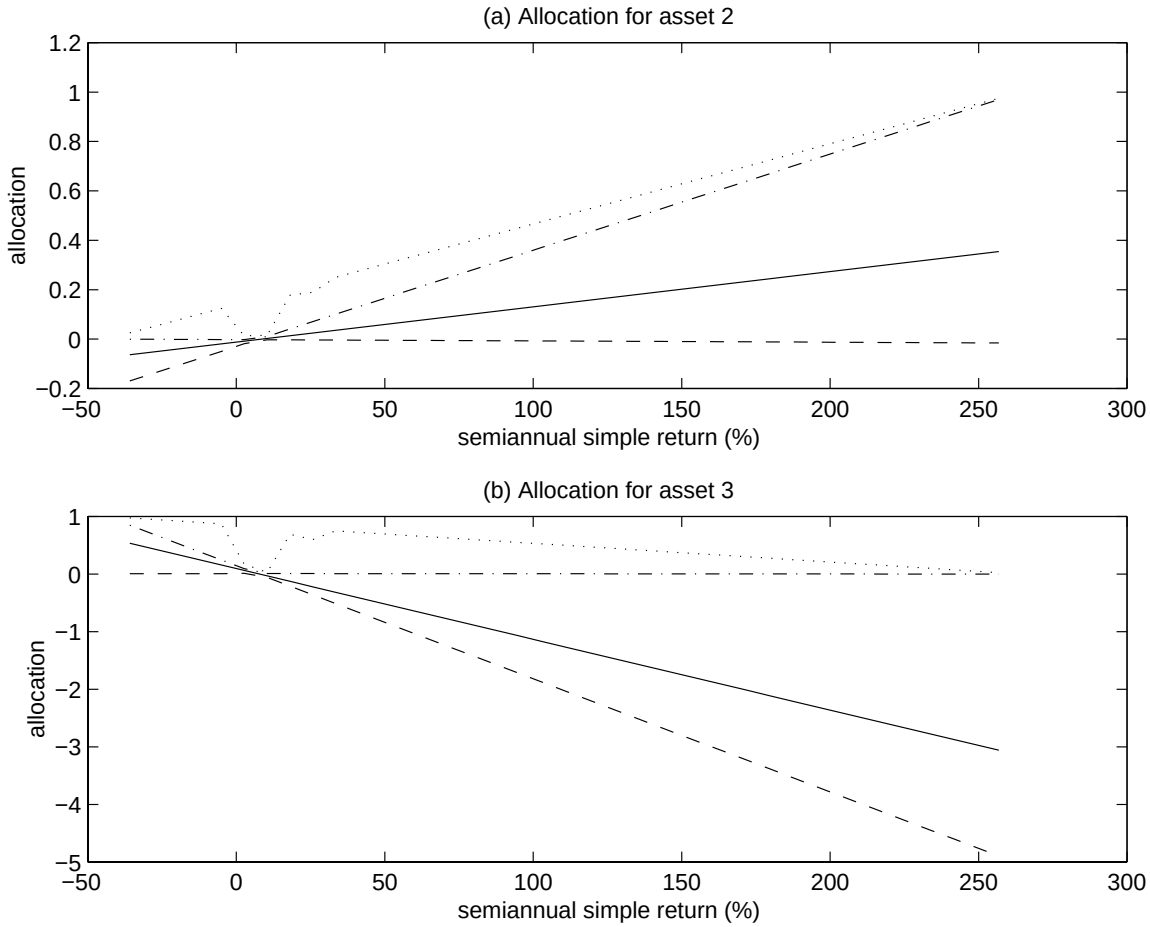

Figure 3-11: Weights of an optimal portfolio. Put-Call data. (no risk-free asset, shortsales allowed)

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 2: Put. asset 3: Call.

Example where there is no mutual fund separation as no risk-free asset is available; only the mean-variance optimal weights generate the mutual fund separation.

obtained by the four methods is quite similar. In 3-13, the optimal weights are plotted for only two of the stocks, Anheuser-Busch and Caterpillar. The mutual fund separation is apparent for the mean-variance and the mean-semivariance methods; the mean-shortfall and the mean-VaR deviate from the linearity, particularly the latter, which appears to be very noisy. Again, the mean-shortfall method appears to be superior than the mean-VaR method. In this case, as practitioners often assume for real stocks, the mean-variance method is preferred, given that quadratic optimization algorithms are very efficient.

Example 3.5.7 *No shortsales allowed, no risk-free asset.* The mutual fund separation does not hold when no shortsales are allowed (as mentioned by [32]). Figure 3-14, shows the optimal weights for this particular case. It is clear that the plotted optimal weights follow a non-linear curve and we see that even the mean-variance optimal weights are not linear. The additional constraints make the analysis of the behavior of optimal portfolios difficult, but correspond to practical conditions required by some investors.

3.6 Numerical optimization with noise

In practical cases, it is usual to estimate the gradient an to have an uncontrollable error $\mathbf{v}$. In the following case, we solve the following minimization problem:

$$\min \rho(\mathbf{x}) + \mathbf{v}'\mathbf{x} \quad \text{s.t.} \quad E[\tilde{W}(\mathbf{x})] = W_p.$$

(as explained in section A.3.2). Equation (3.19) becomes

$$E[\tilde{\mathbf{b}}] - b_f\mathbf{1} = \frac{\nabla_{\mathbf{x}}\rho(\mathbf{x}^*) + \mathbf{v} - c_f\mathbf{1}}{\rho(\mathbf{x}^*) + \mathbf{v}'\mathbf{x} - c_f\mathbf{1}'\mathbf{x}^*}(b_p - b_f).$$

where $b_p = W_p/W_0$, while equation (3.23), the efficient frontier, becomes

$$\rho(\mathbf{x}^*) = \left(\frac{\frac{\partial\rho(\mathbf{x}^*)}{\partial x_j} - c_f + v_i}{E[\tilde{b}_j] - b_f}\right)(W_p - W_0 b_f) + c_f\mathbf{1}'\mathbf{x}^* - \mathbf{v}'\mathbf{x}^*.$$

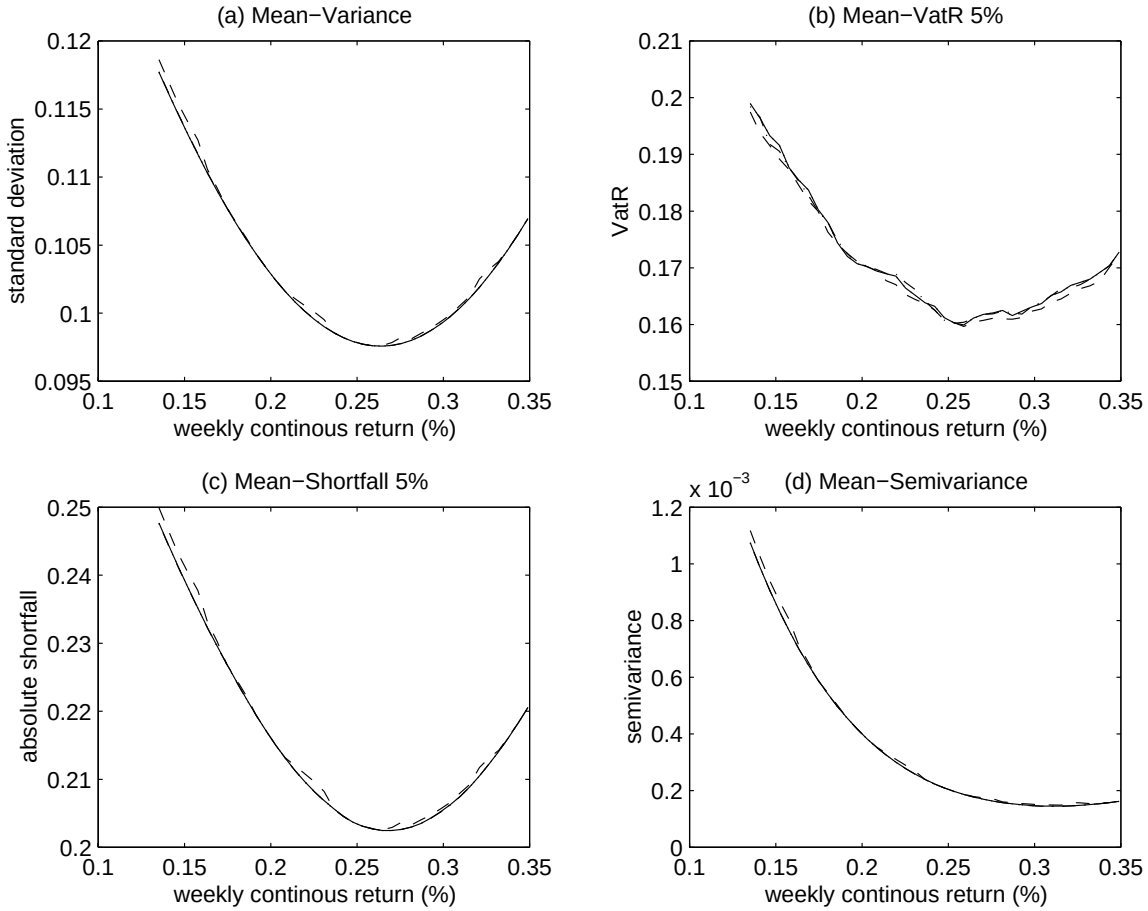

Figure 3-12: Efficient frontiers. Stock data. (no risk-free asset, shortsales allowed)

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

Example where the efficient frontiers are built using historical stock data. The result is very similar to the Gaussian case.

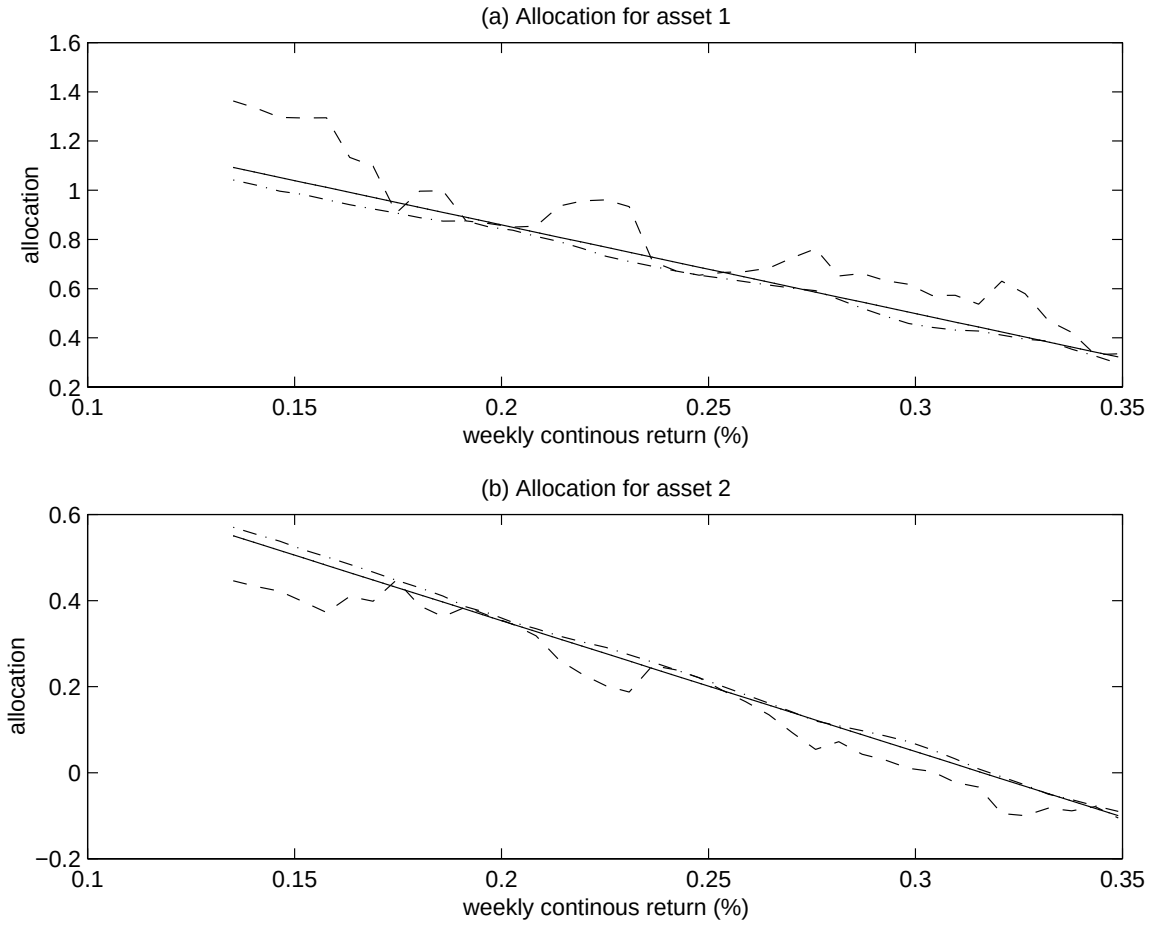

Figure 3-13: Weights of optimal portfolios. Stock data. (no risk-free asset, shortsales allowed)

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 1: Anheuser-Busch stock returns. asset 2: Caterpillar stock returns.

Example where the mutual fund separation is built using historical stock data. The deviation from a line is probably due to numerical errors, as in the Gaussian case.

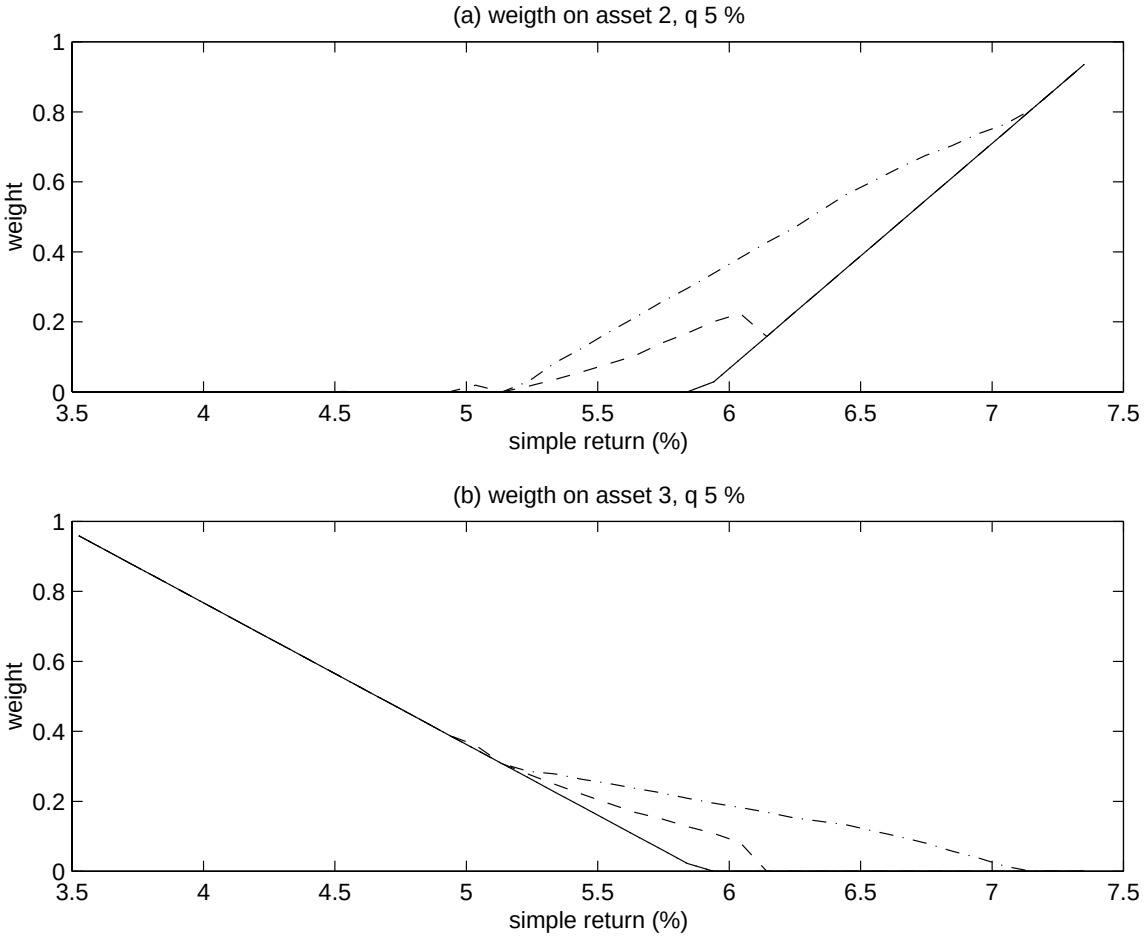

Figure 3-14: Weights of optimal portfolios. Option-based strategies data (no shortsales).

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 2: "long-put" at 50 %. asset 3: "write-call" at 100 %.

Example of violation of the mutual fund separation theorem, in which the weights are not allowed to be negative.

If we assume the risk measure $\rho(\mathbf{x})$ is convex, we can generate efficient portfolios by solving the unconstrained problem

$$\min \rho(\mathbf{x}) - \lambda(\mathbf{x}'E[\tilde{\mathbf{b}}] + -\mathbf{x}'\mathbf{1}b_f),$$

which becomes

$$\min \rho(\mathbf{x}) - \lambda\left(\mathbf{x}'E[\tilde{\mathbf{b}}] - \mathbf{x}'\mathbf{1}b_f - \frac{1}{\lambda}\mathbf{x}'\mathbf{v}\right).$$

Noise in the optimization has some significant impact when λ is small. However, in the risk-free case, the impact is less pronounced as the solution is $\mathbf{x}_f = \mathbf{0}$ (i.e., all the initial wealth is invested in the risk-free asset, see section 3.5.1) as λ approaches 0.

Chapter 4

Risk Gradient: definitions, properties and estimation

In this chapter we fully analyze the characteristics of the gradients of quantiles from their derivation to their estimation, and address their application to risk measures based on quantiles. In section 4.1 we review the use of the gradient of the risk as an analysis tool, and justify the use of an α-level specific analysis relying on empirical evidence of higher order models of the CAPM.

In section 4.2 we derive a general formula for the gradient of quantile functions when the random variable is the result of a linear combination of a random vector. This is a new result and will be very helpful for the analysis and optimization of portfolios (described in Chapter 5).

In many practical problems the distribution of the random vector is unknown, or very difficult to parameterize; thus we will use a set of n i.i.d. data samples, either from past observations or from Monte Carlos simulations. In section 4.3 we review different estimation techniques of the gradient of a quantile, such as the parametric method and the finite difference method (section 4.3), and also analyze the characteristics of the estimation errors. The error induced by the gradient estimators is particularly significant because during the numerical optimization the bias introduced by the estimator will also lead to errors in the optimization problem solution.

We will propose and analyze a technique based on local polynomial regression which estimates the gradient of a quantile function based on the formula developed in 4.2. An alternative representation (F-transformations) that is computationally more attractive is discussed in section 4.4.1. In section 4.4.2 we interpret the output of the local polynomial regression, and analyze the error introduced when the technique is applied to a gradient estimation; some practical considerations are also reviewed. An alternative interpretation of the F-transformation is analyzed in section 4.4.3. The different methods to estimate the gradient are analyzed from a complexity point of view in section 4.4.4. Finally, an estimator of the gradient of the empirical shortfall is introduced in section 4.4.5.

4.1 The risk gradient

The gradient of a $risk(\mathbf{x})$ measure with respect to the portfolio $\mathbf{x}$ is denoted as $\nabla_{\mathbf{x}} risk(\mathbf{x})$. For example, for the VaR_α we have the closed form solution:

$$\nabla_{\mathbf{x}} VaR(\mathbf{x}) = E[\tilde{\mathbf{b}}] - E\{\tilde{\mathbf{b}} | \mathbf{x}'\tilde{\mathbf{b}} = q_\alpha(\mathbf{x})\}. \tag{4.1}$$

The previous equation was derived from equation (3.3), the definition of the VaR, and equation (4.7), the gradient of a quantile function, a formula to be derived in section 4.2.

The *risk* gradient with respect to $\mathbf{x}$ has been proposed as a tool for practitioners [28]. Using $\nabla_{\mathbf{x}} risk(\mathbf{x})$, it is possible to obtain an approximation of the $risk(\mathbf{x}+\mathbf{a})$, or the risk of a new portfolio $\mathbf{x}$ perturbed by a small portfolio $\mathbf{a}$ of total initial wealth A_0. Assuming that $A_0 << W_0$, the following holds:

$$risk(\mathbf{x}+\mathbf{a}) - risk(\mathbf{x}) \approx \mathbf{a}'\nabla_{\mathbf{x}} risk(\mathbf{x}); \tag{4.2}$$

$\mathbf{a}'\nabla_{\mathbf{x}} risk(\mathbf{x})$ returns the variation in *risk*. Each component of $\nabla_{\mathbf{x}} risk$ can be thought of as a decomposition of the *risk*. The basic idea is to understand the effect of a trade $\mathbf{a}$ on the overall *risk* of the portfolio. Using this technique, traders can analyze the effect of a single trade on a portfolio. $\nabla_{\mathbf{x}} risk$ is also an important component of the generalized β

that we defined in equation (3.20), and can be used extensively to analyze the sensitivities of portfolios not only to trades, but also to the asset prices.

Example 4.1.1 *Gradient of VaR.* For the case when the Gaussian data described in C.4.1 is used with the normalized portfolio $\mathbf{x} = [0.0667, 0.6000, 0.3333]'$ and $\alpha = 0.05$, the parametric gradient using equation ((A.9)) is:

$$\nabla VaR(\alpha, \mathbf{x}) = \begin{bmatrix} 0.0026 \\ 0.0143 \\ 0.0124 \end{bmatrix}.$$

A practitioner would consider the first asset as the least risky for the portfolio, while the second asset appears as the most risky (as defined by the VaR). Hence, the practitioner could decide a certain trading strategy based on this observation, e.g., increasing the amount invested in the first asset and decreasing the amount of the second asset.

Risk decomposition The homogeneity of risk measures allows the decomposition of risk component-wise, since for a homogeneous measure of risk $\rho(\mathbf{x})$:

$$\rho(\mathbf{x}) = \sum_{j=1}^{m} x_j \frac{\partial}{\partial x_j} \rho(\mathbf{x}), \tag{4.3}$$

so that the contribution of the asset j to the portfolio's risk $\rho(\mathbf{x})$ is $x_j \frac{\partial}{\partial x_j} \rho(\mathbf{x})$.

Generalized $\breve{\beta}$. The generalized $\breve{\beta}_j(\mathbf{x})$ defined in equation (3.20) can be obtained for homogeneous measures of risk such as the standard deviation, the VaR and the shortfall. In the case of the last two, the formula becomes a function of the level α, e.g., for the VaR we have

$$\breve{\beta}_{j\alpha}(\mathbf{x}) = \frac{\frac{\partial VaR_\alpha(\mathbf{x})}{\partial x_j}}{VaR_\alpha(\mathbf{x})} \qquad \text{for} \qquad j = 1, \ldots, m, \tag{4.4}$$

and for the shortfall:

$$\breve{\beta}_{j\alpha}(\mathbf{x}) = \frac{\frac{\partial e_\alpha(\mathbf{x})}{\partial x_j}}{e_\alpha(\mathbf{x})} \qquad \text{for} \qquad j = 1, \ldots, m. \tag{4.5}$$

Generalized $\breve{\beta}$ and higher moments CAPM models. From a practical point of view, it indeed makes sense to consider risk-level-specific $\breve{\beta}_\alpha$, i.e., to allow β to vary at different risk levels α. Such risk-level-specific β is the quantitative tool to measure the empiric phenomena that the market components (or portfolio stocks) become more dependent on the market as it gets more volatile, and less dependent as it stabilizes.

This interpretation agrees with the research on non-normality of financial assets (section 2.2.1), which is reflected in the cubic market model (equation (2.4)). The higher moment market model can be used to estimate the effect of a realized movement in the market r_k on the return of a particular stock, by disregarding the noise variable ϵ_j;

$$\hat{r}_j = \alpha_j + \beta_j r_k - \gamma_j r_k^2 + \delta_j r_k^3 \qquad j = 1, \cdots, m \tag{4.6}$$

Although β has been traditionally defined in terms of covariances and variances, it is interesting to note that a *generalized* $\breve{\beta}_j$ can also be defined in terms of the derivative of $\frac{\partial \hat{r}_j}{\partial r_k}$. For the case of the higher moment market model, the generalized $\breve{\beta}_j$ would be

$$\breve{\beta}_j(r_k) = \beta j - 2\gamma_j r_k + 3\delta_j r_k^2.$$

Since each possible outcome r_k can be associated with a probability α such that $\Pr[\tilde{r}_k \leq r_k] = \alpha$, at each level α it is possible to compute $\breve{\beta}_{j\alpha}$ by finding the r_k corresponding to α (using the quantile function). For higher moment models, it is evident that the value of $\breve{\beta}_{j\alpha}$ will be a function of α, unlike the classical CAPM model in which $\breve{\beta}_j = \beta_j$ for all levels of r_k.

This phenomena cannot be captured by either normal or elliptically symmetric distributions, which keep $\beta_{j\alpha}$ constant over α.

4.2 The gradient of quantile functions

In appendix A.2.2 is derived a closed form formula (A.8) for the gradient of a quantile function. Still, it is still possible to derive a non-parametric expression for the gradient of

the quantile.

Theorem 4.2.1 *Assuming* $\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t|\tilde{b}_j = b_j)$ *is differentiable with respect to* x_j*, for* $j = 1, \ldots, m$*, and the conditions of theorem A.1.1 are enforced (see remark (i) at the end of the proof), the following expression for the gradient of* $q_\alpha(\mathbf{x})$ *is true:*

$$\nabla_{\mathbf{x}}\ q_\alpha(\mathbf{x}) = E(\tilde{\mathbf{b}}|\mathbf{x}'\tilde{\mathbf{b}} = q_\alpha(\mathbf{x})). \tag{4.7}$$

Proof. Differentiating the identity

$$\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq q_\alpha(\mathbf{x})) = \alpha$$

with respect to one component of $\mathbf{x}$, e.g. x_1, we get, denoting $f_{\mathbf{x}'\tilde{\mathbf{b}}}(\cdot)$ the density of $\mathbf{x}'\tilde{\mathbf{b}}$,

$$\Big[\frac{\partial}{\partial x_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t)\Big]_{t=q_\alpha(\mathbf{x})} + f_{\mathbf{x}'\tilde{\mathbf{b}}}(q_\alpha(\mathbf{x}))\frac{\partial q_\alpha(\mathbf{x})}{\partial x_1} = 0. \tag{4.8}$$

Now, conditioning on b_1, we can write

$$\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t) = E_{b_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t|b_1),$$

and, if $\Pr(\mathbf{x}'\mathbf{b} \leq t|\tilde{b}_1 = b_1)$ is differentiable with respect to x_1, then differentiation and expectation can be interchanged, by theorem A.1.1:

$$\frac{\partial}{\partial x_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t) = \frac{\partial}{\partial x_1}E_{b_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t|b_1) = E_{b_1}\frac{\partial}{\partial x_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t|b_1).$$

The last derivative should be computed as follows

$$\frac{\partial}{\partial x_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t|\tilde{b}_1 = b_1) = \lim_{\Delta x_1 \to 0}\frac{\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t - \Delta x_1 b_1|\tilde{b}_1 = b_1) - \Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t|\tilde{b}_1 = b_1)}{\Delta x_1}$$

$$= f_{\mathbf{x}'\tilde{\mathbf{b}}}(t|b_1)(-b_1),$$

so that

$$\frac{\partial}{\partial x_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t) = E_{\tilde{b}_1}[f_{\mathbf{x}'\tilde{\mathbf{b}}|\tilde{b}_1}(t|\tilde{b}_1)(-\tilde{b}_1)] = -\int b_1 f_{\mathbf{x}'\tilde{\mathbf{b}}|\tilde{b}_1}(t|b_1) f_{\tilde{b}_1}(b_1) db_1$$

$$= -\int b_1 f_{\mathbf{x}'\tilde{\mathbf{b}},\tilde{b}_1}(t, b_1) db_1.$$

Plugging this result in (4.8), we get

$$\frac{\partial q_\alpha(\mathbf{x})}{\partial x_1} = E(\tilde{b}_1|\mathbf{x}'\tilde{\mathbf{b}} = q_\alpha(\mathbf{x})),$$

and repeating the argument for each component of $\mathbf{x}$ we get (4.7). □

Remark (i) The theorem A.1.1 assumes

$$\frac{\partial}{\partial x_1}\Pr(\mathbf{x}'\tilde{\mathbf{b}} \leq t|\tilde{b}_1 = b_1)$$

is continuous for x_1 in a closed interval; this is not be valid for discrete distributions, since they can introduce discontinuities. However, as long as the derivative is continuous within the interval of interest, the theorem will be valid.

Remark (ii) The same expression could be obtained as a special case of a theorem formulated by Pflug [49], although he introduces special constraints to generalize his theorem.

Remark (iii): The gradient of the shortfall can be derived from the gradient of the quantile:

$$\nabla_{\mathbf{x}} e_\alpha(\mathbf{x}) = E\{\tilde{\mathbf{b}}|\mathbf{x}'\tilde{\mathbf{b}} = q_\alpha(\mathbf{x})\} - \frac{1}{\alpha}\int_0^\alpha E\{\tilde{\mathbf{b}}|\mathbf{x}'\tilde{\mathbf{b}} = q(\omega, \mathbf{x})\} d\omega.$$

4.3 Estimation of gradients

We know that the estimator $\hat{q}_{\alpha,n}$ is only asymptotically unbiased. We can model q_α as the sum of the estimator plus a random error η; $q_\alpha = \hat{q}_{\alpha,n} + \eta,$. The error $\eta = \eta(n, \alpha, \mathbf{x})$ depends on the number of samples, the value α and the vector $\mathbf{x}$, as described in equation (A.13), and only asymptotically unbiased when the estimator has zero mean.

Parametric formulas In the particular case when it is known that the distribution of $\tilde{\mathbf{b}}$ has an elliptic form, all we need is to estimate the matrix $\mathbf{V}$ and the mean vector μ respectively as $\hat{\mathbf{V}}_n$ and $\hat{\mu}_n$, using the n samples available. The estimators found are substituted for the real (unknown) values in the formulas described in A.2.2. The use of finite samples introduce an error, but as n grows by the Law of Large Numbers the gradient estimator converges to the theoretical gradient [49].

Finite differences

The case of finite differences for unbiased estimates of a function is discussed thoroughly in [49]. However, the discussion assumes unbiased estimates of the function are available, although in the case of quantiles we can only assume that the estimates will be asymptotically unbiased. In that case, we will have to assume that n is sufficiently large to generate estimates with a negligible bias, so that the error η is almost zero mean.

The finite difference approximation for the i-th component of the gradient vector is (assuming α is constant)

$$D_{fd,i}(\mathbf{x}, c, n) = \frac{1}{2c}(\hat{q}_{\alpha,n}(\mathbf{x} + c\mathbf{e}_i) - \hat{q}_{\alpha,n}(\mathbf{x} - c\mathbf{e}_i)).$$

The bias error is

$$b_{fd,i}(\mathbf{x}, c, n) = \frac{q_\alpha(\mathbf{x} + c\mathbf{e}_i) - q_\alpha(\mathbf{x} - c\mathbf{e}_i)}{2c} - \frac{\partial q_\alpha(\mathbf{x})}{\partial x_i},$$

and will be small if c is small. The zero mean random error is

$$w_{fd,i}(\mathbf{x}, c, n) = \frac{\eta_{i,1} - \eta_{i,2}}{2c}$$

(recall that η is a function of $\mathbf{x}$ and n); the zero mean error has an unbounded variance if c tends to zero and $\eta_{i,1}$ is independent from $\eta_{i,2}$. The vector $\mathbf{D}_{fd}(\mathbf{x}, c, n)' = [D_{fd,1}(\mathbf{x}, c, n), \cdots, D_{fd,m}(\mathbf{x}, c, n)]$ denotes the finite difference gradient estimator, which is a biased estimator;

$$\nabla_{\mathbf{x}} q_\alpha(\mathbf{x}) = \mathbf{D}_{fd}(\mathbf{x}, c, n) + \mathbf{v}_{fd}(\mathbf{x}, c, n).$$

The vector error $\mathbf{v}_{fd}(\mathbf{x}, c, n)$ is formed by the bias

$$\mathbf{b}_{fd}(\mathbf{x}, c, n))' = [b_{fd,1}(\mathbf{x}, c, n), \cdots, b_{fd,m}(\mathbf{x}, c, n)],$$

and the zero mean errors $\eta'_j = [\eta_{1,j}, \cdots, \eta_{m,j}]$ for $j = 1, 2$:

$$\mathbf{v}_{fd}(\mathbf{x}, c, n) = \mathbf{b}_{fd}(\mathbf{x}, c, n) + \frac{\eta(\mathbf{x}, n)_1 - \eta(\mathbf{x}, n)_2}{2c}. \tag{4.9}$$

Asymptotically, as n goes to infinity and c goes to 0, the vector error will tend to zero. However, in small samples the error $\mathbf{v}$ is quite large. A confidence interval for the size of the errors η_j for $j = 1, 2$ can be obtained by using the equation $(A.14)$. A good choice of the parameter c is essential: a large value of c increases the bias, but a small value increases the variance of the zero mean error.

4.4 Estimation of gradients using local polynomial regression

In section 4.2 we derived a theoretical formula of the gradient. In practice, there are situations when a set of independent samples from a multivariate distribution (i.e., historical data, or Monte-Carlo simulation) is available; we assume they are n i.i.d. samples of the random vector $\tilde{\mathbf{b}}$. As described in A.2.4, we can obtain M-order statistics $\mathbf{b}_{i:n}$, $i = 1, \cdots, n$ associated in a particular linear transformation $\tilde{W}(\mathbf{x})$. For the rest of this section we assume that we are given a fixed vector $\mathbf{x}$.

The estimation of equation (4.7) is similar to the estimation of a regression function for $W \equiv \tilde{W}(\mathbf{x})$, which is explained in more detail in appendix A.4.1. We use the techniques described there to find an estimator $\hat{\nabla}_{\mathbf{x}} q_\alpha(\mathbf{x})$ of $\nabla_{\mathbf{x}} q_\alpha(\mathbf{x})$; because it is a function of the bandwidth parameter h, we denote the estimator as $\mathbf{D}_{lpr}(\mathbf{x}, h)$.

Consider now one component of the gradient vector. We can express the partial of $q_\alpha(\mathbf{x})$

with respect to one component x_i of $\mathbf{x}$ as

$$\frac{\partial q_\alpha(\mathbf{x})}{\partial x_i} = E\{\tilde{b}_i | \mathbf{x}'\tilde{\mathbf{b}} = q_\alpha(\mathbf{x})\},$$

where $\tilde{b}_i$ denotes the i-th component of the vector $\tilde{\mathbf{b}}$. We can interpret the regression function for one component, $m_W(\tilde{W}(\mathbf{x}) = q_\alpha(\mathbf{x})) = E(\tilde{b}_i|\tilde{b}(\mathbf{x}) = q_\alpha(\mathbf{x}))$ as if the data were being generated from the model

$$\tilde{b}_i = m_W(\tilde{W}(\mathbf{x})) + \sigma_W(\tilde{W}(\mathbf{x}))\omega,$$

where $E(\omega) = 0$, $\sigma_W(\omega) = 1$, and $\tilde{\mathbf{b}}$ and ω are independent. Because $m_W(\tilde{W}(\mathbf{x}) = q_\alpha)$, the value of the regression function evaluated at q_α, is equal to $\partial q_\alpha(\mathbf{x})/\partial dx_i$, we can use the local polynomial regression to estimate the gradient.

With finite samples, we can compute an estimate of $m_W(\tilde{W}(\mathbf{x}))$ by pairing each vector $\mathbf{b}_{i:n}$ with a scalar $\mathbf{x}'\mathbf{b}_{i:n}$. If we use the estimator $\hat{q}_{\alpha,n}$, we introduce an additional error to the estimation of the gradient; we are estimating $\hat{m}_W(\tilde{W}(\mathbf{x}) = q_\alpha + \eta)$ instead of $\hat{m}_W(\tilde{W}(\mathbf{x}) = q_\alpha)$,

Defining the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ as

$$\mathbf{X}_1 = \begin{bmatrix} 1 & \mathbf{x}'\mathbf{b}_{1:n} - \hat{q}_{\alpha,n} \\ \vdots & \\ 1 & \mathbf{x}'\mathbf{b}_{n:n} - \hat{q}_{\alpha,n} \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} 1 & \mathbf{x}'\mathbf{b}_{1:n} - \hat{q}_{\alpha,n} & (\mathbf{x}'\mathbf{b}_{1:n} - \hat{q}_{\alpha,n})^2 \\ \vdots & & \\ 1 & \mathbf{x}'\mathbf{b}_{n:n} - \hat{q}_{\alpha,n} & (\mathbf{x}'\mathbf{b}_{n:n} - \hat{q}_{\alpha,n})^2 \end{bmatrix},$$

or if necessary, matrices $\mathbf{X}_3$ or larger, following the notation in the appendix section $A.4.1$; also

$$\mathbf{W}_h = \text{diag}\left(\frac{1}{h}K\left(\frac{\mathbf{x}'\mathbf{b}_{j:n} - \hat{q}_{\alpha,n}}{h}\right), j = 1, \cdots, n\right),$$

where h is known as the bandwidth of the kernel, and

$$\mathbf{Y} = \begin{bmatrix} \mathbf{b}'_{1:n} \\ \vdots \\ \mathbf{b}'_{n:n} \end{bmatrix}, \mathbf{Y}_j = \begin{bmatrix} b_{j,1:n} \\ \vdots \\ b_{j,n:n} \end{bmatrix}, j = 1, \cdots, m, \tag{4.10}$$

where $b_{j,i:n}$ is the j-th component of the M-order statistic vector $\mathbf{b}_{i:n}$. In some cases there is no need to do a local polynomial regression; for instance, the linear CAPM model (section 2.2.1) suggests that a global linear regression is sufficient.

Experiments (see section 4.4.6) indicate that this method of estimation tends to be the most accurate. However, it is necessary to compute the matrices $\mathbf{X}_i$ and $\mathbf{W}_h$ every time $\mathbf{x}$ changes. This can be a drawback for some applications that require the computation of the gradient a large number of times, as in portfolio optimization, In the next section we introduce a slightly different way of posing the estimation, which does not require the updating of the matrices $\mathbf{X}_i$ and $\mathbf{W}_h$ (unless h changes).

4.4.1 The F-transformation

Another estimation technique of the conditional expectation can be derived from

$$E\{\tilde{\mathbf{b}}|\mathbf{x}'\tilde{\mathbf{b}} = q_\alpha(\mathbf{x})\} = E\{\tilde{\mathbf{b}}|F(q_\alpha(\mathbf{x})) = \alpha\}.$$

Expressing the partial of $q_\alpha(\mathbf{x})$ with respect to one component x_i of $\mathbf{x}$ as a conditional expectation (from the equation (4.7));

$$\frac{\partial q_\alpha(\mathbf{x})}{\partial x_i} \quad = \quad E\{\tilde{b}_i|F(q_\alpha(\mathbf{x})) = \alpha\}.$$

Defining $z \equiv F(q_\alpha(\mathbf{x}))$, we will have a different implicit model for $E(\tilde{b}_i|z = \alpha)$; the new model is

$$y = m_F(z) + \sigma_F(z)\omega$$

Similarly to the previous section, $m_F(\alpha)$ corresponds to $\partial q_\alpha(\mathbf{x})/\partial dx_i$. The price pay for using this model is the use of a more complex modeling function $m_F(\cdot)$.

With finite samples we can compute an estimate of $m_F(z)$ by pairing each vector $\mathbf{b}_{i:n}$ with a scalar z_i. F can be empirically estimated as F_n using equation (A.12); since we are already using order statistics each vector $\mathbf{b}_{i:n}$ is paired with the scalar $z_i = i/n$.

We can construct the matrices $\mathbf{X}_j$, $j = 1, 2$ (or larger)

$$\mathbf{X}_1 = \begin{bmatrix} 1 & z_1 - \alpha \\ \vdots & \\ 1 & z_n - \alpha \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} 1 & z_1 - \alpha & (z_1 - \alpha)^2 \\ \vdots & & \\ 1 & z_n - \alpha & (z_n - \alpha)^2 \end{bmatrix},$$

(the matrix $\mathbf{Y}$ is the same as in (4.10), and

$$\mathbf{W}_h = \text{diag}\left(\frac{1}{h} K\left(\frac{z_j - \alpha}{h}\right), j = 1, \cdots, n\right),$$

Although the underlying regression model seems more complex, and is not linear, in practice it is easier to implement since we are not required to estimate q_α, and since the matrices $\mathbf{X}_i$ and $\mathbf{W}_h$ are independent from the data. Still, we need to find an adequate h.

4.4.2 Local polynomial regression

The following procedure applies regardless of the underlying model used. Once the matrices $\mathbf{X}_i$ and $\mathbf{W}_h$ are available, the weighted least squares problem (A.33) for the j-th column of the gradient can be written as

$$\min_{\beta_j} (\mathbf{Y}_j - \mathbf{X}\beta_j)' \mathbf{W}_h (\mathbf{Y}_j - \mathbf{X}\beta_j),$$

with $\beta_j = [\beta_{j,0}, \cdots, \beta_{j,p}]'$, $\mathbf{Y}_j$ is the j-th column of the matrix $\mathbf{Y}$, and $\mathbf{X}$ could be any of the two matrices $\mathbf{X}_1$ or $\mathbf{X}_2$. The solution vector is provided by the ordinary least squares theory and is given by

$$\hat{\beta}_j = (\mathbf{X}' \mathbf{W}_h \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}_h \mathbf{Y}_j.$$

The estimator of the partial derivative for the j-th component of the gradient is

$$\frac{\partial q_\alpha(\mathbf{x})}{\partial x_j} \approx \hat{\beta}_{j,0}, \tag{4.11}$$

where h sets the bandwidth of the local regression.

We denote the estimated gradient in local polynomial regression as the vector

$$\mathbf{D}_{lpr}(\mathbf{x}, h, n)' = [\hat{\beta}_{1,0}, \cdots, \hat{\beta}_{m,0}].$$

Depending on the kernel, the bandwidth decides how many order statistics to use. Procedures to estimate the bandwidth are detailed in appendix A.4.6.

Each component of the gradient should be estimated using the corresponding optimal bandwidth if we were to rigorously estimate the gradient.

Error vector for finite samples The gradient estimator using local polynomial regression is also a biased estimator of the real gradient:

$$\nabla_{\mathbf{x}} q_\alpha(\mathbf{x}) = \mathbf{D}_{lpr}(\mathbf{x}, h, n) + \mathbf{v}_{lpr}(\mathbf{x}, h, n).$$

The error vector $\mathbf{v}_{lpr}(\mathbf{x}, h, n)$ behaves very differently from the error vector $\mathbf{v}_{fd}(\mathbf{x}, c, n)$; the bias $\mathbf{b}_{lpr}(\mathbf{x}, h, n)$ due to modeling error increases as h increases, *but* the zero mean error vector $\mathbf{w}_{lpr}(\mathbf{x}, h, n)$ decreases. For finite samples, the choice is between a biased estimator or a zero mean noisy estimator. As n increases, by the law of large numbers, the gradient estimator improves.

Practical considerations In practice (e.g. for nonlinear optimization), it may be necessary to compute the estimator in the fastest way possible. It might be useful to compute all the components of the gradient at once, with a crude pilot bandwidth h which is the same for all the components:

$$\hat{\beta} = (\mathbf{X}_1' \mathbf{W}_h \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{W}_h \mathbf{Y}, \tag{4.12}$$

with the matrix $\hat{\beta} = [\hat{\beta}_0, \cdots, \hat{\beta}_p]'$, and the estimator

$$\hat{\nabla}_{\mathbf{x}} q_\alpha(\mathbf{x}) = \hat{\beta}_0. \tag{4.13}$$

In the appendix section A.4 the complete description of the local polynomial regression is described, as developed in [25].

4.4.3 Alternative interpretation of the F-transformation

For the local linear regression in the case of the F-transformation we have an alternative interpretation. To simplify the notation, let us define for a given fixed $\mathbf{x}$ the function

$$\mathbf{G}(\alpha) \equiv E\{\tilde{\mathbf{b}} \mid F(q_\alpha(\mathbf{x})) = \alpha\};$$

a Taylor's expansion of first order $\mathbf{G}(\alpha + \varepsilon)$ around α yields

$$\mathbf{G}(\alpha + \varepsilon) \approx \mathbf{G}(\alpha) + \varepsilon \frac{d\mathbf{G}(\omega)}{d\omega} \mid_{\omega=\alpha} .$$

This approximation is true for small values of ε; the accuracy as a function of ε depends on the characteristics of the function $\mathbf{G}(\alpha)$. We want to estimate $\mathbf{G}(\alpha)$ using an estimator $\hat{\mathbf{G}}(\alpha)$. A very crude estimator to use is the M-order statistic:

$$\hat{\mathbf{G}}(\alpha) = \mathbf{b}_{k:n}.$$

The vector $\mathbf{b}_{k:n}$ is a sample of the conditional distribution $f_{\tilde{\mathbf{b}}|\tilde{W}(\mathbf{x})}(\tilde{\mathbf{b}}|\hat{q}_{\alpha,n})$ where $\hat{q}_{\alpha,n} = q_\alpha + \eta)$. Hence this estimator has a bias vector γ which satisfies $\mathbf{x}'\gamma = \eta$;

$$\mathbf{G}(\alpha) = \hat{\mathbf{G}}(\alpha) + \gamma + \kappa,$$

where $\mathbf{x}'\kappa$ is a zero mean error with a variance of the distribution $f_{\tilde{\mathbf{b}}|\tilde{W}(\mathbf{x})}(\tilde{\mathbf{b}}|\hat{q}_{\alpha,n})$. However, due to the Taylor's expansion, estimators for $\mathbf{G}(\alpha)$ that use values of $\mathbf{G}(\omega)$ for ω close to α are of the form

$$\mathbf{G}(\alpha) \approx \mathbf{G}(\alpha + \varepsilon) - \varepsilon \frac{d\mathbf{G}(\omega)}{d\omega} \mid_{\omega=\alpha};$$

knowing $\varepsilon \frac{d\mathbf{G}(\omega)}{d\omega} \mid_{\omega=\alpha}$ we can compensate for the bias introduced by η. Furthermore, we can use other samples close to α, e.g. $\mathbf{b}_{w:n}$ for $w = k \pm l$, for a small integer l, and average them to get a better estimator $\hat{\mathbf{G}}\alpha$.

That is precisely what we are doing while fitting linearly an F-model. A one sample estimator will correspond to a local fit with a very small bandwidth, such that only one M-

order statistic is used, and as we increase the bandwidth the number of M-order statistics is increased, hopefully *reducing the bias and variance* of the estimator $\hat{\mathbf{G}}(\alpha)$.

4.4.4 Algorithm operation counts

Let us assume the following case: we are given n vectors of size m (which we assume are i.i.d), and the vector $\mathbf{x}$. We can form the matrix $\mathbf{Y}$ as in (4.10). To compute a gradient from "scratch", we require for the following three cases:

Parametric case:

i. $O(m^2 n)$ operations to compute the covariance matrix $\mathbf{\Sigma}$.

ii. $O(m^2)$ operations to compute $\mathbf{x}'\mathbf{\Sigma}\mathbf{x}$.

Some implementations [28] assume that those values are already computed elsewhere, and therefore the implementation issues are negligible.

Finite difference

i. $O(m^2 n)$ operations to compute $2m$ "perturbed" vectors $\mathbf{x}_p$.

ii. $O(m \log n)$ operations to estimate the quantile of all perturbed vectors.

Forward differences can be used to reduce the number of operations needed, although that will increase the error of the gradient.

Local polynomial regression

For a given initial bandwidth g:

i. $\mathbf{n}$ "kernel operations" to compute the matrix $\mathbf{W}_g$. However, the kernel operations might be expensive to compute; we denote as c the fraction of elements of $\mathbf{W}_g$ which are of significant order.

ii. $O(nm)$ arithmetic operations to compute $\mathbf{Y}\mathbf{x}$ and $O(n \log n)$ sort it.

iii. $O(mc + c^2)$ operations to compute the local polynomial regression.

To find the optimal bandwidth, we require to

i. do all the previous computations Q times, the number of bandwidths (h) in the grid; (see A.4.6).

The interesting aspect for the local polynomial regression is that is it not of order $O(m^2)$, so its computation will not grow quadratically as the m grows. However, it requires the computation of an optimal bandwidth; we see that it is necessary to perform all the computations Q times, one for each different bandwidth tested. As h increases, the number of samples to be used (c) increases and $O(mc + c^2)$ becomes very expensive to compute. Under certain practical circumstances a crude bandwidth might be sufficient (e.g. the rule of thumb). Also, if c is significantly smaller than n, it might be very fast to compute.

The use of F-based models also offers many advantages in the practical implementation of local polynomial regressions: because all the data points z are located on a grid, the number of kernel operations can be computed and stored in memory [25].

4.4.5 Gradient Estimator for the empirical shortfall

Instead of estimating directly the shortfall, we need to estimate the gradient of the empirical shortfall:

$$\hat{\nabla}_{\mathbf{x}} e_n(\alpha) = \frac{1}{\alpha} \sum_{i=1}^{k-1} \left(\hat{\nabla} q_\alpha(\mathbf{x}) - \hat{\nabla} q_{i/n}(\mathbf{x}) \right),$$

where $k = \alpha n$, and each quantile gradient estimators can be obtained from the local polynomial regression technique, or we can use finite differences to estimate the gradient of the empirical shortfall (using formulas similar to the ones employed in section 4.3).

4.4.6 Experiments

For illustration purposes, we developed three examples that exemplify the computation of the gradient of a quantile. In the first two examples we use multivariate random variables

Table 4.1: Results for the multivariate Gaussian case.

Portfolio: $\mathbf{x} = [0.0667, 0.6000, 0.3333]'$; Gaussian kernel.
Methods: par. = parametric, LPR = local polynomial regression (LPR_F with F transformation), f.d. = finite differences. h = bandwidth for l.p.r., c = perturbation for f.d. ROT = rule of thumb, RSC = RSC squares criterion. a. = asset.

a.	method:	par.	$LPR_{F,ROT}$	$LPR_{F,RSC}$	LPR_{ROT}	LPR_{RSC}	f.d.	f.d.
	h,c		0.07207	0.02184	0.002559	0.01743	0.001	0.1
(1)	bias (%)	0.2135	8.997	5.302	1.328	2.366	11.74	9.283
	σ (%)	22	38.04	54.29	48.96	25.92	192.8	90.52
	h,c		0.05585	0.0625	0.004061	0.0625	0.001	0.1
(2)	bias (%)	-0.07373	6.461	6.872	0.402	1.202	0.4432	-1.14
	σ (%)	6.266	9.233	9.245	9.943	8.911	25.87	15.37
	h,c		0.1291	0.04989	0.00482	0.0625	0.001	0.1
(3)	bias (%)	0.03035	8.564	8.211	2.76	1.338	2.618	5.323
	σ (%)	7.703	11.85	12.74	12.44	9.508	50.31	23.96

where a parametric form is known, while in the third example we use investment assets with nonlinear payoffs, which yield joint returns where it is difficult to identify a predefined parametric form.

For these examples, we generated 200 sets of 200 samples of the random vectors.

Experiment 4.4.1 *Elliptic Multivariate Returns.* For the case when the t and the Gaussian data (described in C.4.1) are used with the normalized portfolio $\mathbf{x} = [0.0667, 0.6000, 0.3333]'$ and $\alpha = 0.05$, the parametric gradient that follows using equation (A.9) is:

$$\nabla VaR(\alpha, \mathbf{x}) = \begin{bmatrix} 0.0026 \\ 0.0143 \\ 0.0124 \end{bmatrix}.$$

Multivariate t distributions can exhibit "fat" tails, which some researchers have proposed as an alternate model for return distributions. The parametric gradient for the portfolio

$\mathbf{x} = [0.0667, 0.6000, 0.3333]'$ is:

$$\nabla VaR(\alpha, \mathbf{x}) = \begin{bmatrix} 0.0025 \\ 0.0117 \\ 0.0105 \end{bmatrix}.$$

The results of different gradient estimators are listed in the table 4.1 for the multivariate Gaussian case and in the table 4.2 for the non-normal case. The finite difference method performs very poorly, compared with any other method based on the variance around the mean value: gradient estimators are very noisy. The parametric method is (obviously) the best, although a gradient estimator using LPR_{RSC} with a bandwidth selected using the RSC bandwidth selector (see section A.4.5) performs very well too. Unfortunately, computing the RSC bandwidth selector is quite expensive, compared with other LPR computations.

We notice that the estimation error for the first component of the gradient, when $x_1 = .0667$ is the largest. This indicates that when there are many assets and each of their weights are small, the performance of the estimators will be poor, a consequence of the "curse of dimensionality". In figure 4-1 there are histograms and e.d.f.'s for the daily returns of the portfolio, for both the Gaussian and the t cases. In the t case, it is interesting to notice the "fat" tails; the probability of having "extreme" samples is higher than in the Gaussian case, and the shape of the histogram of a Monte-Carlo simulation depends on the number of samples used. As the size of samples is increased the tails of the empirical distribution become "fatter".

Experiment 4.4.2 *Non-parametric Returns.* See the section C.4.2 for the procedure used to obtain the data. For a normalized portfolio of $\mathbf{x} = [1/3, 1/3, 1/3]'$ and a quantile value $\alpha = 0.05$, the results are in the table 4.3. The bias is not directly measurable, so we do not present it. Still, all the components (in this case only 3) have approximately the same order; only the variance is smaller for the F modeling with the RSC bandwidth selection. The quantile at 5% is -0.0294.

Table 4.2: Results for the multivariate t case.

Portfolio: $\mathbf{x} = [0.0667, 0.6000, 0.3333]'$; Gaussian kernel.
Methods: par. = parametric, LPR = local polynomial regression (LPR_F with F transformation), f.d. = finite differences. h = bandwidth for l.p.r., c = perturbation for f.d.
ROT = rule of thumb, RSC = residual squares bandwidth selector.

asset	method:	par.	$LPR_{F,ROT}$	$LPR_{F,RSC}$	LPR_{ROT}	LPR_{RSC}	f.d.	f.d.
	h, c		0.0061	0.04	0.069	0.022	0.001	0.10
(1)	bias (%)	0.83	27.81	8.32	4.24	4.69	3.35	7.0015
	σ(%)	66.55	73.93	82.96	58.91	40.11	291.05	151.21
	h, c		0.0066	0.015	0.11	0.063	0.001	0.10
(2)	bias (%)	-5.40	26.68	24.95	3.16	3.30	3.64	2.19
	σ(%)	17.45	22.39	21.03	15.18	14.58	43.96	26.54
	h, c		0.0039	0.04	0.16	0.063	0.0010	0.10
(3)	bias (%)	-3.87	20.13	22.63	3.27	2.63	2.52	3.94
	σ(%)	26.60	27.14	28.26	28.03	17.83	89.04	42.45

Table 4.3: Results for the Non parametric case.

Portfolio: $\mathbf{x} = [1/3, 1/3, 1/3]'$; Gaussian kernel.
Methods: par. = parametric, LPR = local polynomial regression (LPR_F with F transformation), f.d. = finite differences. h = bandwidth for l.p.r., c = perturbation for f.d.
ROT = rule of thumb, RSC = RSC bandwidth selector.

	method:	$LPR_{F,ROT}$	$LPR_{F,RSC}$	LPR_{ROT}	LPR_{RSC}	f.d.	f.d.
	cp. 1	0.079	0.075	0.073	0.074	0.07	0.073
	cp. 2	0.107	0.107	0.105	0.105	0.105	0.105
	cp. 3.	0.069	0.068	0.06	0.0598	0.062	0.059
(1)	h, c	0.038	0.015	0.075	0.022	0.001	0.10
	σ	0.88	1.012	0.94	1.06	2.76	1.56
(2)	h, c	0.034	0.02	0.077	0.063	0.001	0.10
	σ	0.72	0.73	0.73	0.83	2.41	1.56
(3)	h, c	0.027	0.015	0.094	0.063	0.001	0.10
	σ	1.32	1.34	1.55	1.72	4.37	2.41

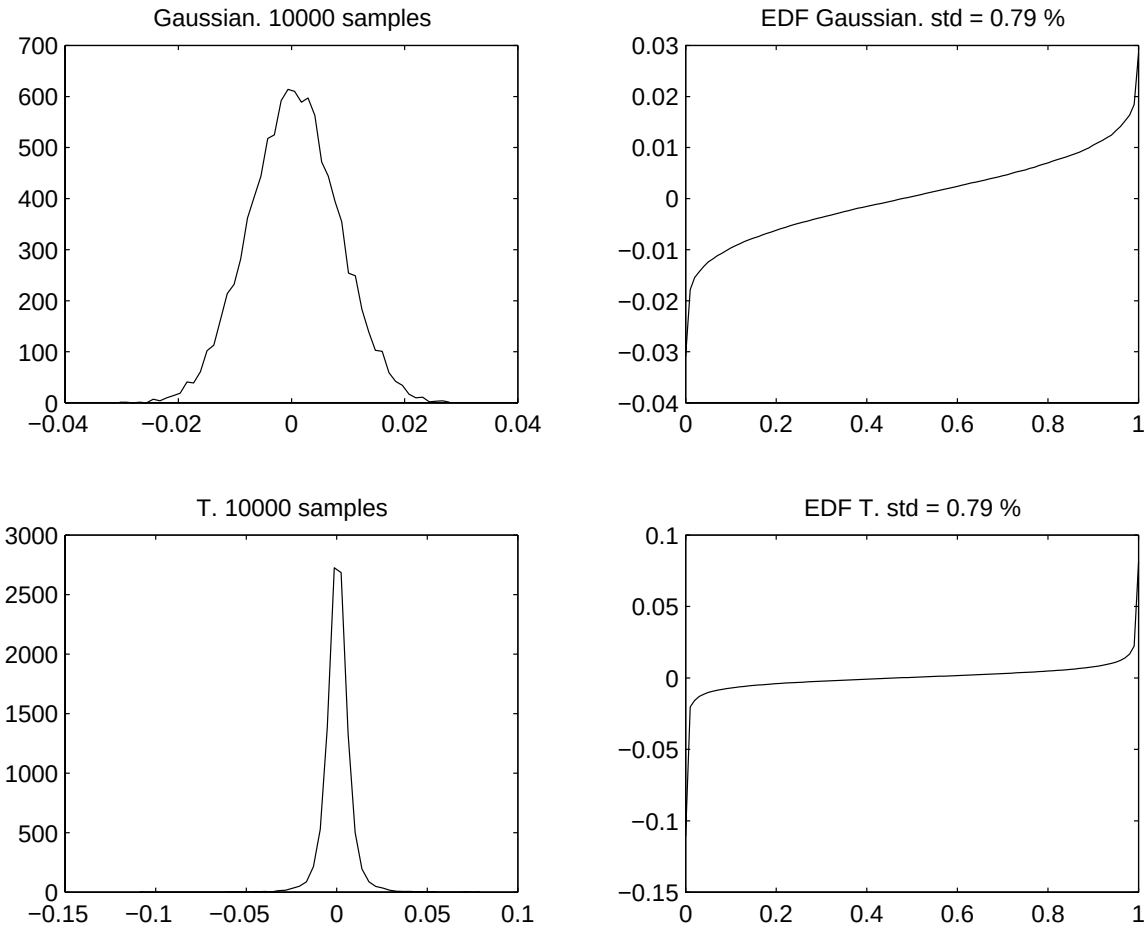

Figure 4-1: Parametric portfolio returns

Chapter 5

Optimization with quantile-based functions

In optimization problems, whether the goal is practical or theoretical, it is important to analyze cases where the optimized function is convex. Section 5.1 reviews the optimization of problems involving quantile-based functions, including the case where parametric formulas of these functions are available. In very specific cases, the optimization of quantile based functions can be done with non-gradient based algorithms. The Linear Programming (LP) and the Mixed Integer Programming (MIP) approaches are reviewed in section 5.2, including combinatorial problems such as the Brute force method (section 5.2.1), the MIP approach (section 5.2.1), and the *greedy* method (section 5.2.3).

Based on the gradient estimators derived in chapter 4, we propose a new alternative for the practical optimization of quantile based functions, which uses gradient-based methods (section 5.3), and analyzes the effect of the estimation error on the optimization. For completeness, we review other gradient-based methods, such as a recursive approach (discussed in section 5.3.1); we also review methods that can handle biased or unbiased gradient estimators (sections 5.3.2 and 5.3.2 respectively). We offer a brief comparison of the different methods in section 5.4.

In section 5.5 we present the outcome of applying the gradient-based method to the specific case of portfolio optimization. In this section we also analyze the behavior of the

optimal portfolios with respect to the α parameter.

Concavity

Concavity, or in some cases, pseudo-concavity (see [60]), is a very important issue to consider in optimization problems. Since quantiles appear in at least two important risk measures used in finance, it is of interest for Economists to understand in which cases quantile functions will be concave, both to ensure that numerical methods do not converge to local minima, and to establish the economic properties of the market.

We have seen in section 3.2.2 that some quantile-based functions could be non-concave for linear combinations of random variables with some particular distribution function. In other cases the quantile function is concave with respect to the weights, such as when the distribution function is elliptical (see the appendix section A.2.2). It would be very useful to characterize the distribution functions that imply a concave (or at least pseudo-concave) quantile function.

5.1 General quantile-based optimization

In this section we introduce the optimization of a quantile of a linear combination of random variables. We used this concept extensively in the previous chapters, but it is worth to mention that this technique could have other applications in which quantile constraints would be imposed.

For the linear case $\tilde{W}(\mathbf{x}) = \mathbf{x}'\tilde{\mathbf{b}}$, we propose some optimization methods involving linear combinations of quantile functions:

$$Q(\mathbf{x}) = \sum_{i=1}^{k} \lambda_i q(\alpha_i, \mathbf{x}) + H(\mathbf{x}), \tag{5.1}$$

where Q is the weighted sum of k quantiles for different values of α_i and positive weights λ_i, $i = 1, \cdots, k$ an arbitrary concave function $H(\mathbf{x})$. We introduce the function Q, since it can represent different quantile-based risk measures (such as the VaR and the empirical shortfall, see section A.2.4).

If each one of the quantiles $q(\alpha_i, \cdot)$ is concave, [54] then the function Q($\mathbf{x}$) will also be concave. The gradient of Q is

$$\nabla_{\mathbf{x}} Q(\mathbf{x}) = \sum_{i=1}^{k} \lambda_i \nabla_{\mathbf{x}} q(\alpha_i, \mathbf{x}) + \nabla_{\mathbf{x}} H(\mathbf{x}). \tag{5.2}$$

We analyze two cases:

i. The optimization of $Q(\mathbf{x})$ over a convex set C (or in particular, over a polyhedron $\mathcal{P}$ defined by a finite number of equalities and inequalities; the latter will be a problem similar to the classical one described in the appendix, equation (A.20)):

$$\max Q(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in C.$$

ii. The optimization of a convex function $G(\mathbf{x})$ over a convex set C intersected with the convex set $Q(\mathbf{x}) \geq L$, where L is a predefined constant;

$$\min G(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in C, \mathbf{x} \in \{Q(\mathbf{x}) \leq L\}.$$

The appendix A.3.1 gives a characterization of stochastic optimization problems. Stochastic optimization problems have to be approximated. In this section we review different approximation techniques that involve quantile-based functions.

Parametric approach If the distribution function is known to be of a certain parametric form, we only need to compute the parameters of $f(\tilde{\mathbf{b}})$ from the matrix $\mathbf{Y}$. For example, for elliptic distributions we will be able to use the equation (A.8), and plug in the estimated parameters. The optimization (either maximizing $Q(\mathbf{x})$, or with $Q(\mathbf{x})$ as a constraint) can then be performed with any constrained nonlinear deterministic method.

Where one of the constraints of the polyhedron $\mathcal{P}$ is $\mathbf{x}' E[\tilde{\mathbf{b}}] = b_p$, and we want to to optimize only the quantile of α (see $A.8$),

$$\max_{\mathbf{x} \in \mathcal{P}} \mathbf{x}' E[\tilde{\mathbf{b}}] - p_\alpha \sqrt{\mathbf{x}' \mathbf{V} \mathbf{x}} \quad \Rightarrow \quad \min_{\mathbf{x} \in \mathcal{P}} \mathbf{x}' \mathbf{V} \mathbf{x}, \tag{5.3}$$

the optimization problem can be related to a constrained quadratic optimization, which can be easily optimized, due to its special characteristics.

5.2 Non-gradient-based optimization methods

There are several ways to solve the linear case where $\tilde{W}(\mathbf{x}) = \mathbf{x}'\tilde{\mathbf{b}}$ and $E[\tilde{W}(\mathbf{x})]$ is to be maximized using both quantile constraints and constraining the vector $\mathbf{x}$ to belong to the polyhedron $\mathcal{P}$; the first two methods described next give a global solution, although they run in exponential time. The last two algorithms run in polynomial time, but they are only approximations, and will only give a sub-optimal result.

5.2.1 The Brute force method

[1] When a predefined value Q_p is given, the problem constraint $q_\alpha(\mathbf{x}) = Q_p$ can be approximated by letting $\hat{q}_{\alpha,n}(\mathbf{x}) = Q_p$. We know that $\hat{q}_{\alpha,n}(\mathbf{x})$ is the k-th order statistic $\mathbf{x}'\mathbf{b}_{k:n}$ (for $k = \alpha n$). In other words, given n samples of the optimal vector $\mathbf{x}^*$, $k-1$ samples will be less than a predefined value Q_p. Finding the optimal portfolio can be posed as a combinatorial problem which solves $\binom{n}{k-1}$ LP subproblems. A LP subproblem will be

$$\begin{aligned} W_{\mathcal{A}} &= \max E[\tilde{W}(\mathbf{x})] \\ \text{s.t.} \quad \mathbf{Y}_{\mathcal{A}}\mathbf{x} &\geq Q_p\mathbf{1}, \\ \mathbf{x} &\in \mathcal{P}, \end{aligned} \tag{5.4}$$

where the set $\mathcal{A}$ can be defined by its complement: $\mathcal{A}'$ is a subset with $k-1$ samples from the set of n samples; $\mathbf{1}$ is a vector composed of ones, and

$$\mathbf{Y}_{\mathcal{A}} = (\mathbf{b}'_i), \qquad \forall i \in \mathcal{A}. \tag{5.5}$$

The optimal vector value $\mathbf{x}^*$ is the vector which maximizes $W_{\mathcal{A}}$ for all $\binom{n}{k-1}$ subsets $\mathcal{A}$. The solution of this problem can only be obtained for small sizes of n and k. Since this method

[1] From personal communication with David Gay.

is assured to return the global maxima, we use it as a benchmark to compare the solutions returned by the other methods. Unfortunately, this is also a combinatorial method, worse than an exponential method.

5.2.2 Mixed integer programming

The previous problem can also be posed as a MIP problem (see [40, 31]); the naive implementation would be

$$\begin{aligned} & \max E[\tilde{W}(\mathbf{x})] \\ \text{s.t.} \quad \mathbf{Y}_{\mathcal{U}}\mathbf{x} + c \cdot \mathbf{p} \geq & \quad Q_p \cdot \mathbf{1}, \\ \mathbf{x} \in & \quad \mathcal{P}, \\ \mathbf{p}'\mathbf{1} = & \quad k-1, \\ p_i = & \quad \{0 \text{ or } 1\} \text{ for} \\ & \quad i = 1, \cdots, n, \end{aligned} \tag{5.6}$$

where $\mathcal{U}$ is the complete set of samples. The selection of c must be done such that the solution of (5.6) is equal to that of (5.4). In the special case where the constraints $x_i \geq 0$ for $i = 1, \cdots, m$ and $\mathbf{x}'\mathbf{1}$ are enforced, the MIP problem can be simplified. In that case, then $c = b_{min} + Q_p + \delta$, where b_{min} is the minimum value of all the components from $\mathbf{Y}_{\mathcal{U}}$, and δ is a positive value. Here, the number of binary variables can be reduced, including only the cases when $\mathbf{x}'\mathbf{B}_i$ can be less than Q_p. The cases where the maximum value $b_{max,i}$ of a sample $\mathbf{b}_i$ are less than Q_p can be eliminated, since we know that those samples should not belong to the set; the binary variable corresponding to those samples is fixed to be $p_i = 1$.

When negative values of the components $\mathbf{x}$ are allowed, c must be chosen in such a way that the MIP still represents the combinatorial problem.

The algorithm runs in exponential time, and although it is a better method than the brute force method, it is computationally very intense when the number of samples is large.

5.2.3 The Greedy linear programming

A greedy heuristic is a very fast method to obtain and approximate results: let $\mathcal{B}_k$ be a sequence of sets; each set $\mathcal{B}_k$ includes samples on which we enforce the bound. We want to

enforce the bound on at least $k^* = \alpha n$ samples. Let's define the LP subproblem based on $\mathcal{B}_k$ as

$$\begin{aligned} W_{\mathcal{B}_k} &= \max E[\tilde{W}(\mathbf{x})] \\ \text{s.t.} \quad \mathbf{Y}_{\mathcal{B}_k}\mathbf{x} &\geq Q_p\mathbf{1}, \\ \mathbf{x} &\in \mathcal{P}, \end{aligned} \tag{5.7}$$

where $\mathbf{Y}_{\mathcal{B}_k}$ is defined as in (5.5).

i. Start with $\mathcal{B}_0 =$ all the available samples, and while $\mathcal{B}_0$ is too large, solve the problem (5.7) and remove from $\mathcal{B}_0$ the sample with the most negative dual value, ($\mathcal{B}_1 = \mathcal{B}_0 -$ the sample with the most negative dual value).

ii. if $k \leq k^* - 1$, solve the subproblem 5.7 for $\mathcal{B}_k$ and make $\mathcal{B}_{k+1} = \mathcal{B}_k -$ the sample with the most negative dual value, iterate (ii) as needed.

For example, when $\alpha = 0.01$ and 200 daily returns are used, the algorithm requires 1 solve; if we increase the number of daily returns to 300 daily returns, then 2 solves are required. Because it solves a finite number of LP programs at each iteration, it runs in polynomial time.

Combinations Combinations of the former methods could be developed; but the existence of a gradient estimator allows us to derive a very fast method which is explained next.

In practice, the MIP problem is only able to handle very few samples (e.g., 300 samples) to give an "optimal" portfolio. Due to the limited size of the samples used, the order statistic obtained introduces a significant error with respect to the true quantile (see section A.2.4). For that reason, MIP techniques are not very useful to optimize portfolios in practice.

5.3 Gradient-Based Optimization methods

We could pose the problem

$$\text{maximize } Q(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in P \tag{5.8}$$

which is similar to the stochastic problem (A.16). In the non-recursive approach, all the n samples are used to get estimators of $Q(\mathbf{x})$ and $\nabla_{\mathbf{x}} Q(\mathbf{x})$. Because no new samples are introduced, the estimator $\hat{\nabla}_{\mathbf{x}} Q(\mathbf{x})$ will add a *constant* error $\mathbf{v}$ which is the sum of the gradient estimator bias plus the zero mean random error due to finite sampling.

If the non-recursive method converges to a point $\mathbf{x}^*$, as explained in section A.3.2, the point $\mathbf{x}^*$ will be a solution of the problem

$$\max Q(\mathbf{x}) + \mathbf{v}'\mathbf{x} \quad s.t. \quad \mathbf{x} \in \mathcal{P}.$$

It is very important that the error $\mathbf{v}$ be kept as small as possible. The analysis of the effect of the bias error on the optimization problem should be done by taking advantage of the particular characteristics of each problem, as in section 3.6.

Finite differences estimation error. The error vector $\mathbf{v}_{fd}$ due to finite differences is:

$$\mathbf{v}_{fd}(\mathbf{x}, c, n) = \mathbf{b}_{fd}(\mathbf{x}, c, n) + \frac{\eta(\mathbf{x}, n)_1 - \eta(\mathbf{x}, n)_2}{2c}.$$

as derived in (4.9). The finite differences method introduces $2m$ error terms η for each perturbed vector $\mathbf{x}+ c \cdot \mathbf{e_j}$ used to compute an approximation of the partial derivative. If c is increased to reduce the error due to the zero mean errors η_j, then the error due to the bias increases. For relatively small finite samples, the stochastic counterpart that uses finite differences does not generally converge.

Local polynomial regression estimation error. The error $\mathbf{v}_{lpr}$

$$\mathbf{v}_{lpr}(\mathbf{x}, h, n) = \mathbf{b}_{lpr}(\mathbf{x}, h, n) + \mathbf{w}_{lpr}(\mathbf{x}, h, n)$$

due to modeling increases as h increases, but the zero mean error vector $\mathbf{w}_{lpr}(\mathbf{x}, h, n)$ decreases. Whenever h is selected to make the zero mean error negligible, a fast deterministic

algorithm can be selected to obtain the solution of the constrained problem

$$\text{maximize } Q(\mathbf{x}) + \mathbf{b}'_{lpr}\mathbf{x} \quad \text{subject to} \quad \mathbf{x} \in P$$

(denoting $\mathbf{b}_{lpr} = \mathbf{b}_{lpr}(\mathbf{x}, h, n)$). Therefore, the solution of a stochastic counterpart that uses the estimated gradient can only be an approximation. The bias and variance can be estimated in practice using formulas (A.50) and (A.51).

Examples of optimization problems Other possible representations of problems that arise in practice are

$$\text{maximize } E[U(\tilde{W}(\mathbf{x}))] \quad \text{subject to} \quad q_\alpha(\mathbf{x}) \geq Q_p, \mathbf{x} \in \mathcal{P},$$

$$\text{or : maximize } q_{0.5}(\mathbf{x}) \quad \text{subject to} \quad q_\alpha(\mathbf{x}) \geq Q_p, \mathbf{x} \in \mathcal{P}.$$

They can also be approximated in practice using the gradient estimators; but their validity depends heavily on whether or not the quantile functions used are "well-behaved", e.g. concave or not. The bias introduced by the new optimization problem should always be taken into account; e.g. for the former optimization problem:

$$\text{maximize } E[U(\tilde{W}(\mathbf{x}))] \quad \text{subject to} \quad q_\alpha(\mathbf{x}) + \mathbf{x}'\mathbf{b}_{lpr} \geq Q_p, \mathbf{x} \in \mathcal{P}.$$

The optimization can be done if the error is known to be small.

In practice, when we implemented the median optimizer using a gradient based algorithm, the convergence behavior turned out to depend heavily on the distribution of the random vector $\tilde{\mathbf{b}}$; in the case of multi-modal data (which arises in optioned strategies), convergence to an optimal was not always ensured.

5.3.1 The recursive approach

So far we have assumed that n samples are available. The finite number of samples can be generated from a Monte-Carlo simulation, or can be n historical i.i.d. samples. The latter is more difficult to ascertain, since stochastic processes can often be non-stationary and the samples might not be i.i.d.

With the Monte-Carlo simulations, we do not have a priori limitations on the size of the sample. Monte-Carlo simulations may be required whenever the stochastic factors of a system are known. However, there might be a nonlinear transformation which makes difficult the derivation of parametric equations for the quantiles. In those cases, recursive approaches are useful to obtain approximate solutions. Recursive approaches are not very practical to implement, but it is interesting from a theoretical point of view since it assure the a.s. convergence of the algorithm, as long as $Q(\mathbf{x})$ has local optima.

5.3.2 Optimization with biased gradient estimators

Because in the estimation of $Q(\mathbf{x})$ as $\hat{Q}_n(\mathbf{x})$, we can use asymptotic zero mean estimators (the order statistics), we can use the Kiefer-Wolfowitz optimization procedure which uses finite differences (although the convergence will be even slower, since we are using asymptotic zero mean estimators; described in A.3.2) to obtain a stochastic approximation of the optimal vector $\mathbf{x}^*$. The technique has assured convergence in the asymptotic case, but in practice we terminate the algorithm when some a priori conditions are fulfilled; for that reason, in practice only up to n samples are used to compute an approximate optimal vector $\mathbf{x}^*$. An important point is that the recursive method requires i.i.d. samples of the quantiles: a quantile estimator requires at least k/α samples. Therefore, if we are limited to n samples, k/α different samples are used in the recursive case, and up to $n\alpha/k$ optimization iterations can be done.

The speed of convergence is also detailed in section A.3.2; however, the number of iterations can be reduced if the estimated gradient has a smaller variance and bias than the finite difference estimator, as in the case when we estimate the gradient with a local polynomial regression.

If $n\alpha/k$ is not an integer, or it is greater than 1, then we can bootstrap n^* subsamples of size $n^* = k\alpha$ to generate a series of quantile estimators. However, we must take into consideration that the quantile to be optimized belongs to the discrete distribution of the boot-strapped samples, not the original distribution; this optimization might turn into a non-convex problem.

5.3.3 Optimization with unbiased gradient estimators

The local polynomial regression offers an (almost) unbiased estimator. We can select the unbiased estimator using the M-order statistic $\tilde{\mathbf{b}}_{k:n}$ as estimator of the gradient. It is not a pure unbiased estimator, because it is not an unbiased estimator of the function $f_{\tilde{\mathbf{b}}|\tilde{W}(\mathbf{x})}(\mathbf{b}|q_\alpha)$ but rather of $f_{\tilde{\mathbf{b}}|\tilde{W}(\mathbf{x})}(\mathbf{b}|\hat{q}_{\alpha,n})$. The previous functions will be almost equivalent (unless there are discontinuities in the functions), and it is possible to use the Robbins-Monro optimization approach, which needs unbiased estimators of the gradient (as described in section A.3.2).

5.4 Comparison of methods

In experiments with the same conditions:

- maximization of expected return $\tilde{W} = \mathbf{x}'\tilde{\mathbf{b}}$,
- constraining $1 - \alpha n$ samples above a predefined level Q_p ($\hat{q}_{\alpha,n} = Q_p$).
- no shortsales allowed ($\mathbf{x} \geq 0$).

The Brute-force method, the MIP and the gradient based method using local polynomial regression (LPR) consistently give almost the same solution. The Greedy method almost always gives the same solution as the brute force, with some exceptions when it only gives approximated solutions.

The main drawback of the brute-force method is the computational effort required to use it. The MIP method is faster, but it is limited to cases when the shortsales are bounded to a predefined value (such that the relaxation variable introduced does or does not select certain samples). Still, both methods are impractical for large numbers of n (larger than 200). The

Greedy method is quite fast and acts in polynomial time, although the answer given is not always the optimal. Gradient based methods with finite difference gradients, interestingly, sporadically converge to the optimal solution if n is small (e.g., less than 200).

There are several alternative to improve the precision of the optimal weights (if the mean and the VaR obtained are not within acceptable tolerance levels). For instance, once the gradient based method with LPR has converged, then new samples can be used to improve the optimal value (in a recursive fashion). Also, the smoothness of the efficient frontier offers a way to improve the optimal weights, by using the optimal solutions with nearby performance measures as initial values for an iterative algorithm.

5.5 Mean-VaR and shortfall portfolio optimization

We use mainly the non-recursive stochastic optimization described in section A.3.1, which uses a non-differentiable Penalty-Based algorithm with a non-parametric gradient estimator. This method gives practically equivalent results to those generate by the MIP-based algorithm of section 5.2.2. The optimal value can be refined by the addition of samples.

The mean-variance and the mean-semivariance have already been very well researched. For the VaR_α and the shortfall, given that they depend on the level α required, it is interesting to analyze how in practice variations in α affect the optimal portfolio weights. We analyze four relevant cases:

i. Gaussian data.

ii. Real stock data.

iii. Option-based strategies data.

iv. Put-Call data.

Experiment 5.5.1 *Gaussian data.* As a control case, we want to see how our optimization method behaves with synthetic Gaussian data. We use the simulated returns of three assets, as described in appendix C.4.1. In theory, given the quadratic nature of the optimization problem, We compute the optimal portfolio for the case where no shortsales are allowed, and

no risk-free asset is available. The desired expected return is the average of the returns of the three assets. In the case of the mean-VaR and the mean-shortfall, an additional variable α is needed; in figure 5-1 we plot the variations of the optimal allocations for the Gaussian case, with respect to the variable α. In this case the optimal portfolio should be independent of α, but the noise of the gradient estimator in the experiments influenced the optimal weight for the cases of the mean-VaR and the mean-shortfall. As a consequence, the noise introduced by the gradient estimators has to be considered carefully in the following experiments that test the variation of the optimal weights with respect to α.

Experiment 5.5.2 *Stock data.* Given all the available financial literature on the non-normality of returns in real financial data, (reviewed in section 2.2.1), we decide to study the effects of variations in α when computing optimal portfolios with real historical data. For the experiment we use the daily returns of four stocks, as described in the appendix section C.4.3. In figure 5-2 we plot the variations of the optimal weights for two of the assets, (Anheuser-Busch and Caterpillar) with respect to α.

For this experiment, although the optimal weights obtained with the mean-VaR and the mean-shortfall algorithms differ from those obtained with the mean-variance and the mean-semivariance methods, the deviations seem to be due to the noise of the gradient estimation, (as in the previous examples).

Experiment 5.5.3 *Option-based strategy data.* Figure 5-3 represents the variations of the optimal portfolio with respect to α, when option-based strategies data is used. In the experiment we use three option-based strategies, described in appendix C.4.2. The optimizations in this example allows shortsales.

For quantile-based risk measures, as the VaR_α and the shortfall, the slope of the optimal portfolio seems to be smooth (rather than noisy as in previous cases), which may indicate that the optimal weight is in fact a function of the level α. This leads to and interesting conclusion about the economic investors' behavior; investors could have a different optimal portfolio depending on how they chose the parameter α, even for the same desired expected return. In this example, as α increases, an investor using quantile risk measures such as

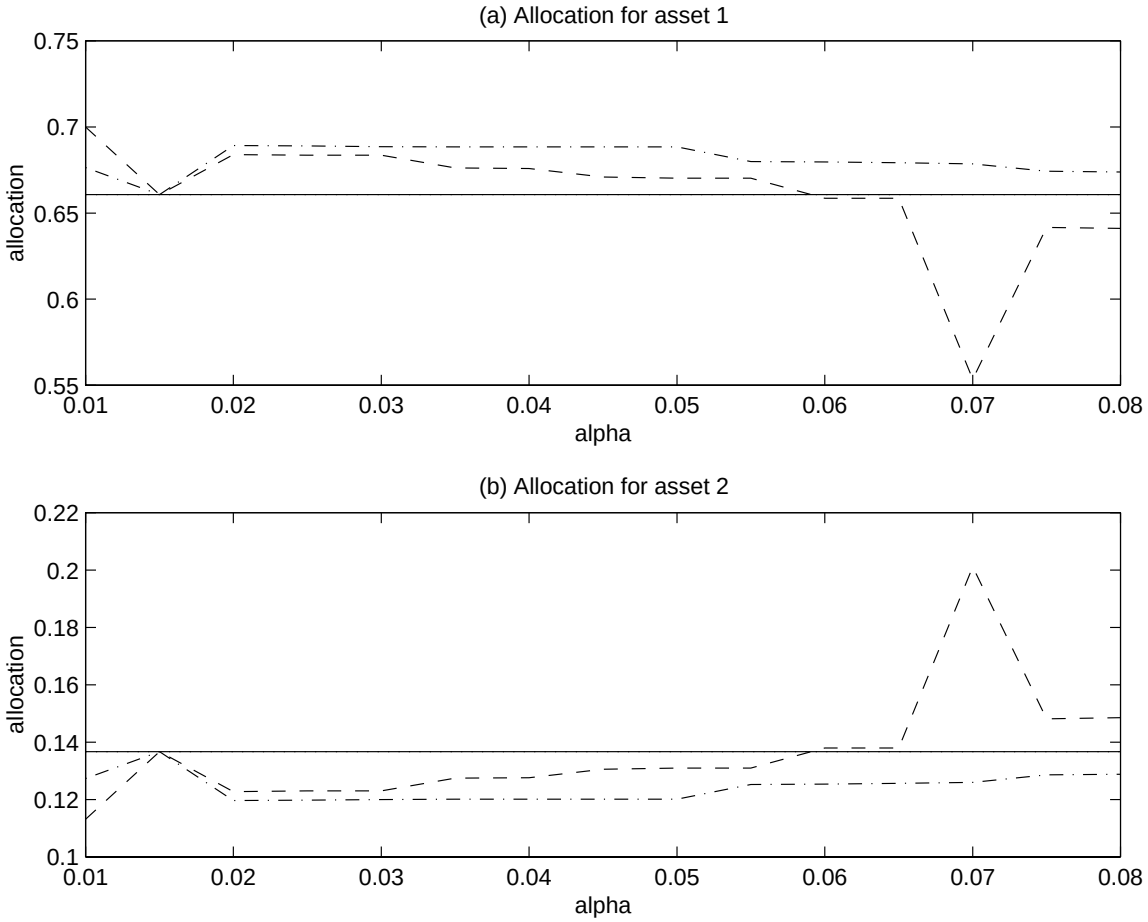

Figure 5-1: Weight variation with respect to α. Gaussian data.

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$

With Gaussian data, there should be no variations of the optimal weight with respect to α. The deviations of the mean-VaR and the mean-shortfall with respect to the mean-variance optimal weights are due to numerical errors.

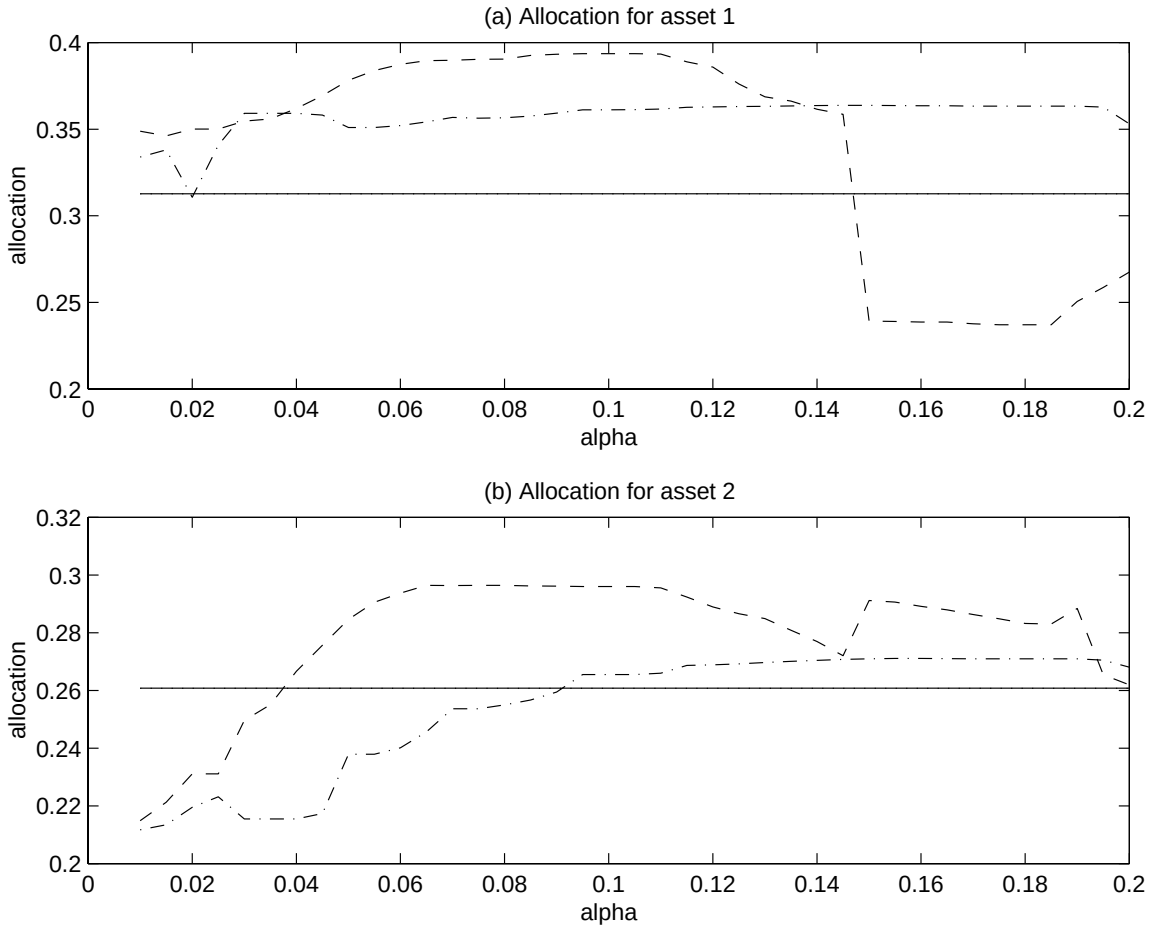

Figure 5-2: Weight variation with respect to α. Stock data.

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 1: Anheuser-Busch stock. asset 2: Caterpillar.

The variations of optimal weights with respect to α are due to numerical errors.

the VaR and the shortfall allocates more weight to the "write-call" strategy, and less to the "long-put" strategy. Intuitively, this behavior indicates that as α increases, the mean-VaR method will favor "riskier" assets. In this case, two risk-averse investors having different α levels could have different optimal portfolios. This could be true for assets with asymmetric returns, and may explain why some investors with the same apparent risk-averseness have different portfolios. However, we should note that our algorithm is less reliable when α is smaller, due to noise introduced by a reduced sample size. Hence, one should be careful when making conclusion about investors' behavior.

Experiment 5.5.4 We use the "Put-Call"data described in appendix C.4.4 to show the drastic change in optimal portfolios resulting from variations in α. We optimize portfolios (allowing shortsales) using α ranging from 0.01 to 0.05. The options are designed so that the $VaR_{0.05}$ risk measure (for $\alpha = 0.05$) cannot measure the risk (or loss of money). The results are shown in table 5.1. When the risk measure has an α of 0.05, the $VaR_{0.05}$ (-0.16) appears even to outperform the risk-free asset. However, when the risk of the same portfolio is measured using an α of 0.01, the $VaR_0.01$ is incredible large (5.3663). This experiment amplifies the problem of using the "wrong" risk measure; even if the risk measure seems to be correct, it is very sensitive to changes in α.

It is also interesting to compare the risk and reward measures of the mean-variance optimal portfolio with those of to the other optimal portfolios; the mean-variance has the highest median, while the optimal portfolio with the $VaR_{0.01}$ loses money most of the time. The mean-$VaR_{0.01}$ optimal portfolio behaves very similarly to the sub-optimal trading strategy of the experiment 6.5.1, which again reinforces the idea of using reward measures other than the mean.

Experiment 5.5.5 *Put-Call data.* The put-call data (figure 5-4) is a case where the variation in α changes the optimal portfolio drastically. As for the experiment 5.5.4, this example is generated to "deceive" the risk when the confidence level is 0.95; the optimal weights for the VaR_α risk measure when α is greater than 0.05 does not really reflect the risk of losing money. The optimization is done assuming that shortsales are allowed. The radical change

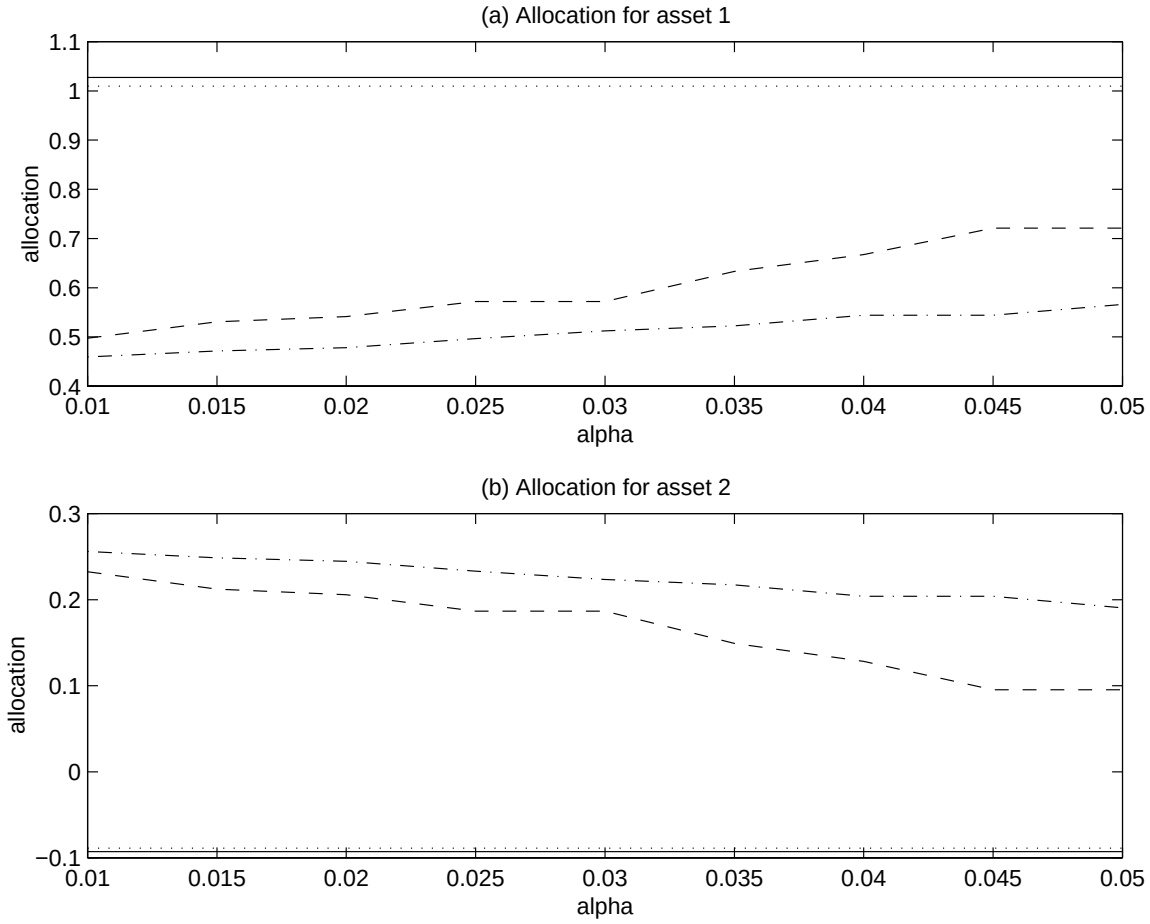

Figure 5-3: Weight variation with respect to α. Option-based strategies data

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 1: "write-call" at 50 %. asset 2: "long-put" at 50 %.

The variation of optimal weights with respect to α appears to be significant; assets 1 is riskier than asset 2.

Table 5.1: Results for the Optimization.

Optimization Method: μ vs. $risk$

risk:	σ	$VaR_{0.05}$	$VaR_{0.01}$
Stock	2.2843	0.0001	-0.0000
call	-0.0073	0.0000	0.0926
put	0.0060	-0.3126	-0.0000
Risk free	-1.2830	1.3125	0.9074
$1+\mu$	1.1558	1.1558	1.1558
1+median	1.1476	1.3159	0.9097
σ	0.3143	1.0014	1.5081
$VaR_{0.05}$	0.5125	-0.1600	0.2462
$VaR_{0.01}$	0.7128	5.3663	0.2462

of weights when the VaR is used can be appreciated here. This experiment is designed in such a way that arguments about the numerical reliability of the optimization algorithm cannot be used. It should also be noted that the mean-shortfall algorithm does not change as a function of α, which can be considered as an advantage of the shortfall over to the VaR. The weights of the mean-semivariance algorithm do not coincide those of the mean-variance algorithm, which indicates that a more risk-averse portfolio can be obtained using the mean-semivariance method (or the mean-shortfall method, which in this example gives optimal weights close to those of the mean-semivariance method).

The mean-semivariance and the mean-shortfall optimal portfolio which indicates that a more conservative optimal portfolio can be obtained by using the mean-semivariance algorithm, or using the mean-shortfall algorithm (whose optimal weights are very similar to the mean-semivariance ones).

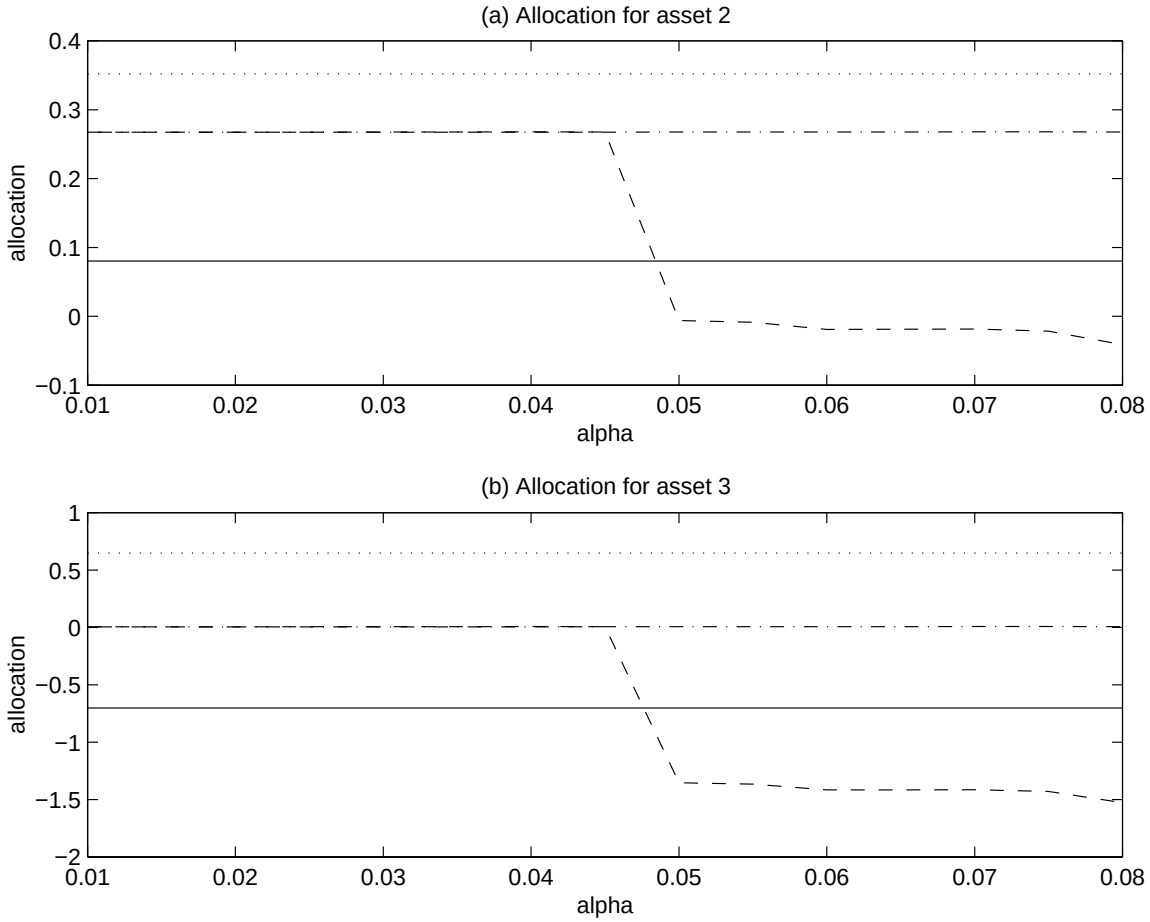

Figure 5-4: Weight variation with respect to α. Put-Call data.

(solid) mean-variance, (dotted) mean-semivariance,
(dash-dot) mean-shortfall 5%, (dashed) mean-$VaR_{0.05}$
asset 1: Put. asset 2: Call

Example when there is a noticeable change of optimal weight as a function of α. The optimal weight using the mean-VaR optimization exhibits a drastic "drop" at 5 %.

Chapter 6

Dynamic optimization

In this chapter we analyze a simplified discrete time portfolio strategy which yields an efficient final return, in the risk-reward sense. The main goal is to understand what are the effects of dynamic trading considering that an investor prefers investments which belong to the risk-reward efficient frontier.

The dynamic case of portfolio optimization is reviewed in section 6.1, which introduces the notation commonly used to describe multiperiod asset allocations, and reviews some of the previous attempts to solve the dynamic case (such as the expected utility maximization in the dynamic case, the continuous-time analysis, and the dynamic option replication, sections 6.2, 6.2 and 6.2 respectively). The notation follows closely the following books: [50, 46] and [14].

We introduce the concept of *consistency* of a trading strategy (section 6.4.1), and define the concept of risk-reward optimal trading strategies in the portfolio optimization framework (section 6.3). We focus on the absolute VaR_α measure (which uses the quantile function as a risk measure), and in section 6.4 we develop one possible way to analyze the simplified problem in which the trading strategy is discrete in time and wealth is allocated between one risk-free asset and one risky asset. The analysis, similar to dynamic programming, starts with one period and generalizes to the case with T periods. A sub-optimal solution for an α-consistent trading strategy (a trading strategy, constrained to be consistent, and to satisfy the quantile constraint) is presented in section 6.5.

6.1 Multiperiod asset allocation

The multiperiod portfolio optimization problem can be seen as an extension of the static case, however, we take advantage of the possibility to re-balance the portfolio at every intermediate trading period. Some of the single period techniques (such as the mean-variance) cannot be readily generalized to the case where an investment takes place over several periods of time (known also as *trading periods*). There are several approaches available for multiperiod strategies: the optimization of expected utilities, the dynamic replication of options, as well as our own portfolio optimization with quantile constraints. In this section we analyze a simple case that takes advantage of inter-period trading.

Let us assume that there are $T+1$ trading periods, $i = 0, 1, \ldots, T$, and there are m assets where an investor can allocate her wealth $\tilde{W}_i$ at the i-th period, $i = 1, 2, \cdots, T$; W_0 is a known value. The price of the j-th asset, for $j = 1, 2, \cdots, m$, follows the stochastic process $\{\tilde{P}\}_j = \{\tilde{P}_{i,j};\ i = 1, 2, \cdots, T\}$ where $\tilde{P}_{i,j}$ is a random variable that represents the price of the j-th asset at time i; $P_{0,j}$ represents the initial known price of the j-th asset. At times $i = 1, 2, \cdots, T$, the random vector $\tilde{\mathbf{b}}_i$ is composed of the simple gross return of m assets; $\tilde{\mathbf{b}}_i' = [\tilde{b}_{i,1}, \tilde{b}_{i,2}, \cdots, \tilde{b}_{i,m}]$, where $\tilde{b}_{1,j} = \tilde{P}_{1,j}/P_{0,j}$ and $\tilde{b}_{i,j} = \tilde{P}_{i,j}/\tilde{P}_{i-1,j}$ for $i = 2, 3, \cdots, T$ and $j = 1, 2, \cdots, m$. The random vector $\tilde{\mathbf{b}}_i$ has a multivariate cumulative distribution F (which is a simplifying assumption; in practice $\tilde{\mathbf{b}}_i$ could be non-stationary).

A *trading strategy* $\{\mathbf{y}\}$ is the vector of stochastic processes $\{\mathbf{y}\} = \{\mathbf{y}_i; i = 0, 1, \cdots, T-1\}$, where $\mathbf{y}_i' = [y_{i,1}, y_{i,2}, \ldots, y_{i,m}]$. Note that $\mathbf{y}_T$ is not specified; this is because $y_{i,j}$ should be interpreted as the percentage of units that the investor owns (i.e. carries forward) of the j-th asset from time i to time $i+1$. A trading strategy is modeled as a stochastic process because trading strategies are rules (i.e., functions) that specify the investor's position in each security at each point in time, for each possible wealth available. We assume that investors neither make additions to nor withdrawals from their invested wealth, generating what is known as a *self-financing* trading strategy. In this case, the wealth is constrained to evolve according to the following dynamic equation over time:

$$\tilde{W}_1 = W_0 \mathbf{y}_0'(W_0)\tilde{\mathbf{b}}_1,$$

$$\tilde{W}_{i+1} = \tilde{W}_i \mathbf{y}_i'(\tilde{W}_i)\tilde{\mathbf{b}}_{i+1}, \qquad i = 1, 2 \cdots, T. \tag{6.1}$$

The wealth (for a self-financing strategy) can be characterized by the stochastic process $\{\tilde{W}\} = \{\tilde{W}_i; i = 1, 2, \cdots, T\}$. Again, we use self-financing strategies in order to simplify the analysis of dynamic strategies, since a certain amount of wealth could be "consumed" at the i-th period. We assume that at the trading period i the outcome of the wealth $\tilde{W}_i$ is known (which is denoted as W_i), but the future value of the wealth (i.e., $\tilde{W}_{i+1}$) is still a random variable. For some trading strategies, each allocation vector $\mathbf{y}_i$ at the period i could be a vector function of the wealth at the i-th period, $\mathbf{y}_i = \mathbf{y}(W_i)$ (i.e., a rule that specifies the investor's position in each security as a function of the wealth at period i).

An *admissible self-financing trading strategy* is a trading policy in which the values of $\mathbf{y}_i$ are constrained to belong to a predefined set,

$$\mathcal{Y} = \{\mathbf{y}_i'\mathbf{1} = 1; \forall i = 0, 1, \ldots, T-1\}, \tag{6.2}$$

and where the wealth $\{\tilde{W}\}$ follows the dynamics of equation (6.1), therefore being self-financed. Other constraints (such as no short-selling) can also be included in the set $\mathcal{Y}$. If techniques such as dynamic programming are used, it is important to stress the assumption of *independence* of the random vectors $\tilde{\mathbf{b}}_i$ across time.

As in single period optimization, it is necessary to define a preference relation among different trading strategies. The preference relation can be based on the final expected wealth utility, or, as in the static case, on the performance vector combining the risk and reward of the final wealth. The goal of a multiperiod portfolio optimization is to find an *optimal trading strategy* $\{\mathbf{y}^*\}$ with respect to the criteria selected (usually the maximization of a certain utility function). The optimal trading strategy is composed of the optimal allocation rules at each trading period, $\{\mathbf{y}^*\} = \{\mathbf{y}_i^*(W_i); i = 0, 1, \cdots, T-1\}$.

6.2 The expected utility in a multiperiod case

In the expected utility framework, for tractability reasons, most researchers assume the existence of *time additive* utility functions, or functions U_i such that

$$U(W_0, W_1, \ldots, W_T) = \sum_{i=0}^{T} U_i(W_i). \tag{6.3}$$

Defining

$$J^0_{\{\mathbf{y}\}}(W_0) \equiv E\left\{\sum_{i=0}^{T} U_i(\tilde{W}_i)\right\}, \tag{6.4}$$

the formulation of the problem can be posed as the maximization of

$$\max_{\{\mathbf{y}\}\in\mathcal{Y}} J^0_{\{\mathbf{y}\}}(W_0). \tag{6.5}$$

Given this utility function, it is natural to consider the problem at any time j of maximizing the "remaining utility," given current wealth W_j. Consider the subproblem in which we are at the period j with a wealth W_j and we wish to maximize the "cost-to-go" $J^j_{\{\mathbf{y}\}}(W_j)$ from time j to time T

$$J^j_{\{\mathbf{y}\}}(W_j) = E\left\{\sum_{i=j}^{T} U_i(\tilde{W}_i)\right\}. \tag{6.6}$$

Bellman's approach is to find the optimal sub-trading strategy of this subproblem ($\{\mathbf{y}^*_i; i = j, j+1, \cdots, T-1\}$), and assume that the total optimal trading strategy will contain it. Then, to obtain the total optimal trading strategy, the dynamic programming algorithm states that it is necessary to start with the last period and proceed backward in time.

At the heart of the dynamic programming algorithm, lies the Markovian property which implies that the conditional probability distributions $\Pr\{W_{j+1}|W_j\}$ depend only on the trading strategy $\{\mathbf{y}\}$; that is one of the reasons why the random vectors $\tilde{\mathbf{b}}_i$ should be independent in time, otherwise the analysis would become very complex.

One particular approach which uses expected utilities, assumes the investor maximizes the expected utility of his final wealth, such that the total utility on time is a function of only the utility at the final period: $U(W_0, W_1, \ldots, W_T) = U_T(W_T)$. In this case, it has been

shown that if

$$\frac{1}{R_A} = -\frac{U'(W_T)}{U''(W_T)} = \quad a + bW_T, \quad \forall W_T, \tag{6.7}$$

then there is a closed form solution (where R_A is the measure of risk aversion defined in section B.3.1).

Myopic and partially myopic trading strategies

If $A = 0$ in equation (6.7), (i.e. for utility functions of the form $\ln(W)$ and $(1/(b-1))(bW)^{1-1/b}$ for $b \neq 0$ or $b \neq 1$.), when the investor has the opportunity to reinvest sequentially her wealth, she will use a trading strategy similar to that of the single period case. At each stage j she will behave as if she were faced with a *single period* investment characterized by the objective function $E\{U_T(\tilde{W}_{j+1})\}$. This trading strategy is called a *myopic trading strategy.*

When $A \neq 0$, assuming a risk-free asset with a rate of return $r_{f,i}$ at the i-th period is available, then the optimal trading strategy at the period j is the same as the investor would chose if faced with a single period problem whereby he would maximize over $\mathbf{y}_j$,

$$E\{U((1 + r_{f,T-1}) \dots (1 + R_{f,j+1})\tilde{W}_{j+1})\}.$$

In other words, the investor maximizes the expected utility of wealth if a portfolio $\mathbf{y}_j$ is used in the j period and the resulting wealth W_{j+1} is subsequently invested exclusively in the risk-free asset during the remaining periods $j+1, \dots, T-1$. This type of trading strategy is called a *partially myopic trading strategy.* As the horizon keeps expanding, the trading strategy on the initial stages approaches a myopic trading strategy.

Continuous time finance

The special case when there is a a very high frequency of trading, e.g., that an investor is trying to maximize her expected utility at some point in the distant future, or it is allowed to make many trades in a short period of time, can be approached from a continuous time point

of view. Merton in [46] describes in a very detailed manner most of the portfolio optimization techniques in continuous time. Using a martingale representation approach, Cox and Huang derived an alternative method for solving optimal consumption-portfolio problems. The derivation of their techniques includes the restrictions that consumption and wealth must be nonnegative. A fixed trading strategy that is globally optimal or "asymptotically optimal" for some utility functions is known as a *growth optimal portfolio*. For deeper coverage of these topics, see [46, 20, 32, 35].

Relation to mean-variance In an economy where all asset prices are log-normally distributed and there is one risk-free asset, Merton [46] shows that the proportions of each asset in a portfolio are derived by finding the locus of points in the instantaneous mean-standard deviation space of composite returns which minimize the variance for a given mean (i.e. the efficient risky-asset frontier).

Risk sensitive control

Another interesting case is when

$$E\left\{\beta \exp\left(\alpha\left(\sum_{i=0}^{T} U_i(\tilde{W}_i)\right)\right)\right\},$$

where α and β are given scalars. This type of cost functional can be called a *risk sensitive cost functional* since it corresponds to an exponential utility function expressing the risk aversion or risk preference (depending on the sign of β) of the decision maker.

Other risk sensitive cost functionals include (as used in [51])

$$\lim_{T\to\infty} \inf \tfrac{1}{\gamma} T^{-1} \ln E\{\tilde{W}_T\}^{\gamma}, \quad \gamma < 1, \quad \gamma \neq 0 \ .$$

Letting $\gamma \to 0$, this cost functional at the limit will tend to maximize the portfolio's long-run expected growth rate. Letting $\theta = -\gamma/2$, after some algebra and doing a Taylor series

expansion about $\theta = 0$, [51] the following expected utility is used;

$$J_\theta \equiv E\{\ln \tilde{W}_T\} - \frac{\theta}{4}\text{var}\{\ln \tilde{W}_T\} + O(\theta^2).$$

The variance term will be a penalty that depends on the term θ, called a *risk aversion parameter*.

Dynamic option replication

Another interesting use of trading strategies is the dynamic option replication. Under certain conditions an option's payoff can be exactly replicated by a particular dynamic investment strategy involving only the underlying stock and risk-free debt. This particular strategy may be constructed to be *self-financing*, i.e., requiring no cash infusions except at the start and allowing no cash withdrawals until the option expires; since the strategy replicates the option's payoff at expiration, the initial cost of this self-financing investment strategy must be identical to the option's price, otherwise an arbitrage opportunity will arise. The pricing formula of an option can only be expressed implicitly as the solution of a parabolic partial differential equation (PDE). However, a new kind of self-financing portfolios which can replicate option's payoff has been proposed recently. Using DP, the optimal trading strategy of

$$\min_{\{\mathbf{y}\}} E\{(\tilde{W}_T - F(\tilde{P}_T, \mathbf{z}_T)^2\}.$$

will replicate a self financing portfolio, such that the final payoff $F(\tilde{P}_T, \mathbf{z}_T)$ is achieved as close as possible (using a quadratic function for tractability). $\tilde{P}_T$ will be the final price of the risky asset, while $\mathbf{z}_T$ is a state vector. The main argument regarding the use of discrete strategies versus discretized continuous ones, is that in reality the trading can only be made at discrete intervals of time, hence the analysis of real-life problems should also be addressed in a discrete time scenario.

6.3 Risk-reward optimal trading strategies

As we reported in the previous sections, the dynamic strategies analyzed in the literature will optimize a dynamic trading strategy over a certain period of time, so that the final expected utility is maximized. From the risk-reward perspective, that method does not take explicitly the risk into account. In this section we set the foundations of trading strategy optimization within a risk-reward scenario, for the particular cases of the VaR and the expected return maximizations. However, the use of multiple criteria, as well as other trading strategy restrictions (such as consistency) will complicate the solution of the dynamic problem.

All the previous multiperiod alternatives assume the existence of a utility function which can be maximized. Little work has been done on Pareto's optimal trading strategies, even if this problem can be expressed as a fixed end stochastic control problem. In optimal control problems, state variables evolve with time and form paths; the payoff depends on which path is used. In the static case the distribution function of the final wealth is a function of the portfolio weights; in the dynamic case, it is a function of the whole trading strategy (see section 6.1 for a review of the notation used). Thus, the multicriteria, multiperiod portfolio optimization control consists in finding the *efficient trading strategies* $\{\mathbf{y}^*\} = \{\mathbf{y}_i^*; i = 0, 1, \cdots, T-1\}$ which make the performance vector $\mathbf{h}^*(\tilde{W}_T)$ an E-point. In a risk-reward framework, the performance vector $\mathbf{h}$ will be composed of the $risk(\tilde{W}_T)$ and $reward(\tilde{W}_T)$. Using theorem B.1.5, we could find an efficient trading strategy by maximizing one component of the vector $\mathbf{h}(\tilde{W}_T(\cdot))$ while keeping the others constrained (although this method is considered "naive").

Assuming we have a two dimensional performance vector, (e.g., risk-reward) we would like to find efficient trading strategies that are solution of the problem:

$$\begin{aligned} & \max reward(\tilde{W}_T) \\ \text{such that} \quad & risk(\tilde{W}_T) \leq L_p \\ & W_0: \text{ given}, \qquad \{\mathbf{y}\} \in \mathcal{Y} \end{aligned} \tag{6.8}$$

where W_T is the final wealth based on a self-financing strategy which generated a wealth

process $\{W\}$ that evolves according to equation (6.1), L_p is a predefined "risk" value and $\mathcal{Y}$ is a set of restriction for the strategy (see section 6.1). The solution of the following problem is also an efficient trading strategy:

$$\begin{aligned} &\min risk(\tilde{W}_T) \\ \text{such that} \quad &reward(\tilde{W}_T) \geq R_p \\ &W_0: \text{ given}, \qquad \{\mathbf{y}\} \in \mathcal{Y}, \end{aligned} \tag{6.9}$$

and where R_p is a predefined "reward" value. Methods to obtain efficient trading strategies for the particular case of the mean-VaR_α are analyzed in the following sections. As far as the dynamic case is concerned, we must be very careful about the trading strategies producing the "best" performance vector. There could be trading strategies denoted as *gambling* where the proportion of wealth invested in the risky assets is increased as the wealth decreases. To avoid gambling, we prefer trading strategies which are always consistent, (a concept that is explained in the next section). If gambling arises, it is necessary to enforce consistency.

6.4 Trading strategies with one risky asset

To further simplify the analysis of the trading strategies, we will consider two assets only: one of them risky with a random simple gross return $\tilde{b}_i = 1 + \tilde{r}_i$, the other risk-free; i.e., the return is non random and has a simple gross return of $g_i = 1 + r_{i,f}$. Defining the variable u_i (interpreted as the *total* amount of wealth invested in the risky asset), a trading strategy will be of the form $\{u\} = \{u_i; i = 0, 1 \cdots, T-1\}$. An admissible trading strategy belongs to the set $\mathcal{U}$

$$\mathcal{U} = \{u_i \in \Re, i = 0, \ldots, T-1\},$$

(although we can impose a restriction in short-selling and budget constraints as well). The wealth evolution (6.10) for a self-financing strategy $\{u\}$ is defined as

$$\tilde{W}_1 = W_0 g_1 + u_0(\tilde{b}_1 - g_1),$$

$$\tilde{W}_{i+1} = \tilde{W}_i g_{i+1} + u_i(\tilde{b}_{i+1} - g_{i+1}) \qquad i = 1, 2, \cdots, T. \tag{6.10}$$

The wealth dynamics are completely dependent on the chosen trading strategy $\{u\}$.

6.4.1 The consistency concept

Before explaining the concept of consistency, we need to understand a problem that arises when dynamic programming techniques with semivariance utility functions (equation (3.9)) are used. Due to the piecewise nature of the utility function, if the wealth falls below a certain level (i.e., the wealth moves from the linear part of the utility function to the quadratic), the optimal strategy requires the investor to "gamble" to recover her losses, i.e., so as to increase the amount invested in the risky assets. Similar problems happen with dynamic option replication strategies, as explained in section 6.2.

We will now need to introduce the concept of consistency. Our goal is to obtain portfolio strategies which improve the performance measure without having to resort to gambling trading strategies. To do so, we borrow the wealth elasticity concept (section B.3.1). from the expected utility theory.

For a trading strategy with only one risky asset, the trading strategy $\{u\}$ is *consistent* if, for $i = 0, 1, \cdots, T-1$, the derivative

$$\frac{du_i(W_i)}{dW_i}$$

always keeps the same sign (is either non-positive or non-negative). The interpretation of this constraint is simple: because u_i is the "investing rule" assigned to the i-th period, function of the available wealth W_i at period i-th, we do not allow rules that change the behavior of an investor based on the level of wealth. The investors' behavior can be categorized depending on their wealth elasticity. In section B.3.1 we described how the wealth elasticity $\nu > 0$ can be interpreted in the expected utility theory as elastic, in which the fraction of wealth invested in the risky asset (i.e., u_i) is directly proportional to W_i. For reasons explained above, if the wealth decreases, gambling strategies are not allowed (i.e., be inversely proportional to W_i). Hence, the derivative $\frac{du_i(W_i)}{dW_i}$ should always be non-negative, so whenever the wealth

decreases, the amount invested in the risky asset does not increase.

In economic terms, we define a *consistent* investor as one whose wealth elasticity ν remains unchanged whether his demand for the risky asset is elastic, unitary elastic to inelastic, as function of W_i. A *consistent trading strategy* is defined as a trading strategy in which the wealth elasticity as function of the wealth at period i ($\nu(W_i)$) is either $\nu(W_i) >= 1$ or $\nu(W_i) < 1$, for all $i = 0, \ldots, T-1$. This means that if an investor is non-increasing risk averse, she will always decrease the amount invested in the risky asset as her wealth decreases, but there will not be a case when she suddenly decides to increase the dollar amount invested in the risky asset as her wealth decreases. This concept is very useful in the multiperiod case.

In the expected utility case, if we want to avoid non-consistent strategies, we have to impose non-consistency to functions where $dR_A(W)/dW \leq 0 \ \forall W$. An example of a *non-consistent* utility function is the semivariance utility function. Because the semivariance is inelastic when $W \leq \tau$, the trading strategy tends to increase the proportion of wealth invested in the risky asset as the wealth *decreases*; which can be interpreted as a "gambling" trading strategy that a risk averse investor is not interested in pursuing.

6.4.2 Quantile constraints

Like the static period optimization, our goal is to develop a methodology that will allow us to find preferred trading strategies in the Pareto sense. The basic idea of the multiperiod optimization with quantile constraints relies on the use of a self-financing strategy (see 6.2) with a finite horizon. A trading strategy which maximizes the expected return while keeping a quantile constrained to a fixed value, is equivalent to finding the optimal mean-VaR_α trading strategy. The investor is assumed to invest her initial wealth W_0, reinvest it at some intermediate trading periods i, and decide about optimality of the strategy based on the predefined probability α the final wealth $\tilde{W}_T$ falls below a disaster level L. Upholding the quantile constraint of portfolio optimization, we do not require the maximization of a complex utility function: to maximize the expected wealth $E[\tilde{W}_T]$ or the $E[\ln \tilde{W}_T]$ at period T suffices. But our goal also requires the trading strategy to be consistent; we denote a trading strategy $\{u^*\}$ as an efficient trading strategy from the set of consistent trading

strategies.

Our goal is to find an efficient trading strategy from the set of consistent trading strategy $\{u\}$ that solves the following problem at time 0:

$$\begin{aligned} &\max reward(\tilde{W}_T) \\ &\text{such that} \qquad q_\alpha(\tilde{W}_T) \geq L \qquad \{u\} \in \mathcal{U} \end{aligned} \tag{6.11}$$

where L is a predefined risk level. Unfortunately, the constraint $q_\alpha(\tilde{W}_T)$ is non-separable (i.e., the problem cannot be separated into an independent subproblem for each time period, see [11]), which makes the solution of this problem non-trivial.

We divide the problem in two parts:

- first, we identify the set of consistent trading strategies that satisfy the quantile constraint $q_\alpha(\tilde{W}_T)$,
- and then we find the efficient trading strategy. (The one that maximizes the *reward*, since the trading strategy satisfies the quantile constraint).

A consistent trading strategy which satisfies the constraint $q_\alpha(\tilde{W}_T) \geq L$ is denoted an α-*consistent trading strategy* . The set of α-consistent trading strategies is denoted as $\mathcal{U}_\alpha$, where

$$\mathcal{U}_\alpha(L) = \{u_i \in \Re, i = 0, \ldots, T-1, q_\alpha(\tilde{W}_T) \geq L\}.$$

Due to the quantile constraint, we cannot use the classical dynamic programming techniques. However, we use a similar approach: first we solve the case for the last period, and then we propagate our results backwards in time.

6.4.3 With a single period

The main purpose of the disaster level constraint consists in quantifying a preference relation which defines an investor's characteristics. However, we need to make additional assumptions about the investors' behavior, to help us differentiate among possible options whenever the disaster level constraint does not distinguish alternatives:

i. The risky asset has a probability distribution such that $g_i > q_\alpha$. If not, then the risky asset would be *always* preferred to the risk-free asset when the disaster level q_α is used to select investments; $u_i = 1$. (i.e., no *arbitrage opportunities*, otherwise, the outcome of the risky asset is greater than the risk-free return). This case would appear if α is not correctly chosen and the absolute risk measure is unable to quantify the risk of an asset (see example 5.5.4).

ii. $E[\tilde{b}_i] > g_i$. If this situation is violated, the risk-free asset has a better expected return than the risky one, and the investor will always choose the risk-free asset, $u_i = 0$.

iii. If $W_{T-1} < L/g_T \equiv L_{T-1}$, and all the previous assumptions are true, then we will assume the investor is *consistent* (i.e., once the future value of her current wealth is below the disaster level, the investor will decide to reinvest exclusively in the risk-free asset; $u_{T-1} = 0$, see B.3). Therefore, the trading strategy will not be allowed to "recover the losses" by gambling. Using the risk aversion concept, a consistent investor will be one that keeps the same risk aversion regardless the amount of wealth.

The single period case with one risky asset has a sub-trading strategy of $\{u\} = \{u_{T-1}\}$. We need to find the set of $\{u\} \in \mathcal{U}_\alpha(L)$, hence we need to enforce the disaster constraint:

$$\Pr\left[W_{T-1}g_T + u_{T-1}(\tilde{b}_T - g_T) \leq L\right] = \alpha. \tag{6.12}$$

Equation (6.12) admits two solutions for u_{T-1}: (i) for $u_{T-1} > 0$ and (ii) for $u_{T-1} < 0$. We assume that solution (i) will always be preferred (otherwise $E[\tilde{W}_T]$ will be negative), hence the α-consistent solution is:

$$u^*_{T-1} = \frac{W_{T-1}g_T - L}{g_T - q_\alpha}, \tag{6.13}$$

which is valid as long as $W_{T-1} \geq L/g_T$. There is no feasible solution when $W_{T-1} < L/g_T$. For the solution to be consistent for all possible non-negative values of W_{T-1}, the derivative $\frac{du^*_{T-1}(W_{T-1})}{dW_{T-1}}$ should not become negative. A solution that reflects this condition is

$$\check{u}_{T-1}(W_{T-1}, \alpha) = \max\left(\frac{W_{T-1} - L/g_T}{1 - q_\alpha/g_T}, 0\right), \tag{6.14}$$

where $\tilde{u}_{T-1}(W_{T-1}, \alpha)$ will be the investment value that will allow consistent strategies. If $W_{T-1} < L/g_T$, a trading strategy cannot be α-consistent in a single period, but (6.14) will allow us to derive the set of α-consistent trading strategies for two or more periods.

6.4.4 With two periods

The new sub-trading strategy is $\{u\} = \{u_{T-2}, u_{T-1}\}$, and the final wealth $\tilde{W}_T$ will be a function of u_{T-1}, u_{T-2}, α, W_{T-2} and the random returns:

$$\tilde{W}_T = \tilde{W}_{T-1} g_T + u_{T-1}(\tilde{b}_T - g_T), \tag{6.15}$$

$$\tilde{W}_{T-1} = W_{T-2} g_{T-1} + u_{T-2}(\tilde{b}_{T-1} - g_{T-1}). \tag{6.16}$$

To find the α-consistent trading strategy, using (6.14) (which is a piecewise function), we must use the total probability property:

$$\begin{aligned} \Pr[\tilde{W}_T \leq L] = \; & \Pr\{\tilde{W}_T \leq L | \tilde{W}_{T-1} \leq L_{T-1}\} \Pr[\tilde{W}_{T-1} \leq L_{T-1}] + \\ & \Pr\{\tilde{W}_T \leq L | \tilde{W}_{T-1} > L_{T-1}\} \Pr[\tilde{W}_{T-1} > L_{T-1}], \end{aligned} \tag{6.17}$$

where $L_{T-1} \equiv L/g_T$ (the present value of the disaster level at the period $T-1$). Substituting (6.14) in (6.15), we can have different outcomes:

i. when W_{T-1}, the outcome of $\tilde{W}_{T-1}$ is $W_{T-1} < L_{T-1}$, the amount to be invested in the risky asset is $u_{T-1} = 0$ (by 6.4.3). The final wealth is totally invested in the risk-free asset, $\tilde{W}_T = W_{T-1} g_T$, and $\tilde{W}_T$ falls below the disaster level with probability 1:

$$\Pr\{\tilde{W}_T \leq L | \tilde{W}_{T-1} \leq L_{T-1}\} = 1. \tag{6.18}$$

ii. when $W_{T-1} \geq L_{T-1}$ the formula (6.13) applies. It is true that a $u^*_{T-1}(W_{T-1}, \beta_{T-1})$ can be chosen to fix $\Pr\{\tilde{W}_T \leq L | \tilde{W}_{T-1} > L_{T-1}\} = \beta_{T-1}$.

We define the following probabilities and conditional probabilities appearing in (6.17) as:

$$\gamma_{T-1} \equiv \quad \Pr[\tilde{W}_{T-1} \leq L_{T-1}], \tag{6.19}$$

$$\beta_T \equiv \Pr\{\tilde{W}_T \leq L | \tilde{W}_{T-1} > L_{T-1}\}. \tag{6.20}$$

If definitions (6.19), (6.20), and equation (6.18) are substituted in equation (6.17), solving for γ_{T-1} we have:

$$\gamma_{T-1} = \frac{\alpha - \beta_T}{1 - \beta_T}, \quad \beta_T \in [0, \alpha]. \tag{6.21}$$

Solving equation (6.19) for u_{T-2} and assuming the investor is consistent in both periods we can propose the following consistent u_{T-2}

$$\check{u}_{T-2}(W_{T-2}, \gamma_{T-2}) = \max\left(\frac{W_{T-2} - \frac{L}{g_T g_{T-1}}}{1 - \frac{q_{\gamma_{T-2}}}{g_{T-1}}}, 0\right). \tag{6.22}$$

The solution of equation (6.20) is

$$u_{T-1}(W_{T-1}, \beta_T) = \frac{W_{T-1} - L/g_T}{1 - q_{\beta_T}/g_T},$$

hence, the sub-trading strategy $\mathbf{u}$ is also α-consistent as long as $W_{T-2} \geq L/(g_T g_{T-1})$.

6.4.5 With T periods

For T periods, to enforce $\Pr[\tilde{W}_T \leq L] = \alpha$ we need to expand the total probability $T-1$ times. Defining the variable L_j (which is useful for the expansion of the total probability formula) for the T periods, $i = 1, \ldots, T$

$$L_j \equiv \begin{cases} L\left(\prod_{k=j+1}^{T} g_k\right)^{-1}, & j = 0, 1, \ldots, T-1, \\ L, & j = T. \end{cases} \tag{6.23}$$

We have, for each period, the following probability formula:

$$\begin{aligned} \Pr[\tilde{W}_{i+1} \leq L_{i+1}] = \quad & \Pr\{\tilde{W}_{i+1} \leq L_{i+1} | \tilde{W}_i \leq L_i\} \Pr[\tilde{W}_i \leq L_i] + \\ & \Pr\{\tilde{W}_{i+1} \leq L_{i+1} | \tilde{W}_i > L_i\} \Pr[\tilde{W}_i > L_{i+1}]. \end{aligned} \tag{6.24}$$

Again, we define the following probabilities and conditional probabilities (6.24):

$$\gamma_i \equiv \quad \Pr[\tilde{W}_i \leq L_i], \tag{6.25}$$

$$\beta_{i+1} \equiv \quad \Pr\{\tilde{W}_{i+1} \leq L_{i+1} | \tilde{W}_i > L_i\}. \tag{6.26}$$

To satisfy the consistency requirement, we chose trading strategies that make true the following statement: $\Pr\{\tilde{W}_{i+1} \leq L_{i+1} | \tilde{W}_i \leq L_i\} = 1$, which, in turn, also makes true the following:

$$\gamma_i = \quad \frac{\gamma_{i+1} - \beta_{i+1}}{1 - \beta_{i+1}}, \quad \beta_{i+1} \in [0, \gamma_{i+1}], \tag{6.27}$$

$$\gamma_T = \quad \alpha. \tag{6.28}$$

To enforce consistency, at every period $i = 1, 2, \ldots, T-1$ we solve equation (6.26), with solution:

$$\check{u}_i(W_i, \beta_{i+1}) = \frac{W_i - L_i}{1 - q_{\beta_{i+1}}/g_{i+1}}. \tag{6.29}$$

For the initial period, $i = 0$ we use the following investment rule

$$\check{u}_0(W_0, \gamma_1) = \max\left(\frac{W_0 - L_0}{1 - q_{\gamma_1}/g_1}, 0\right), \tag{6.30}$$

where γ_1 is defined by equation (6.28). The whole trading strategy is α-consistent, as long as $W_0 > L_0$ (otherwise, it is only consistent).

We now have a set of α-consistent trading strategies $\{u\}$, as we wanted at the beginning of this section. However, now we have $T-1$ variables that form a new trading strategy: $\{\beta\} = \{\beta_i; i = 2, 3, \cdots, T\}$. These are the new control variables, while the γ_i, u_i and $\tilde{W}_i$ are the new state variables. Maximizing the reward of the final wealth $\tilde{W}_T$ derived from this set of α-consistent strategies comes as the next step.

6.5 Mean maximization

The problem (6.12) now can be posed as the maximization

$$\max reward(\tilde{W}_T) \tag{6.31}$$

subject to

$$\begin{aligned} W_0, \alpha, L, \gamma_T = \alpha \quad &: \text{given}, \\ \{\beta\} \in \{\beta_i \in [0, \gamma_i], \quad &i = 2, \ldots, T\}, \end{aligned} \tag{6.32}$$

where the state variable γ_i evolves according to

$$\gamma_i = \quad (\gamma_{i+1} - \beta_{i+1})/(1 - \beta_{i+1}), \quad i = 1, \ldots, T-1,$$

the trading strategy $\{u\}$ evolves as

$$\begin{aligned} u_i(W_i, \beta_{i+1}) = &\quad (W_i - L_i)/(1 - q_{\beta_{i+1}}/g_{i+1}), \quad i = 1, \ldots, T-1, \\ u_0(W_0, \gamma_1) = &\quad (W_0 - L_0)/(1 - q_{\gamma_1}/g_1), \end{aligned}$$

and the wealth evolves as

$$W_{i+1} = \quad W_i g_{i+1} + u_i(\tilde{b}_{i+1} - g_{i+1}) \quad i = 0, 1, \ldots, T-1.$$

The optimal control problem is to find the trading strategy $\{\beta\} = \{\beta_i; i = 2, 3, \cdots, T\}$, that maximizes the value of $\tilde{W}_T$. Unfortunately, this is not a simple optimal control problem, particularly because the set of admissible trading strategies depends on the state variables γ_i

Sub-optimal trading strategy Instead of solving explicitly (6.31), we propose and analyze an admissible sub-optimal trading strategy. We make all β_i equal to a scalar value β; hence, the α that corresponds to this trading strategy can be derived from the recursive

formulas:

$$\begin{aligned}
\gamma_1 &= \beta, \\
\gamma_i &= \gamma_{i-1}(1-\beta)+\beta, \quad i = 2, \ldots, T-1, \\
\alpha &= \gamma_{T-1}(1-\beta)+\beta.
\end{aligned}$$

The main attraction of this method is its simplicity. Experimentally, it can be seen that it usually over-performs a static trading strategy, i.e., one which only changes the composition of the investment once in an investment horizon, as will be shown in the next experiment.

Experiment 6.5.1 *Dynamic trading strategy.* The previous suboptimal technique was tested on a synthetic example. Using dynamics of the stock price (see appendix C.1), we assumed that exists a financial asset of initial price $S_0 = 1$, with instantaneous expected return $\mu = 0.15$ and instantaneous variance $\sigma = 0.3$. The time horizon for the portfolio optimization was 1 day (or 1/250 business years), the optimization parameters are $\alpha = 0.05$, $L = 0.9$ and the initial wealth is $W_0 = 1$. We used cash as the risk-free asset, assuming a gross simple return of 1.

The suboptimal trading strategy was generated for 3 different scenarios:

i. a optimal trading strategy for a single trade period.

ii. a optimal trading strategy, assuming that the horizon could be divided in 5 trades per horizon.

iii. a optimal trading strategy, assuming that the horizon could be divided in 10 trades per horizon.

The results can be seen in figure 5-1; both the cumulative distribution functions and the histograms are plotted. Since the quantile was constrained to the same value, we can compare the trading strategies using the expected return that each trading strategy generates. It can be seen that, apparently, trading strategies that allow several trading periods during the same time horizon have a better expected return. In the example, for for 1 trade per period the mean is 1.0008, for 5 trades per period the mean is 1.0011, while for 10 trades per

period the mean is 1.0077. However, analyzing the distribution functions, we can notice that increasing the number of trades per period favors "lottery" payoffs. Although the expected return is higher as the number of trades is increased, the median return decreases; for 1 trade per period the median is 0.9978, for 5 trades per period the median is 0.96298 while for 10 trades is 0.94334. For the single period case, the distribution is log-normal, but as the trading periods are increased to 5 and 10, the "weight" of the final distribution shifts from the left tail to the right.

In the example, the dynamic strategy optimizes the performance measures defined by us (the mean and the quantile) while keeping a consistent trading strategy that almost vanished the downside exposure. However, in this case, the reward measure used (the mean) does not really describe the real preference of an investor.

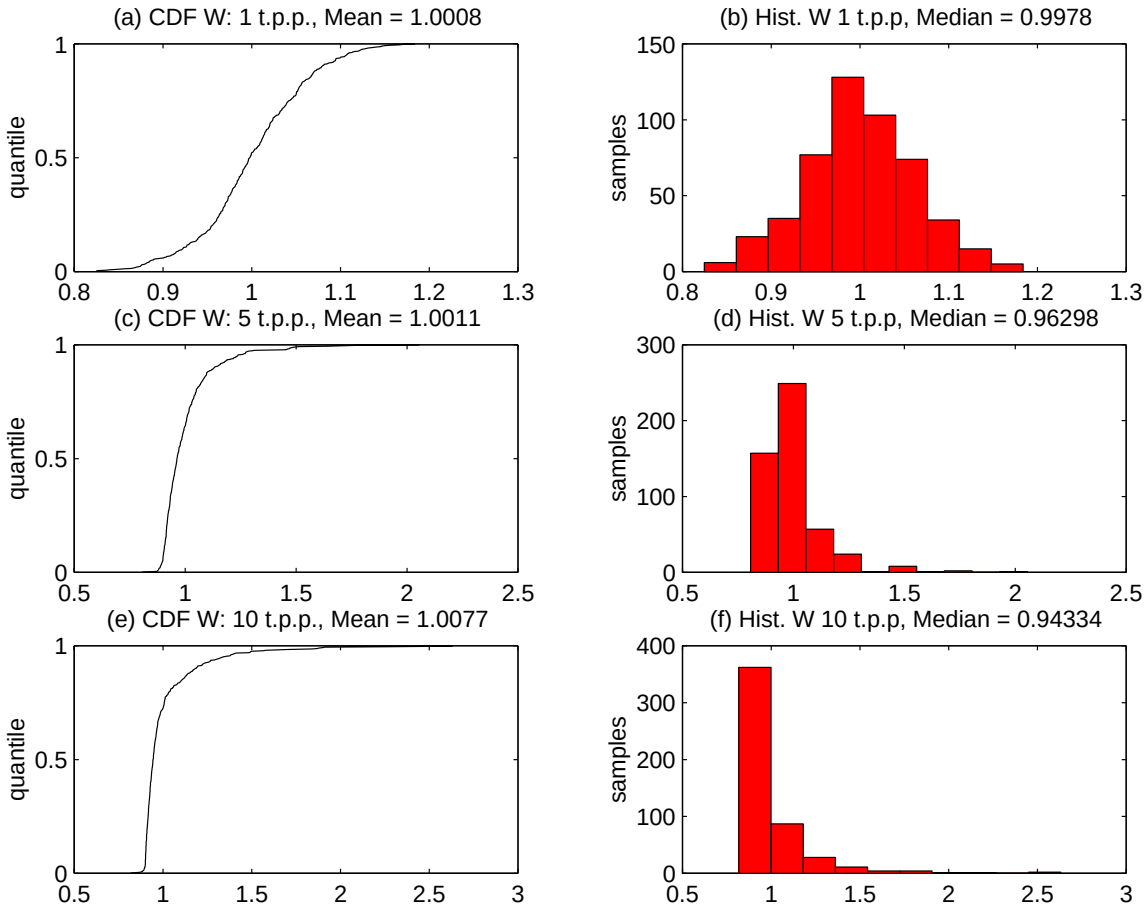

Figure 6-1: Mean-$VaR_{0.05}$ dynamic case.

x-axis: gross return. CDF: cumulative distribution function. t.p.p.: trades per period.

Chapter 7

Summary and conclusions

The expected utility is a special case of the risk-value theory. Although the expected utility theory has been the solid foundation of Economic theories, practitioners tend to be "loss-averse" rather than utility maximizers; the Allais paradox is the most basic example which contradicts the utility theory. Risk-reward theory was considered at first as a special case of the utility theory. In this thesis we showed that the risk-reward theory applies to a wider set of cases including those explained by the expected utility theory. In the particular cases where the risk and value measures are respectively convex and concave, it is possible to assign an ordinal value function that ranks financial assets with random returns. The value function can coincide with the expected utility function, but not with quantile-based risk measures, where it is sometimes possible to find a value function even though no utility function exists.

When value functions exist, Economic theories hold. It is only the very special case where the risk is not convex or the reward is not concave, that there will be no value functions. Theoretically, this would present some interesting contradictions to the current Economic theories. Concepts such as the no-arbitrage condition should be used in the future to restrict the risk and reward measures (e.g., to be at least pseudo-convex and pseudo-concave respectively, see [60]).

The mean and pseudo-coherent risk optimization methods have interesting theoretical properties. We analyzed the mathematics of optimal portfolios and efficient

frontiers for pseudo-coherent risk measures, even if they are not concave Pseudo-coherent measures of risk (as the VaR and the shortfall) have a CAPM-like formula for optimal portfolios, which can be used to further analyze the returns of some financial assets. It is interesting to note that other asymmetric risk measures as the semivariance and the LPM do not share this property, which may explain why they are seldom used in practice. If a risk-free asset is available for investment, then, even if the risk measure is not concave, the efficient frontier will be linear, and there will be mutual fund separation; that might not be the case if no risk-free asset is available.

The gradient of quantile-based risk measures can be estimated and applied to static portfolio optimizations. The gradients of risk measures play a very important role; they can be used to compute a "generalized" β that could be used by practitioners for many financial applications, for example to identify the risk of a particular trade or to optimize portfolios. An important contribution of this thesis is the derivation of an analytical formula for the quantile functions gradient. This gradient is used to derive analytical formulas for gradients of quantile-based risk measures. Then, the use of the local polynomial regression allows us to estimate efficiently these gradients. Depending on the application one can either chose the most precise (but computationally expensive), or the fastest (F-transformation, less precise) method; both methods outperformed the finite differences method in our experiments.

Although some non-gradient based algorithms for the optimization of the VaR exist, they tend to be very slow when the number of samples is increased (which is necessary to calculate the quantile with more precision). Gradient-based algorithms perform better, but the nature of the quantile estimator make finite difference methods very unreliable. So far, the best optimization algorithm we used is the gradient-based optimization algorithm which uses local polynomial regression for the estimation of the gradient.

Comparing results between the shortfall and the VaR, we have noticed that the dependence of the optimized portfolio with respect to the α parameter can be drastic in a few extreme cases (particularly with the "Put-Call" data and the VaR measure). In our experiments, the efficient frontiers of the VaR, as well as the efficient portfolios did not change as

smoothly as they should have, due to the sampling noise. Still, the shortfall shares many of the interesting properties of the VaR, and in experiments it tends to be more robust to noise and to extreme cases such as the "Put-Call" data. For that reason we believe the shortfall is a better measure of risk.

Risk-reward optimization can be performed in the multiperiod case. We introduced an additional characteristic of the risk-aversion concept, which we called *consistency*. We assumed risk-averse investors would avoid changing their behavior after a loss, and would not start gambling their remaining wealth. The expected utility theory does not prevent gambling; i.e., if we try to maximize the semivariance utility function in the multiperiod case, we will have a "non-consistent" optimal policy, due to the piecewise nature of the risk-measure. We show in our thesis that the risk-reward concept can be extended to the multiperiod case, deriving a sub-optimal trading strategy for the case of the mean-absolute VaR optimization. Although the trading strategy indeed improves the performance vector, the resulting probability distribution function of the final wealth might "mislead" the investors: although it is true that the investment becomes less and less risky (in an absolute VaR sense) as the number of trading periods increase, the investment does not become more "valuable": most of the time the investor will lose money. That points out the inefficacy of the mean as reward measure. Other measures of reward (preferably coherent or pseudo-coherent, as the median) could be defined in the future, but such that efficient computational algorithms can be derived for them.

7.1 Limitations

The local polynomial regression requires the computation of an adequate bandwidth, and it will introduce a bias, albeit small. Numerical optimization methods require the computation of the gradient at each iteration; since the computation of the optimal bandwidth for each gradient estimator can be very computationally expensive, the optimization might not be adapted to practical purposes. Sub-optimal bandwidths can be used to save computational time, but unfortunately, the optimal solution will be affected by the estimator bias.

In practice, the optimization algorithm becomes very inefficient when the number of assets is increased. This suggests the need to first come up with an adequate selection of financial assets through either a manual classification of assets or through a heuristic algorithm.

We must remember that the algorithm assumes that N sample data points are available, and that they are i.i.d.; hence the optimization will be as good as the available data is. Historical data provides a challenge, as it is scarce, and financial processes are thought to be non-stationary.

In mathematical Finance it is easy to forget the interaction with the real life investor. Many assumptions are made to simplify the mathematics, even if they cannot be made in real life. For that reason, we must remember that practitioners may not always behave according to what is mathematically simpler. Although the risk-reward theory apparently models investors' behavior in a more realistic way (using two measures to form a performance vector), the measures used so far might in some cases mislead the investor, e.g.;

- the mean can be very high, but only due to the very "fat" tail in the right side of the distribution,
- the $VaR_{0.05}$ may not measure correctly the possible amount of losses, when compared with the $VaR_{0.01}$ measure (high sensitivity of the risk measure with respect to the α parameter).

7.2 Future research

- A very interesting topic that still needs to be developed consists in the characterization of probability distribution classes that make quantile-based risk measures concave or pseudo-concave. Most of the previous analysis tends to assume that the risk measure is concave, to simplify the theoretical analysis. In the VaR case, the analysis should be reversed: the classes of distributions that give a concave VaR measure must be characterized, first taking into consideration that arbitrage opportunities do not exist in financial applications.
- Many financial instruments are known to have non-normal distributions. Using the

techniques developed in this thesis, a complete analysis of the characteristics of optimal portfolios including those financial instruments should be quite interesting.

- The optimization algorithm developed for the static case appeared to be quite reliable when the number of assets is limited (less that 20). In practice, portfolios may hold hundreds of assets; it is imperative to further optimize the programming code and analyze ways to solve large-scale problems.

- We used well-known measures of risk and reward. Our analysis of the dynamic strategy points out that better measures of reward should be used to accurately reflect the preference of an investor. From a theoretical point of view, coherent measures of reward (or even pseudo-coherent, as the median) are very attractive and will inherit the CAPM-like equation for the optimal portfolio. However, for new reward measures efficient ways of computing the optimal solution should be found.

Appendix A

Background

A.1 Miscellaneous functions and definitions

Definition A.1.1 (Generalized inverse) *Suppose h is a non-decreasing function on $\Re$. The* generalized inverse *of h is defined as*

$$h^{-1}(t) = \inf\{x|h(x) \geq t, x \in \Re\}.$$

By convention, the infimum of an empty set is ∞.

Remark: If h is right-continuous, then h is increasing if and only if h^{-1} is continuous.

Definition A.1.2 (Unit step function) *The* unit step function *is the formal integral of the Dirac delta function and is given by*

$$u_s(x) = \begin{cases} 1, & x > 0, \\ 0, & x < 0, \end{cases} \tag{A.1}$$

Definition A.1.3 (Homogeneous functions) *A first-order homogeneous function will be a function that satisfies, for any real $\alpha > 0$, the property $h(\alpha\mathbf{x}) = \alpha h(\mathbf{x})$.*

Remark: It can easily be shown that homogeneous functions of first-order will also satisfy the following properties:

- $\mathbf{x}'\nabla h(\alpha\mathbf{x}) = h(\mathbf{x})$.
- $\nabla h(\mathbf{x}) = \nabla h(\alpha\mathbf{x})$.
- $\nabla^2 h(\mathbf{x})\mathbf{x} = 0$.

Definition A.1.4 *Let $H(\mathbf{x})$ be a real valued function on R^m. H is called* differentiable *at the point $\mathbf{x}$, if there is a vector $\mathbf{x} \in \Re^m$ with the property that for the Euclidean norm $\|\cdot\|$*

$$\frac{|H(\mathbf{x}) - H(\mathbf{y}) - \mathbf{z}'(\mathbf{x} - \mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|} \to 0 \quad \text{as} \quad \mathbf{y} \to \mathbf{x}. \tag{A.2}$$

The row vector $\mathbf{z}$ appearing in (A.2) is called the gradient *of H at $\mathbf{x}$ and is denoted by*

$$\nabla_{\mathbf{x}} H(\mathbf{x}) = \left[\frac{\partial H(\mathbf{x})}{\partial x_1}, \cdots, \frac{\partial H(\mathbf{x})}{\partial x_m}\right]'.$$

If there is no danger of confusion the symbol $\nabla H(\mathbf{x})$ will also be used.

Remark: The partial derivatives of a function $H(\mathbf{x})$ can be approximated by using a *Forward Difference Formula,*

$$\frac{\partial H(\mathbf{x})}{\partial x_i} \approx \frac{1}{c}(H(\mathbf{x} + c \cdot \mathbf{e}_i) - H(\mathbf{x})),$$

where c is a small positive scalar, and $\mathbf{e}_i$ is the i-th unit vector. A more accurate approximation (although computationally more intense) will use the *Central Difference Formula:*

$$\frac{\partial H(\mathbf{x})}{\partial x_i} \approx \frac{1}{2c}(H(\mathbf{x} + H\mathbf{e}_i) - H(\mathbf{x} - c \cdot \mathbf{e}_i)).$$

An error ϵ due to finite precision arithmetic results in error of ϵ/c for the forward difference formula, or $2\epsilon/c$ for the central difference formula.

If we write $H(x_1, \cdots, x_n)$ in place of $H(\mathbf{x})$, the notation $D_j H$ is often used to denote the derivative of H with respect to x_j, keeping the other variables fixed. $D_j H$ is called a *partial derivative.*

Theorem A.1.1 (Differentiation of integrals.) *Suppose ψ is a function of two variables; it will be convenient to use the notation $\psi^t(x) = \psi(x, t)$. Thus ψ^t is, for each t, a function of one variable. Suppose*

i. $\psi(x,t)$ *is defined for* $a \le x \le b$, $x \le t \le d$;

ii. α *is an increasing function on* $[a,b]$;

iii. $\psi^t \in \Re(\alpha)$ *for every* $t \in [c,d]$;

iv. $c < s < d$, *and to every* $\epsilon > 0$ *corresponds a* $\delta > 0$ *such that*

$$|(D_2\psi)(x,t) - (D_2\psi)(x,s)| < \epsilon$$

for all $x \in [a,b]$ *and for all* $t \in (s-\delta, s+\delta)$.

Define

$$f(t) = \int_a^b \psi(x,t)d\alpha(x) \qquad (x \le t \le d). \tag{A.3}$$

then $(D_2\psi)^s \in \Re(\alpha)$, $f'(s)$ *exists, and*

$$f'(s) = \int_a^b (D_2\psi)(x,s)d\alpha(x). \tag{A.4}$$

Note that (iii) simply asserts the existence of integrals (A.3) for all $t \in [c,d]$. *Note also that (iv) certainly holds whenever* $D_2\psi$ *is continuous on the rectangle on which* ψ *is defined.*

Proof. See Rudin [55].

Definition A.1.5 *Let* (X_n) *be a sequence of random variables defined on some probability space* $(\Omega, \mathcal{A}, P)$.

- (X_n) *converges* in probability *(notation:* $X_n \xrightarrow{IP} X$*), if* $P(\|X_n - X\| > \epsilon) \to 0$ *for all* $\epsilon > 0$.
- (X_n) *converges* almost surely *(notation:* $X_n \to X$ *a.s.), if* $P(X_n \to X) = 1$.

Theorem A.1.2 Strong Law of Large numbers. *Let* (ξ_i) *be a sequence of independent identically distributed random variables with* $E(\|\xi_1\|) < \infty$. *Then*

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \to E(\xi_1) \qquad \text{a.s.}$$

This theorem may also be stated as follows: let $\hat{\rho}_n$ be the empirical measure

$$\hat{\rho}_n = \frac{1}{n}\sum_{i=1}^{n}\delta(i)$$

and suppose that $H(\xi)$ is integrable. Then

$$\frac{1}{n}\sum_{i=1}^{n}H(\xi_i) = \int H(u)d\hat{\rho}_n(u) \to \int H(u)d\rho(u) = E(H(\xi_1))$$

Proof. See Etemadi [22].

Definition A.1.6 *A subset C of $\Re^m$ is said to be* convex *if $(1-\lambda)\mathbf{a}+\lambda\mathbf{b} \in C$ whenever $\mathbf{a} \in C$, $\mathbf{b} \in C$ and $0 < \lambda < 1$.*

Definition A.1.7 *Let g be a function from C to $(-\infty, +\infty]$, where $C \subset \Re^m$ is a convex set. Then g is convex on C if and only if*

$$g((1-\lambda)\mathbf{a}+\lambda\mathbf{b}) \leq (1-\lambda)g(\mathbf{a})+\lambda g(\mathbf{b}),\ 0 < \lambda < 1,$$

for every $\mathbf{a}$ and $\mathbf{b}$ in C. g will be concave *if its negative is convex.*

A.2 Quantile functions

Two important measures of risk, the VaR (defined in section 3.2.2) and the shortfall (defined in section A.2.3), use the quantile function as a key element. For that reason, in section A.2 we will review the general properties of quantile functions as described in the modern literature, in particular for linear combinations of random variables (section A.2.1), elliptic distributions (section A.2.2), and other quantile based functions, such as the shortfall (section A.2.3). In section A.2.4 we review current techniques to estimate quantiles, quantiles of linear combination of random vectors (section A.2.4), and the shortfall (section A.2.4). Part of this chapter has been adapted from [48, 8, 53, 17, 37].

Definition A.2.1 Quantile function: *Suppose F is the cumulative distribution function of a real valued random variable $W \in \Re$. Given a threshold probability $\alpha \in (0,1)$, the*

α-quantile *(or population quantile)* q_α *of* F *is defined as:*

$$q_\alpha \equiv F^{-1}(\alpha) = \inf\{w \in \Re : F(w) \geq \alpha\}. \tag{A.5}$$

When we want to explicitly express the quantile as a function of α, we will use the notation $q_\alpha \equiv q(\alpha)$ to denote the α-quantile.

A function $q(\alpha)$ will be monotonically non-decreasing, given that all cumulative distributions functions are non-decreasing. If the cumulative distribution is strictly increasing, then $q(\alpha)$ will be continuous.

A.2.1 Quantile of a linear function of random vectors

Let $\tilde{\mathbf{b}}$ be a m-dimensional random vector with cdf F, and let $\mathbf{x}$ be a deterministic vector $\mathbf{x} \in \Re^m$. We can define the linear combination $\tilde{W}(\mathbf{x}) = \mathbf{x}'\tilde{\mathbf{b}}$. The α-quantile of the linear combination will be denoted as $q_\alpha(\mathbf{x})$, or $q(\alpha, \mathbf{x})$, and comes directly from the definition A.2.1 (see [48, 8, 53, 17]).

The quantile function will be *homogeneous of order 1* (see definition A.1.3), a property that can be derived from the definition of quantile (see A.5):

$$q_\alpha(t\mathbf{x}) = tq_\alpha(\mathbf{x}) \qquad \forall t > 0. \tag{A.6}$$

It is also easy to show (by differentiating the previous equation with respect to t, and rearranging) that the following property of homogeneous functions will be true:

$$q_\alpha(\mathbf{x}) = \mathbf{x}'\nabla_{\mathbf{x}} q_\alpha(\mathbf{x}). \tag{A.7}$$

A.2.2 Quantile of an elliptic distribution

Definition A.2.2 *The* $m \times 1$ *random vector* $\tilde{\mathbf{b}}$ *is said to have an elliptical distribution with*

parameters μ $(m \times 1)$ *and* $\mathbf{V}$ $(m \times m)$ *if its density function is of the form*

$$\tilde{\mathbf{b}} \sim \mathrm{E_m}(\mu, \mathbf{V}) \equiv c_m |\mathbf{V}|^{-1/2} h\left((\tilde{\mathbf{b}} - \mu)' \mathbf{V}^{-1} (\tilde{\mathbf{b}} - \mu)\right),$$

for some function h*, where* $\mathbf{V}$ *is positive definite [37].*

The normalizing constant c_m could be absorbed into the function h. If $\tilde{\mathbf{b}}$ has an m-dimensional elliptical distribution we will write that $\tilde{\mathbf{b}}$ is $\mathrm{E_m}(\mu, \mathbf{V})$. In the special case when the random vector $\tilde{\mathbf{b}}$ has an elliptic distribution of the form $\mathrm{E_m}(\mathbf{0}, \mathbf{I}_m)$, where $\mathbf{I}_m$ is the $m \times m$ identity matrix, $\tilde{\mathbf{b}}$ will have a *spherical distribution*. Also, if $\mathbf{D}$ has an m-variate spherical distribution with a density function and the random vector $\tilde{\mathbf{b}} = \mathbf{CD} + \mu$, where $\mathbf{C}$ is a nonsingular $m \times m$ matrix, then $\tilde{\mathbf{b}}$ has an elliptic distribution $\mathrm{E_m}(\mu, \mathbf{V})$ with $\mathbf{V} = \mathbf{CC}'$.

The characteristic function $\phi(\mathbf{t}) = E[\exp(i\mathbf{t}'\tilde{\mathbf{b}})]$ has the form $\phi(\mathbf{t}) = \exp(i\mathbf{t}\mu)\psi(\mathbf{t}'\mathbf{V}\mathbf{t})$ for some function ψ. Provided they exist, $E[\tilde{\mathbf{b}}] = \mu$ and $Cov(\tilde{\mathbf{b}}) = \gamma\mathbf{V}$ for some constant γ. In terms of the characteristic function this constant is $\gamma = -2\psi'(0)$.

If $\tilde{W}(\mathbf{x}) = \mathbf{x}'\tilde{\mathbf{b}}$, the random variable $\tilde{W}$ is a linear combination of the elliptic distribution, and will have a symmetric univariate distribution that can be standardized. Lets define the standard variable $p_\alpha(\mathbf{x}) = (q_\alpha(\mathbf{x}) - E[\tilde{W}])/v_W$, where $E[\tilde{W}] = \mathbf{x}'\mu$ and $v_W = \sqrt{\mathbf{x}'\mathbf{V}\mathbf{x}}$. Then,

$$q_\alpha(\mathbf{x}) = \mathbf{x}'\mu - p_\alpha\sqrt{\mathbf{x}'\mathbf{V}\mathbf{x}}. \tag{A.8}$$

Because $\mathbf{V}$ is a positive semidefinite matrix, then $q_\alpha(\mathbf{x})$ is concave The formula for $\nabla_{\mathbf{x}} q_\alpha(\mathbf{x})$ is

$$\nabla_{\mathbf{x}} q_\alpha(\mathbf{x}) = \mu - p_\alpha \frac{\mathbf{V}\mathbf{x}}{\sqrt{\mathbf{x}'\mathbf{V}\mathbf{x}}}. \tag{A.9}$$

Two very important examples of elliptic distributions are the Multivariate Gaussian and t distributions, which are described below.

Multivariate Gaussian Distributions

For the Multivariate Gaussian distributions, the linear combination $\tilde{W}(\mathbf{x})$ has a univariate distribution

$$\tilde{W}(\mathbf{x}) \sim \mathrm{N}(\mathbf{x}'\mu, \sqrt{\mathbf{x}'\mathbf{V}\mathbf{x}}),$$

and the matrix $\mathbf{V}$ is also the covariance matrix

$$\mathbf{V} = \mathbf{\Sigma} \equiv E[(\tilde{\mathbf{b}} - \mu)(\tilde{\mathbf{b}} - \mu)'].$$

In the case of the Gaussian distribution, the parameter p_α will be called z_α.

Multivariate t distributions

The spherical multivariate t distribution (of order d) is

$$f(\mathbf{y}) = \frac{\Gamma\left(\frac{1}{2}(d+m)\right)}{\Gamma\left(\frac{1}{2}d\right)(d\pi)^{m/2}} \frac{1}{\left(1 + \frac{1}{d}\mathbf{y}'\mathbf{y}\right)^{(d+m)/2}},$$

where m is the size of the vector $\mathbf{y}$. If $\mathbf{y} = \mathbf{V}^{1/2}(\tilde{\mathbf{b}} - \mu)$, the previous equation can be converted to a multivariate elliptic distribution. The relationship between $\mathbf{V}$ and $\mathbf{\Sigma}$ is $\mathbf{V} = \mathbf{\Sigma}(d-2)/d$, for $d \geq 3$. The univariate distribution of a linear combination of $\tilde{\mathbf{b}}$, $\tilde{W}(\mathbf{x})$ is a univariate t distribution of order d. The parameter p_α for the particular case of elliptic distributions will be called $t_{\alpha,d}$.

A.2.3 Shortfall

Definition A.2.3 (Shortfall function) *Let $\tilde{W}$ be a random variable; then*

$$e(\alpha) = E\{q_\alpha - \tilde{W} | \tilde{W} < q_\alpha\} \tag{A.10}$$

is called the shortfall function *of $\tilde{W}$.*

As for the quantile function, the previous definition can easily be extended to handle random variables generated by the linear combination of the components of a random vector $\tilde{W} = \mathbf{x}'\tilde{\mathbf{b}}$, then:

$$e_\alpha(\mathbf{x}) = E\{q_\alpha(\mathbf{x}) - \tilde{W}(\mathbf{x}) | \tilde{W}(\mathbf{x}) < q_\alpha(\mathbf{x})\}. \tag{A.11}$$

From the definition $e_\alpha(\mathbf{x})$ can be represented in the following format:

$$e_\alpha(\mathbf{x}) = \frac{1}{\alpha}\int_{-\infty}^{q_\alpha(\mathbf{x})} F_W(w)dw = q_\alpha(\mathbf{x}) - \frac{1}{\alpha}\int_0^\alpha q_\omega(\mathbf{x})d\omega.$$

From the last representation of the shortfall function, we can see that the shortfall is also a homogeneous function of order 1.

A.2.4 Estimation of an α-quantile

To estimate q_α (see [8]), lets suppose we have available a set $\mathcal{Y}_n$ with n i.i.d. samples from a real valued distribution F; $\mathcal{Y}_n = \{Y_1, \cdots, Y_n\}$. An estimator will use the *empirical distribution function*:

$$F_n(y) = \frac{1}{n}\sum_{t=1}^{n} 1_{\{y > Y_i\}}. \tag{A.12}$$

We may estimate $q_\alpha = F^{-1}(\alpha)$ as $\hat{q}_{\alpha,n} \equiv F_n^{-1}(\alpha)$. The properties of this estimator can be expressed in terms of ordered set of i.i.d samples, or the *order statistics*:

$$Y_{1:n} = \min(\mathcal{Y}_n) \leq Y_{2:n} \leq \cdots \leq Y_{n:n} = \max(\mathcal{Y}_n).$$

Then, for $k = 1,\cdots,n$

$$\hat{q}_{\alpha,n} = Y_{k:n}, \quad \tfrac{k-1}{n} < \alpha \leq \tfrac{k}{n},$$

where $\hat{q}_{\alpha,n}$ is known as the estimator of q_α or as the *sample quantile*.

In the continuous case, if the population quantile q_α has a cumulative distribution F with density f such that $f(q_\alpha) \neq 0$, $k = n\alpha + 1 + o(n^{1/2})$, and $0 < \alpha < 1$, then the distribution of the sample quantile when $n \to \infty$ is asymptotically normal (see [17]). For certain distributions, the estimator $\hat{q}_{\alpha,n}$ is an unbiased estimator, with zero mean error η;

although that might not be true for arbitrary distributions, in the asymptotic case it can be considered that $q_\alpha = \hat{q}_{\alpha,n} + \eta$ where the error η behaves as

$$\eta \sim \mathrm{N}\left(q_\alpha, \frac{\alpha(1-\alpha)}{nf^2(q_\alpha)}\right) \quad \text{if} \quad n \to \infty. \tag{A.13}$$

Using the i.i.d. assumption, in the continuous case it is possible to obtain, for $r < s$, distribution free confidence intervals:

$$\Pr[Y_{r:n} \leq q_\alpha \leq Y_{s:n}] = \sum_{i=1}^{s-1} \binom{n}{i} \alpha^i (1-\alpha)^{n-i} \equiv \pi(r, s, n, \alpha). \tag{A.14}$$

The previous equation can be used to derive confidence intervals when n is not too large. In the discrete case, however,

$$\Pr[Y_{r:n} \leq q_\alpha \leq Y_{s:n}] \geq \pi(r, s, n, \alpha).$$

For some discrete random variables, it could happen that there is no error when estimating the quantile.

Example A.2.1 *Confidence interval.* In figure A-1 we plotted the empirical distribution function of the random return of the equally weighted portfolio $\mathbf{x} = [1, 1, 1]'/3$ on the option-based data of example 2.2.1. The dashed lines represent the 95 % confidence interval that the quantile will be within that region. For example, there is a 95 % probability that the 5%-quantile belongs to the interval $[-0.04801, -0.013306]$.

Estimator of α-quantiles for linear combinations of a random vector

Denote by $\mathbf{b}_1, \cdots, \mathbf{b}_n$ a random sample of an m-dimensional random vector $\tilde{\mathbf{b}}$, and by $\mathbf{x}$ an m-dimensional deterministic vector. Applying the linear transformation $W_i = \mathbf{x}'\mathbf{b}_i$, $i = 1, \cdots, n$ to the data, we will be able to assign a *marginal ordering* (M-ordering) to the m dimensional random sample. To do so, we compute the order statistics of the linear

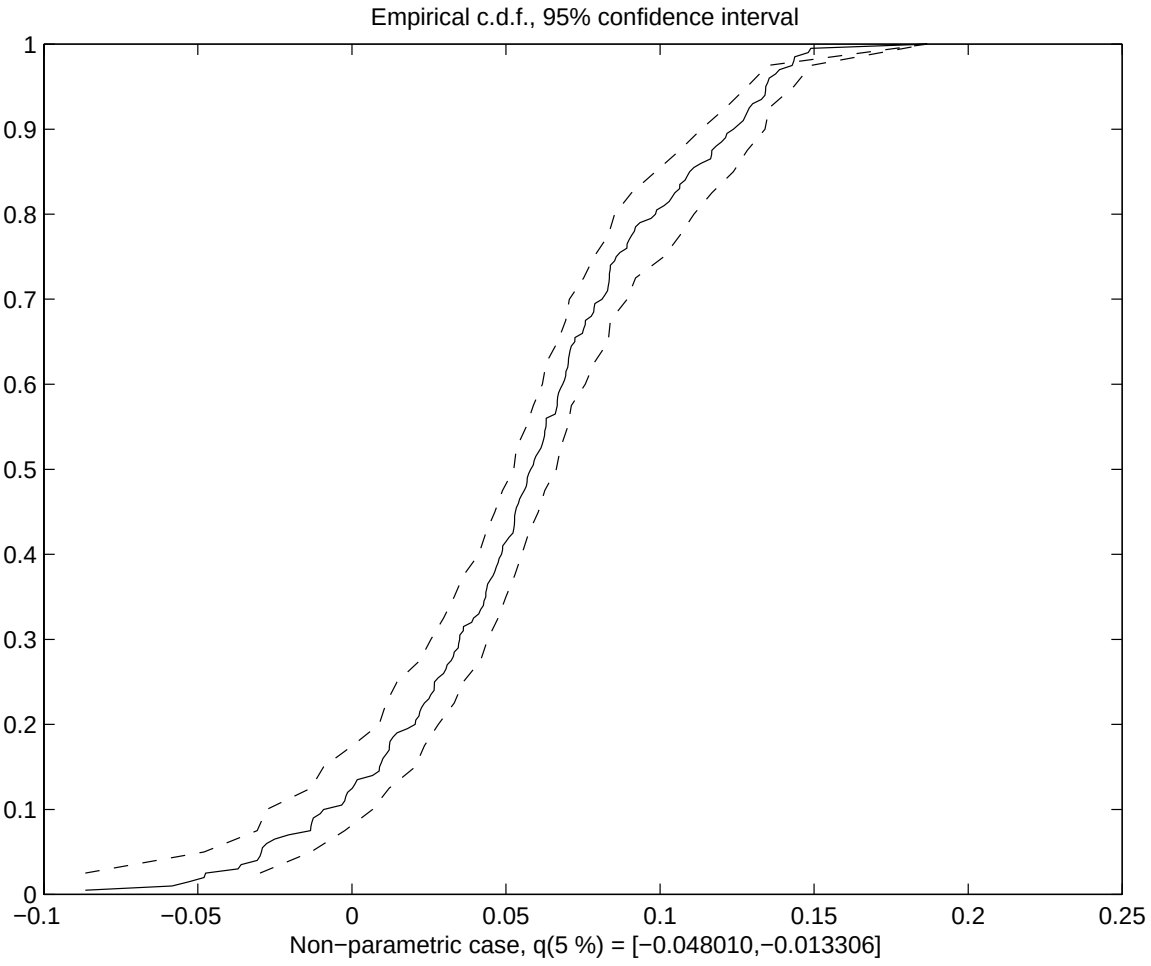

Figure A-1: Non-normal portfolio returns.

transformation $W_i \equiv \mathbf{x}'\mathbf{b}_i$ following the procedure defined in the section A.2.4)

$$W_{1:n} \leq W_{2:n} \leq \cdots \leq W_{n:n}.$$

Each i-th order statistic $W_{i:n}$ will have associated with a sample random vector which we will call $\mathbf{b}_{i:n,\mathbf{x}}$ or $\mathbf{b}_{i:n}$, if $\mathbf{x}$ is constant in a section.

The vector $\mathbf{b}_{i:n}$ is known in the literature as the *M-order statistic* (see [7, 53]).

Empirical shortfall

Suppose that $W_1, \cdots, W_n$ are i.i.d. with d.f. F, let F_n denote the empirical cumulative distribution function and $\Delta_n(\alpha) = \{i : i = 1, \cdots, n; W_i < \hat{q}_\alpha\}$. The *empirical shortfall function* is

$$\hat{e}_n(\alpha) = \frac{1}{\alpha}\int_{-\infty}^{\hat{q}_{\alpha,n}} F_n(w)dw = \frac{1}{\text{card}\Delta_n(\alpha)} \sum_{i\in\Delta_n(\alpha)} (\hat{q}_{\alpha,n} - W_i), \tag{A.15}$$

with the convention that $0/0 = 0$.

A.3 Constrained optimization

For a comprehensive review of this topic, see [49, 11, 9]. The general form of an optimization problem is

$$\text{minimize } H(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \mathcal{S} \tag{A.16}$$

The value that minimizes the problem (A.16) will exist if $\mathcal{S}$ is nonempty. Every point $\mathbf{x}^*$ $\in \arg\min\{H(\mathbf{x}) : \mathbf{x} \in \mathcal{S}\}$ is called a *minimizer* of H.

Optimization over convex sets One of the classical nonlinear programming problems is the constrained problem over a convex set. In the particular case when the convex can be defined with r equalities and t inequalities the optimization problem can be posed as:

$$\text{minimize} \qquad H(\mathbf{x}) \tag{A.17}$$

$$\text{subject to} \qquad \mathbf{g}(\mathbf{x}) = \mathbf{0}, \tag{A.18}$$

$$\mathbf{d}(\mathbf{x}) \leq \mathbf{0}, \tag{A.19}$$

$$\mathbf{x} \in \Re^m, \tag{A.20}$$

where the vector $\mathbf{g}(\mathbf{x})' = [g_1(\mathbf{x}), \ldots, g_r(\mathbf{x})]$, represents the r equality constraints, and the vector $\mathbf{d}(\mathbf{x})' = [d_1(\mathbf{x}), \ldots, d_t(\mathbf{x})]$ represents the t inequality constraints. The scalar functions $H(\mathbf{x})$, $g_i(\mathbf{x})$ and $d_i(\mathbf{x})$ are usually assumed to be continuously differentiable. A large arsenal of techniques has been developed to solve this instance of the problem (A.16). To simplify the analysis, we can discard the inequality constraint (for the complete analysis refer to [11]) and propose the optimization problem

$$\text{minimize } H(\mathbf{x}) \quad \text{subject to} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0}.$$

The first order necessity constraints of an optimal solution, also known as Kuhn-Tucker condition, are

$$\nabla_x L(\mathbf{x}^*, \lambda^*) = \nabla_x H(\mathbf{x}^*) + \nabla_{\mathbf{x}} \mathbf{g}(\mathbf{x}^*)' \lambda^* = \mathbf{0}, \tag{A.21}$$

$$\nabla_\lambda L(\mathbf{x}^*, \lambda^*) = \mathbf{g}(\mathbf{x}^*) = \mathbf{0}. \tag{A.22}$$

Where L is the Lagrangian function

$$L(\mathbf{x}, \lambda) = H(\mathbf{x}) + \lambda' \mathbf{g}(\mathbf{x}).$$

Penalty-based constrained optimization

There are several methods available to solve this problem. One of the approaches is based on elimination of constraints through the use of *penalty functions* $P(\mathbf{x})$. For example, the quadratic penalty function method consists of sequential unconstrained minimization of the form:

$$\min_{x \in \Re^m} H(\mathbf{x}) + \frac{1}{2} c_k \|\mathbf{g}(\mathbf{x})\|^2,$$

where $\{c_k\}$ is a positive scalar sequence with $c_k < c_{k+1}$ for all k and $c_k \to \infty$; the penalty function $P(\mathbf{x})$ is $\frac{1}{2} c_k \|\mathbf{g}(\mathbf{x})\|^2$. However, this method has many disadvantages such as slow

convergence and ill-conditioning for large values of c_k. Since each iteration is an unconstrained minimization, for analysis purposes it can be used to understand how convergence is affected when the function $H(\mathbf{x})$ and its gradient are replaced by noisy estimators.

Non-differentiable penalty function

The non-differentiable exact penalty function is one of the favorites techniques used for the optimization of constrained problems. It can be shown that the (non-differentiable) unconstrained problem has the same local minimum x^* as in the constrained case,

$$\min_{x \in \Re^m} H(\mathbf{x}) + cP(\mathbf{x}),$$

where $c > 0$ and P is the non-differentiable penalty function defined by

$$\begin{aligned} P(\mathbf{x}) &= \max_{i=1,\ldots,r} |g_i(\mathbf{x})|, \\ c &> \sum_{i=1}^{m} |\lambda_i^*|. \end{aligned}$$

This method also is applied when there are inequalities, and is already implemented in several optimization packages such as Matlab [29]. It has been proven to be very efficient, and its also very fast, since it is a quasi-Newton method. A detailed analysis of this method can be found in [11, 29]. Differentiable exact penalty functions could also be used in the case when the Hessian of the function H is available.

Another simple penalty function could be $P(\mathbf{x}, \lambda) = \|\nabla_x L(\mathbf{x}, \lambda)\|^2 + \|\mathbf{g}(\mathbf{x})\|^2$, but it has some drawbacks since it does not discriminate between local minima and maxima.

A.3.1 Stochastic optimization

Stochastic optimization problems are characterized by the fact that not all decision-relevant data are exactly known at the time when the decision has to be made. Mathematically, uncertainty is described by random variables, which appear in the optimization model (A.16). The random variables appearing in the cost function may or may not depend on the decision

$\mathbf{x}$.

It is typical for a stochastic optimization problem that the objective function $H(\mathbf{x})$ is not explicitly known. A stochastic optimization problem has to be approximated. The typical way of approximation is by simulation: the uncertain random quantities in the original problem (A.16) are either replaced by artificially generated random variables or bootstrapped from a set of historical samples, and the optimization is based on them. Since the generated quantities are random, the approximate problems are random optimization problems.

The approximation of a stochastic optimization problem is based on the validity of *law of large numbers*. For example, we could optimize the expectation of a random process assuming that for each decision value $\mathbf{x}$ we may generate stochastic sequences $\xi_1, \xi_2, \cdots, \xi_n$ such that the empirical measure

$$\hat{\rho}_n = \frac{1}{n}\sum_{i=1}^{n} \delta(\xi - \xi_i), \qquad \text{converges weakly to } \rho_{\mathbf{x}}. \tag{A.23}$$

here $\delta(u - u_0)$ denoted the point mass (Dirac mass) at the point u_0.

The non-recursive method

Also called the "stochastic counterpart", or "sample path optimization". We generate a sequence $\xi_1, \xi_2, \cdots, \xi_n$ of random variables such that the empirical measure defined in (A.23) converges weakly to ρ. If the function H in (A.16) represents the expectation of a random process, then inserting the empirical measure $\hat{\rho}$ instead of ρ in problem (A.16) we get the approximate problem (A.24):

$$\text{minimize } H_n(\mathbf{x}) \quad \text{subject to} \quad \mathbf{x} \in \mathcal{S}. \tag{A.24}$$

The solution of (A.24) is used as an approximative solution of the original problem (A.16). This solution is not restricted to empirical expectations of random processes; we will apply this method to the maximization of quantiles in section 5.

The recursive method

As a basis, one may take any algorithm for deterministic optimization which requires either the function values $H(\mathbf{x})$ or the gradients $\nabla H(\mathbf{x})$. Whenever the algorithm makes a function call to H one replaces this unknown value by an estimate

$$\hat{H}(\mathbf{x}) := G(\mathbf{x}, \xi), \text{ where } \xi \sim \rho.$$

If the algorithm needs gradients $\nabla H(\mathbf{x})$, one uses *stochastic gradients*, i.e. unbiased estimates $\hat{\nabla} H(\mathbf{x})$ of $\nabla H(\mathbf{x})$. Sometimes we may not find an unbiased estimate and we must accept *approximative gradients* (stochastic quasi-gradients), i.e. estimates of $\nabla H(\mathbf{x})$ with a small bias.

The recursive methods produces one random sequence of approximative solutions $(\mathbf{x}_k)$, where $\mathbf{x}_{k+1}$ depends on $\mathbf{x}_k$ and on k^* random observations:

$$\mathbf{x}_{k+1} = T_k(\mathbf{x}_k, \xi_1^{(k)}, \cdots, \xi_{k^*}^{(k)}).$$

For some algorithms, $k^* = 1$ is sufficient or k^* depends only on the dimensions of $\mathbf{x}$; in other cases k^* tend to infinity with k. Convergence results may be established for $k \to \infty$. The weakest form of such a result stated that the distance between $\mathbf{x}_k$ and $\arg\min H$ tends to zero in probability.

Optimization stochastic counterparts

We get the *stochastic counterpart* of a deterministic algorithm, if we replace every call to

the function value	$H(\mathbf{x})$	by an unbiased estimate	$\hat{H}(\mathbf{x})$,
the gradient value	$\nabla H(\mathbf{x})$	by an unbiased estimate	$\hat{\nabla} H(\mathbf{x})$,
the Hessian value	$\nabla^2 H(\mathbf{x})$	by an unbiased estimate	$\hat{\nabla}^2 H(\mathbf{x})$.

The condition of unbiasedness can be weakened by a condition about convergence of the bias to zero. A stochastic algorithm must have a structure such that the stochastic error

terms cancel themselves out in the long run by virtue of the law of large numbers (LLN). However, in any case one may enforce the effect of the LLN by taking repeated observations at each search point.

A.3.2 Gradient methods with errors

This appendix deals with algorithms for the optimization of simulated systems. In particular we study stochastic variants of the gradient algorithm

$$\mathbf{x}_{n+1} = \mathbf{x}_k - a_k \nabla H(\mathbf{x}_k), \tag{A.25}$$

which is useful to solve the problem

$$\min H(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \Re^m,$$

where H is bounded from below.

The stochastic version of (A.25) is needed in the case where the objective function $H(\mathbf{x})$ or its gradient $\hat{\nabla} H(\mathbf{x}) = \nabla H(\mathbf{x})$ can be observed only by computer simulation. Suppose for each $\mathbf{x} \in \Re^m$ one may get an estimate $\hat{\nabla} H(\mathbf{x})$ of $\nabla H(\mathbf{x})$ which contains a deterministic error $\mathbf{b}(\mathbf{x})$ and a zero-mean random error $\mathbf{w}(\mathbf{x})$

$$\hat{\nabla} H(\mathbf{x}) = \nabla H(\mathbf{x}) + \mathbf{b}(\mathbf{x}) + \mathbf{w}(\mathbf{x}),$$

The *systematic error* $\mathbf{b}(\mathbf{x})$ contains the *bias* of $\hat{\nabla} H(\mathbf{x})$ in situations where an unbiased estimate of $\nabla H(\mathbf{x})$ is impossible.

The stochastic generalization of the gradient method is based on the recursion

$$\mathbf{x}_{n+1} = \mathbf{x}_k - a_k(\nabla H(\mathbf{x}_k) + \mathbf{b}_k + \mathbf{w}_k), \tag{A.26}$$

where $\mathbf{b}_k = \mathbf{b}(\mathbf{x}_k)$ and $\mathbf{w}_k = \mathbf{w}(\mathbf{x}_k)$. The *uncontrollable error* is $\mathbf{v}_k = \mathbf{b}_k + \mathbf{w}_k$. We will let $\mathbf{y}_k = \hat{\nabla} H(\mathbf{X}_k)$.

For simplicity, we can focus on the steepest descent method with errors, for unconstrained

problems. Several cases can arise:

i. The error is small relative to the gradient, that is $\|\mathbf{v}_k\| < \|\nabla H(\mathbf{x}_k)\|$, $\forall n$. Then convergence is assured, since $\mathbf{y}_k$ will always make an angle less than 90 degrees with respect to $\nabla H(\mathbf{x}_k)$.

ii. $\{\mathbf{v}_k\}$ is bounded, that is $\|\mathbf{v}_k\| \leq \epsilon$, $\forall n$, where ϵ is some scalar. Then, the method operates like a descent method within the region $\{x|\|\nabla H(\mathbf{x})\| > \epsilon\}$. In the complementary region, where $\|\nabla H(\mathbf{x})\| \leq \epsilon$ the behavior of the method will depend on the nature of the error. For example, if the errors $\mathbf{v}_k$ are constant, say $\mathbf{v}_k \equiv \mathbf{v}$, then since $\mathbf{y}_k = \nabla H(\mathbf{x}_k) + \mathbf{v}$, the method will essentially be trying to minimize $H(x) + \mathbf{v}'\mathbf{x}$ and will typically converge to a point $\bar{\mathbf{x}}$ with $\nabla H(\bar{\mathbf{x}}) = -\mathbf{v}$. If the errors $\mathbf{v}_k$ vary substantially, the method will tend to oscillate within the region where $\|\nabla H(\mathbf{x})\| \leq \epsilon$. The precise behavior will also depend on whether a constant or diminishing step-size is used.

iii. $\{\mathbf{v}_k\}$ is proportional to the step-size, that is $\|\mathbf{v}_k\| \leq ka_k$,$\forall n$, where k is some scalar. If the step-size is constant, we come under case (ii), while if the step-size is diminishing, the behavior described in case (ii) applies, but with $\epsilon \to 0$, so the method will tend to converge to a stationary point of H.

iv. $\{\mathbf{v}_k\}$ are independent zero mean random vectors with finite variance. The steepest descent method will converge because the effects of the error term are "averaged out". With a diminishing step-size, the occasional use of a bad direction $\mathbf{y}_k$ cannot deteriorate the cost enough for the method to oscillate.

Cases (i) and (ii) are typical of non-recursive algorithms. In the beginning of the algorithm the noise due to the finite sample does not affect the convergence of the algorithm, as in case (i.), but it will converge as if the algorithm were trying to optimize the function $H(x) + \mathbf{v}'\mathbf{x}$ (case ii.). In the vicinity of a fixed point $\mathbf{x}^*$ the finite sample error $\mathbf{v}$ behaves as a constant error; in non-recursive algorithms only approximated solutions exist.

Recursive algorithms will profit from the behavior described in cases (iii) and (iv), as explained in the next section.

Convergence and asymptotic distribution

In this section we will introduce one theorem that establishes convergence in stochastic optimization using a martingale approach. The theorem will be stated for the slightly more general situation of constrained optimization. Suppose that the set of constraints C is a closed convex set. Pflug [49] uses the projection operator π_C in the *projected stochastic gradient algorithm*

$$\mathbf{x}_{k+1} = \pi_C(\mathbf{x}_k - a_k \mathbf{y}_k), \tag{A.27}$$

and proves convergence using martingale methods.

Theorem A.3.1 *Consider the recursion equation (A.27) where*

$$\mathbf{y}_k = \nabla H(\mathbf{x}_k) + \mathbf{b}_k + \mathbf{w}_k.$$

Let $\mathcal{H}_k$ be an increasing sequence of σ-algebras such that $\mathbf{X}_k$ and $\mathbf{b}_k$ are $\mathcal{H}_k$-measurable and $E\{\mathbf{w}_k|\mathcal{H}_k\} = 0$. If

i. there is a $\mathbf{x}^* \in \Re^m$ such that $\inf\{\nabla H(\mathbf{x})'(\mathbf{x} - \mathbf{x}^*)|\ \epsilon \le \|\mathbf{x}\| \le \epsilon^{-1}\} > 0$ for all $\epsilon > 0$,

ii. $\|\nabla H(\mathbf{x})\| \le K_1\|\mathbf{x} - \mathbf{x}^*\| + K_2$,

iii. $\sum a_k\|\mathbf{b}_k\| < \infty$ a.s.,

iv. $\sum a_k^2 E\{\|\mathbf{w}_k\|^2|\mathcal{H}_k\}) < \infty$ a.s.,

v. $a_k \ge 0$, $a_k \to 0$, $\sum a_k = \infty$, $\sum_k a_k^2 < \infty$,

then $\mathbf{x}_k$ converges to $\mathbf{x}^*$ almost surely.

Proof. See Pflug ([49]).

The Kiefer-Wolfowitz procedure

This procedure applies in situation when there are unbiased estimates $\hat{H}(x)$ of $H(x)$ (in the univariate case) available, but not such estimates of $\nabla H(x)$. Using the central finite

difference estimate for the univariate case:

$$D(x,c) \equiv \frac{\hat{H}(x+c) - \hat{H}(x-c)}{2c} = \nabla H(x) + b(x) + w(x),$$

The systematic error of $D(x,c)$ is

$$b(x,c) = \frac{H(x+c) - H(x-c)}{2c} - \nabla H(x), \tag{A.28}$$

which is small if c is small. The zero-mean random error

$$w(x,c) = \frac{\hat{H}(x+c) - H(x+c) - \hat{H}(x-c) + H(x-c)}{2c}, \tag{A.29}$$

has unbounded variance if c tends to zero and $\hat{H}(x+c)$ is independent from $\hat{H}(x-c)$. The right choice for c as a sequence of constants tending to zero (but not too fast) is crucial for the KW-procedure. Coupling $\hat{H}(x+c)$ and $\hat{H}(x-c)$ is another method of controlling the variance.

In the multivariate case, suppose that for each parameter point $\mathbf{x}$ one may observe a random variable $\hat{H}(\mathbf{x})$ with expectation $H(\mathbf{x})$. The Kiefer-Wolfowitz (KW)-procedure uses divided differences to estimate the gradient of H. Let $(\mathbf{e}_i)_{i=1,\cdots,m}$ be the unit vectors in $\Re^m$. The KW procedure is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \sum_{i=1}^{m} \left(\frac{\hat{H}(\mathbf{x}_k + c_k\mathbf{e}_i) - \hat{H}(\mathbf{x}_k - c_k\mathbf{e}_i, c_k)}{2c_k} \right) \mathbf{e}_i. \tag{A.30}$$

The a.s. convergence if this procedure is a consequence of theorem A.3.1. The speed of convergence is given in the following remark. It is assumed that all error variables are independent (the complete derivation can be found in [49]).

Remark A.3.1 *Specializing a_k and c_k to $a_k = a/k^\alpha$ and $c_k = c/k^\gamma$, where $\alpha \leq 1$, $\alpha+\gamma > 1$, $2\alpha - 2\gamma > 1$, after some calculation the following result results:*

$$E\{\|\mathbf{x}_k - \mathbf{x}^*\|^2\} = \begin{cases} O(k^\gamma), & \gamma < \alpha/3, \\ O(k^{2\gamma-\alpha}), & \gamma \geq \alpha/3. \end{cases}$$

The best choice for α and γ is $\alpha = 1$ and $\gamma = 1/3$ which results in

$$E\{\|\mathbf{x}_k - \mathbf{x}^*\|^2\} = O(k^{-1/3}).$$

As can be seen, the convergence rate is quite poor. In practical cases this method should only be used when there is no other alternative.

The Robbins-Monro procedure

The Robbins-Monro procedure requires the existence of unbiased estimates of the gradient ∇H of H. It is assumed that for each x there is a random variable $\mathbf{Y}(\mathbf{x})$ such that $E[\mathbf{Y}(\mathbf{x})] = \nabla H(\mathbf{x})$ and consider the recursion

$$\mathbf{x}_{k+1} = \mathbf{x}_k - a_k \mathbf{Y}(\mathbf{x}_k). \tag{A.31}$$

Remark: Suppose that

i. There are constants C_0, C_1, such that

$$C_0\|\mathbf{x} - \mathbf{x}^*\|^2 \leq \nabla H(\mathbf{x})'(\mathbf{x} - \mathbf{x}^*) \leq k_1\|\mathbf{x} - \mathbf{x}^*\|^2,$$

ii. $\|\nabla H(\mathbf{x})\| \leq K_1\|x\| + K_2$, which implies that $\|\nabla H(\mathbf{x})\|^2 \leq K_3\|x\|^2 + K_4$, where $K_3 = 2K_1^2$ and $K_4 = 2K_2^2$,

iii. $Var(\mathbf{Y}(\mathbf{x}_j)) \leq \sigma_j^2$,

iv. $\sum_k a_k \to \infty$, $\sum_k a_k^2 < \infty$.

The method will converge, since it is a special case of the theorem A.3.1. Now specialize a_k to $a_k = a/k^\alpha$, and assumelets that $\sup_j \sigma_j^2 < \infty$. If $1/2 < \alpha < 1$, then

$$E\{\|\mathbf{x}_k - \mathbf{x}^*\|^2\} = O(n^{-\alpha}).$$

If $\alpha = 1$, then

$$E\{\|\mathbf{x}_k - \mathbf{x}^*\|^2\} = O(n^{-1}),$$

provided that $2aV_0 > 1$. Consequently, the best choice is $\alpha = 1$.

The convergence rate of the Robbins-Monro procedure is better than the KW procedure. Clever implementations (as the "two pass" method described in [9]) can reduce the convergence rate in the order of a Stochastic Newton method.

Stopping times

The convergence result tells us that we will get to the desired point, if we let the procedure run for sufficiently long time. Such a statement is evidently unsatisfactory for practical purposes. In practice, what is done usually is to stop the algorithm once the accuracy of the solution (denoted as the fraction $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|/\|\mathbf{x}_{k+1}\|$) reaches a pre-established accuracy level (e.g. 10^{-4}).

A.4 Local polynomial regression

A.4.1 Introduction

The sections that concern local polynomial regression are adapted from Fan and Gijbels, [25]; the presentation given here is simply an overview to introduce the concepts used in the estimation of gradient of quantile functions.

Consider the bivariate data (X_1, Y_1), ..., (X_n, Y_n), which can be thought as a realization from a population (X, Y). Of interest is often to estimate the regression function $m(x_0)$ $= E(Y|X = x_0)$ and its derivatives $m'(x_0)$, $m''(x_0)$, ..., $m^{(p)}(x_0)$. To help us understand the estimation methodology, we can regard the data as being generated from the model

$$Y = m(X) + \sigma(X)\omega,$$

where $E(\omega) = 0$, $\sigma(\omega) = 1$, and X and ω are independent. We always denote the conditional variance of Y given $X =x_0$ by $\sigma^2(x_0)$ and the marginal density of X, i.e., the *design density*, by $f(\cdot)$.

Supposed that the $(p+1)^{\text{th}}$-derivative of $m(x)$ at the point x_0 exists. We then approximate

the unknown regression function $m(x)$ locally by a polynomial of order p. A Taylor series expansion gives, for x in a neighborhood of x_0,

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + \frac{m''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \qquad \text{(A.32)}$$

This polynomial is fitted locally by weighted least squares regression problem: minimize

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 K_h(X_i - x_0), \qquad \text{(A.33)}$$

where h is a bandwidth controlling the size of the local neighborhood, and $K_h(\cdot) = K(\cdot/h)/h$ with K a kernel function assigning weights to each datum point.

Denote by $\hat{\beta}_j$, $j = 0, \ldots, p$, the solution to the least squares problem (A.33). It is clear from the Taylor's expansion in (A.32) that $\hat{m}_\nu(x_0) = \nu!\hat{\beta}_\nu$ is an estimator for $m^{(\nu)}(\cdot)$, $\nu = 0, \ldots, p$. To estimate the entire function $m^{(\nu)}(\cdot)$ we solve the above weighted least squares problem for all points x_0 in the domain of interest.

It is more convenient to work with matrix notations. Denote by $\mathbf{X}$ the design matrix of problem (A.33):

$$\mathbf{X} = \begin{bmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{bmatrix}$$

and put

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{bmatrix},$$

Further, let $\mathbf{W}$ be the $n \times n$ diagonal matrix of weights:

$$\mathbf{W} = \operatorname{diag}\left\{ K_h\left(X_i - x_0\right), i = 1, \ldots, n \right\},$$

the weighted least squares problem (A.33) can be written as

$$\min_{\beta}(\mathbf{y}-\mathbf{X}\beta)'\mathbf{W}(\mathbf{y}-\mathbf{X}\beta), \tag{A.34}$$

with $\beta = [\beta_0, \ldots, \beta_p]'$. The solution vector is provided by ordinary least squares theory and is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \tag{A.35}$$

There are several important issues which have to be discussed. First of all is the choice of the bandwidth parameter h, which plays a crucial role. Too large a bandwidth underparametrizes the regression function, causing a large modeling bias, while too small a bandwidth over-parameterizes the unknown function and result in noisy estimates. Another issue in local polynomial fitting is the choice of the order of the local polynomial. Fan and Gijbels recommend using the lowest odd order, i.e. $p = \nu + 1$, or ocassionally $p = \nu + 3$.

Another question concerns the choice of the kernel function K. Fan and Gijbels show that for all choices of p and ν the optimal weight function is $K(z) = \frac{3}{4}(1-z^2)_+$, the Epanechnikov kernel.

Also, according to research by Fan and Gijbels, other kernel estimators suffer from some drawbacks; e.g. undesirable form of the bias, or pay a price in variance for random design models. Particularly, there is absence of boundary effects for local polynomial regression: the bias at the boundary stays automatically of the same order as in the interior, without use of specific boundary kernels.

A.4.2 Bias and variance

The conditional bias and variance of the estimator $\hat{\beta}$ given a finite set of i.i.d samples can be derived from its definition in (A.35):

$$E(\hat{\beta}|\mathbf{X}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{m} = \beta + (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}, \mathbf{X}'\mathbf{W}\mathbf{r}, \tag{A.36}$$

$$Var(\hat{\beta}|\mathbf{X}) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{\Sigma}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}, \tag{A.37}$$

where $\mathbf{m} = [m(X_1),\cdots,m(X_n)]'$, $\mathbf{r} = \mathbf{m} - \mathbf{X}\beta$, the vector of residuals of the local polynomial approximation, $\mathbf{\Sigma} = \text{diag}\ \{K_h^2(X_i - x_0)\sigma^2(X_i)\}$ and $\mathbb{X} = (X_1, \cdots, X_n)$.

These exact bias and variance expressions are not directly accessible, since they depend on unknown quantities: the residual $\mathbf{r}$ and the diagonal matrix $\mathbf{\Sigma}$, although asymptotic expansions can be found. The following notation will be used: the moments of K and K^2 are denoted respectively by

$$\mu_j = \int u^j K(u)du \quad \text{and} \quad \nu_j = \int u^j K^2(u)du.$$

Some matrices and vectors of moments are appearing in the asymptotic expressions:

$$\begin{aligned}
&\mathbf{S} = (\mu_{j+l})_{0\le j,l\le p}, \qquad c_p = [\mu_{p+1}, \cdots, \mu_{2p+1}]', \\
&\tilde{\mathbf{S}} = (\mu_{j+l+1})_{0\le j,l\le p}, \quad \tilde{c}_p = [\mu_{p+2}, \cdots, \mu_{2p+2}]', \\
&\mathbf{S}^* = (\nu_{j+l})_{0\le j,l\le p}.
\end{aligned}$$

Further, we consider the unit vector $\mathbf{e}_{\nu+1} = [0, \cdots, 0, 1, 0, \cdots, 0]'$, with 1 in the $(\nu+1)^{\text{th}}$-position. Theoretical results explain why what most of the time the use of $p - \nu$ odd is preferred in practice; the theoretical expression for the bias has a simpler structure.

A.4.3 Equivalent kernels

From the notation

$$\mathbf{S}_{n,j} = \sum_{i=1}^{n} K_h(X_i - x_0)(X_i - x_0)^j \tag{A.38}$$

lets define $\mathbf{S}_n \equiv \mathbf{X'WX}$, the $(p+1) \times (p+1)$ matrix $(S_{n,j+l})_{0\le j,l\le p}$. The estimator $\hat{\beta}_\nu$ can be written as

$$\begin{aligned}
\hat{\beta}_\nu = \mathbf{e}'_{\nu+1}\hat{\beta} &= \mathbf{e}'_{\nu+1}\mathbf{S}_n^{-1}\mathbf{X'Wy} \\
&= \sum_{i=1}^{n} W_\nu^n\left(\frac{X_i - x_0}{h}\right) Y_i,
\end{aligned} \tag{A.39}$$

Table A.1: The equivalent kernel functions $K^*_{p,\nu}$.

ν	p	Equivalent Kernel Function $K^*_{\nu,p}(t)$
0	1	K(t)
0	3	$(\mu_4 - \mu_2^2)(\mu_4 - \mu^2 t^2)K(t)$

where $W^n_\nu(t) = \mathbf{e}'_{\nu+1}\mathbf{S}_n^{-1}[1, th, \cdots, (th)^p]'K(t)/h$. The weights satisfy the following discrete moment conditions:

$$\sum_{i=1}^{n}(X_i - x_0)^q W^n_\nu\left(\frac{X_i - x_0}{h}\right) = \delta_{\nu,q}, \qquad 0 \leq \nu, q \leq p. \tag{A.40}$$

A direct consequence of this relation is that finite sample bias when estimating polynomials up to order p is zero (while for other methods the zero bias only hold true asymptotically).

An expression for the *equivalent kernel* is

$$K^*_\nu(t) = \mathbf{e}'_{\nu+1}\mathbf{S}^{-1}[1, t, \cdots, t^p]'K(t) = \left(\sum_{l=0}^{p}\mathbf{S}^{\nu l}t^l\right)K(t) \tag{A.41}$$

which satisfies the following moment conditions

$$\int u^q K^*_\nu(u)du = \delta_{\nu,q}, \qquad 0 \leq \nu, q \leq p, \tag{A.42}$$

which are asymptotic version of the discrete moments condition presented in (A.40). Thus, this weighting scheme does not only correct bias up to a polynomial of order p, but also adapts automatically to all design densities.

The conditional bias and variance of the estimators $\hat{m}_\nu(x_0)$ can be expressed in terms of the equivalent kernel K^*_ν, leading to the asymptotic expressions

$$\begin{aligned} Bias\{\hat{m}_\nu(x_0)|\mathbf{X}\} &= \left(\int t^{p+1}K^*_\nu(t)dt\right)\frac{\nu!}{(p+1)!}m^{(p+1)}(x_0)h^{p+1-\nu} + \\ &\quad o_P\left(h^{p+1-\nu}\right) \end{aligned} \tag{A.43}$$

Table A.2: The constants $C_{\nu,p}(K)$.

ν	p	Gaussian	Uniform	Epanechnikov	Biweight	Triweight
0	1	0.776	1.351	1.719	2.036	2.312
0	3	1.161	2.813	3.243	3.633	3.987

and its asymptotic variance equals

$$Var\{\hat{m}_\nu(x_0)|\mathbb{X}\} = \int K^{*2}_\nu(t)dt \frac{\nu!^2\sigma^2(x_0)}{f(x_0)nh^{1+s\nu}} + o_P\left(\frac{1}{nh^{1+2\nu}}\right). \tag{A.44}$$

A.4.4 Ideal bandwidth choice

A theoretical optimal local bandwidth for estimating $m^{(\nu)}(x_0)$ is obtained by minimizing the conditional Mean Squared Error (MSE) given by

$$(Bias\{\hat{m}_\nu(x_0)|\mathbb{X}\})^2 + Var\{\hat{m}_\nu(x_0)|\mathbb{X}\}.$$

Fan and Gijbels derive an asymptotically optimal bandwidth given by

$$h_{opt} = C_{\nu,p}(K)\left(\frac{\int \sigma^2(x)w(x)/f(x)dx}{\int (m^{(p+1)}(x))^2 w(x)dx}\right)^{1/(2p+3)} n^{-1/(2p+3)}, \tag{A.45}$$

where n is the number of data points, $w(\cdot) > 0$ is some weighting function; it is understood that the integrals are finite and that the denominator does not vanish; and

$$C_{p,\nu}(K) = \left(\frac{(p+1)!^2(2\nu+1)\int K^*_\nu(t)dt}{2(p+1-\nu)\left(\int t^{p+1}K^*_\nu(t)dt\right)^2}\right)^{1/(2p+3)}. \tag{A.46}$$

The latter constants are easy to calculate, and some are tabulated in the table A.2.

Commonly used kernels:

$$\text{Gaussian}: \quad K(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2), \tag{A.47}$$

$$\text{Epanechnikov}: \quad K(z) = \frac{3}{4}(1-z^2)_+, \tag{A.48}$$

$$\text{Uniform}: \qquad K(z) = 1_{[-0.5,+0.5]}(z), \tag{A.49}$$

A.4.5 Estimated bias and variance

The bias can be estimated by

$$\mathbf{S}_n^{-1} \begin{bmatrix} \hat{\beta}_{p+1} S_{n,p+1} + \cdots \hat{\beta}_{p+a} S_{n,p+a} \\ \vdots \\ \hat{\beta}_{p+1} S_{n,2p+1} + \cdots \hat{\beta}_{p+a} S_{n,2p+a} \end{bmatrix}, \tag{A.50}$$

where $\hat{\beta}_{p+1}, \cdots, \hat{\beta}_{p+a}$ are the estimated regression coefficients obtained by fitting a polynomial of degree $p+a$ locally. For this $(p+a)^{\text{th}}$-order fit, one needs a pilot bandwidth h^*.

An estimator for the conditional variance is provided by

$$(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^2\mathbf{X})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\hat{\sigma}^2(x_0). \tag{A.51}$$

where the quantity $\hat{\sigma}^2(x_0)$ is an estimator of the unknown quantity $\sigma^2(x_0)$;

$$\hat{\sigma}^2(x_0) = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 K_{h^*}(X_i - x_0)}{\text{tr}\left(\mathbf{W}^* - \mathbf{W}^*\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{W}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{W}^*\right)} \tag{A.52}$$

which results from the $(p+a)^{\text{th}}$-order polynomial fit using the pilot bandwidth h^*. Here $\mathbf{X}^*$ and $\mathbf{W}^*$, similar to $\mathbf{X}$ and $\mathbf{W}$, denote respectively the design matrix and the weight matrix for this local $(p+a)^{\text{th}}$-order polynomial fit.

With the estimated conditional bias and variance, given in (A.50) and (A.51), we obtain the following estimator for the Mean Squared Error of $\hat{\beta}_\nu = \hat{m}(x_0)/\nu!$

$$\hat{MSE}_{p,\nu}(x_0; h) = \hat{\beta}^2_{p,\nu}(x_0) + \hat{V}_{p,\nu}(x_0), \tag{A.53}$$

where $\hat{b}_{p,\nu}(x_0)$ denotes the $(\nu+1)^{\text{th}}$ element of the estimated bias vector in (A.50). Further, the $(\nu+1)^{\text{th}}$ diagonal element of the matrix in (A.51) is denoted by $\hat{V}_{p,\nu}(x_0)$.

Table A.3: Adjusting constants for the Epanechnikov and Gaussian kernel.

Epanechnikov Kernel					
	p	1	2	3	4
$p-\nu$					
1		.8941	.7643	.7776	.7827
3				.8718	.8324
Gaussian Kernel					
	p	1	2	3	4
$p-\nu$					
1		1.00	.8403	.8205	.8085
3				.9554	.8975

RSC constant bandwidth selector The *residual squares criterion* is defined as

$$RSC(x_0; h) = \hat{\sigma}^2(x_0)(1 + (p+1)V) \tag{A.54}$$

where $\hat{\sigma}^2(\cdot)$ is the normalized weighted residual sum of squares after fitting locally a p^{th} order polynomial, and V is the first diagonal element of the matrix

$$(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}^2\mathbf{X})(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}.$$

By finding h^*, the minimizer of the integrated version of the residual squares criterion

$$IRSC(h) = \int_{[a,b]} RSC(y; h)dy,$$

we can obtain the *RSC constant bandwidth selector* as

$$\hat{h}^{RSC}_{\nu,p} = adj_{\nu,p}h^*. \tag{A.55}$$

A.4.6 Pilot bandwidth selection

A rule of thumb is very suitable to get a quick idea about how large the amount of smoothing should be. It is somewhat crude, but possess simplicity and requires very little programming effort that other methods are hard to compete with. A simple way to do so is by fitting a polynomial of order $p+3$ *globally* to $m(x)$, leading to the parametric fit

$$\check{m}(x) = \check{\alpha}_0 + \cdots + \check{\alpha}_{p+2} x^{p+3}. \tag{A.56}$$

the standardized residual sum of squares from this parametric fit is denoted by $\check{\sigma}^2$. Taking $w(x) = f(x)W_0(x)$ for some specific function w_0, (e.g. the indicator function $w_0(x) = 1_{[a,b]}$), a simple bandwidth selector is:

$$\check{h}_{ROT} = C_{p,\nu}(K) \left(\frac{\check{\sigma}^2 \int w_0(x)dx}{\sum_{i=1}^{n} \left(\check{m}^{(p+1)}(X_i)\right)^2 w_0(X_i)} \right)^{1/(2p+3)}. \tag{A.57}$$

This simple bandwidth selector $\check{h}_{ROT}$ is derived under certain conditions. However, even in situations where these conditions are not strictly fulfilled this bandwidth selector can be applied in order to get an initial idea of the amount of smoothing to be used. In practice, the indicator function over an interval has been used as the weighting function $w_0(\cdot)$.

Another simple estimator can be found using an improved Akaike information criterion [34]. Denoting $\hat{\mathbf{y}} = \mathbf{H}\,\mathbf{y}$, with $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}_h\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_h$,

$$AIC_C = \log(\hat{\sigma}^2) + 1 + \frac{2(\mathrm{tr}(\mathbf{H}) + 1)}{n - \mathrm{tr}(\mathbf{H}) - 2}. \tag{A.58}$$

The vector $\hat{\mathbf{y}}$ represents the estimate of $\mathbf{y}$ obtained using the fitted polynomial model. Given an initial pilot estimate h^*, we can find the optimal bandwidth that minimizes the previous equation.

Another option to find the pilot would be to minimize over a grid the following statistic [19]:

$$S = \hat{\sigma}_j^2(g)\left(1 + 18Q_{11}^*(g)\right).$$

Appendix B

Asset ranking theories

In this appendix we briefly review all the finance nomenclature and definitions required in the previous chapters, following closely classic books such as [35, 32] for static portfolio optimization, and [59] for the multicriteria introduction.

Section B.1 reviews the basic theory of preference relations, which is used in Chapter 3 for the comparison of our generalization of the risk-reward theory against the expected utility theory. In the same section, the multicriteria optimization theory is reviewed, a concept used to define efficient portfolios, efficient frontiers and their characteristics in a generalized risk-reward context (sections B.1.2).

The characteristics of the performance space will be a function of the decision variable chosen (the linear weights) (section B.1.3), which will affect the convexity of the efficient frontier. The multicriteria point of view is used in Chapters 3 to analyze and compare the risk-reward framework against existing methods, such as stochastic dominance (section B.2) and utility theory, (reviewed in sections B.3, as well as B.3.1 and B.3.2). The concepts reviewed in section B.3 are used in previous chapters, as we compare the risk-reward framework against established methods of decision under uncertainty, particularly against expected utility theory.

B.1 Preference relations

Let us assume for now that an investor is asked to choose between two possible investment opportunities with uncertain outcomes $\tilde{A}$ and $\tilde{B}$, i.e., $\tilde{A}$ and $\tilde{B}$ are random variables with respective (cumulative) probability distribution functions $F_A(a) = \Pr[\tilde{A} \leq a]$ and $F_B(b) = \Pr[\tilde{B} \leq b]$ which belong to a certain class $\mathcal{F}$ of distributions, not necessarily the class defined in (2.3). If an investor is asked to select only one of the two assets, one and only one of the following can occur:

i. the investor is convinced that $\tilde{A}$ is better than (or is preferred to) $\tilde{B}$, denoted by $\tilde{A} \succ \tilde{B}$ (or $F_{\tilde{A}} \succ F_{\tilde{B}}$, the distribution $F_{\tilde{A}}$ is preferred to the distribution $F_{\tilde{B}}$);

ii. the investor is convinced that $\tilde{A}$ is worse (or less preferred) than $\tilde{B}$, denoted by $\tilde{A} \prec \tilde{B}$ (or $F_{\tilde{A}} \prec F_{\tilde{B}}$); or

iii. the investor is *indefinite* or *indifferent* between $\tilde{A}$ and $\tilde{B}$, thus the preference relation can be defined as $\tilde{A} \sim \tilde{B}$ (or $F_{\tilde{A}} \sim F_{\tilde{B}}$).

Each of the previous statements involves a comparison or *relation* between a pair of outcomes. The symbols "$\prec$", "$\succ$" and "$\sim$" are *operators* defining the comparisons and relations. For each one of the operators we can define a preference; e.g., the preference relation associated with the operator $\prec$ is $\{\prec\}$ (see [59, 35] for a complete coverage of this topic).

It is possible to define the relation $\{\succeq\}$, where the comparison $\tilde{A} \succeq \tilde{B}$ is read "$\tilde{A}$ is weakly preferred to $\tilde{B}$".

Some characteristics of preference relations that very important; depending on some of these characteristics, we will see, in the next section that it may be possible to assign a *multi-objective value function* (value function) to the relation. A value function is a function which assigns a scalar to each possible distribution function so that a complete and transitive ordering of all the distributions can be achieved by ordering the assigned scalar. If the value function exists, the optimization problem becomes a unidimensional problem. The utility theory is a specific case of a value function.

Some assumptions that preference relations (e.g. $\{\succeq\}$) might satisfy are:

i. *completeness.* For every pair of random variables $\tilde{A}$, $\tilde{B}$ with respective cumulative distributions functions $F_{\tilde{A}}, F_{\tilde{B}} \in \mathcal{F}$, either $\tilde{A} \succeq \tilde{B}$ or $\tilde{B} \succeq \tilde{A}$.

ii. *reflexivity.* For every random variable $\tilde{A}$, with $F_A \in \mathcal{F}$, $\tilde{A} \succeq \tilde{A}$.

iii. *transitivity.* If $\tilde{A} \succeq \tilde{B}$ and $\tilde{B} \succeq \tilde{C}$, then $\tilde{A} \succeq C$.

B.1.1 Performance space

From now on we will assume each investor can at least be represented by a transitive and reflexive preference relation $\{\succeq\}$ (although economists would prefer the relation to also be complete, to allow the use of the expected utility theory [35]). If it is possible to express this preference relation numerically, then it may be possible to use the arsenal of techniques developed by researchers in Operations Research to find the optimal portfolio. Although preference relations can be expressed in terms of risk and reward, Economists prefer to use the expected utility function theory to for its many very useful properties that make it possible to analyze the behaviors of many investors.

Depending on the characteristics of the preference relation, there are several ways to express it numerically. However, it should be noted that few researchers agree on the basic numerical representation of preference (Economics is still a social science); hence, even if a numerical method can be efficiently developed to optimize a certain numerical expression, there is no guarantee that the global optimum will be found.

Value Function: A lot of research has been done assuming the preference relation $\{\succeq\}$ is also complete. In this particular case, we will be able to assign a scalar *value function* $v(F_{\tilde{W}}) : F_W \to \Re^1$ to each distribution function $F_{\tilde{W}}$, (or to simplify notation, $v(\tilde{W}) = v(F_{\tilde{W}})$), so that $v(\tilde{A}) \geq v(\tilde{B})$ implies that $\tilde{A} \succeq \tilde{B}$. Therefore, the decision problem is reduced to evaluating the value function $v(\cdot)$. Unfortunately, in practice, such a value function proves difficult to obtain, and also assumes that the relation is complete. Several textbooks give detailed explanations of how to find the value function if those conditions are true.

An alternative to the value function approach, consists in assigning a vector of size Q to the distribution function, i.e., $\mathbf{h}(F_{\tilde{W}}) : F_{\tilde{W}} \to \Re^Q$ (or, to simplify the notation, $\mathbf{h}(\tilde{W})$

$= \mathbf{h}(F_{\tilde{W}})$, $\mathbf{h}(\tilde{W}) = [h_1(\tilde{W}),\ h_2(\tilde{W}),\ \ldots,\ h_Q(\tilde{W})]'$), and then using the concept of Pareto preference to establish a preference relation among the two vectors $\mathbf{h}(\tilde{A})$ and $\mathbf{h}(\tilde{B})$. A Pareto preference is defined as follows:

Pareto Preference : For each component $h_i(\cdot)$, let greater values be more preferred, and assume no other information on the preference is available or established. The Pareto preference is defined by $\tilde{A} \succ \tilde{B}$ if and only if $\mathbf{h}(\tilde{A}) \geq \mathbf{h}(\tilde{B})$, i.e., component-wise $h_i(\tilde{A}) \geq h_i(\tilde{B})$, $i = 1,\ldots,Q$ and $\mathbf{h}(\tilde{A}) \neq \mathbf{h}(\tilde{B})$. A Pareto preference is not complete, and lacks a representation with value functions, unless very specific conditions hold (which will be introduced in a future section B.1.2).

If we assume that we will be able to assign a finite number of parameters to each distribution function, the preference relation will be incomplete whenever two different members of the class share the same parameters. We can see that the value function approach reduces to the expected utility theory when $Q = 1$.

The Risk-Reward theory, as proposed by Markowitz, will assign two scalars to each distribution, ($Q = 2$), and compare the distributions using them. The utility theory can be regarded as a special case of this framework, in which one of the scalars is the utility of a distribution, and the other is always set to a fixed value. Of course, more than 2 scalars can be used, and economists have proposed the use of higher moments than 2, [56, 57]. However, practitioners so far have shown a tendency to work within the realms of the risk-reward framework, given that they prefer to have an objective representation of the risk of certain outcomes.

Lets define the *performance vector* $\mathbf{p} = \mathbf{h}(\tilde{W})$ which belongs to the *performance space* $\mathcal{P}$ $\equiv \{\mathbf{p}|\mathbf{p} = \mathbf{h}(\tilde{W}),\ F_{\tilde{W}} \in \mathcal{F}\}$. In some cases, the representation of the investors' preference relation can be more accurate using the performance space rather than the expected utility theory.

B.1.2 Efficient frontiers

The efficient frontier is a concept widely used in Finance, and in this section we review the basic definitions, as well as some theorems which will be useful to determine algorithms for

the computations of efficient portfolios and efficient frontiers.

A vector $\mathbf{p}^*$ will be an *efficient vector* (E-vector), if there is no $\mathbf{p} \in \mathcal{P}$ such that $\mathbf{p} \neq \mathbf{p}^*$ and $\mathbf{p} \geq \mathbf{p}^*$. Depending on $\mathcal{P}$, there could be one, multiple, infinite E-vectors, or none. The set of all the E-vectors will be called the *efficient frontier*. The necessary and/or sufficient conditions for a vector $\mathbf{p}$ to be an E-vector are the following:

Theorem B.1.1 *$\mathbf{p}^*$ is an E-vector if and only if for any $i \in \{1, 2, \ldots, Q\}$, $\mathbf{p}^*$ uniquely maximizes p_i for all $\mathbf{p} \in \mathcal{P}_i(\mathbf{p}^*) = \{\mathbf{p} | p_k \geq p_k^*,\ k \neq i,\ k = 1, \ldots, Q\}$. That is, $p_i^*(\cdot) > p_i$ for all $p \in \mathcal{P}_i(p^*)$, $\mathbf{p} \neq \mathbf{p}^*$.*

Theorem B.1.2 *if $\mathbf{p}^* \in \mathcal{P}$ maximizes $\lambda'\mathbf{p}$, for some $\lambda \in \{\mathbf{d} \in \Re^Q | \mathbf{d} > 0\}$, over $\mathcal{P}$, then $\mathbf{p}^*$ is an E-vector.*

Theorem B.1.3 *if $\mathbf{p}^* \in \mathcal{P}$* uniquely *maximizes $\lambda'\mathbf{p}$, for some $\lambda \in \{\mathbf{d} \in \Re^Q | \mathbf{d} \geq 0\}$, over $\mathcal{P}$, then $\mathbf{p}^*$ is an E-vector.*

Theorem B.1.4 *if $\mathcal{P}$ is $\Lambda^{\leqq}$-convex, then a necessary condition for $\mathbf{p}^*$ to be an E-vector is that $\mathbf{p}^*$ maximizes $\lambda'\mathbf{p}$ over $\mathcal{P}$ for some $\lambda \in \{\mathbf{d} \in \Re^Q |\ \mathbf{d} \geq 0\}$. $\mathcal{P}$ is $\Lambda^{\leqq}$-convex if $\mathcal{P} + \Lambda^{\leqq}$ is a convex set, where $\Lambda^{\leqq} = \{\mathbf{d} \in \Re^Q | \mathbf{d} \overset{\leq}{=} 0\}$.*

Proofs: The proofs for these theorems can be found in [59].

Remark: The theorems B.1.1 and B.1.4 state sufficient and necessary conditions, while B.1.2 and B.1.3 only state sufficient conditions. Theorems B.1.1, B.1.2 and B.1.3 are valid for any set $\mathcal{P}$, including non-convex and discrete sets, sets of any shape, although B.1.2 and B.1.3 do not assure the existence of a vector λ for each E-vector (e.g. if the set $\mathcal{P}$ is not convex). If the functions $h_i(\cdot)$ are not convex, the maximizing a linear combination of the measures will not return the complete efficient frontier.

These theorems will be very important for chapter 3, where we will analyze the characteristics of the efficient frontier for pseudo-coherent measures of risk. When the performance space is limited to be two-dimensional (risk-reward), the theorems will imply that the computation of the efficient frontier can be obtained by either maximizing the reward while letting

the risk be constant, or minimizing the risk and letting the reward be constant. However, if both the reward and risk measures are convex, the efficient frontier can be obtained by maximizing a linear combination of the measures. This could have interesting implications in portfolio optimization, since it might imply mutual fund separation; i.e., that all efficient portfolios can be obtained as the linear combination of a finite set of efficient portfolios.

B.1.3 Performance space and decision variables.

For asset allocation problems, the class of distribution functions is limited by the equation (2.3) which will depend on the decision vector $\mathbf{x}$ and the joint distribution function of the m financial assets. The characteristics of the performance set $\mathcal{P}$ (as defined in section B.1.1) will depend on the joint distribution of financial assets considered:

$$\mathcal{P} = \{\mathbf{p}|\mathbf{p} = \mathbf{h}(\tilde{W}(\mathbf{x}))|\tilde{W}(\mathbf{x}) = W_0\mathbf{x}'\tilde{\mathbf{b}}, \tilde{\mathbf{b}} \sim F\}.$$

The vector function $\mathbf{h}(\tilde{W}(\mathbf{x}))$ will assign a vector of parameters to each distribution function. As explained in the previous chapter, for the risk-reward framework that vector is composed of the risk and reward measures. Since the distribution functions belong to the class $\mathcal{F}$ and can be paired with the vector $\mathbf{x}$, we can also use the notation $\mathbf{h}(\mathbf{x})$. When we are assigning a finite vector to each random variable, points belonging to the performance set $\mathcal{P}$ may not have a one to one correspondence to all possible distribution functions (unless the class is restricted). This is a point of consideration: even if the performance set $\mathcal{P}$ is convex, and would allow the representation of preference relations with a scalar value function (such as an expected utility function, see section B.1.1), the Pareto preference will not be able to order distributions which share the same performance vector $\mathbf{p}$; for example, if the set of possible distributions has to be represented with more than two moments, a mean-variance performance vector will not distinguish the portfolios which have the same expected return and variance. [26]

The following theorem is valid for any convex set $\mathcal{X}$. Let $\mathcal{F} = \{F(\mathbf{x})|\mathbf{x} \in \mathcal{X}\}$; then $\mathcal{P} = \{\mathbf{h}(\mathbf{x})|\mathbf{x} \in \mathcal{X}\}$;

Theorem B.1.5 *If each* $h_i(\mathbf{x})$, $i = 1, \ldots, Q$ *is concave on a convex set* $\mathcal{X}$, *then* $\mathcal{P} = \{\mathbf{h}(\mathbf{x})|\mathbf{x} \in \mathcal{X}\}$ *is* $\Lambda^{\leqq}$*-convex, i.e., if* $\mathcal{P} + \Lambda^{\leqq}$ *is a convex set, where* $\Lambda^{\leqq} = \{\mathbf{d} \in \Re^Q | \mathbf{d} \stackrel{\leqq}{=} 0\}$.

The proof can be found in [59]. The implication of the last theorem is quite interesting; in some circumstances, given the appropriate vector λ, the preference relation can be represented by the scalar $v(\tilde{W}) = \lambda' \mathbf{h}(\tilde{W})$; or as a *weighted sum* of components. If that is the case, the function $v(\tilde{W})$ can be regarded as some sort of value function; and will be denoted a *weighting function*. We will refer to the relationship of the weighting functions in the expected utility theory, mean-variance and mean-semivariance in section 3.3.

At this point, assuming we find a vector function $\mathbf{p} = \mathbf{h}(\tilde{W})$ that properly characterizes and investor's preference relation, we can say that an investor will always select an investment $\tilde{W}(\mathbf{x}^*)$ that makes $\mathbf{p}^* = \mathbf{h}(\tilde{W}(\mathbf{x}^*))$ an E-point. In that case, a decision vector $\mathbf{x}^*$ in the decision space $\mathcal{X}$ is an *efficient portfolio (E-portfolio)* if and only if $\mathbf{p}^* = \mathbf{h}(W(\mathbf{x}^*))$ is an E-point in the performance space $\mathcal{P}$ (see section B.1.1).

B.2 Stochastic dominance

In many cases, we only have limited information about the behavior of the investor; for example we only know that she is risk averse and non-satiable. Still, it is possible to determine some conditions in which unambiguously one risky asset will be preferred over another, even if it is not possible to establish a complete order among risky assets.

We will say that an asset with risky payoff $\tilde{A}$ dominates an asset with risky payoff $\tilde{B}$ in the sense of *First Degree Stochastic Dominance*, if all individuals having continuous and increasing wealth utility functions (non-satiable investors) prefer $\tilde{A}$ to $\tilde{B}$ or are indifferent between $\tilde{A}$ and $\tilde{B}$. The following statements are equivalent:

i. $\tilde{A} \stackrel{FSD}{\succeq} \tilde{B}$,

ii. $F_{\tilde{A}}(y) \leq F_{\tilde{B}}(y), \forall y \in \Re$.

If we only know that the investor is risk averse, i.e., that they have concave utility functions, then we will say that the risky payoff $\tilde{A}$ dominates $\tilde{B}$ in the sense of *Second*

Table B.1: Common utility functions

$U(W)$	name:
$\ln(W)$	Lognormal utility,
$W - \frac{b}{2}W^2, b > 0$	Quadratic utility
$\frac{1}{B-1}W^{1-\frac{1}{B}}, W > 0, B > 0$	Power utility.

Degree Stochastic Dominance, if all risk averse individuals having utility functions whose first derivatives are continuous except on a countable subset of $[1, 2]$ prefer $\tilde{A}$ to $\tilde{B}$. Then the following statements will be equivalent:

i. $\tilde{A} \overset{SSD}{\succeq} \tilde{B}$,

ii. $s(y) = \int_{-\infty}^{y}(F_{\tilde{A}}(t) - F_{\tilde{B}}(t))dt \leq 0$, $\forall y \in \Re$.

Henceforth, whenever one of the above two conditions is satisfied, we say that payoff $\tilde{B}$ is *more risky* that $\tilde{A}$. As a remark, it should be noticed that $\tilde{A}$ will be preferred if $E[\tilde{A}] > E[\tilde{B}]$ (since the linear function is also a valid utility function).

From the previous statements, we can see that the stochastic dominance corresponds to a Pareto preference relation, but which has a performance vector of infinite size.

B.3 The expected utility theory

The most accepted framework used to establish preference relations is based on the concept of expected utility. This concept is mathematically quite attractive, although sometimes it is difficult to assign simple utility functions to investors. When the number of possible values of $\tilde{W}$ is very large, it is convenient to define a function U (a utility function) that allows comparisons between the investment alternatives, so that the preference relation can be represented as an *expected utility*; the value function is equivalent to $v(\tilde{W}) = E\{U(\tilde{W})\}$. However, the expected utility representation will not be able to handle all possible preference relations (see section B.3.2, and [32, 10] for some examples). Different functions $U(W)$ can be selected [32, 60]; as the ones found in table B.1.

B.3.1 Risk aversion for expected utility

In the utility function framework, an individual is said to be *risk-averse* if he is unwilling to accept, or is indifferent to, any random return with expected payoff of zero. Consider the gamble that has positive return h_1, with probability p and a negative return, h_2, with probability $(1-p)$. An expected payoff of zero implies $ph_1 + (1-p)h_2 = 0$. In the expected utility framework, it can be demonstrated that risk aversion implies concavity of the utility function U. Hence, risk-averse investors who maximize their utility function maximize will only have one optimal portfolio.

The measure of *absolute risk aversion* is defined as $R_A(\cdot) = -U''(\cdot)/U'(\cdot)$. A utility function is said to display *increasing absolute risk aversion* if $dR_A(W)/dW < 0$. In particular, we are interested in the *non-increasing risk aversion* concept. When only two assets are available, one of them risk-free and the other risky, and the investor has to decide what portion of his initial capital (W_0) to invest in the risky asset (u, leaving $W_0 - u$ to be invested in the riskless asset), an investor is said to be non-increasing risk averse when:

$$\frac{dR_A(W)}{dW} \leq 0 \Rightarrow \nu = \frac{du}{dW_0} \geq 0 \tag{B.1}$$

where W is the realization of $\tilde{W}$ at the end of the next period, and ν is known as the *wealth elasticity*. This condition assumes that an investor is going to increase the dollar amount invested in the risky asset as her initial wealth increases (an increasingly risk averse person will decrease the dollar amount invested in the stock as her wealth increases).

The *wealth elasticity* ν for the demand of the risky asset, can be expressed in terms of η; The demand for the risky asset can be elastic, if $\nu > 0$; inelastic if $\nu < 0$ or unitary elastic, if $\nu = 0$. If an investor has an elastic ν, then the fraction of wealth invested in the risky asset will increase as her initial wealth increases, and conversely, the fraction of wealth invested in the risk-free asset increases as her initial wealth decreases.

B.3.2 The Allais paradox

We introduce the *Allais paradox* mainly because it contradicts the utility theory, while in section 3.3.2 we will show that it does not contradict the more general risk-reward theory. Let us consider the following four lotteries:

(p1) Lottery $p1$ guarantees 1 million for the gambler.

(p2) Lottery $p2$ gives 5 million with 0.1 probability, 1 million with 0.89 probability, and 0 with 0.01 probability.

(p3) Lottery $p3$ gives 5 million with 0.1 probability, and 0 with 0.9 probability.

(p4) Lottery $p4$ gives 1 million with 0.11 probability, and 0 with 0.89 probability.

Most individuals choose lottery $p1$ over $p2$, and $p3$ over $p4$. This behavior is inconsistent with the expected utility theory, as described in [32].

Appendix C

Data

C.1 Dynamics of stock prices

The instantaneous return of the common stock can be described by the stochastic differential equation

$$\frac{dS}{S} = \mu dt + \sigma dz, \tag{C.1}$$

where S is the instantaneous price of the stock, μ is the instantaneous expected return of the common stock (the compounded return), σ^2 is the instantaneous variance of the return, and dz is a standard Gauss-Wiener return (see [46, 33]). The discrete time version of the model is

$$\frac{\Delta S}{S} = \mu \Delta t + \sigma \epsilon \sqrt{\Delta t}.$$

The variable ΔS is the change in the stock price in a small interval of time Δt; and ϵ is a random sample from a standardized normal distribution. The previous equation shows that $\Delta S/S$ is normally distributed with mean $\mu \Delta t$ and standard deviation $\sigma\sqrt{\Delta t}$. In other words

$$\frac{\Delta S}{S} \sim \mathrm{N}[\mu \Delta t, \sigma \sqrt{\Delta t}].$$

The model of the stock price behavior implies that

$$\ln S_T \sim \mathrm{N}\left[\ln S + \left(\mu - \frac{\sigma^2}{2}\right)(T-t), \sigma\sqrt{T-t}\right], \tag{C.2}$$

where S_T is the stock price at the future time T; S is the stock price at the current time, t; and $\mathrm{N}(m, s)$ denotes a normal distribution with mean m and standard deviation s.

The distribution of the rate of return Equation (C.2) implies that

$$\ln \frac{S_T}{S} \sim \mathrm{N}\left[\left(\mu - \frac{\sigma^2}{2}\right)(T-t), \sigma\sqrt{T-t}\right], \tag{C.3}$$

thus the *continuously* compounded rate of return is normally distributed with mean $\mu - \sigma^2/2$ and standard deviation $\sigma/\sqrt{T-t}$. The *simple* rate of return will be log-normally distributed.

C.2 Options

Definition C.2.1 A *call option* is a financial contract that gives the holder the right to buy the underlying asset by a certain date for a certain price. A *put option* gives the holder the right to sell the underlying asset by a certain date for a certain price. The price in the contract is known as the *exercise price* or *strike price*; the date in the contract is known as the *expiration date, exercise date* or *maturity. European options* can only be exercised on the expiration date itself.

If K is the strike price and S_T is the final price of the underlying asset, the payoff from holding a European call option is $\max(S_T - K, 0)$. The payoff to the holder of a European put option is $\max(K - S_T, 0)$.

The Black-Scholes pricing formula

Denoting the current stock price as S, the time to expiration as $T - t$, the volatility of the stock price as σ^2, the continuously compounded risk-free interest rate as r_f, and assuming that the stock price follows a geometric Brownian motion as described in (C.1), the value of a European call is

$$c = SN(d_1) - Ke^{-r_f(T-t)}N(d_2), \tag{C.4}$$

where

$$\begin{aligned} d_1 &= \frac{\ln(S/K) + (r_f + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}, \\ d_2 &= d_1 - \sigma\sqrt{T-t}, \end{aligned}$$

and $N(x)$ is the cumulative distribution function for a standardized normal variable.

The value of a European put can be computed from the *put-call parity*

$$p = c + Ke^{-r_f(T-t)} - S,$$

which yields

$$p = Ke^{-r_f(T-t)}N(d_2) - SN(-d_1)$$

C.3 Option-based portfolio strategies

We present two common option-based strategies, which were analyzed in [12].

C.3.1 Writing covered call options

The strategy (MSG1, or W.C.) consists of the following:

i. Buy 1 stock at time t; the current price of the stock is S; the price of the stock in the future will be S_T; $B_T = S_T/S$.

ii. Sell (write) γ call options per stock purchased, $0 \leq \gamma \leq 1$; the time to maturity $\tau = T - t$; the strike price is equal to the current price of the stock, $K = S$, and the current price of the call is c.

The simple return of the portfolio at time T will be:

$$R_{wc} = \frac{B_T + \gamma \min(K/S - B_T, 0)}{(1 - \gamma c)} - 1.$$

C.3.2 Buying covered put options

The strategy (MSG1, or L.P.) consists of the following:

i. Buy 1 stock at time t; the current price of the stock is S; the price of the stock in the future will be S_T; $B_T = S_T/S$.

ii. Buy (long) γ put options per stock purchased, $0 \leq \gamma \leq 1$; the time to maturity $\tau = T - t$; the strike price is equal to the current price of the stock, $K = S$, and the current price of the call is p.

The simple return of the portfolio will be:

$$R_{lp} = \frac{B_T + \gamma \max(K/S - B_T, 0)}{(1 + \gamma p)} - 1.$$

C.4 Simulated and historical data

C.4.1 Elliptic returns

The characteristics of the data are the following: (taken from [47])

The yearly expected return of each one of the 3 assets is

$$E[\mathbf{R}] = [10.54, 13.67, 17.73]'\%,$$

(where $\mathbf{R} = \mathbf{B}_T - 1$) and the standard deviation is

$$\sigma = [11.26, 17.87, 19.33]'\%.$$

The correlation matrix is

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0.237 & 0.211 \\ 0.237 & 1 & 0.454 \\ 0.211 & 0.454 & 1 \end{bmatrix}.$$

We can generate either *Gaussian* random variables with those characteristics, or a Multivariate t joint distribution with 3 degrees of freedom, as defined in section A.2.2. In the

Table C.1: Option-based strategies data.

Optioned portfolio	Coverage	Asset number
Write-Call,	50%	1
Long-Put,	50%	2
Write-Call,	100%	3

latter case, the covariance matrix Σ is the same one and the corresponding $\mathbf{V}$ matrix is computed as $\mathbf{V} = \mathbf{\Sigma}/3$.

C.4.2 Option-based strategies

Non-parametric returns were generated by following three different option-based strategies (as described in section C.3 [12]; other references for option-based strategies can be found in [45, 33]):

i. A "write-call" (W.C.) strategy covering 50% of the underlying asset.

ii. A "long-put" (L.P.) strategy covering 50% of the underlying asset.

iii. A "write-call" strategy covering 100% of the underlying asset.

the histograms and empirical distributions for Monte Carlo simulations with 10000 samples are depicted in figure 2-1. The returns are very asymmetric, and it should be noted that the L.P. histogram is also multi-modal. In that case, which in practice is possible, no parametric methods are available. For this example we generated 200 sets of 200 Monte-Carlo simulations, and the returns simulate returns in a 6 months period.

The characteristics of the underlying data are the following: The semiannual expected return of each one of the 3 assets is

$$E[\mathbf{R}] = [5.9843, 7.8722, 3.4079]'\%,$$

and the standard deviation is (semiannual)

$$\sigma = [4.3258, 8.5034, 3.4771]'\%.$$

Table C.2: Stock Data

Dates: January 4, 1982 to March 20, 1997	
NYSE Symbol	Asset number
BUD (Anheuser-Busch)	1
CAT (Caterpillar)	2
EK (Eastman-Kodak)	3
PEP (PepsiCo.)	4

The correlation matrix is

$$\mathbf{\Phi} = \begin{bmatrix} 1.0000 & 0.8221 & 0.5519 \\ 0.8221 & 1.0000 & 0.4512 \\ 0.5519 & 0.4512 & 1.0000 \end{bmatrix}.$$

Those numbers correspond to the correlations and expected returns computed from the real stock data (section C.4.3).

C.4.3 Historical stock returns

The returns of four stocks were used to perform experiments on real data (which can be found in table C.2). The returns of the stocks were weekly continuously compounded returns.

C.4.4 Put-call returns

We created an example based on the out of the money options (a put and a call), proposed in [6] which are used to show the non-convexity of the VaR_α. In our synthetic example we used a common stock with an annualized continuous return $r_{stock} = 0.15$ and an annualized volatility $\sigma_{stock} = 0.2$. The stock is assumed to follow a log-normal diffusion process, and the initial price is normalized to 1. Two far-out-of-the-money options are assumed to be available: one European put with exercise price of $K_{put} = 0.08445$ and one European call with exercise price $K_{call} = 1.3337$; the exercises prices were chosen such that the option will be exercised only 5% of the time. The continuous annualized risk-free return is $r_f = 0.05$, and time to maturity it $\tau = 0.5$ years. The price of the options were computed using the

classic Black-Scholes valuation formula, resulting in a call price of $c = 0.002$, and a put price of $p = 0.005$. [1] The empirical cdf of the synthetic common stock and the European options can be seen in figure 2-2.

[1] In real life, however, the returns of the underlying asset can be skewed, and the Black-Scholes formula will misprice the far out-of-the-money options.

Bibliography

[1] R. Aggarwal, R. Ramesh, and H. Takato. Skeweness and kurtosis in japanese equity returns: Empirical evidence. *Journal of Financial Research*, 12:253–260, Fall 1989.

[2] R. Aggarwal and R. Rao. Institutional ownership and distribution of equity returns. *Financial Review*, 25:211–229, May 1990.

[3] R. Aggarwal and J. D. Schtzberg. Day of the week effect, information seasonality, and higher moments of security returns. *Journal of Economics and Business*, 49:1–20, 1997.

[4] R. Aggarwal and L. Soenen. The nature and efficiency of the gold markets. *Journal of Portfolio Management*, 14:18–21, Spring 1988.

[5] F. Arditti. Another look at mutual fund performances. *Journal of Financial and Quantitative Analysis*, 6:909–912, June 1971.

[6] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath. Thinking coherently. *Risk*, 10(11):68–70, November 1997.

[7] V. Barnett. The ordering of multivariate data. *Journal of the Royal Statistics Society*, (139):318–344, 1975. Ser. A.

[8] F. Bassi, P. Embrechts, and M. Kefetzaki. A survival kit on quantile estimation. Talk given for Olsen & Associates, 1993.

[9] D. Bersekas and J. N. Tsitsiklis. *Neuro-dynamic programming.* Athena Scientific, Belmont, Mass., 1996.

[10] D. Bertsekas. *Dynamic Programming and Stochastic Control.* Academic Press, 1976.

[11] D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1994.

[12] R. Bookstaber and R. Clarke. Option portfolio strategies: Measurement and evaluation. *Journal of Business*, 57(4), 1984.

[13] R. A. Brealey and S. C. Myers. *Principles of Corporate Finance*. McGraw-Hill, 1991.

[14] J. Y. Campbell, A. W. Lo, and A. C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, 1997.

[15] G. Chamberlain. A characterization of the distributions that imply mean-variance utility functions. *Journal of Economic Theory*, 29(1):185–201, February 1983.

[16] C. J. Corrado and T. Su. Skewness and kurtosis in s&p 500 index returns implied by option prices. *The Journal of Financial Research*, XIX(2):175–192, Summer 1996.

[17] H. A. David. *Order Statistics*. John Wiley & Sons, New York, 2nd edition, 1981.

[18] L. de Haan, D. W. Jansen, K. Koedijk, and C. G. de Vries. Safety first portfolio selection, extreme value theory and long run asset risks. *Extreme Value Theory and Applications*, 1994.

[19] K. Doksum, D. Peterson, and A. Samarov. On badwidth selection in local polynomial regression. *submitted to the Journal of the Royal Statistical Society, series B*, May 1998.

[20] D. Duffie. *Dynamic Asset Pricing Theory*. Princeton University Press, second edition, 1996.

[21] J. S. Dyer. *Multiple Criteria Decision Making and Risk Analysis*, volume 56 of *NATO ASI Series F: Computer and Systems Sciences*, chapter The Effects of Risk on Decision Making. Springer-Verlag, 1987.

[22] N. Etemadi. An elementary proof of the strong law of large numbers. *Z. Wahrscheinlichkeitsthorie verw. Geb.*, 1(55):119–122, 1981.

[23] E. Fama and K. French. The cross-section of expected stock returns. *The Journal of Finance*, XLVII(2):427–465, June 1992.

[24] E. F. Fama. The behavior of stock markets prices. *Journal of Business*, 38:34–105, January 1965.

[25] J. Fan and I. Gijbels. *Local polynomial modeling and its applications*. Monographs on statistics and applied probability. Chapman & Hall, 1996.

[26] H. Fang and T.-Y. Lai. Co-kurtosis and capital asset pricing. *The Financial Review*, 32(2):293–307, May 1997.

[27] M. S. Feldstein. Mean-variance analysis in the theory of liquidity preference and portfolio selection. *Review of Economic Studies*, 36(105):5–12, January 1969.

[28] M. Garman. Improving on var. *Risk*, May 1996.

[29] A. Grace. *Optimization Toolbox*. The MathWorks Inc., 1995.

[30] W. V. Harlow. Asset allocation in a downside-risk framework. *Financial Analysts Journal*, September 1991.

[31] F. S. Hillier and G. J. Lieberman. *Introduction to Mathematical Programming*. McGraw-Hill, Inc., 1995.

[32] C.-F. Huang and R. H. Litzenberger. *Foundations for Financial Economics*. Prentice-Hall, 1988.

[33] J. C. Hull. *Options, futures and other Derivatives Securities*. Prentice-Hall, 1993.

[34] C. M. Hurvich, J. S. Simonoff, and C.-L. Tsai. Smoothing parameter selection in non-parametric regression using an improved akaike information criterion. *Journal R. Statist. Soc. B*, 60:271–293, 1998.

[35] J. E. Ingersoll. *Theory of financial decision making*. Rowman & Littlefield studies in financial economics. Rowman & Littlefield, 1987.

[36] B. Jacobs and K. Levy. Residual risk: How much is too much. *The Journal of Portfolio Management*, Spring 1996.

[37] N. L. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*, volume 4. John Wiley & Sons, 1969.

[38] P. Jorion. *Value at risk : the new benchmark for controlling market risk.* Irwin Professional Pub., Chicago, 1997.

[39] J. Encarnación Jr. Portfolio choice and risk. *Journal of Economic Behavior and Organization*, 16:347–353, 1991.

[40] A. Kaufman and A. Henry-Labordère. *Integer and Mixed Programming, Theory and Applications*, volume 137 of *Mathematics in Science and Engineering.* Academic Press, New York, 1977.

[41] H. Leland. Existence and diversification of optimal portfolio policies. *Journal of Economic Theory*, (4):35–44, 1972.

[42] J. C. Mao. Models of capital budgeting, e-v vs e-s. *Journal of Financial and Quantitative Analysis*, pages 657–75, January 1970.

[43] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7, 1952.

[44] H. M. Markowitz. The general mean-variance portfolio selection problem. *Philosophical Transactions: Physical Sciences and Engineering*, November 1993.

[45] H. S. Marmer and F. K. Louis Ng. Mean-semivariance analysis of option-based strategies: A total asset mix perspective. *Financial Analysts Journal*, May 1993.

[46] R. C. Merton. *Continuous-time finance.* B. Blackwell, 1990.

[47] NEOS. Neos demo: Stock market investment. http://www.mcs.anl.gov/otc/Guide/CaseStudies/port/demo.html, 1998.

[48] C. Klüpperberg P. Embrechts and T. Mikosch. *Modeling Extremal Events*, volume 33 of *Applications of Mathematics.* Springer, 1997.

[49] G. P. Pflug. *Optimization of Stochastic Models.* Kluwer Academic Publishers, Boston, 1996.

[50] S. Pliska. *Introduction to Mathematical Finance*. Blackwell Publishers, 1997.

[51] S. Pliska. Risk sensitive portfolio management with underlying stochastic factors. In H. Frydman and P. Lakner, editors, *Workshop on Statistical Aspects of Financial Models*, New York, May 1997. NYU.

[52] D. H. Pyle and S. J. Turnovsky. Safety-first and expected utility maximization in mean-standard deviation portfolio analysis. *The Review of Economics and Statistics*, LII(1), February 1970.

[53] R. D. Reiss. *Approximate distribution of order statistics*. Springer Series in Statistics. Springer-Verlag, 1989.

[54] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[55] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, third edition, 1976.

[56] P. Samuelson. The fundamental approximation theorem on portfolio analysis in terms of means, variances and higher moments. *Review of Economic Studies*, 37:537–542, October 1970.

[57] R. Scott and P. Horvath. On the direction of preference for moments of higher order than variance. *Journal of Finance*, 35:915–919, September 1980.

[58] Various. *Riskmetrics - Technical Document*. JP Morgan, 1995.

[59] P.-L. Yu. *Multiple-Criteria Decision Making*. Plenum Press, 1985.

[60] W.T. Ziemba and R.G. Vickson. *Stochastic Optimization Models in Finance*. Academic Press, 1975.