Essays on Risk Measurement and Fund Separation

Fang Liu
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Business Commons

Recommended Citation
https://openscholarship.wustl.edu/art_sci_etds/422
WASHINGTON UNIVERSITY IN ST. LOUIS

Olin Business School

Dissertation Examination Committee:
Philip Dybvig, Co-Chair
Ohad Kadan, Co-Chair
Radhakrishnan Gopalan
Asaf Manela
John Nachbar

Essays on Risk Measurement and Fund Separation
by
Fang Liu

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2015
St. Louis, Missouri
# Table of Contents

List of Figures ........................................... iii
Acknowledgement ....................................... iv

1 On Investor Preferences and Mutual Fund Separation .......... 1
   1.1 Introduction ......................................... 1
   1.2 Literature Review .................................. 5
   1.3 Mutual Fund Separation ............................. 7
       1.3.1 Setup ........................................ 7
       1.3.2 Some Examples ............................... 9
       1.3.3 General Characterization of $K$-Fund Separation . 14
       1.3.4 One-Fund Separation .......................... 19
       1.3.5 Two-Fund Separation .......................... 20
       1.3.6 From Low-Degree to High-Degree Separation .... 22
       1.3.7 Range of Optimal Consumption ................. 24
       1.3.8 Machina Preferences .......................... 25
   1.4 Primal Utility Function ............................. 31
       1.4.1 SAHARA Utility ................................ 31
       1.4.2 GOBI Utility .................................. 33
   1.5 Money Separation .................................... 34
       1.5.1 Money Separation and Constant Investment Weight . 34
       1.5.2 Money Separation and Shifts in Utility .......... 37
   1.6 Strict Concavity ..................................... 40
   1.7 Conclusion .......................................... 45

2 Generalized Systematic Risk .......................... 47
   2.1 Introduction ......................................... 47
   2.2 Related Literature .................................. 52
   2.3 Risk Measures and Their Properties .................. 54
       2.3.1 Examples of Risk Measures ..................... 57
   2.4 Systematic Risk in an Equilibrium Setting .......... 61
       2.4.1 Model Setup ................................... 61
       2.4.2 A Generalized CAPM ........................... 64
3 Recovering Conditional Return Distributions by Regression: Estimation and Applications

3.1 Introduction .......................... 84
3.2 Literature Review ...................... 91
3.3 Setup .................................. 94
3.4 Estimation Methodology ................. 97
  3.4.1 Extracting Risk-Neutral Marginal Distributions from Option Prices .. 98
  3.4.2 Estimating Conditional Distributions by Constrained Regression ... 100
3.5 Discussions and Extensions .......... 103
  3.5.1 Elaboration on Key Assumptions 103
  3.5.2 Multi-Factor Framework 105
  3.5.3 Continuous Security Returns . 106
  3.5.4 Alternative Econometric Models 107
3.6 Application I: Systematic Disaster Risk Premium .......... 109
  3.6.1 Measure of Systematic Disaster Risk 110
  3.6.2 Data ................................. 111
  3.6.3 Empirical Strategy ................. 113
  3.6.4 Results ............................. 116
3.7 Application II: Bank Systemic Exposure .......... 120
  3.7.1 Measure of Systemic Exposure 122
  3.7.2 Data and Estimation ................. 123
  3.7.3 Results ............................. 125
3.8 Application III: Ross Recovery for Individual Assets .... 130
  3.8.1 Ross Recovery Theorem ............. 131
  3.8.2 Extension to Individual Assets .... 135
3.9 Conclusion ............................. 136

Bibliography .......................... 138

Appendix A: Proofs for Chapter 1 146

Appendix B: Proofs, Derivations and Discussions for Chapter 2 155

Appendix C: Proofs and Discussions for Chapter 3 175
List of Figures

1.1 Possible Terms in $I$ of a Separating Preference .................................... 18
2.1 Portfolio Opportunity Set and Efficient Frontier ...................................... 65
2.2 Graphical Illustration of the Proof of Theorem 41 .................................... 68
3.1 Second Moment Properties in Joint Return Behavior .............................. 86
3.2 Nonlinearity in Joint Return Behavior ...................................................... 86
3.3 Summary Statistics of Disaster Risk Variables .......................................... 118
3.4 Pairwise Correlations of Stock Characteristics ......................................... 118
3.5 Systematic Disaster Risk Premium for Entire Sample .............................. 119
3.6 Systematic Disaster Risk Premium and Market Disaster Risk ................... 121
3.7 Systemic Exposure of Banks ....................................................................... 126
3.8 Systemic Exposure and Bank Characteristics ............................................ 129
Acknowledgement

I am most grateful to my advisors Dr. Philip Dybvig and Dr. Ohad Kadan for their continuous encouragement, guidance and patience over the years. I would have never finished my research without their persistent support. I would also like to thank Dr. Radhakrishnan Gopalan, Dr. Asaf Manela and Dr. John Nachbar for being on my committee and for their helpful comments and suggestions.

I would also like to express my most sincere appreciation to my family and friends for being unconditionally understanding and supportive. This dissertation would not have been possible without their ongoing encouragement.

Finally, I want to thank the Olin Business School for providing generous financial support over the past few years.

Fang Liu

Washington University in St. Louis

May 2015
Chapter 1

On Investor Preferences and Mutual Fund Separation

This chapter extends Cass and Stiglitz’s analysis of preference-based mutual fund separation. We show that high degrees of fund separation can be constructed by adding inverse marginal utility functions exhibiting lower degrees of separation. However, this method does not allow us to find all utility functions satisfying fund separation. In general, we do not know how to write the primal utility functions in these models in closed form, but we can do so in the special case of SAHARA utility defined by Chen et al. and for a new class of GOBI preferences introduced here. We show that there is money separation (in which the riskless asset can be one of the funds) if and only if there is a fund (which may not be the riskless asset) with a constant allocation as wealth changes.

1.1 Introduction

Mutual fund separation is an important concept in portfolio selection. It means that all investors’ optimal portfolio choice can be constructed as the linear combination of a set of mutual funds regardless of the initial wealth level, where a mutual fund can be any portfolio of tradable assets in the market. In other words, under mutual fund separation investors should be able to achieve the same level of utility from the individual assets as if they were

\footnote{This chapter is joint work with Philip Dybvig.}
only offered a set of mutual funds. The term “separation” comes from the fact that every investor can “separate” his portfolio choice into two steps. First, the investor chooses a small set of funds that spans optimal portfolios of all wealth levels. Second, the investor determines the optimal mixture of the separating funds based on his current wealth level.

Mutual fund separation has been studied in the literature from two perspectives, the distribution perspective and the preference perspective. In particular, Cass and Stiglitz (1970) characterizes the class of investor preferences exhibiting mutual fund separation for “all” distributions of asset returns, and Ross (1978) studies the distributions of asset returns that support mutual fund separation for “all” utility functions. While Cass and Stiglitz (1970) focus on preference-based separation, they mostly restrict attention to one- and two-fund separation, and claim “there is nothing intrinsically interesting in the generalization (K-fund separation) not already contained in the argument previously given (two-fund separation).” We disagree!

In this paper, we extend Cass and Stiglitz’s analysis of preference-based mutual fund separation, with a special focus on high-degree separation. We show that high-degree separating preferences can be constructed by adding inverse marginal utility functions exhibiting lower degrees of separation. However, this method does not allow us to find all utility functions satisfying fund separation. In general, we do not know how to write the primal utility function of a separating preference in closed form, but we show that this can be achieved for two special classes of preferences, both of which exhibit three-fund separation. We also study money separation in which the riskless asset can be chosen as one of the separating funds, and show that money separation holds if and only if there is a fund (which may not be the riskless asset) with a constant allocation as wealth changes.

The study of mutual fund separation has important implications. In practice, there are a huge number of assets in the market available for trading, and it is impossible for individual investors to examine each and every asset before setting up their portfolios. Even if a
complete analysis of all assets is possible, it would involve a large amount of time and effort, let alone the substantial transaction costs that need to be incurred to trade these assets. If we have reasons to believe that $K$-fund separation holds, where $K$ is much smaller than the number of assets, then instead of having to consider all available assets, it suffices to restrict attention to the $K$ separating funds. The resulting optimal portfolio would deliver exactly the same level of utility as the one constructed from the individual assets. In particular, this suggests that we can set up the $K$ separating funds as index funds, and that these funds are all that an investor would ever need to trade.

It is useful to study mutual fund separation especially with high degrees also because it helps motivate new preferences with tractable functional forms. In many important finance problems such as portfolio selection and pricing, a nice feature that ensures tractability is fund separation. Traditionally, the only preferences with fund separation that are well studied are restricted to the one- and two-fund separating classes. In fact, in an economy with the presence of a risk-free asset, two-fund separation implies that there must a unique portfolio of risky assets held by all investors in equilibrium. In other words, all investors’ optimal consumption bundles are homogeneous up to leverage. In comparison, with a higher degree of separation, different investors can hold different portfolios of risky assets. This allows for a larger extent of heterogeneity among investors when modeling an economy with fund separation and analytical tractability.

We start with a one-period setting, where investors invest at the beginning of the period and consume at the end. If a utility function exhibits $K$-fund separation, then its inverse marginal utility can be spanned by $K$ mutual funds, with the associated weights being function of the initial wealth. Solving this equation then allows us to characterize the class of separating preferences in terms of the inverse marginal utility. This characterization shows that one can construct high-degree separating preferences by adding low-degree ones in the inverse marginal utility. However, this method does not allow us to find all utility functions
satisfying fund separation, because high-degree separation may feature separating funds that never show up in low-degree separation.

We then ask whether we are able to recover the primal utility functions of separating preferences from the inverse marginal utility characterization. A natural way to do this is to first invert the inverse marginal utility to obtain the marginal utility, and then to integrate the marginal utility to obtain the primal utility. Unfortunately, this generally does not yield a closed-form expression, but there are a few cases for which a closed-form primal utility can be analytically obtained. One such case is the SAHARA preferences recently proposed by Chen, Pelsser and Vellekoop (2011), and another case is the GOBI preferences to be introduced in this paper. We will show that both classes exhibit three-fund separation, and they have not only a simple form in the inverse marginal utility, but a closed-form expression in the primal utility.

We then turn to a special case of fund separation, in which the risk-free asset can be chosen as one of the separating funds. We follow Cass and Stiglitz (1970) and call this case money separation. We show that money separating holds if and only if we can choose a separating fund whose optimal investment weight is constant and in particular does not depend on the initial wealth level. Interestingly, the constant weight can be assigned to either the risk-free asset (e.g., quadratic utility) or a risky fund (e.g., CARA utility). In addition, we also show that money separation is closely related to shifts in the utility. In particular, starting with a non-money separating preference, we can easily construct money separation by introducing a non-zero shift in the utility function.

Finally, concavity imposes additional constraints on our separation characterization. We show that strict concavity is maintained if and only if the inverse marginal utility is monotonically decreasing. This enables us to check strict concavity by directly looking at the inverse marginal utility. It is intrinsically very hard to derive the necessary and sufficient conditions for strict concavity in terms of the parameter values. But, this condition does allow us to
refine the class of separating preferences by ruling out parameter values and forms of the inverse marginal utility that cannot exist.

The rest of the paper proceeds as follows. Section 1.2 reviews the literature. Section 1.3 defines mutual fund separation and characterizes the class of separating preferences in terms of the inverse marginal utility. We also demonstrate how low-degree separation can be used to construct high-degree separating preferences. Section 1.4 derives the primal utility for the SAHARA and GOBI preferences, both of which exhibit three-fund separation. Section 1.5 studies the special case of money separation. Section 1.6 examines conditions for strict concavity and discuss how they can be used to refine the separating class. Section 1.7 concludes the paper.

1.2 Literature Review

The first results of mutual fund separation are developed under the mean-variance framework. Markowitz (1952) and Tobin (1958) show that when investors only care about the mean and variance of the return distribution and in the presence of a risk-free asset, the optimal portfolio can be constructed in two stages. The first stage is to set up a risky portfolio by finding the right weight assigned to each risky asset, and the second stage is to determine the division of the entire wealth between the risky portfolio and the risk-free asset. Using an equilibrium approach, Sharpe (1964) and Lintner (1965a) later show that this benchmark risky portfolio is in fact the market portfolio. Black (1972) then suggests that even in the absence of the risk-free asset, similar two-fund separation results still hold with both separating funds being risky portfolios. In addition, Merton (1972) analytically solves the portfolio selection problem.

While the mean-variance framework nicely supports two-fund separation, its appropriateness in describing investor preferences has been increasingly challenged by subsequent studies. In particular, research in asset pricing shows that investors have preferences over
high distribution moments of the portfolio returns. For instance, it is suggested that investors favor right skewness (e.g., Kraus and Litzenberger (1976), Jean (1971), Kane (1982), and Harvey and Siddique (2000)), but are averse to tail-risk and rare disasters (e.g., Barro (2009) and Gabaix (2008, 2012)). When the mean-variance preference does not hold, additional conditions are needed to support mutual fund separation. Such conditions can be roughly classified into two types: those in terms of investor preferences and those in terms of the distributions of asset returns.

In terms of investor preferences, Pye (1967) and Hakansson (1969) find that the HARA class exhibits two-fund separation with one of the separating funds being the risk-free asset. Cass and Stiglitz (1970) further characterize the class of preferences that permits mutual fund separation, regardless of the distributions of asset returns. More recently, Rockafellar, Uryasev, and Zabarankin (2006b) and Kadan, Liu, and Liu (2015) extend the mean-variance preference into a mean-risk framework to capture a wide variety of risk attributes, and they provide sufficient conditions on the risk measure that guarantee two-fund separation.

On the other side of the research, efforts have been made to delineate the class of stochastic processes that supports separation for all utility functions (see Fama (1965), Feldstein (1969), and Merton (1971)). In particular, Ross (1978) derives necessary and sufficient conditions on the stochastic structure of asset returns such that mutual fund separation can be sustained, independent of investor preferences. Further, Russell (1980) presents a unified approach to the two-fund separation problem that incorporates both Cass and Stiglitz (1970) and Ross (1978).

More recently, people have turned to separability under the dynamic portfolio optimization framework (see Schmedders (2007) and Canakoglu and Ozekici (2010)) and portfolio separation with heterogeneous beliefs and attitudes towards risk (see Chabi-Yo, Ghysels and Renault (2008)). All of this contributes to the theory of mutual fund separation.
1.3 Mutual Fund Separation

In this section, we study necessary and sufficient conditions for preference-based $K$-fund separation.

1.3.1 Setup

A group of investors exhibits $K$-fund separation if the optimal portfolio choice of all of them can be constructed as the linear combination of the same set of $K$ mutual funds, regardless of the initial wealth level.

Following Cass and Stiglitz (1970)'s analysis, we consider a one-period model, in which investors invest at the beginning of the period and consume at the end. Assume that there exists a unique stochastic discount factor (also known as the state-price density) $\rho > 0$ that takes all positive values with $E(\rho) < \infty$. As an informality, we also use $\rho$ to represent realizations of the random stochastic discount factor throughout the paper. Also assume that all investors have von Neumann-Morgenstern utility $u(\cdot)$ defined on any open interval $D \subset \mathbb{R}$, which is twice differentiable with $u' > 0$ and $u'' < 0$. We allow for both positive and negative consumption levels. We denote the set of utility functions of all investors by $\mathcal{U}$. Then, an investor with utility function $u \in \mathcal{U}$ and initial wealth $w_0 \in \mathbb{R}$ solves the following utility maximization problem.

**Problem 1** Choose consumption $x$ to

$$
\max_x E[u(x)]
$$

subject to the budget constraint

$$
E[x\rho] \leq w_0.
$$

---

2 One way to guarantee the existence of a unique stochastic discount factor is by assuming that the market is complete enough such that all assets needed to produce optimal risk sharing are available and that all investors agree on pricing.

3 While negative consumption may seem absurd on its face, what we call consumption might be the net trade or it can be justified by a promise to do work to cover any negative amount. Also, even if consumption is not literally negative, it can be a useful modelling device.
We denote the set of solutions to Problem 1 by $S(u,w_0)$. By the strict concavity of $u$, $S(u,w_0)$ is either an empty set or a singleton. Assume that for all utility functions under consideration, there exists an open interval for the initial wealth such that an optimum to Problem 1 exists, i.e., $S(u,w_0) \neq \emptyset$. Now we define $K$-fund separation if there are no fewer than $K$ mutual funds whose random payoffs span the optimal consumptions of all investors whenever an optimum exists, regardless of the initial wealth level.

**Definition 2** We have $K$-fund separation if $K$ is the smallest positive integer such that there exists $K$ mutual funds $\{f_k(\rho)\}_{k=1,\ldots,K}$, which satisfies that for all $u \in \mathcal{U}$ and $w_0 \in \mathbb{R}$, if $S(u,w_0) \neq \emptyset$, then we can find $\{\alpha_k(u,w_0)\}_{k=1,\ldots,K}$ such that $\sum_{k=1}^{K} \alpha_k(u,w_0)f_k(\rho) \in S(u,w_0)$.

Several comments are worth pointing out. First, the optimal consumptions and the separating funds are both identified in terms of payoff, whereas the associated portfolio compositions may not be uniquely determined in the presence of redundant assets. Second, whenever $K$-fund separation holds for $K \geq 2$, the set of separating funds is not unique. Indeed, having one set of separating funds, we can easily construct another by taking linear combinations of the original set of funds, and the resulting investment weights are also linear combinations of the original weights. Finally, while $K$-fund separation is defined for a set of utility functions, we are often interested in $K$-fund separation for a single utility as a special case, which is obtained when $\mathcal{U}$ contains one utility function only.

One special form of mutual fund separation obtains when we can choose the risk-free asset as one of the separating funds. We follow Cass and Stiglitz (1970) and refers to this special case as money separation. In other words, money separation holds as long as the risk-free asset is in the linear span of the separating funds. Formally, we have the following definition.

**Definition 3** We have $K$-fund money separation if $K$-fund separation holds and we can
choose \( f_1 (\rho) = 1 \).\(^4\)

To characterize utility functions exhibiting mutual fund separation, we start by solving Problem 1. Suppose that a solution exists, then the first order condition implies that the optimal consumption portfolio is given by

\[
x^* = I (\lambda \rho) ,
\]

where \( \lambda > 0 \) is the shadow price whose value depends on the initial wealth level \( w_0 \), and \( I = (u')^{-1} \) is the inverse marginal utility function. Since \( u'' < 0 \), it is apparent that \( I \) exists and is unique.

If the utility function \( u \) satisfies \( K \)-fund separation, then the optimal consumption (1.1) can be written as the weighted sum of \( K \) mutual funds, with the associated weights depending on the initial wealth \( w_0 \) and thus on the shadow price \( \lambda \), i.e.,

\[
I (\lambda \rho) = \sum_{k=1}^{K} \alpha_k (\lambda) f_k (\rho) .
\]

Notice that to ensure non-degeneracy, we must have that the separating funds \( f_k (\rho) \)'s are linearly independent, and that the associated investment weights \( \alpha_k (\lambda) \)'s are also linearly independent. Otherwise, the degree of separation can always be reduced by combining two or more funds to form a larger separating fund. In addition, for tractability, we only consider the case in which the \( \alpha_k (\lambda) \)'s are locally analytical.

### 1.3.2 Some Examples

Before we formally characterize the set of separating preferences, let us first look at a few examples. Some of the following examples involve very well-known preferences, while others are less so. One might wonder how we come up with these examples. In fact, these examples

\(^4\)The payoff to the risk-free asset can take any constant value. Without loss of generality, we normalize it to be equal to 1.
can be easily constructed from the general characterization of separating preferences to be introduced in the next section.

**Example 4 (CRRA Utility)** Consider the CRRA utility function

\[ u(x) = \begin{cases} 
\frac{x^{1-a}}{1-a}, & a > 0 \text{ and } a \neq 1 \\
\log x, & a = 1 
\end{cases} \]

defined on all \( x \in (0, +\infty) \), where \( a \) is the coefficient of relative risk aversion. It is easy to verify that the inverse marginal utility can be written as

\[ I(\xi) = \xi^{-\frac{1}{a}}. \]

Since

\[ I(\lambda \rho) = (\lambda \rho)^{-\frac{1}{a}} = \lambda^{-\frac{1}{a}} \rho^{-\frac{1}{a}}, \]

by (1.2) we have that the CRRA utility function exhibits one-fund separation with the separating fund

\[ f(\rho) = \rho^{-\frac{1}{a}}, \]

and the corresponding investment weight given by

\[ \alpha(\lambda) = \lambda^{-\frac{1}{a}}. \]

This indicates that an investor with CRRA utility would always find it optimal to invest his entire wealth into a single mutual fund \( \rho^{-\frac{1}{a}} \), regardless of the initial wealth level.

**Example 5 (Quadratic Utility)** Suppose we have the following quadratic utility function

\[ u(x) = -x^2 + 2bx, \]

where \( x < b \). We can easily verify that the inverse marginal utility is given by

\[ I(\xi) = b - \frac{1}{2} \xi. \]
Since

\[ I (\lambda \rho) = b - \frac{1}{2} \lambda \rho, \]

the quadratic utility function exhibits two-fund separation with one of the separating funds being the risk-free asset

\[ f_1 (\rho) = 1, \]

and the other a risky portfolio

\[ f_2 (\rho) = \rho. \]

The corresponding investment weights are given by

\[ \alpha_1 (\lambda) = b, \]
\[ \alpha_2 (\lambda) = -\frac{1}{2} \lambda. \]

This implies that an investor with quadratic utility would optimally invest a fixed amount \( b \) into the risk-free asset and take a wealth-dependent short position in the risky fund \( \rho \).

**Example 6 (SAHARA Utility)** The SAHARA preferences are introduced in Chen, Pelsser and Vellekoop (2011). They show that the inverse marginal utility of a SAHARA utility function with scale parameter \( b > 0 \) and risk aversion parameter \( a > 0 \) can be written as

\[ I (\xi) = \frac{1}{2} \left( \xi^{- \frac{1}{a}} - b^2 \xi^\frac{1}{a} \right). \]

Since

\[ I (\lambda \rho) = \frac{1}{2} \left( (\lambda \rho)^{- \frac{1}{a}} - b^2 (\lambda \rho)^{\frac{1}{a}} \right) = \frac{1}{2} \lambda^{\frac{1}{a}} \rho^{- \frac{1}{a}} - \frac{1}{2} b^2 \lambda^{\frac{1}{a}} \rho^{\frac{1}{a}}, \]

this suggests that the SAHARA utility function exhibits two-fund separation with the two separating funds defined as

\[ f_1 (\rho) = \rho^{- \frac{1}{a}}, \]
\[ f_2 (\rho) = \rho^{\frac{1}{a}}. \]
The corresponding wealth dependent investment weights are given by

\[ \alpha_1 (\lambda) = \frac{1}{2} \lambda^{-\frac{1}{\alpha}}, \]
\[ \alpha_2 (\lambda) = -\frac{1}{2} b^2 \lambda^{\frac{1}{\alpha}}. \]

Namely, it is optimal for an investor with a SAHARA utility function to take a long position in fund \( \rho^{-\frac{1}{\alpha}} \) and a short position in fund \( \rho^{\frac{1}{\alpha}} \).

Up to this point, one may notice that a common feature of the above three examples is that their inverse marginal utility functions can all be viewed as linear combinations of power terms \( \xi^\gamma \). Specifically, in the CRRA example there is only one power term \( \xi^{-\frac{1}{\alpha}} \) with \( \gamma = -\frac{1}{\alpha} \), in the quadratic case we have two power terms 1 and \( \xi \) corresponding to \( \gamma = 0, 1 \), and for the SAHARA utility we again have two power terms \( \xi^{\frac{1}{\alpha}} \) and \( \xi^{-\frac{1}{\alpha}} \) where \( \gamma = \pm \frac{1}{\alpha} \). While one may suspect that the power terms are the only form permitted in a separating preference, the following example demonstrates that it is actually not the case.

**Example 7 (CARA Utility)** Consider the CARA utility function

\[ u (x) = -e^{-ax}, \]

with the coefficient of absolute risk aversion \( a > 0 \). The inverse marginal utility can be expressed as

\[ I (\xi) = \frac{1}{a} (\log a - \log \xi). \]

Since

\[ I (\lambda \rho) = \frac{1}{a} (\log a - \log (\lambda \rho)) = \frac{1}{a} (\log a - \log \lambda) - \frac{1}{a} \log \rho, \]

the CARA utility function exhibits two-fund separation with one of the separating funds being a risk-free asset

\[ f_1 (\rho) = 1, \]
and the other a risky portfolio

\[ f_2(\rho) = \log \rho. \]

The corresponding investment weights are given by

\[
\alpha_1(\lambda) = \frac{1}{a} (\log a - \log \lambda), \\
\alpha_2(\lambda) = -\frac{1}{a}.
\]

This suggests that an investor with CARA utility would always find it optimal to invest a wealth-dependent amount into the risk-free asset and take a constant short position in the risky fund \( \log \rho \).

Notice that in the above example, the inverse marginal utility consists of two terms: a degenerate power term \( \xi^0 = 1 \) and an additional term \( \log \xi \). This suggests that \( \log \xi \) may also show up in a separating preference.

In all four examples above, the power terms \( \xi^\gamma \) in the inverse marginal utility feature real power values. However, this does not have to be the case. In particular, when we have a pair of complex power values \( \gamma \pm bi \), \( \xi^{\gamma \pm bi} \) can be transformed into \( \cos (b \log \xi) \xi^\gamma \) and \( \sin (b \log \xi) \xi^\gamma \). The following example illustrates that these terms can also appear in a separating preference.

Example 8 Consider a utility function \( u \), whose inverse marginal utility is given by

\[
I(\xi) = [\cos (\log \xi) + \sin (\log \xi) + b] \xi^{-a},
\]

where \( a > 1 \) and \( b > 2 \). It can be verified that

\[
I(\lambda \rho) = [\cos (\log (\lambda \rho)) + \sin (\log (\lambda \rho)) + b] (\lambda \rho)^{-a}
= [\cos (\log \lambda) + \sin (\log \lambda)] \lambda^{-a} \cos (\log \rho) \rho^{-a}
+ [\cos (\log \lambda) - \sin (\log \lambda)] \lambda^{-a} \sin (\log \rho) \rho^{-a} + b \lambda^{-a} \rho^{-a}.
\]
Hence, we have three-fund separation with the following separating funds

\[
\begin{align*}
    f_1 (\rho) &= \cos (\log \rho) \rho^{-a}, \\
    f_2 (\rho) &= \sin (\log \rho) \rho^{-a}, \\
    f_3 (\rho) &= \rho^{-a}.
\end{align*}
\]

The corresponding investment weights are given by

\[
\begin{align*}
    \alpha_1 (\lambda) &= [\cos (\log \lambda) + \sin (\log \lambda)] \lambda^{-a}, \\
    \alpha_2 (\lambda) &= [\cos (\log \lambda) - \sin (\log \lambda)] \lambda^{-a}, \\
    \alpha_3 (\lambda) &= b \lambda^{-a}.
\end{align*}
\]

### 1.3.3 General Characterization of $K$-Fund Separation

In this section, we provide a general characterization of preferences exhibiting $K$-fund separation. Our characterization is stated in terms of the inverse marginal utility function $I$. We show that a separating preference can only have the following terms in its inverse marginal utility function: $C$ (constant), $\xi^a$, $(\log \xi)^a$, $\xi^a (\log \xi)^a$, $\cos (b \log \xi)$, $\sin (b \log \xi)$, $\xi^a \cos (b \log \xi)$, $\xi^a \sin (b \log \xi)$, $(\log \xi)^a \cos (b \log \xi)$, $(\log \xi)^a \sin (b \log \xi)$, $\xi^a (\log \xi)^a \cos (b \log \xi)$, and $\xi^a (\log \xi)^a \sin (b \log \xi)$. Indeed, we have seen many of these terms in the examples above.

The following theorem provides the necessary and sufficient conditions for $K$-fund separation, where $K \geq 1$ can be any positive integer. This characterization is similar to Theorem 7.1 in Cass and Stiglitz (1970), although it is difficult to see whether that characterization is equivalent to ours. Theorem 7.1 of Cass and Stiglitz (1970) is stated without proof and contains terms that should not be there. However, the remark to the theorem describes an additional restriction which rules out at least some of the extra terms and may make their result equivalent to ours.
Theorem 9  A utility function $u$ (with $u' > 0$ and $u'' < 0$) exhibits $K$-fund separation if and only if the inverse marginal utility function $I = (u')^{-1}$ can be expressed as

$$I(\xi) = \sum_{k=1}^{J} \xi^\gamma_k P_{k,1}(\log \xi) \cos (b_k \log \xi) + \sum_{k=1}^{J} \xi^\gamma_k P_{k,2}(\log \xi) \sin (b_k \log \xi),$$  \hspace{1cm} (1.3)$$

where

1. The ordered pairs $(\gamma_k, b_k)$ are distinct for each $k$ with $b_k \geq 0$;
2. For $i = 1, 2$, $P_{k,i}(\log \xi)$ is a polynomial function of $\log \xi$ of degree $d_{k,i} \geq 0$, i.e. $P_{k,i}(\log \xi) = \sum_{l=0}^{d_{k,i}} C_{k,i,l}(\log \xi)^l$, where the leading coefficient $C_{k,i,d_{k,i}} \neq 0$;
3. If $b_k = 0$ (the sin terms disappear, but the cos terms do not), then $d_{k,2} = 0$; and
4. $\sum_{k=1}^{J} (d_k + 1) (1 + 1_{b_k \neq 0}) = K$, where $d_k = \max_{i=1,2} (d_{k,i})$, and $1_{b_k \neq 0}$ is an indicator function that takes a value of 1 when $b_k \neq 0$ and 0 otherwise.

The separating funds can be chosen as follows: $\forall k = 1, 2, \cdots, J$ and $\forall l = 0, 1, \cdots, d_k$,

$$f_{k,l}(\rho) = \rho^{\gamma_k}(\log \rho)^l,$$  \hspace{1cm} (1.4)$$

when $b_k = 0$, and

$$f_{k,1,l}(\rho) = \rho^{\gamma_k} \cos (b_k \log \rho) (\log \rho)^l,$$  \hspace{1cm} (1.5)$$

$$f_{k,2,l}(\rho) = \rho^{\gamma_k} \sin (b_k \log \rho) (\log \rho)^l,$$  \hspace{1cm} (1.6)$$

when $b_k \neq 0$.

The associated investment weights are given by

$$\alpha_{k,l}(\lambda) = \lambda^{\gamma_k} \sum_{j=l}^{d_{k,1}} C_{k,1,j} \binom{j}{l} (\log \lambda)^{j-l},$$  \hspace{1cm} (1.7)$$
when $b_k = 0$, and

$$
\alpha_{k,1,l}(\lambda) = \sum_{j=1}^{d_{k,1}} C_{k,1,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l} \quad (1.8)
$$

$$
+ \sum_{j=1}^{d_{k,2}} C_{k,2,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l},
$$

$$
\alpha_{k,2,l}(\lambda) = \sum_{j=1}^{d_{k,2}} C_{k,2,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l} \quad (1.9)
$$

$$
- \sum_{j=1}^{d_{k,1}} C_{k,1,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l},
$$

when $b_k \neq 0$.

**Proof of Theorem 9 (sketch):** Here is a sketch of the proof. The formal proof is delegated to the Appendix.

Utility function $u$ exhibits $K$-fund separation if and only if (1.2) holds whenever a solution to Problem 1 exists. Taking derivatives of (1.2) with respect to $\lambda$ yields

$$
\begin{pmatrix}
\rho I'(\lambda \rho) \\
\vdots \\
\rho^K I^{(K)}(\lambda \rho)
\end{pmatrix} = M_0(\lambda) \begin{pmatrix} f_1(\rho) \\
\vdots \\
f_K(\rho) \end{pmatrix},
$$

where $I^{(k)}$ denotes the $k^{th}$ derivative of $I$, and $M_0(\lambda)$ is defined as

$$
M_0(\lambda) = \begin{pmatrix} 
\alpha_1'(\lambda) & \cdots & \alpha_K'(\lambda) \\
\vdots & \ddots & \vdots \\
\alpha_1^{(K)}(\lambda) & \cdots & \alpha_K^{(K)}(\lambda) 
\end{pmatrix}. \quad (1.10)
$$

Assume for now that $M_0(\lambda)$ is non-singular for some $\lambda$, i.e., $\exists \lambda$ such that $(M_0(\lambda))^{-1}$ exists. We show in the appendix that a simple trick allows us to tackle the singularity case for which similar results obtain. When $M_0(\lambda)$ is not singular, we have

$$
\begin{pmatrix} f_1(\rho) \\
\vdots \\
f_K(\rho) \end{pmatrix} = (M_0(\lambda))^{-1} \begin{pmatrix} \rho I'(\lambda \rho) \\
\vdots \\
\rho^K I^{(K)}(\lambda \rho) \end{pmatrix}. \quad (1.11)
$$
Plugging (1.11) back into (1.2) gives us

\[
I(\lambda \rho) = \left( \begin{array}{c}
\alpha_1(\lambda) \\
\vdots \\
\alpha_K(\lambda)
\end{array} \right)^T (M_0(\lambda))^{-1} \left( \begin{array}{c}
\rho I'(\lambda \rho) \\
\vdots \\
\rho^K I^{(K)}(\lambda \rho)
\end{array} \right).
\]  

(1.12)

Without loss of generality, assume that \((M_0(\lambda))^{-1}\) exists when \(\lambda = 1\). Evaluating (1.12) at \(\lambda = 1\) and rearranging yield a differential equation of the form

\[
A_K I^{(K)}(\xi) \xi^K + \cdots + A_1 I'(\xi) \xi + I(\xi) = 0,
\]

(1.13)

where \(A_1, A_2, \cdots, A_K\) are constants. To ensure non-degenerate \(K\)-fund separation, we must have \(A_K \neq 0\). Then, (1.13) is a \(K^{th}\)-order homogeneous Euler’s equation.

To solve this differential equation, we conjecture \(I(\xi) = \xi^\delta\) and plug this into (1.13). This gives us the following \(K^{th}\)-order polynomial equation

\[
A_K \delta (\delta - 1) \cdots (\delta - K + 1) + \cdots + A_2 \delta (\delta - 1) + A_1 \delta + 1 = 0,
\]

(1.14)

with \(K\) roots. Some of these \(K\) roots may be repeated, thus reducing to \(J \leq K\) distinct roots \(\{\gamma_k + b_k i\}_{k=1}^J\), each of which can be either real \((b_k = 0)\) or complex \((b_k \neq 0)\). If a real root \(\gamma_k\) is not repeated, then it yields a power term \(\xi^{\gamma_k}\) in the solution of \(I(\xi)\). If \(\gamma_k\) is repeated for \(d_k + 1\) times, then it gives rise to \(d_k + 1\) terms \(\{\xi^{\gamma_k} (\log \xi)^l\}_{l=0}^{d_k}\) in \(I(\xi)\), which can be combined as \(\xi^{\gamma_k} P_{k,1}(\log \xi)\). Similarly, if a pair of complex roots \(\gamma_k \pm b_k i\) is not repeated, it gives rise to two terms \(\xi^{\gamma_k} \cos (b_k \log \xi)\) and \(\xi^{\gamma_k} \sin (b_k \log \xi)\) in \(I(\xi)\). If the pair \(\gamma_k \pm b_k i\) is repeated for \(d_k + 1\) times, then it generates \(\xi^{\gamma_k} P_{k,1}(\log \xi) \cos (b_k \log \xi)\) and \(\xi^{\gamma_k} P_{k,2}(\log \xi) \sin (b_k \log \xi)\). To ensure that the total number of roots is equal to \(K\), we must have

\[
K = \sum_{k=1}^J (d_k + 1) (1 + 1_{b_k \neq 0})
\]

Combining all the above terms, it then follows that the solution to (1.14) is given by (1.3).

By the theory of ordinary differential equations (see, for example, Birkhoff and Rota (1962), Lemma IV.3.2 and the discussion after Theorem IV.2.2 on how to convert Euler’s
Terms in $I$ & $\gamma_k$ & $b_k$ & $d_k$ \\
--- & --- & --- & --- \\
$C$ (constant) & 0 & 0 & 0 \\
$\xi^\gamma$ & $\gamma$ & 0 & 0 \\
$\{ (\log \xi)^t \}_{t}$ & 0 & 0 & $\geq 0$ \\
$\{ \xi^\gamma (\log \xi)^t \}_{t}$ & $\gamma$ & 0 & $\geq 0$ \\
$\cos (b \log \xi), \sin (b \log \xi)$ & 0 & $b$ & 0 \\
$\xi^\gamma \cos (b \log \xi), \xi^\gamma \sin (b \log \xi)$ & $\gamma$ & $b$ & 0 \\
$\{ (\log \xi)^t \cos (b \log \xi), (\log \xi)^t \sin (b \log \xi) \}_{t}$ & 0 & $b$ & $\geq 0$ \\
$\{ \xi^\gamma (\log \xi)^t \cos (b \log \xi), \xi^\gamma (\log \xi)^t \sin (b \log \xi) \}_{t}$ & $\gamma$ & $b$ & $\geq 0$ \\

Figure 1.1: Possible Terms in $I$ of a Separating Preference

differential equation to an equation with fixed coefficients), we know that (1.4)-(1.6) are linearly independent and they together form a solution basis for the Euler’s equation (1.13). Thus, they can be chosen as a set of separating funds.

While the expression of (1.3) seems complicated, it is indeed a concise way to incorporate all possible terms listed at the beginning of this section. Figure 1.1 summarizes the correspondence of various parameter values in (1.3) to different possible terms in $I$.

The characterization of $K$-fund separation for a class of preferences $U$ follows almost immediately from Theorem 9. The inverse marginal utility of each $u \in U$ must be the sum of terms as in (1.3), and each of these terms must appear with non-zero coefficient for some utility function $\hat{u} \in U$ to ensure non-degeneracy. Formally, we have the following corollary.

**Corollary 10** A class of preferences $U$ exhibits $K$-fund separation if and only if there exist $J$ distinct ordered pairs $\{(\gamma_k, b_k)\}_{k=1}^J$ with $b_k \geq 0$ and non-negative integers $\{D_k\}_{k=1}^J$ that satisfy $\sum_{k=1}^J (D_k + 1) (1 + 1_{b_k \neq 0}) = K$ such that $\forall u \in U$, the inverse marginal utility function $I = (u')^{-1}$ can be expressed as (1.3), where

1. For $i = 1, 2$, $P_{k,i}(\log (\xi))$ is a polynomial function of $\log (\xi)$ of degree $d_{k,i} \geq -1$. When $d_{k,i} = -1$, $P_{k,i}(\log (\xi))$ is an empty sum, which we take to be uniformly equal to zero;
(2) If $b_k = 0$ (the sin terms disappear, but the cos terms do not), then $d_{k,2} = -1$;

(3) $\forall k = 1, 2, \cdots, J, \max_{i=1,2} d_{k,i} \leq D_k$;

(4) $\forall k = 1, 2, \cdots, J, \exists \hat{u} \in U$ such that $\max_{i=1,2} \hat{d}_{k,i} = D_k$.

The separating funds can be chosen to be $(1.4)-(1.6)$.

According to the above corollary, mutual fund separation for a class of preferences is very similar in spirit to that for a single utility function. Hence, we do not make formal distinctions between these two cases in the analyses below.

### 1.3.4 One-Fund Separation

It is immediate from Theorem (9) that a utility function $u$ exhibits one-fund separation if and only if its inverse marginal utility can be written as

$$I(\xi) = C\xi^\gamma,$$  \hspace{1cm} (1.15)

for some constant $C$. This corresponds to the case where (1.14) has a single real root. We will show in Section 1.6 that strict concavity entails

$$C\gamma < 0.$$

The unique separating fund (up to multiplication by a scalar) and the associated investment weight are given by

$$f(\rho) = \rho^\gamma,$$

and

$$\alpha(\lambda) = C\lambda^\gamma.$$

One can verify that the primal utility function for the one-fund separating preferences is

$$u(x) = \begin{cases} 2C \xi^{\frac{1}{\gamma+1}}, & \gamma \neq -1 \text{ and } \gamma \neq 0, \\ C\log \left( \frac{x}{\xi} \right), & \gamma = -1 \end{cases},$$  \hspace{1cm} (1.16)
where \( x \in (0, +\infty) \) when \( \gamma < 0 \) and \( x \in (-\infty, 0) \) when \( \gamma > 0 \). Notice that when \( C > 0 \) and \( \gamma < 0 \), this corresponds to the CRRA utility function, which is defined on positive consumption levels. In fact, setting \( C = 1 \) and \( \gamma = -\frac{1}{a} \), we obtain the standard form as given in Example 4.

When \( C < 0 \) and \( \gamma > 0 \), we define this utility function as a mirror version of the CRRA preference. Without loss of generality, let \( C = -1 \) and \( \gamma = \frac{1}{a} \) with \( a > 0 \), and then (1.16) is reduced to

\[
u(x) = \frac{(-x)^{1+a}}{1+a},
\]

for all \( x \in (-\infty, 0) \). Note that the mirror CRRA utility has a very similar form to that of the standard CRRA utility with a coefficient of relative risk aversion \( a \), except that now it is defined on negative wealth levels.

### 1.3.5 Two-Fund Separation

If a utility function \( u \) exhibits two-fund separation, then Theorem (9) suggests that three different cases are possible. These three cases correspond to the scenarios in which (1.14) have two distinct real roots, two repeated real roots, and a pair of complex roots, respectively. We now discuss each of these cases separately.

**Case 1:** When (1.14) has two distinct real roots, (1.3) is reduced to

\[
I(\xi) = C_1 \xi^{\gamma_1} + C_2 \xi^{\gamma_2},
\]

where \( \gamma_1 \neq \gamma_2 \) and \( C_1, C_2 \) are arbitrary constants. It can be easily verify that the two separating funds and the associated investment weights are given by

\[
f_k(\rho) = \rho^{\gamma_k},
\]

and

\[
\alpha_k(\lambda) = C_k \lambda^{\gamma_k},
\]

20
for $k = 1, 2$.

Examples of this case include the quadratic utility function obtained when $\gamma_1 = 0$ and $\gamma_2 = 1$ (see Example 5), and the SAHARA utility function obtained when $\gamma_1 = -\frac{1}{a}$ and $\gamma_2 = \frac{1}{a}$ (see Example 6). Another example, which will later be introduced in Section 1.4.2, is the GOBI preference, whose inverse marginal utility takes the form

$$I(\xi) = C_1 \xi^\gamma + C_2 \xi^{2\gamma},$$

with $\gamma \neq 0$.

An interesting observation is that in this case, the inverse marginal utility of a two-fund separating preference can be viewed as the linear combination of that of two different one-fund separating preferences. In fact, we will soon discuss that taking the linear combination of distinct separating preferences in the inverse marginal utility leads to another separating preference with a higher degree.

**Case 2:** When (1.14) has two repeated real roots, (1.3) can be written as

$$I(\xi) = C_1 \xi^\gamma + C_2 \xi^{2\gamma} \log \xi,$$

for some constants $C_1$ and $C_2$. We will show in Section 1.6 that strict concavity entails $\gamma = 0$, so we must have

$$I(\xi) = C_1 + C_2 \log \xi,$$

which is exactly the CARA utility (see Example 7).

In this case, the two separating funds are

$$f_1(\rho) = 1,$$
$$f_2(\rho) = \log \rho,$$

with associated investment weights

$$\alpha_1(\lambda) = C_1 + C_2 \log \lambda,$$
$$\alpha_2(\lambda) = C_2.$$
While the previous case with two distinct real roots can be constructed by taking linear combinations of two one-fund separating preferences in the inverse marginal utility, it is obvious that the case with repeated real roots does not obtain this way. This is because the log ξ term, which yields one separating fund in a two-fund separating preference, does not show up in one-fund separation. Indeed, we will see later that the log terms are not the only ones that cannot be constructed from low-degree separations. Also absent in low-degree separations are the cos and sin terms.

Case 3: When (1.14) has a pair of complex roots, the inverse marginal utility takes the form of

\[ I(\xi) = C_1 \xi \cos(b \log \xi) + C_2 \xi \sin(b \log \xi), \]

with constants \( C_1 \) and \( C_2 \). We will show in Section 1.6 that this form cannot exist under strict concavity.

1.3.6 From Low-Degree to High-Degree Separation

As have been discussed in Section 1.3.5, taking linear combinations of two distinct one-fund separating preferences in the inverse marginal utility allows us to construct a large class of two-fund separating preferences, such as the SAHARA and GOBI utility. More generally, this approach can be used broadly to generate higher-degree separating preferences from those with lower degrees. We now formalize this idea in the following theorem.

**Theorem 11** Consider \( N \) utility functions \( \{u_n\}_{n=1}^N \) with corresponding inverse marginal utility given by \( \{I_n\}_{n=1}^N \). Suppose that each \( u_n \) exhibits \( K_n \)-fund separation. If we have another utility function \( u \), whose inverse marginal utility is given by

\[ I(\xi) = \sum_{n=1}^{N} t_n I_n(\xi), \] (1.17)
for some non-zero constants $a_1, a_2, \ldots, a_n$, then $u$ satisfies $K$-fund separation with

$$K \leq \sum_{n=1}^{N} K_n,$$

where the equality holds when the separating funds of the $N$ utility functions are linearly independent.

**Proof of Theorem 11:** Since $u_n$ exhibits $K_n$-fund separation, we must have

$$I_n(\lambda \rho) = \sum_{k=1}^{K_n} \alpha_{n,k}(\lambda) f_{n,k}(\rho).$$

By (1.17), we obtain

$$I(\lambda \rho) = \sum_{n=1}^{N} t_n I_n(\lambda \rho) = \sum_{n=1}^{N} \sum_{k=1}^{K_n} t_n \alpha_{n,k}(\lambda) f_{n,k}(\rho),$$

which is a linear combination of the $f_{n,k}(\rho)$’s with corresponding weights depending on $\lambda$ only. This suggests that $u$ satisfies mutual fund separation.

The degree of separation depends on whether the $f_{n,k}(\rho)$’s are linearly independent. If so, then $u$ satisfies $K$-fund separation with $K = \sum_{n=1}^{N} K_n$. Otherwise, $K < \sum_{n=1}^{N} K_n$, because multiple funds can be cancelled against each other or combined to form a larger fund, which lowers the degree of separation. \vspace{1em}

Theorem 11 proposes a simple way of constructing high-degree separating preferences by taking linear combinations of low-degree ones in the inverse marginal utility. For example, if we start with $K$ distinct one-fund separating preferences, taking the linear combination in the inverse marginal utility generates a $K$-fund separating utility with

$$I(\xi) = \sum_{k=1}^{K} C_k \xi^k,$$

for some constants $C_1, C_2, \ldots, C_K$.

It is tempting to think that this approach allows us to construct any high-degree separating preference from low-degree ones, but it is actually not the case. In fact, we have already
seen a counterexample in Section 1.3.5. Specifically, we are not able to construct the CARA utility, which exhibits two-fund separation, from one-fund separating preferences. This is because the CARA class has a \( \log \xi \) term in the inverse marginal utility, which only shows up when (1.14) has repeated roots and thus never exists in one-fund separation. Similarly, separation of even higher degrees can involve \((\log \xi)^2\), \((\log \xi)^3\) and so forth, implying that we can always obtain new terms when the degree of separation becomes larger. Hence, it is impossible to find a set of separating preferences that can be used to construct all higher-degree separation.

Notice that also absent from one- and two-fund separation are the \( \cos \) and \( \sin \) terms, which obtain when (1.14) has complex roots. As will be shown later in the paper, while the condition of strict concavity prevents these terms from showing up in two-fund separation, they do appear in separating preferences of higher degrees. One such case is given in Example 8.

1.3.7 Range of Optimal Consumption

This section examines the range of optimal consumptions for separating preferences. Previous research typically defines utility functions on positive wealth only, implicitly assuming that consumptions cannot be negative. In this paper, we take a more general view and allow consumptions to be either positive or negative. It would then be interesting to ask under what conditions an investors optimally consumes a positive amount in all states of the world. The following theorem addresses this issue.

**Theorem 12** Consider a utility function \( u \) exhibiting mutual fund separation. The following statements are equivalent:

1. For all initial wealth levels such that an optimum exists, an investor with utility function \( u \) optimally consumes a positive amount of wealth in all states of the world.
2. For all $\xi > 0$, we have $I(\xi) > 0$.

3. The equation $I(\xi) = 0$ has no positive solution and $I(\xi^*) > 0$ for some $\xi^*$.

**Proof of Theorem 12:** Since the optimal consumption is $I(\lambda \rho)$, the investor consumes a positive amount in all states at all wealth levels if and only if $I(\lambda \rho) > 0$ for all $\lambda > 0$ and $\rho > 0$. This is equivalent to $I(\xi) > 0$ for all $\xi > 0$. Hence, statements 1 and 2 are equivalent.

On the other hand, if $u$ satisfies mutual fund separation, its inverse marginal utility function takes the form of (1.3), which is clearly a continuous function in $\xi$. It is then immediate that statements 2 and 3 are equivalent.

The following theorem characterizes the range of optimal consumptions for all utility functions exhibiting mutual fund separation.

**Theorem 13** For a utility function $u$ satisfying mutual fund separation, the range of optimal consumptions $\{I(\xi) : \xi > 0\}$ is

1. $(-\infty, 0)$ or $(0, +\infty)$ when $K = 1$;

2. an open unbounded interval, and it can be any open unbounded interval when $K \geq 2$.

Notice that Theorem 13 equips us with a simple way of identifying some of the utility functions that do not satisfy mutual fund separation by examining the range of optimal consumptions without having to know the exact form of the utility. In particular, if we have a utility function with a bounded range of optimal consumptions, then we know for sure that such a preference cannot exhibit mutual fund separation.

**1.3.8 Machina Preferences**

Our analyses so far have focused on von Neumann-Morgenstern utility functions. More generally, investor preferences may not take the expected utility form, in which case our
previous results do not necessarily apply. Machina (1982) shows that expected utility can be viewed as a special case of a larger class of preferences, which we call “Machina preferences” hereafter, and many properties and results in expected utility theory obtain similarly for Machina preferences. In this section, we ask whether our fund separation results for von Neumann-Morgenstern utility can be extended to the Machina preferences. As in the case of von Neumann-Morgenstern utility, Machina assumes that investor preferences depend on the distribution of consumptions only and thus are not state-dependent. Unlike von Neumann-Morgenstern utility, Machina preferences do not assume the form of expected utility. Indeed, Machina (1982) shows that when the utility function is smooth in the sense of Fréchet differentiability, preferences can be modeled locally as expected utility.\footnote{We are being informal about the topology used to define the Fréchet derivative if $\Omega$ is not bounded. In Machina’s original work (as in many derivations of von Neumann-Morgenstern preferences), it was assumed that consumption is bounded. To formalize what we are doing for unbounded random variables, we would have to specify the topology over distribution functions to define the sense of approximation.}

Consider a Fréchet differentiable utility function $V(\cdot)$ defined over distributions of consumption portfolios. Let $x^*$ and $x$ denote two random consumption portfolios with the corresponding distribution functions given by $F^*$ and $F$. Suppose that $F^*$ and $F$ lies very close to each other. We then have

$$V (F) - V (F^*) \approx \int U (z; F^*) (dF - dF^*) = E [U (x; F^*)] - E [U (x^*; F^*)],$$

or equivalently,

$$V (F) \approx V (F^*) + E [U (x; F^*)] - E [U (x^*; F^*)],$$

(1.18)

where $U (z; F)$ is the local utility function over consumption level $z$ evaluated at distribution $F$. Assume that $U (\cdot; F)$ is strictly concave for all $F$.

With the Machina preferences, investors face the following utility maximization problem.

**Problem 14** Choose consumption $x$ to

$$\max_x V (F)$$
subject to the budget constraint

\[ \mathbf{E}[x \rho] \leq w_0. \]

Since \( U(\cdot;F) \) is strictly concave for all \( F \), if an optimum exists, it must be unique. Suppose that \( x^* \) solves Problem 14. Then, it must maximize (1.18) for local utility function \( U(\cdot;F^*) \) evaluated at \( F^* \). Notice that once we fix \( F^* \), both \( V(F^*) \) and \( \mathbf{E}[U(x^*;F^*)] \) are constants. Therefore, \( x^* \) must maximize \( \mathbf{E}[U(x;F^*)] \). Formally, the optimal consumption portfolio \( x^* \) must be the solution to the following problem.

**Problem 15** Choose consumption \( x \) to

\[
\max_x \mathbf{E}[U(x;F^*)]
\]

subject to the budget constraint

\[ \mathbf{E}[x \rho] \leq w_0. \]

It seems that we are faced with a similar problem as in the baseline case of von Neumann-Morgenstern utility. Apparently, if the local utility function \( U(\cdot;F) \) satisfies mutual fund separation with the same separation funds at all \( F \), then \( V(\cdot) \) exhibits mutual fund separation globally. In fact, this condition is much stronger than what is needed. The only thing we need is for the all optimal consumption portfolios corresponding to different initial wealth levels to be spanned by the same set of separating funds. Since different initial wealth gives rise to different optimal consumption portfolios, which in turn feature different local utility functions, for each local utility function \( U(\cdot;F) \) we only need fund separation to hold at the particular wealth level supporting \( F \) as the optimal consumption portfolio.

**Theorem 16** Consider a Machina utility function \( V(\cdot) \), whose associated local utility function at any consumption distribution \( F \) is given by \( U(\cdot;F) \). If \( U(\cdot;F) \) satisfies mutual fund separation at all \( F \) with respect to the same set of separation funds, then \( V(\cdot) \) exhibits mutual fund separation.
To illustrate mutual fund separation for Machina preferences, we now provide two examples. The first example is the mean-variance preference, which has been shown to satisfy two-fund money separation in Tobin (1958).

**Example 17** Consider the mean-variance preference

\[
V(F) = E(x) - \frac{a}{2} \text{Var}(x) = E(x) - \frac{a}{2} \left[ E(x^2) - (E(x))^2 \right],
\]

where \( a > 0 \) represents the level of risk aversion. The first order condition is given by

\[
1 - a [x^* - E(x^*)] = \lambda \rho, \tag{1.19}
\]

which yields

\[
x^* = E(x^*) + \frac{1}{a} - \frac{\lambda}{a} \rho. \tag{1.20}
\]

This implies that \( V(F) \) exhibits two-fund money separation with separating funds

\[
f_1(\rho) = 1,
\]

\[
f_2(\rho) = \rho.
\]

The corresponding investment weights for the two separating funds are given by

\[
\alpha_1(w_0) = E(x^*) + \frac{1}{a},
\]

\[
\alpha_2(w_0) = -\frac{\lambda}{a}.
\]

To determine the value of \( \lambda \), taking the expectation of (1.20) and rearranging lead to

\[
E[x^* - E(x^*)] = E \left( 1 - \frac{\lambda \rho}{a} \right).
\]

The left hand side is apparently equal to 0, which implies

\[
E \left( \frac{1 - \lambda \rho}{a} \right) = 0,
\]
and hence
\[
\lambda = \frac{1}{\mathbb{E}(\rho)}.
\] (1.21)

This suggests that fixing the distribution of \( \rho \), \( \lambda \) is a constant and does not change with wealth. This is different from the case of Von Neumann-Morgenstern utility, where the initial wealth level and \( \lambda \) has a one-to-one map.

On the other hand, the budget constraint implies

\[
w_0 = \mathbb{E}(x^* \rho)
= \mathbb{E} \left[ \left( \mathbb{E}(x^*) + \frac{1 - \lambda \rho}{\alpha} \right) \rho \right]
= \mathbb{E} \left[ \mathbb{E}(x^*) \rho + \frac{1}{\alpha} \left( 1 - \frac{\rho}{\mathbb{E}(\rho)} \right) \rho \right]
= \mathbb{E}(x^*) \mathbb{E}(\rho) + \frac{1}{\alpha} \mathbb{E}(\rho) - \frac{1}{\alpha} \mathbb{E}(\rho^2),
\]

where the second and third equalities follow (1.20) and (1.21), respectively. Solving for \( \mathbb{E}(x^*) \) leads to

\[
\mathbb{E}(x^*) = \frac{\mathbb{E}(\rho^2)}{a[\mathbb{E}(\rho)]^2} + \frac{w_0}{\mathbb{E}(\rho)} - \frac{1}{a},
\]

which varies with the wealth level.

Therefore, investors with the mean-variance preferences find it optimal to short a fixed amount in the risky fund \( \rho \) and to invest the rest of their wealth in the risk-free asset.

The mean-variance preferences essentially assume that the only risk that investors are concerned with is the variance of consumptions. As another example, we incorporate another dimension of risk into the utility function, which is the downside risk \( \mathbb{E} \left[ (|x - \mathbb{E}(x)|^-)^2 \right] \), where \(|:|^- = \min(0, \cdot)\). This downside risk measure is originally introduced in Bawa and Lindenberg (1977).

**Example 18** Consider a utility function

\[
V(F) = \mathbb{E}(x) - \frac{a}{2} \mathbb{E} \left[ (x - \mathbb{E}(x))^2 \right] - \frac{b}{2} \mathbb{E} \left[ (|x - \mathbb{E}(x)|^-)^2 \right],
\]
with \( a, b > 0 \).\(^7\) The first order condition is given by

\[
1 - a [x^* - \mathbb{E}(x^*)] - b [x^* - \mathbb{E}(x^*)]^- = \lambda \rho. \tag{1.22}
\]

This implies that when \( \rho \leq \frac{1}{\lambda} \),

\[
x^* = \mathbb{E}(x^*) + \frac{1 - \lambda \rho}{a} \geq \mathbb{E}(x^*),
\]

and when \( \rho > \frac{1}{\lambda} \),

\[
x^* = \mathbb{E}(x^*) + \frac{1 - \lambda \rho}{a + b} < \mathbb{E}(x^*). \tag{1.23}
\]

These can be combined and expressed as

\[
x^* = \mathbb{E}(x^*) + \max \left( \frac{1 - \lambda \rho}{a}, \frac{1 - \lambda \rho}{a + b} \right).
\]

As in Example 17, one can show that fixing the distribution of \( \rho, \lambda \) is a constant, whereas \( \mathbb{E}(x^*) \) depends on the value of the initial wealth \( w_0 \). Therefore, \( V(F) \) satisfies two-fund money separation with separating funds

\[
f_1(\rho) = 1,
\]

\[
f_2(\rho) = \max \left( \frac{1 - \lambda \rho}{a}, \frac{1 - \lambda \rho}{a + b} \right).
\]

The corresponding investment weight for the risk-free asset is a wealth-dependent amount

\[
\alpha_1(w_0) = \mathbb{E}(x^*),
\]

and the risky fund is always assigned a constant weight

\[
\alpha_2(w_0) = 1.
\]

---

\(^7\)Kadan, Liu and Liu (2014) use another approach to show that this class of preferences exhibits two-fund money separation.
1.4 Primal Utility Function

So far we have characterized the set of separating preferences in terms of the inverse marginal utility function. We then ask whether we are able to derive the associated primal utility functions. The most natural way to do this is to first invert $I$ to obtain $u'$, and then integrate $u'$ to get $u$. However, this does not yield a closed form solution in most cases. Examples for which the primal utility exists in closed form are mostly limited to one- and two-fund separation (CARA, CRRA, quadratic, etc.). In this section, we introduce two classes of three-fund separating preferences (SAHARA and GOBI), for which we are able to obtain a closed form expression in the primal utility.

1.4.1 SAHARA Utility

The SAHARA class is proposed by Chen, Pelsser and Vellekoop (2011). The standard SAHARA utility has an inverse marginal utility function of the form

$$I(\xi) = \frac{1}{2} \left( \xi^{\frac{1}{a}} - b^2 \xi^{\frac{1}{a}} \right), \quad (1.24)$$

with $a, b > 0$. As mentioned in Example 6, the standard form exhibits two-fund separation, for which an investor would optimally take a long position in one risky fund and a short position in another.

More generally, one could introduce a shift into the SAHARA utility, which is obtained by adding a constant term to the inverse marginal utility, i.e.,

$$I(\xi) = \frac{1}{2} \left( \xi^{\frac{1}{a}} - b^2 \xi^{\frac{1}{a}} \right) + C_0. \quad (1.25)$$

The constant $C_0$ is referred to as the default point in Chen, Pelsser and Vellekoop (2011), and we can easily obtain the standard form by setting $C_0 = 0$. When $C_0 \neq 0$, the shifted SAHARA utility exhibits three-fund money separation with a long position

$$\alpha_1(\lambda) = \frac{1}{2} \lambda^{-\frac{1}{a}}$$
in the first risky fund
\[ f_1 (\rho) = \rho^{-\frac{1}{a}}, \]
a short position
\[ \alpha_2 (\lambda) = -\frac{1}{2} b^2 \lambda^{\frac{1}{a}} \]
in a second risky fund
\[ f_2 (\rho) = \rho^{\frac{1}{a}}, \]
and a fixed investment
\[ \alpha_3 (\lambda) = C_0 \]
in the risk-free asset
\[ f_3 (\rho) = 1. \]

One nice property of the SAHARA utility is that the two power terms \( \xi^{-\frac{1}{a}} \) and \( \xi^{\frac{1}{a}} \) in the inverse marginal utility are reciprocals of each other. As such, by multiplying \( \xi^{\frac{1}{a}} \) on both sides of (1.25), we can easily rewrite it as a quadratic equation of \( \xi^{\frac{1}{a}} \), which further allows us to invert \( I \) so as to recover the primal utility function. Explicitly, multiplying \( \xi^{\frac{1}{a}} \) on both sides of (1.25) and setting \( x = I (\xi) \) and \( u' (x) = \xi \) yield
\[ b^2 (u' (x))^{\frac{2}{a}} + 2 (x - C_0) (u' (x))^{\frac{1}{a}} - 1 = 0. \]
Since \( u'(x) > 0 \), solving this quadratic equation gives us
\[ u' (x) = \left( \frac{\sqrt{x^2 + b^2} + \hat{x}}{b^2} \right)^a, \]
where \( \hat{x} = x - C_0 \). We then integrate \( u'(x) \) to obtain the primal utility function
\[ u (x) = \begin{cases} 
\frac{1}{1-a} \left( \sqrt{x^2 + b^2} + \hat{x} \right)^{-a} (a \sqrt{x^2 + b^2} + \hat{x}), & a \neq 1 \\
\frac{1}{2} \left[ \log \left( \sqrt{x^2 + b^2} + \hat{x} \right) + \frac{\hat{x}}{b^2} \left( \sqrt{x^2 + b^2} - \hat{x} \right) \right], & a = 1 
\end{cases}. \]
Notice that having a constant term \( C_0 \) in the inverse marginal utility is equivalent to introducing a shift in the primal utility.
1.4.2 GOBI Utility

Motivated by the SAHARA class, we then turn to another class of utility functions whose inverse marginal utility takes the form

\[ I(\xi) = C_1 \xi^\gamma + C_2 \xi^{2\gamma} + C_0, \]

(1.26)

where \( \gamma, C_1, C_2 \neq 0 \). We name this class the GOBI preferences. We will show in Section 1.6 (Proposition 22) that strict concavity implies that \( C_1 \gamma < 0 \) and \( C_2 \gamma < 0 \). As before, we include a constant term \( C_0 \) to allow for potential shifts.

It is easy to verify that when \( C_0 \neq 0 \) we have three-fund separation. The optimal portfolio of an investor with GOBI utility consists of wealth-dependent investments

\[ \alpha_1(\lambda) = C_1 \lambda^\gamma, \]

\[ \alpha_2(\lambda) = C_2 \lambda^{2\gamma}, \]

in two risky funds

\[ f_1(\rho) = \rho^\gamma, \]

\[ f_2(\rho) = \rho^{2\gamma}, \]

and a fixed position

\[ \alpha_3(\lambda) = C_0 \]

in the risk-free asset

\[ f_3(\rho) = 1. \]

When \( C_0 = 0 \), we have a degenerate case with two-fund separation, in which the two separating funds are given by \( f_1(\rho) \) and \( f_2(\rho) \), respectively.

Similar to the SAHARA class, (1.26) can be viewed as a quadratic equation of \( \xi^\gamma \). Setting \( x = I(\xi) \) and \( u'(x) = \xi \), (1.26) becomes

\[ C_2 (u'(x))^{2\gamma} + C_1 (u'(x))^\gamma - (x - C_0) = 0, \]

(1.27)
which can be solved to obtain \( u'(x) \). Since \( u'(x) > 0 \), (1.27) must have a positive solution. This requires \( C_2(x - C_0) > 0 \), which imposes a constraint on the domain of the utility function. In particular, when \( C_2 < 0 \) (which by strict concavity happens when \( \gamma > 0 \)), the utility function is defined on \( x \in (-\infty, C_0) \). When \( C_2 > 0 \) (which by strict concavity happens when \( \gamma < 0 \)), the utility function is defined on \( x \in (C_0, +\infty) \).

Since \( u'(x) > 0 \), solving (1.27) yields

\[
u'(x) = \left( \frac{G(x) - C_1}{2C_2} \right)^{\frac{1}{\gamma}} \]

where \( G(x) = \sqrt{C_1^2 + 4C_2(x - C_0)} \). Integrate \( u'(x) \) then gives us the primal utility function

\[
u(x) = \begin{cases} \frac{\gamma}{(2C_2)^{\frac{1}{\gamma}}(2\gamma + 1)}(G(x) - C_1)^{\frac{1}{\gamma} + 1} \left( G(x) + \frac{\gamma}{\gamma + 1}C_1 \right), & \gamma \neq -\frac{1}{2} \text{ and } \gamma \neq -1 \\ \frac{1}{3} \left( \frac{2}{C_2} \right)^{\frac{2}{\gamma}}(G(x) - C_1)^{\frac{2}{\gamma}}(G(x) + 2C_1), & \gamma = -\frac{1}{2} \\ G(x) + C_1 \log(G(x) - C_1), & \gamma = -1 \end{cases}
\]

Again, having a constant term \( C_0 \) in the inverse marginal utility is equivalent to introducing a shift in the primal utility.

### 1.5 Money Separation

Money separation is a special case of mutual fund separation, which obtains when we can choose the risk-free asset as one separating fund. Examples of money separation that we have encountered so far include quadratic (Example 5), CARA (Example 7), SAHARA (Section 1.4.1) and GOBI (1.4.2) preferences. In this section, we discuss money separation and its properties in more detail.

#### 1.5.1 Money Separation and Constant Investment Weight

An interesting observation is that for the quadratic, CARA, SAHARA and GOBI preferences, all of which exhibit money separation, the optimal investment strategy always involves assigning a constant weight (dollar amount) to one of the separating funds, regardless of the...
initial wealth level. In particular, for the quadratic, SAHARA and GOBI cases, the constant weight is assigned to the risk-free asset, whereas a CARA investor optimally assigns a constant weight to a risky fund. A natural question is whether this is merely a coincidence or it actually reveals a property of money separation? The following theorem addresses this issue.

**Theorem 19** A separating utility function exhibits money separation if and only if the optimal investment portfolio can be constructed by assigning a constant dollar investment to one separating fund, regardless of the initial wealth.

**Proof of Theorem 19:** By Theorem 9, if a utility function \( u \) exhibits mutual fund separation, then one set of separating funds must take the following forms

\[
f_{k,l}(\rho) = \rho^{\gamma_k} (\log \rho)^l, 
\]

\[
f_{k,1,l}(\rho) = \rho^{\gamma_k} \cos(b_k \log \rho) (\log \rho)^l, 
\]

\[
f_{k,2,l}(\rho) = \rho^{\gamma_k} \sin(b_k \log \rho) (\log \rho)^l.
\]

Other sets of separating funds can be constructed as linear combinations of the original ones. Therefore, \( u \) satisfies mutual fund separation if and only if (1.28), (1.29), (1.30), or any of their linear combinations equals a constant. Due to the cos and sin terms, (1.29), (1.30) or any linear combination with a non-zero weight on (1.29) or (1.30) can never be a constant. As a result, money separation hold only when (1.28) gives rise to a constant term, which happens if and only if \((\gamma_k, b_k) = (0, 0)\).
For a separating preference, the investment weights associated with separating funds (1.28), (1.29) and (1.30) are given by

\[ \alpha_{k,l}(\lambda) = \lambda^{\gamma_k} \sum_{j=l}^{d_{k,1}} C_{k,1,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l}, \]  
\[ \alpha_{k,1,l}(\lambda) = 1_{l \leq d_{k,1}} \lambda^{\gamma_k} \cos (b_k \log \lambda) \sum_{j=l}^{d_{k,1}} C_{k,1,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l}, \]  
\[ \alpha_{k,2,l}(\lambda) = 1_{l \leq d_{k,2}} \lambda^{\gamma_k} \cos (b_k \log \lambda) \sum_{j=l}^{d_{k,2}} C_{k,2,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l}, \]  
\[ \alpha_{k,2,l}(\lambda) = -1_{l \leq d_{k,1}} \lambda^{\gamma_k} \sin (b_k \log \lambda) \sum_{j=l}^{d_{k,1}} C_{k,1,j} \left( \frac{j}{l} \right) (\log \lambda)^{j-l}, \]  

For other sets of separating funds, the corresponding investment weights are linear combinations of (1.31), (1.32) and (1.33). Therefore, one separating fund receives a constant weight if and only if (1.31), (1.32), (1.33) or any of their linear combinations equals a constant. Again, due to the \( \cos \) and \( \sin \) terms, (1.32), (1.33) or any linear combination with a non-zero weight on (1.32) or (1.33) can never be a constant. As a result, a constant weight can only be obtained from (1.31), which happens if and only if \((\gamma_k, b_k) = (0, 0)\).

Hence, the theorem is proved.

According to Theorem 19, for a money separating preference, the optimal investment strategy involves assigning a wealth-independent amount to one of the separating funds. However the theorem is mute on which fund receives the constant weight. As we have seen so far, the fund with the constant weight can be either the risk-free asset (e.g., quadratic, SAHARA, GOBI utility) or a risky fund (e.g., CARA utility).

To better understand this point, suppose money separation holds with \((\gamma_k, b_k) = (0, 0)\) for some \(k\). This corresponds to the case in which (1.14) has a zero root. Let \(d_{k,1} + 1\) denote
the multiplicity of the zero root. Then the separating fund representing the risk-free asset is

\[ f_{k,0}(\rho) = 1, \]

whose associated investment weight is given by

\[ \alpha_{k,0}(\lambda) = \sum_{j=0}^{d_{k,1}} C_{k,1,j} (\log \lambda)^j. \]

On the other hand, this also gives rise to another separating fund

\[ f_{k,d_{k,1}}(\rho) = (\log \rho)^{d_{k,1}}, \]

which is assigned a constant investment weight

\[ \alpha_{k,d_{k,1}}(\lambda) = C_{k,1,d_{k,1}}. \]

It is thus clear that the risk-free asset may not coincide with the fund with a constant investment weight. Indeed, they coincide with each other only when \( d_k = 0 \), i.e., (1.14) has a non-repeated zero root, as in the case of quadratic, SAHARA and GOBI preferences. Whenever \( d_k > 0 \) representing a repeated zero root, it is a risky fund receiving a wealth-independent investment weight. For example, in the case of the CARA utility, we have \( d_k = 1 \) implying that zero is a twice-repeated root of (1.14).

### 1.5.2 Money Separation and Shifts in Utility

In this section, we show that money separation is closely related to shifts in the utility function. Indeed, we can construct money separating preferences from non-money separation by introducing a shift in the primal utility.

Consider a utility function \( u \) with inverse marginal utility \( I \). Define another utility function

\[ \hat{u}(x) = u(x - C_0). \]
One can verify that the associated inverse marginal utility is given by

\[ \hat{I}(\xi) = I(\xi) + C_0. \]

This indicates that introducing a shift to the primal utility is equivalent to adding a constant term in the inverse marginal utility.

If \( u \) exhibits \( K \)-fund separation, i.e.,

\[ I(\lambda \rho) = \sum_{k=1}^{K} \alpha_k(\lambda) f_k(\rho), \]

then we have

\[ \hat{I}(\lambda \rho) = I(\lambda \rho) + C_0 = \sum_{k=1}^{K} \alpha_k(\lambda) f_k(\rho) + C_0. \]

Suppose that \( u \) does not have money separation, meaning that none of the \( f_k(\rho) \)'s can be chosen as the risk-free asset (a constant). Then, adding a constant \( C_0 \) in the inverse marginal utility introduces the risk-free as an additional separating fund. Hence, \( \hat{u} \) exhibits \((K+1)\)-fund money separation.

The above analysis implies that money separation can be constructed by introducing a shift to a non-money separating utility function. To better see this idea, it is useful to revisit the SAHARA and GOBI preferences discussed in Section 1.4. For both classes, the standard version has two-fund separation with two risky separating funds. Once we include an additional constant term to the inverse marginal utility, which is equivalent to introducing a shift to the primal, the risk-free asset is added as a third separating fund.

As another example, we start with the class of one-fund separating preferences, and we try to construct two-fund money separation by adding a shift to the primal utility. Namely, adding a shift term \( C_0 \neq 0 \) in (1.16) yields

\[ \hat{u}(x) = \begin{cases} \frac{\gamma C}{\gamma + 1} \left( \frac{x-C_0}{C} \right)^{\gamma + 1}, & \gamma \neq -1 \text{ and } \gamma \neq 0, \\ \frac{C}{C} \log \left( \frac{x-C_0}{C} \right), & \gamma = -1 \end{cases}, \]  

(1.34)

where \( x \in (C_0, +\infty) \) when \( C > 0 \) and \( x \in (-\infty, C_0) \) when \( C < 0 \). This is effectively the HARA class with the exclusion of the CARA utility. For instance, the shifted power (log)
utility function is obtained by introducing a shift term to the CRRA utility with any \( \gamma < 0 \), whereas the quadratic utility function is obtained by adding a shift to the mirror CRRA utility with \( \gamma = 1 \). It can be easily verified that any utility of the form (1.34) exhibit two-fund money separation, with one separating fund being the risk-free asset and the other being a risky fund \( \rho^\gamma \).

It is worth noticing that money separating preferences obtained this way always assign a constant weight to the risk-free asset. This, as discussed in Section 1.5.1, is only one of the two possible cases of money separation. The other case in which the constant weight is assigned to a risky fund cannot be obtained in this manner.

The above discussion focuses on the case where the original utility function \( u \) does not exhibit money separation. More generally, it is possible that \( u \) itself satisfies money separation. Then, we may have different cases, which are summarized in the following theorem.

**Theorem 20** Suppose that a utility function \( u \) exhibits \( K \)-fund separation, and define \( \hat{u}(x) = u(x - C_0) \).

1. If money separation does not hold for \( u \), then \( \hat{u} \) exhibits \((K + 1)\)-fund money separation.

2. If money separation holds for \( u \) with \( f_1(\rho) = 1 \) and \( \alpha_1(\lambda) \neq -C_0 \), then \( \hat{u} \) exhibits \( K \)-fund money separation.

3. If money separation holds for \( u \) with \( f_1(\rho) = 1 \) and \( \alpha_1(\lambda) = -C_0 \), then \( \hat{u} \) exhibits \((K - 1)\)-fund non-money separation.

**Proof of Theorem 20:** The discussion in the text has already proved case 1. Now we prove the other two cases.

Case 2. Since \( u \) has money separation with \( f_1(\rho) = 1 \), we have

\[
\hat{I}(\lambda \rho) = I(\lambda \rho) + C_0 = (\alpha_1(\lambda) + C_0) + \sum_{k=2}^{K} \alpha_k(\lambda) f_k(\rho).
\] (1.35)
Because, \( \alpha_1(\lambda) \neq -C_0 \), we have \( \alpha_1(\lambda) + C_0 \neq 0 \), and hence \( \hat{u} \) exhibits \( K \)-fund money separation.

Case 3. For utility function \( u \), we again have (1.35). Since \( \alpha_1(\lambda) = -C_0 \), (1.35) reduces to

\[
\hat{I}(\lambda p) = \sum_{k=2}^{K} \alpha_k(\lambda) f_k(p).
\]

Therefore, \( \hat{u} \) exhibits \( (K - 1) \)-fund non-money separation. ■

### 1.6 Strict Concavity

We have so far assumed that all utility functions under consideration are strictly concave. This condition allows us to characterize the set of separating preferences by (1.3) in terms of the inverse marginal utility. However, not all utility functions satisfying (1.3) are strictly concave, but instead certain restrictions need to be imposed on the parameter values. In this section, we study strict concavity of separating preferences and how it can be used to narrow down our separating class.

**Theorem 21** Consider a utility function \( u \) exhibiting mutual fund separation. The following statements are equivalent:

1. The utility function \( u \) is strictly concave, i.e., \( u'' < 0 \).

2. For all \( \xi > 0 \), we have \( I'(\xi) < 0 \).

3. The equation \( I'(\xi) = 0 \) has no positive solution and \( I'(\xi^*) < 0 \) for some \( \xi^* \).

**Proof of Theorem 21:** Since \( I(\xi) = (u')^{-1}(\xi) \), differentiating yields

\[
I'(\xi) = \frac{1}{u''(I(\xi))}.
\]

This implies that strict concavity holds \( (u'' < 0) \) if and only if \( I'(\xi) < 0 \) for all \( \xi > 0 \). Hence, statements 1 and 2 are equivalent.
On the other hand, if \( u \) satisfies mutual fund separation, its inverse marginal utility \( I(\xi) \) takes the form of (1.3). One can easily verify that \( I'(\xi) \) is a continuous function in \( \xi \). It is then immediate that statements 2 and 3 are also equivalent. ■

Theorem 21 suggests that the primal utility function \( u \) is strictly concave if and only if the corresponding inverse marginal utility \( I \) is monotonically decreasing. This allows us to check strict concavity by directing looking at the inverse marginal utility. For instance, the set of one-fund separating preferences (1.15) satisfies strict concavity if and only if

\[
I'(\xi) = C\gamma \xi^{\gamma - 1} < 0,
\]

which is equivalent to

\[
C\gamma < 0.
\]

While the necessary and sufficient condition for strict concavity turns out to be straightforward for one-fund separation, it can become very complicated, if not impossible, when higher-degree separation is taken into account. To see this, consider as an example a \( K \)-fund separating preference whose inverse marginal utility takes the form

\[
I(\xi) = \sum_{k=1}^{K} C_k \xi^{\gamma_k}.
\]

(1.36)

We learn from Theorem 21 that solving strict concavity in this case is equivalent to solving a polynomial equation. Given the intrinsic difficulty in solving a high-degree polynomial equation, it is no wonder that a necessary and sufficient characterization for strict concavity is in most cases impossible and probably needs to be handled numerically.

Despite the intrinsic challenge in providing a necessary and sufficient characterization, below we seek to identify some conditions on the parameter values that are needed to induce strict concavity. While these conditions are not sufficient, they allow us to narrow down the class of separating preferences by ruling out parameter values and forms of the inverse marginal utility that are not permitted.
The next proposition deals with the special case of fund separation, in which the inverse marginal utility is given by (1.36).

**Proposition 22** Suppose that a utility function exhibits $K$-fund separation, and its inverse marginal utility is given by (1.36) with $\gamma_1 < \gamma_2 < \cdots < \gamma_K$ and non-zero $C_1, C_2, \ldots, C_K$. Then, strict concavity implies the following.

1. If $\gamma_1 \neq 0$, then $C_1 \gamma_1 < 0$;
2. If $\gamma_1 = 0$, then $C_2 \gamma_2 < 0$;
3. If $\gamma_K \neq 0$, then $C_K \gamma_K < 0$;
4. If $\gamma_K = 0$, then $C_{K-1} \gamma_{K-1} < 0$.

The following result further rules out forms of the inverse marginal utility that violate strict concavity.

**Proposition 23** If a separating utility function is strictly concave, then its inverse marginal utility $I$ cannot take the following forms:

1. $I(\xi) = P(\log \xi)$, where the polynomial function $P(\cdot)$ is of an even degree $d \geq 2$;
2. $I(\xi) = \xi^\gamma P(\log \xi)$, where $\gamma \neq 0$ and the degree of $P(\cdot)$ is odd;
3. $I(\xi) = \sum_{k=1}^{J} \xi^{\gamma_k} P_{k,1} (\log \xi) \cos (b_k \log \xi) + \sum_{k=1}^{J} \xi^{\gamma_k} P_{k,2} (\log \xi) \sin (b_k \log \xi)$, where $b_k \neq 0$ for all $k$.

Proposition 23 implies among others that if a two-fund separating utility is concave, then its inverse marginal utility can only take two forms:

$$I(\xi) = C_1 \xi^{\gamma_1} + C_2 \xi^{\gamma_2},$$
with $\gamma_1 \neq \gamma_2$, or
\[ I(\xi) = C_1 + C_2 \log \xi. \]

Examples of the first form include the SAHARA and GOBI preferences, whereas the second form corresponds to the CARA utility. The sin and cos terms, however, can never show up in two-fund separation due to their cyclicality.

While preferences with an inverse marginal utility function of the forms listed in Proposition 23 violate strict concavity, they can actually be saved by including additional terms. We illustrate this using a few examples. These examples also show how one can find parameter values such that strict concavity is satisfied. Example 24 below demonstrates that the cos and sin terms (Case 3 of Proposition 23) can be saved by an extra power term of the same order as the existing one.

**Example 24** Consider a three-fund separating utility, whose inverse marginal utility function is given by
\[ I(\xi) = [C_1 \cos (b \log \xi) + C_2 \sin (b \log \xi)] \xi^{\gamma} + C_3 \xi^\gamma. \]

Differentiating yields
\[ I'(\xi) = \xi^{\gamma - 1} [(C_1 \gamma + C_2 b) \cos (b \log \xi) + (C_2 \gamma - C_1 b) \sin (b \log \xi) + C_3 \gamma] \]
\[ = \xi^{\gamma - 1} \left[ \sqrt{(C_1 \gamma + C_2 b)^2 + (C_2 \gamma - C_1 b)^2} \sin (\theta + b \log \xi) + C_3 \gamma \right], \]

where $\sin \theta = \frac{C_1 \gamma + C_2 b}{\sqrt{(C_1 \gamma + C_2 b)^2 + (C_2 \gamma - C_1 b)^2}}$ and $\cos \theta = \frac{C_2 \gamma - C_1 b}{\sqrt{(C_1 \gamma + C_2 b)^2 + (C_2 \gamma - C_1 b)^2}}$.

Since $\xi^{\gamma - 1} > 0$ and $|\sin (\theta + b \log \xi)| \leq 1$, it is clear that $I'(\xi) < 0$ holds for all $\xi > 0$ if and only if
\[ C_3 \gamma < -\sqrt{(C_1 \gamma + C_2 b)^2 + (C_2 \gamma - C_1 b)^2}. \]

Case 2 of Proposition 23 violates strict concavity when the degree of the polynomial function $P(\cdot)$ is odd. The following example shows that the exact same form with an even degree of $P(\cdot)$ can be permitted.
Example 25  Consider a three-fund separating utility, whose inverse marginal utility function is given by

\[ I(\xi) = \xi^\gamma (C_1 + C_2 \log \xi + C_3 (\log \xi)^2), \]

where \( \gamma \neq 0 \). Differentiating yields

\[ I'(\xi) = \xi^{\gamma-1} ((\gamma C_1 + C_2) + (\gamma C_2 + 2C_3) \log \xi + \gamma C_3 (\log \xi)^2). \]

Since \( \xi^{\gamma-1} > 0 \), in order to have \( I'(\xi) < 0 \), we only need \( \forall \xi > 0, \)

\[ (\gamma C_1 + C_2) + (\gamma C_2 + 2C_3) \log \xi + \gamma C_3 (\log \xi)^2 < 0. \]

Notice that the left hand side can be viewed as a quadratic function of \( \log \xi \). As a result, this inequality holds if and only if

\[ \gamma C_3 < 0, \]

and

\[ (\gamma C_2 + 2C_3)^2 - 4\gamma C_3 (\gamma C_1 + C_2) < 0. \]

Therefore, any set of parameter values satisfying the above two conditions would give rise to a strictly concave separating utility function.

Another way to save Case 2 of Proposition 23 involves adding a power term of a different order as the existing one. We illustrate this in the following example.

Example 26  Consider a three-fund separating utility, whose inverse marginal utility function is given by

\[ I(\xi) = \xi^{\gamma_1} (C_1 + C_2 \log \xi) + C_3 \xi^{\gamma_2}, \]

where \( \gamma_1 \neq \gamma_2 \) and \( \gamma_1 \gamma_2 \neq 0 \). Differentiating yields

\[ I'(\xi) = \xi^{\gamma_1-1} ((\gamma_1 C_1 + C_2) + (\gamma_1 C_2 \log \xi) + C_3 \gamma_2 \xi^{\gamma_2-1} \]

\[ = \xi^{\gamma_2-1} [((\gamma_1 C_1 + C_2) \xi^{\gamma_1-\gamma_2} + \gamma_1 C_2 \xi^{\gamma_1-\gamma_2} \log \xi + C_3 \gamma_2]. \]
We will show that when (i) $\gamma_1 < \gamma_2$, $\gamma_1 C_1 + C_2 < 0$, and $\gamma_1 C_2 > 0$; or (ii) when $\gamma_1 > \gamma_2$, $\gamma_1 C_1 + C_2 < 0$, and $\gamma_1 C_2 < 0$, we can always set $C_3 \gamma_2$ low enough such that $I'(\xi) < 0$ holds for all $\xi > 0$. We will show part (i) in detail only. The analysis for (ii) is parallel.

(i) Since $\gamma_1 C_1 + C_2 < 0$, we have $(\gamma_1 C_1 + C_2) \xi^{\gamma_1 - \gamma_2} < 0$ for all $\xi > 0$. Since $\gamma_1 C_2 > 0$, we know $\gamma_1 C_2 \xi^{\gamma_1 - \gamma_2} \log \xi \leq 0$ for all $\xi \in (0, 1]$. Hence, we only need to show that $\gamma_1 C_2 \xi^{\gamma_1 - \gamma_2} \log \xi$ is bounded from above when $\xi > 1$, and then setting $C_3 \gamma_2$ lower than the negative value of this upper bound is enough to guarantee $I'(\xi) < 0$ for all $\xi > 0$.

Since $\gamma_1 < \gamma_2$, we have

$$\lim_{\xi \to +\infty} \xi^{\gamma_1 - \gamma_2} \log \xi = 0.$$ 

This implies that for any $\bar{C} > 0$, there exists $\bar{\xi}$ such that $\forall \xi > \bar{\xi}$,

$$\xi^{\gamma_1 - \gamma_2} \log \xi < \bar{C}.$$ 

On the other hand, for all $\xi \in (1, \bar{\xi}]$, we have

$$\xi^{\gamma_1 - \gamma_2} \log \xi < \log \xi < \log \bar{\xi}.$$ 

Hence, for all $\xi > 1$ we obtain

$$\xi^{\gamma_1 - \gamma_2} \log \xi < \max \{ \bar{C}, \log \bar{\xi} \}.$$ 

Therefore, strict concavity is guaranteed by setting

$$C_3 \gamma_2 < -\gamma_1 C_2 \max \{ \bar{C}, \log \bar{\xi} \}.$$ 

1.7 Conclusion

This paper extends Cass and Stiglitz (1970) and studies the general preference-based $K$-fund separation. We show that one can construct high-degree separating preferences by adding low-degree ones in the inverse marginal utility function. However, this does not allow us to find all utility functions satisfying fund separation, because there are separating funds
involved in high-degree separations that never appear in low-degree separations. While it is generally very difficult to derive the primal utility of a separating preference, we provide two classes of preferences, SAHARA and GOBI, both exhibiting three-fund separation, for which we can write down the primal utility function in closed form.

We then study the special case of money separation, in which the risk-free asset can be chosen as one of the separating funds. We show that money separating holds if and only if we can choose a separating fund whose optimal investment weight is constant and in particular does not depend on the initial wealth. Somewhat surprisingly, the constant weight can be assigned to either the risk-free asset or a risky fund. We also show that money separation is closely related to shifts in the utility. In particular, starting with a non-money separating preference, we can easily construct money separation by introducing a non-zero shift in the utility function.

The characterization of fund separation in this paper provides us with a broad class of preferences, which can be very useful for theoretical modeling and empirical tests in future research. When modeling an economy with fund separation, high-degree separation allows for a larger extent of heterogeneity among investors relatively to the familiar one- and two-fund separating utility.
Chapter 2

Generalized Systematic Risk

This chapter generalizes the concept of “systematic risk” to a broad class of risk measures potentially accounting for high distribution moments, downside risk, rare disasters, as well as other risk attributes. We offer two different approaches. First is an equilibrium framework generalizing the Capital Asset Pricing Model, two-fund separation, and the security market line. Second is an axiomatic approach resulting in a systematic risk measure as the unique solution to a risk allocation problem. Both approaches lead to similar results extending the traditional beta to capture multiple dimensions of risk. The results lend themselves naturally to empirical investigation.

2.1 Introduction

Risk is a complex concept. The definition of risk and its implications have long been the subject of both academic and practical debate. This issue has gained even more prominence during the recent financial crisis, when markets and individual assets were hit by catastrophic events whose ex-ante probabilities were considered negligible. Indeed, these events demonstrate that “risk” accounts for much more than what is measured by the variance of the returns of an asset. High distribution moments, rare disasters, and downside risk are just some of the different aspects that may be of interest when measuring risk.

1This chapter is joint work with Ohad Kadan and Suying Liu.
In this paper we allow “risk” to take a very general form. We then re-visit the classic notion of “systematic risk,” which reflects the contribution of an asset to the risk of a portfolio. Traditional measures of systematic risk focus on a narrow set of risk attributes. In particular, the most well-known and widely used measure of systematic risk is the beta of the asset, which is the slope from regressing the asset returns on portfolio returns (Sharpe (1964), Lintner (1965a,b), and Mossin (1966)). Beta is the contribution of an asset to the risk of the portfolio as measured by the variance of its return. It sets the foundations for all risk-return analysis as part of the Capital Asset Pricing Model (CAPM). However, the traditional beta ignores all aspects of risk other than the variance, such as high distribution moments and rare disasters.

We offer two different approaches to generalizing systematic risk. First we study an equilibrium framework modifying the traditional CAPM to allow for a broad set of risk attributes. The equilibrium approach allows us to extend classic results such as the geometry of efficient portfolios, the two-fund separation theorem, the efficiency of the market portfolio, and the security market line. Second is an axiomatic approach in which we recast the issue as a risk allocation problem. We then specify desirable properties of systematic risk, leading to a unique solution. Both approaches yield similar results, generalizing the traditional beta to reflect a variety of risk attributes.

We begin with a broad definition of what would constitute a measure of risk. We define a risk measure as any mapping from random variables to real numbers. That is, a risk measure is simply a summary statistic that encapsulates the randomness using just one number. The variance (or standard deviation) is obviously the most commonly used risk measure. However, many other risk measures have been proposed and used. For example, high distribution moments can account for skewness and tail risk, downside risk accounts for the variation in losses, and value at risk is a popular measure of disaster risk. Recently, Aumann and Serrano (2008) and Foster and Hart (2009) offered two appealing risk measures
that account for all distribution moments and for disaster risk.\footnote{See Hart (2011) for a unified treatment of these two measures and Kadan and Liu (2014) for an analysis of the moment properties of these measures.} All of these measures fall under our wide umbrella of risk measures. Moreover, any linear combination of risk measures is itself a risk measure. Thus, one can easily create measures of risk that account for a number of dimensions of riskiness, assigning the required weight to each dimension.

Our first analysis generalizes the classic CAPM to allow for a broad set of risk measures. The idea is simple. In the classic CAPM setting investors are assumed to have mean-variance preferences. That is, their utility is increasing in the expected payoff and decreasing in the variance of their payoffs. In our generalized setting we assume that investors have mean-risk preferences, where the term “risk” stands for a host of potential risk measures. We provide mild sufficient conditions under which these preferences are locally consistent with expected utility in the sense of Machina (1982).

We consider an exchange economy with a finite number of risky assets, one risk-free asset, and a finite number of investors with mean-risk preferences. As usual, in equilibrium each investor chooses a portfolio of assets from the set of efficient portfolios, minimizing risk for a given expected return. However, due to the generality of the risk measure, the geometry of this set is more complicated than in the case where risk is measured by the variance. Nevertheless, we establish sufficient conditions on the risk measure under which the solution to each investor’s problem satisfies Tobin’s (1958) two-fund separation property. That is, each investor’s optimal portfolio of assets can be presented as a linear combination of the risk-free asset and a unique portfolio of risky assets. We demonstrate that a variety of risk measures satisfy these sufficient conditions, where the variance is just one special case. A consequence of two-fund separation is that the equilibrium market portfolio lies on the efficient frontier. Using this we establish a generalization of the classic security market line (SML) to a large class of risk measures. Specifically, in equilibrium, the expected return of
each risky asset \( i \) satisfies

\[
E(\tilde{z}_i) = r_f + B^R_i (E(\tilde{z}^M) - r_f),
\]

where \( \tilde{z}_i \) is the risky return of asset \( i \), \( \tilde{z}^M \) is the risky return of the market portfolio, \( r_f \) is the risk-free rate, and \( B^R_i \) is the systematic risk of asset \( i \) given the risk measure \( R \). Moreover, \( B^R_i \) is given in closed form as the marginal contribution of asset \( i \) to the market risk scaled by the weighted average of such marginal contributions across all assets in the economy.

In the special case in which \( R \) is the variance, \( B^R_i \) coincides with the traditional beta. More generally, we show that our equilibrium setting is versatile enough to allow for a variety of risk attributes such as tail risk, downside risk, and rare disasters, among others. Our setting can also readily account for risk measures that reflect several of these risk attributes, assigning different weights to each of them. We illustrate that in all these cases one can readily derive closed form solutions for the generalized betas. Typically, these betas reflect the covariation of the return of asset \( i \) with some function of the market return. In general, these betas do not take the form of a regression coefficient. Nevertheless, they can be estimated directly from return data and applied in a standard Fama-MacBeth (1973) cross-sectional analysis.

The CAPM equilibrium can be thought of as a special case of the more general problem of risk allocation. Indeed, the CAPM beta measures the contribution of one asset to the risk of the market portfolio. Many other problems of considerable economic import require estimating the contribution of one asset to some specific portfolio of assets (not necessarily the market portfolio). For example, the government is constantly interested in the contribution of particular banks and other financial institutions to the total market risk (known as systemic risk). Banks and other financial institutions may also find it useful to calculate the contribution of different assets on their balance sheet to the total risk of the institution, so that each asset or business unit could be “taxed” appropriately. All of these problems are essentially risk allocation problems in which total risk should be allocated among the
constituents of a portfolio. We broaden the term “systematic risk” to designate solutions to such problems. That is, a systematic risk measure is a vector specifying the portion of the total portfolio risk allocated to each asset in the portfolio. The literature has not yet presented a general solution to this problem for a broad set of risk measures and for arbitrary portfolios.

In the second part of this paper we tackle this problem from an axiomatic point of view. We state desirable properties of systematic risk measures, which we call axioms, and we look for solutions that satisfy these properties. Unlike in the equilibrium setting, here we do not need to impose almost any structure on the risk measure. Moreover, the portfolio of assets is arbitrary and is not limited to the market portfolio.

We state four economically plausible axioms that systematic risk measures are expected to satisfy. We then show that these four axioms imply a unique systematic risk measure which applies to all risk allocation problems. This measure is given by a scaled version of the Aumann-Shapley (1974) diagonal formula, which was developed as a solution concept in cooperative game theory. Essentially, this formula calculates for each asset the average of its marginal contributions to portfolios along a diagonal starting from the origin and ending at the portfolio of interest. In the common case in which the risk measure is homogeneous of some degree, the solution becomes very simple, and it coincides with the generalized beta obtained in the equilibrium setting above. In particular, it assigns to each asset its marginal contribution to total portfolio risk scaled by the weighted average of marginal contributions of all assets. Our proof of the axiomatization result relies on a mapping between risk allocation problems and cost allocation problems studied in Billera and Heath (1982).

The paper proceeds as follows. Section 2.2 discusses the related literature. In Section 2.3 we define the notion of risk measures. Section 2.4 studies the equilibrium setup and offers a generalization of the CAPM. In Section 2.5 we present the axiomatic approach. Section 2.6 concludes. Proofs of the main theorems are in Appendix I, proofs of propositions and
other derivations are in Appendix II, and other technical results are provided in an Internet Appendix.

2.2 Related Literature

Our paper contributes to several strands of the literature. First, the paper adds to the growing literature on high distribution moments, disaster risk, and other risk attributes, as well as their effect on prices. Rubinstein (1973), Kraus and Litzenberger (1976), Jean (1971), Kane (1982), and Harvey and Siddique (2000) argue that investors favor right-skewness of returns, and demonstrate the cross-sectional implications of this effect. In addition, Barro (2006, 2009), Gabaix (2008, 2012), Gourio (2012), Chen, Joslin, and Tran (2012), and Wachter (2013) study the aversion of investors to tail risk and rare disasters. Ang, Chen, and Xing (2006) and Lettau, Maggiori, and Weber (2013) show that downside risk is a good explanatory variable for returns in both equity and currency markets. Our paper adds to this literature by outlining a general approach to measuring systematic risk that can capture the contribution of an asset to a range of risk dimensions such as high distribution moments, downside risk, and rare disasters. Our framework is flexible and can account for either one risk aspect or a combination of several of them.

Our equilibrium approach follows a reduced form, where preferences are described through the aversion to broadly defined risk. Our main results are derived without the need to specify an exact form of the utility function. This is different from the approach in consumption-based asset pricing models (e.g., Bansal and Yaron (2004) and Campbell and Cochrane (1999)). These models rely on the specification of a particular utility function (such as Epstein and Zin (1989) preferences or preferences reflecting past habits). One advantage of our approach is that it provides a parsimonious and simple one-factor model that can capture different aspects of risk in a manner that may lend itself naturally to empirical investigation. Another feature of our approach is that, unlike consumption-based models, it resorts
to prices directly. Thus, one can potentially test our model without relying on consumption data.

The paper also adds to the growing literature on risk measurement. This literature dates back to Hadar and Russell (1969), Hanoch and Levy (1969), and Rothschild and Stiglitz (1970) who extend the notion of riskiness beyond the “variance” framework by introducing stochastic dominance rules. Artzner, Delbaen, Eber, and Heath (1999) specify desirable properties of coherent risk measures, and Rockafellar, Uryasev, and Zabarankin (2006a) introduce the notion of generalized deviation measures. More recently, Aumann and Serrano (2008), Foster and Hart (2009, 2013), and Hart (2011) have come up with appealing risk measures that generalize conventional stochastic dominance rules. Notably, all the risk measures discussed in this literature are idiosyncratic in nature. Our paper contributes to this literature by specifying a method to calculate the systematic risk of an asset for any given risk measure. This in turn allows us to study the fundamental risk-return trade-off associated with a risk measure.

Our paper also adds to the recent literature on systemic risk, which is the risk that the entire economic system collapses. Adrian and Brunnermeier (2011) define the $\Delta CoVaR$ measure as the difference between the value at risk of the banking system conditional on the distress of a particular bank and the value at risk of the banking system given that the bank is solvent. Acharya, Pedersen, Philippon, and Richardson (2010) propose the Systemic Expected Shortfall measure, which estimates the exposure of a particular bank in terms of under-capitalization to a systemic crisis. Huang, Zhou, and Zhu (2009) measure the systemic risk of a financial institution by the price of insurance against financial distress. Our paper takes a general approach to the problem of estimating the contribution of one asset to the risk of a portfolio of assets. We provide an easy-to-calculate and intuitive measure that applies to a wide variety of risk measures, as well as in an array of contexts.

Our paper also contributes to the literature studying conditions for two-fund separation.
The idea of two-fund separation was introduced by Tobin (1958). Since then the literature discussed different sufficient conditions in terms of either agents’ utility (e.g., Cass and Stiglitz (1970) and Dybvig and Liu (2015)) or the distribution of returns (e.g., Ross (1978)). Here we take a somewhat different approach, as we specify sufficient conditions for two-fund separation in terms of properties of the risk measure. This approach is similar to the one taken in Rockafellar, Uryasev, and Zabarankin (2006b), who consider general deviation measures. Our restrictions on risk measures are weaker than theirs as we do not require homogeneity. All of these papers consider two-fund separation only and do not provide any generalization of the notion of systematic risk, which is the focus of our paper.

Finally, the paper adds to an extensive list of studies applying the Aumann-Shapley solution concept in different contexts, e.g., Billera, Heath, and Raanan (1978), Samet, Tauman, and Zang (1984), Powers (2007), and Billera, Heath, and Verrecchia (1981). Tarashev, Borio, and Tsatsaronis (2010) use the Shapley value (Shapley (1953), a discrete version of the Aumann-Shapley solution concept) to measure systemic risk. Our paper offers theoretical foundations for their practical approach.

2.3 Risk Measures and Their Properties

Let \((\Omega, \mathcal{F}, P)\) be a probability space, where \(\Omega\) is the state space, \(\mathcal{F}\) is the \(\sigma\)-algebra of events, and \(P(\cdot)\) is a probability measure. As usual, a random variable is a measurable function from \(\Omega\) to the reals. In the context of investments, we typically consider random variables representing the payoffs or the returns of financial assets. Thus, we often refer to random variables as “investments” or “random returns.” We generically denote random variables by \(\tilde{z}\), which is a shorthanded notation for \(\tilde{z}(\omega), \forall \omega \in \Omega\). We restrict attention to random variables for which all moments exist. We denote the expected value of \(\tilde{z}\) by \(E(\tilde{z})\) and its \(k^{th}\) central moment by \(m_k(\tilde{z}) = E((\tilde{z} - E(\tilde{z}))^k)\), where \(k \geq 2\).

A risk measure is simply a function that assigns to each random variable a single number
summarizing its riskiness. Formally,

**Definition 27** A risk measure is a function mapping random variables to the reals.\(^3\)

We generically denote risk measures by \( R(\cdot) \). The simplest and most commonly used risk measure is the variance \( (R(\tilde{z}) = m_2(\tilde{z})) \). However, many other risk measures have been proposed in the literature, capturing higher distribution moments and other risk attributes. A risk measure \( R(\cdot) \) is homogeneous of degree \( k \), if for any random return \( \tilde{z} \) and positive number \( \lambda > 0 \),

\[
R(\lambda \tilde{z}) = \lambda^k R(\tilde{z}).
\]

A weaker requirement, which is sufficient for most of our results, is that the risk ranking between two investments does not depend on scaling. We say that \( R(\cdot) \) is scaling independent if for all \( \lambda > 0 \) and any two random returns \( \tilde{z}_1 \) and \( \tilde{z}_2 \), \( R(\tilde{z}_1) > R(\tilde{z}_2) \) implies \( R(\lambda \tilde{z}_1) > R(\lambda \tilde{z}_2) \).

The next property of risk measures which will become useful is convexity. Formally, we say that a risk measure \( R(\cdot) \) is convex if for any two random returns \( \tilde{z}_1 \) and \( \tilde{z}_2 \), and for any \( \lambda \in (0, 1) \), we have

\[
R(\lambda \tilde{z}_1 + (1 - \lambda) \tilde{z}_2) \leq \lambda R(\tilde{z}_1) + (1 - \lambda) R(\tilde{z}_2),
\]

with equality holding only when \( \tilde{z}_1 = \tilde{z}_2 \) with probability 1. Notice that \( \lambda \tilde{z}_1 + (1 - \lambda) \tilde{z}_2 \) can be considered as the return of a portfolio that assigns weights \( \lambda \) and \( 1 - \lambda \) to \( \tilde{z}_1 \) and \( \tilde{z}_2 \), respectively. Then the convexity condition says that the risk of the portfolio should not be higher than the corresponding weighted average risk of the constituent investments. Thus, convexity of a risk measure captures the idea that diversifying among two investments lowers the total risk.

---

\(^3\)Strictly speaking, a risk measure is also a function of the underlying probability measure \( P \). However, in our analysis we fix \( P \) throughout, and yet consider different random variables. Thus, it is convenient to think about risk measures as functions of the random variables, viewing the probability measure as a fixed parameter.
Next we would like to formalize a property dealing with the type of assets that are risk-free. We say that a risk measure \( R(\tilde{z}) \) has the risk-free property, if (i) \( R(\tilde{z}) \geq 0 \) for all \( \tilde{z} \); (ii) \( R(\tilde{z}) = 0 \) if and only if \( P(\{\tilde{z} = c\}) = 1 \) for some constant \( c \); and (iii) \( R(\tilde{z}_1 + \tilde{z}_2) = R(\tilde{z}_1) \) whenever \( R(\tilde{z}_2) = 0 \). Namely, \( R \) has the risk-free property if the only assets with zero risk are those that pay a constant amount with probability 1, if all other assets have strictly positive risk, and if adding a zero-risk asset does not change risk. In what follows, we often refer to assets satisfying \( R(\tilde{z}) = 0 \) as risk-free.

Risk measures can be applied to individual random variables or to portfolios of random variables. Formally, assume there are \( n \) random variables represented by the vector \( \tilde{z} = (\tilde{z}_1, ..., \tilde{z}_n) \). A portfolio is a vector \( \mathbf{x} = (x_1, ..., x_n) \in \mathbb{R}^n \), where \( x_i \) is the dollar amount invested in \( \tilde{z}_i \).\(^4\) Then, \( \mathbf{x} \cdot \tilde{z} = \sum_{i=1}^{n} x_i \tilde{z}_i \) is itself a random variable. We then say that the risk of portfolio \( \mathbf{x} \) is simply \( R(\mathbf{x} \cdot \tilde{z}) \). When the vector of random variables is unambiguous, we often abuse notation and denote \( R(\mathbf{x}) \) as a shorthand for \( R(\mathbf{x} \cdot \tilde{z}) \). We say that a risk measure is smooth if for any vector of random returns \( \tilde{z} = (\tilde{z}_1, ..., \tilde{z}_n) \) and for all portfolios \( \mathbf{x} = (x_1, ..., x_n) \) we have that \( R(\mathbf{x} \cdot \tilde{z}) \) is continuously differentiable in \( x_i \) for \( i = 1, ..., n \). We then write \( R_i(\mathbf{x}) \) (or \( R_i(\mathbf{x} \cdot \tilde{z}) \)) for the partial derivative of \( R(\cdot) \) with respect to the amount invested in the \( i^{th} \) asset evaluated at \( \mathbf{x} \).\(^5\)

When restricting attention to homogeneous risk measures, the properties discussed above are maintained when taking convex combinations of different risk measures with the same degree of homogeneity. Thus, we can easily create new risk measures satisfying these properties from existing homogeneous risk measures. That is, let \( s \) be a positive integer, let \( R^1(\cdot), ..., R^s(\cdot) \) be risk measures, and choose \( \theta = (\theta_1, ..., \theta_s) \) with \( \theta_j > 0 \ \forall j \). We can then

\(^4\)Throughout the paper we denote vectors using bold notation (for both numbers and random variables).

\(^5\)Note that we use subscripts to denote both elements of a vector and partial derivatives. For example, \( x_i \) is the \( i^{th} \) element of the vector \( \mathbf{x} \) while \( R_i(\cdot) \) is the partial derivative of \( R(\cdot) \) considered as a function of portfolio amounts. This notation does not result in any ambiguity since the only case in which the subscript should be interpreted as a partial derivative is when applied to the risk measure considered as a function of portfolio amounts.
define a new risk measure by
\[ R^\theta (\tilde{z}) = \sum_{j=1}^{s} \theta_j R^j (\tilde{z}), \]
where \( \theta_j \) reflects the weight assigned to the risk attribute measured by \( R^j \). We then have

**Lemma 28** Assume that each \( R^j \) is homogeneous of degree \( k \), convex, smooth, and satisfies the risk-free property. Then, \( R^\theta \) also satisfies all of these properties.

### 2.3.1 Examples of Risk Measures

Below we present some popular examples of risk measures and discuss their properties. Each of these examples highlights a different aspect of risk that may be of interest in applications. These examples will be crucial later in the paper when we demonstrate how to apply our main results.

**Example 29** Even central moments and normalized even central moments. For any integer \( k \geq 2 \) even, the central moment \( R(\tilde{z}) = m_k (\tilde{z}) \) is a risk measure which is homogeneous of degree \( k \), convex, smooth and satisfies the risk-free property. The normalized central moment \( w_k (\tilde{z}) = (m_k (\tilde{z}))^{\frac{1}{k}} \) is also a risk measure. For example, when \( k = 2 \), \( w_k (\tilde{z}) \) is the standard deviation of \( \tilde{z} \). Normalized central moments satisfy all of the above properties as well (with homogeneity of degree 1). Indeed, homogeneity, smoothness, and the risk-free property are trivial in these cases. Convexity stems from the following result, which shows that \( w_k (\tilde{z}) \) is convex, and thus \( m_k (\tilde{z}) \) is a fortiori convex.

**Proposition 30** For all \( k \geq 2 \) even, \( R(\tilde{z}) = w_k (\tilde{z}) \) is a convex risk measure.

**Example 31** Odd central moments and normalized odd central moments. For any integer \( k \geq 3 \) odd, the central moment \( R(\tilde{z}) = m_k (\tilde{z}) \) is a risk measure which is homogeneous of degree \( k \) and smooth. Similarly, the normalized odd moments \( w_k (\tilde{z}) \) are homogeneous of
degree 1 and smooth. In contrast to the even central moments, neither convexity nor the risk-free property holds in this case.\textsuperscript{6}

Evidently, the feature of odd central moments that prevents them from satisfying convexity and the risk-free property is that they admit negative values. A natural way to fix this is to focus on just one side of the distribution. The next example follows this idea, allowing one to readily incorporate aspects of odd central moments (such as skewness) into risk measures that also satisfy convexity and the risk-free property.

Example 32 Downside risk. When considering risk, investors sometimes restrict attention to the lower outcomes of the distribution, in particular to those which fall below the mean. Such an approach is called downside risk. Formally, for any integer $k \geq 2$, define the downside risk of order $k$ of $\tilde{z}$ as

$$DR_k(\tilde{z}) = (-1)^k \left( \mathbb{E} \left( [\tilde{z} - \mathbb{E}(\tilde{z})]^{-k} \right) \right)^{\frac{1}{k}},$$

where $[t]^\ominus = \min(t, 0)$ for $t \in \mathbb{R}$. Often, this measure is used in the special case of $k = 2$. More generally, for any $k \geq 2$, $DR_k(\tilde{z})$ is a risk measure which is homogeneous of degree 1, smooth, and satisfies the risk-free property. The next proposition establishes that this risk measure is also convex.

Proposition 33 For any $k \geq 2$, $DR_k(\tilde{z})$ is a convex risk measure.

Example 34 Value at risk. A risk measure widely used in financial risk management is the Value at Risk (VaR), designed to capture the risk associated with rare disasters. VaR

\textsuperscript{6}To see the former, consider the simple example of two random returns, $\tilde{z}_1$ and $\tilde{z}_2$, which are independent and have negative third central moments $m_3(\cdot)$. Then, by independence and the homogeneity of central moments,

$$m_3\left(\frac{1}{2} \tilde{z}_1 + \frac{1}{2} \tilde{z}_2\right) = \left(\frac{1}{2}\right)^3 m_3(\tilde{z}_1) + \left(\frac{1}{2}\right)^3 m_3(\tilde{z}_2) > \frac{1}{2} m_3(\tilde{z}_1) + \frac{1}{2} m_3(\tilde{z}_2),$$

since $m_3(\tilde{z}_1) + m_3(\tilde{z}_2) < 0$. To see the latter, note that the third moment can be negative, violating the risk-free property.
measures the amount of loss not exceeded with a certain confidence level. Formally, given some confidence level \( \delta \in (0, 1) \), for any random return \( \tilde{z} \), the VaR measure is defined as the negative of the \( \delta \)-quantile of \( \tilde{z} \), i.e.,

\[
\text{VaR}_\delta(\tilde{z}) = -\inf \left\{ z \in \mathbb{R} : F(z) \geq \delta \right\},
\]

(2.1)

where \( F(\cdot) \) is the cumulative distribution function of \( \tilde{z} \). Notice that we include the minus sign to reflect the fact that a larger loss indicates higher risk. If \( \tilde{z} \) is continuously distributed with a density function \( f(\cdot) \), then (2.1) is implicitly determined by

\[
\int_{-\infty}^{-\text{VaR}_\delta(\tilde{z})} f(z) \, dz = \delta.
\]

(2.2)

This risk measure is homogeneous of degree 1 and smooth. For any risk-free return \( \tilde{z} \) with \( P(\{ \tilde{z} = c \}) = 1 \), we have \( \text{VaR}_\delta(\tilde{z}) = -c \), implying that the VaR of risk-free assets depends on the risk-free return. Hence, the risk-free property is not satisfied. In addition, it is not hard to find examples where convexity is violated for the VaR measure.

**Example 35 Expected shortfall and demeaned expected shortfall.** These measures capture the average loss from disastrous events, defined as those involving a loss larger than the VaR. Formally, assume that \( \tilde{z} \) can be represented by a density \( f(\cdot) \). Given some confidence level \( \delta \in (0, 1) \), for any random return \( \tilde{z} \) the Expected Shortfall (ES) is the negative of the conditional expected value of \( \tilde{z} \) below the \( \delta \)-quantile. That is,

\[
\text{ES}_\delta(\tilde{z}) = -\frac{1}{\delta} \int_{-\infty}^{-\text{VaR}_\delta(\tilde{z})} zf(z) \, dz.
\]

(2.3)

Additionally, when \( \tilde{z} = c \) (a constant) with probability 1 we set \( \text{ES}_\delta(\tilde{z}) = -c \). Similar to VaR, ES is homogeneous of degree 1 and is smooth, but it does not satisfy the risk-free property.

---

7Formally, smoothness follows if a joint density of the random returns in a portfolio exists. This is shown using an application of the implicit function theorem to (2.2). We omit the proof for brevity.

8Expected shortfall is sometimes termed “conditional VaR.”
To ensure that the risk-free property is satisfied it is useful to consider the demeaned version of ES defined as

\[ \text{DES}_\delta(\tilde{z}) = -\frac{1}{\delta} \int_{-\infty}^{-\text{VaR}_\delta(\tilde{z})} (z - \text{E}(\tilde{z})) f(z) \, dz = \text{ES}_\delta(\tilde{z}) + \text{E}(\tilde{z}). \]

Similar to ES, DES also captures the expected loss from a rare disaster. This risk measure is also homogeneous of degree 1, smooth, and it satisfies the risk-free property.\(^9\) Moreover, unlike VaR, both ES and DES satisfy the convexity property as shown in the next proposition.

**Proposition 36** For any \(\delta \in (0, 1)\), \(R(\tilde{z}) = \text{ES}_\delta(\tilde{z})\) and \(R(\tilde{z}) = \text{DES}_\delta(\tilde{z})\) are convex.

**Example 37 The Aumann-Serrano and Foster-Hart risk measures.** Two measures of riskiness have been proposed by Aumann and Serrano (2008, hereafter AS) and Foster and Hart (2009, hereafter FH). These measures generalize the notion of second-order stochastic dominance (SOSD). The AS measure \(R_{\text{AS}}(\tilde{z})\) is given by the unique positive solution to the implicit equation

\[ \text{E} \left[ \exp \left( -\frac{\tilde{z}}{R_{\text{AS}}(\tilde{z})} \right) \right] = 1. \tag{2.4} \]

The FH measure \(R_{\text{FH}}(\tilde{z})\) is given by the unique positive solution to the implicit equation

\[ \text{E} \left[ \log \left( 1 + \frac{\tilde{z}}{R_{\text{FH}}(\tilde{z})} \right) \right] = 0. \tag{2.5} \]

Both these measures are homogeneous of degree 1 and smooth. These two risk measures also satisfy the convexity property.\(^10\) By contrast, these two measures do not satisfy the risk-free property.\(^11\)

All of the risk measures discussed thus far are homogeneous of some degree. However, most of our results do not require homogeneity. The next set of examples illustrates how non-homogeneous risk measures satisfying all of the other properties can be constructed.

---

\(^9\)The risk-free property follows since \(\text{ES}_\delta(\tilde{z}) + \text{E}(\tilde{z}) \geq 0\) for all significance level \(0 < \delta < 1\) with equality if and only if \(\tilde{z}\) is a constant with probability 1.

\(^10\)This follows since these risk measures are subadditive and homogeneous of degree 1.

\(^11\)To see this, note that for any constant \(c > 0\), \(\tilde{z} + c\) first-order stochastically dominates \(\tilde{z}\). Since \(R_{\text{AS}}\) is consistent with first-order stochastic dominance, we have that \(R_{\text{AS}}(\tilde{z} + c) < R_{\text{AS}}(\tilde{z})\). A similar argument applies to \(R_{\text{FH}}\). Also, technically, these two risk measures are not defined for risk-free assets.
Example 38 Let $R$ be a risk measure which is homogeneous of some degree $k$, convex, smooth, and satisfies the risk-free property, and let $h : [0, \infty) \rightarrow \mathbb{R}$ be a strictly increasing, strictly convex, and continuously differentiable function. Define a new risk measure $\hat{R}$ by

$$\hat{R}(\tilde{z}) = h(R(\tilde{z})) - h(0).$$

It is straightforward to verify that $\hat{R}$ is scaling independent, convex, smooth, and satisfies the risk-free property. However, $\hat{R}$ may fail to be homogeneous of any degree. For a concrete example, set $h(x) = e^x$, and let $R(\tilde{z}) = m_k(\tilde{z})$ for $k$ even. Then, $\hat{R}(\tilde{z}) = e^{R(\tilde{z})} - 1$ is not homogeneous of any degree and yet it satisfies all of the other properties.

2.4 Systematic Risk in an Equilibrium Setting

Traditionally, systematic risk is derived from the CAPM equilibrium setting. We will now present a generalized version of this model. We first outline the setup of the model. We then study the geometry of solutions, and present a two-fund separation result implying the efficiency of the market portfolio. Finally, we derive a variant of the security market line, enabling us to obtain a generalization of the traditional beta as a measure of systematic risk.

2.4.1 Model Setup

Investors, Assets, and Timing. Assume a market with $n + 1$ assets $\{0, \ldots, n\}$. Assets $1, \ldots, n$ are risky and pay a random amount denoted by $(\tilde{y}_1, \ldots, \tilde{y}_n)$. Asset 0 is risk-free, paying an amount $\tilde{y}_0$ which is equal to some constant $y_0 \neq 0$ with probability 1. Denote $\tilde{\mathbf{y}} = (\tilde{y}_0, \ldots, \tilde{y}_n)$. There are $\ell$ investors in the market, all of whom agree on the parameters of the model. The choice set of each investor is $\mathbb{R}^{n+1}$, where $\zeta^j \in \mathbb{R}^{n+1}$ represents the number of shares investor $j$ chooses in each asset $i = 0, \ldots, n$, i.e., $\zeta^j$ is a bundle of assets. Negative numbers represent short sales, and we impose no short-sale constraints. The initial endowment of investor $j$ is a non-zero $\mathbf{e}^j \in \mathbb{R}_+^{n+1}$. We assume that $\sum_{j=1}^{\ell} e^j_i > 0$ for $i = 1, \ldots, n$. 
That is, risky assets are in positive net supply. An allocation is an $\ell$-tuple $A = (\zeta^1, ..., \zeta^\ell)$ consisting of a bundle $\zeta^j \in \mathbb{R}^{n+1}$ for each investor. An allocation $A$ is attainable if $\sum_{j=1}^\ell \zeta^j = \sum_{j=1}^n e^j$, that is, if it clears the market. A price system is a vector $p = (p_0, ..., p_n)$ specifying a price for each asset. Similar to the standard CAPM setting, there are two dates. At Date 0, investors trade with each other and prices are set. At Date 1, all random variables are realized.

**Risk and Preferences.** The traditional approach features investors with mean-variance preferences, i.e., they prefer higher mean and lower variance of investments. Instead, we assume that investors have mean-risk preferences. Formally, fix a risk measure $R(\cdot)$. The utility that investor $j = 1, ..., \ell$ assigns to a bundle $\zeta \in \mathbb{R}^{n+1}$ is given by

$$U^j(\zeta) = V^j \left( E(\zeta \cdot \bar{y}), R(\zeta \cdot \bar{y}) \right),$$

where $V^j$ is continuous, strictly increasing in its first argument (expected payoff) and strictly decreasing in its second argument (risk of payoff), and quasi-concave. 

Note that $U^j(\zeta)$ cannot be in general supported as a von Neumann-Morgenstern utility. Nevertheless, in the Internet Appendix we show that if $V^j$ is differentiable and if the risk measure is a differentiable function of a finite number of moments, then $U^j(\zeta)$ is a local expected utility function in the sense of Machina (1982). Namely, comparisons of “close by” investments are well approximated by expected utility. These conditions apply to a wide range of risk measures representing high distribution moments.

An implication of quasi-concavity of $V^j$ is that when plotted in the mean-risk space, the upper contour of each indifference curve is convex. Similar to the standard mean-variance case, we will assume that a risk-free asset cannot be created synthetically from risky assets. That is, there is no redundant risky asset: for any $\zeta = (\zeta_0, \zeta_1, ..., \zeta_n) \in \mathbb{R}^{n+1}$ we have $R(\zeta \cdot \bar{y}) \neq 0$ unless $(\zeta_1, ..., \zeta_n) = (0, ..., 0).^{12}$

---

$^{12}$In the standard mean-variance case this condition corresponds to the variance-covariance matrix of risky assets being positive-definite.
Equilibrium. An equilibrium is a pair \((p, A)\) where \(p \neq 0\) is a price system and \(A = (\zeta^1, \ldots, \zeta^\ell)\) is an attainable allocation, such that for each \(j \in \{1, \ldots, \ell\}\), \(p \cdot \zeta^j = p \cdot e^j\), and if \(\zeta \in \mathbb{R}^{n+1}\) and \(U^j(\zeta) > U^j(\zeta^j)\) then \(p \cdot \zeta > p \cdot e^j\). In words, an equilibrium is a price system and an allocation that clear the market such that each investor optimizes subject to her budget constraint. The next theorem specifies conditions under which an equilibrium exists.

Theorem 39 Suppose that \(R(\cdot)\) is convex, smooth, and satisfies the risk-free property. Then, an equilibrium exists.

It is well known that the CAPM setting can yield negative or zero prices (see for example Nielsen (1992)). The reason for this is that preferences are not necessarily monotone in the number of shares. Specifically, the expected payoff to an investor’s bundle increases as she holds more shares of a risky asset, but so does the risk. It may well be that at some point, the additional expected payoff gained from adding more shares to the bundle is not sufficient to compensate for the increase in risk. If the equilibrium happens to fall in such a region then the asset becomes undesirable, rendering a negative price. For our following results we will need that prices are positive for all assets. The literature has suggested several ways to guarantee such an outcome. In the Internet Appendix we provide one sufficient condition which follows Nielsen (1992). Other (and possibly weaker) sufficient conditions may be obtained, but are beyond the scope of this paper.

From now on we will only consider equilibria with positive prices. Given positivity of prices, naturally, each equilibrium induces a vector of random returns \(\tilde{z}_i = \frac{\tilde{y}_i}{p_i}\), and we can talk about the expected returns and the risk of the returns in equilibrium, as in the usual CAPM setting. In particular, the equilibrium return from the risk-free asset \(\tilde{z}_0\) is equal to some constant \(r_f\) with probability 1. We now study these returns.
2.4.2 A Generalized CAPM

Geometry of Efficient Portfolios

Let \((p, A)\) be an equilibrium. The equilibrium allocation \((\zeta^1, ..., \zeta^\ell)\) naturally induces a portfolio for each investor \(j\) given by \(x^j = (x^j_0, ..., x^j_n)\), where \(x^j_i = p_i \zeta^j_i\) is the amount invested in asset \(i\), and where the vector of portfolio weights of investor \(j\) is denoted by \(\alpha^j\) and given by \(\alpha^j_i = \frac{x^j_i}{\sum_{h=0}^{n} x^j_h}\). Let

\[
\mu^j = \sum_{i=0}^{n} \alpha^j_i E(\tilde{z}_i)
\]

be the expected return obtained by investor \(j\) in equilibrium. The next theorem shows that the standard procedure of “minimizing risk for a given expected return” applies to the equilibrium setting. It relies on the scaling independence and convexity of the risk measure.

**Theorem 40** Suppose that \(R(\cdot)\) is scaling independent and convex. Then, in an equilibrium with positive prices, for all investors \(j \in \{1, ..., \ell\}\), \(\alpha^j\) is the unique solution to

\[
\min_{\alpha \in \mathbb{R}^{n+1}} R(\alpha \cdot \tilde{z})
\]

s.t.

\[
\sum_{i=0}^{n} \alpha_i E(\tilde{z}_i) = \mu^j.
\]

\[
\sum_{i=0}^{n} \alpha_i = 1.
\]

Given this, we can now discuss the geometry of portfolios in the \(\mu-R\) plane where the horizontal axis is the risk of the return of a portfolio \((R)\) and the vertical axis is the expected return \((\mu)\). The locus of portfolios minimizing risk for any given expected return is the boundary of the portfolio opportunity set. This set is convex in the \(\mu-R\) plane whenever \(R(\cdot)\) is a convex risk measure. This follows simply because the expectation operator is linear, implying that the line connecting any two portfolios in the \(\mu-R\) plane lies to the right of the set of portfolios representing convex combinations of these two portfolios. Figure 2.1
Figure 2.1: Portfolio Opportunity Set and Efficient Frontier

illustrates two curves. The blue curve depicts the opportunity set of risky assets only. The red curve depicts portfolios minimizing risk for a given expected return, corresponding to Program (2.7). Both of these are defining convex sets. Unlike in the special case of the standard deviation, we do not, in general, obtain a straight line connecting the risk-free asset and risky portfolios. We say that a portfolio is efficient if it solves Program (2.7) for some $\mu^j \in \mathbb{R}$. Thus, the red curve in Figure 2.1 corresponds to the set of efficient portfolios.

**Two-Fund Separation**

We say that two-fund separation holds if the equilibrium optimal portfolios for all investors can be spanned by the risk-free asset and a unique portfolio of risky assets. That is, there exists a unique portfolio with weights $\alpha^P$ such that $\alpha^P_0 = 0$, and for all investors $j \in \{1, ..., \ell\}$, the solution to Problem (2.7) is a linear combination of $\alpha^P$ and the risk-free asset.

**Theorem 41** Consider an equilibrium with positive prices. Assume that $R(\cdot)$ is scaling independent, convex, and satisfies the risk-free property. Then, two-fund separation holds.
The proof is very intuitive, and we show it here. Let $\alpha^1$ and $\alpha^2$ be solutions to Problem (2.7) for investors $j_1 \neq j_2$, respectively, and without loss of generality assume $j_1 = 1$ and $j_2 = 2$. The case of interest is when both $\alpha^1$ and $\alpha^2$ have non-zero weights in some risky assets.\(^{13}\) By the risk-free property and by the non-redundancy assumption, $R(\alpha^j \cdot \tilde{z}) > 0$ for $j = 1, 2$. Hence, $\mu^j = E(\alpha^j \cdot \tilde{z}) > r_f$ for $j = 1, 2$, since otherwise $\alpha^j$ would be mean-risk dominated by the risk-free asset, and thus would not be optimal.

Now, consider all the linear combinations of these two portfolios with the risk-free asset. Since $R(\cdot)$ is assumed convex, the resulting curves are concave in the $\mu$-$R$ plane as illustrated in Figure 2.2. Note that both $\alpha^1$ and $\alpha^2$ can be presented as a linear combination of the risk-free asset and some portfolios $\alpha^{P_1}$ and $\alpha^{P_2}$ of risky assets only (i.e., $\alpha_0^{P_1} = \alpha_0^{P_2} = 0$).

To show two-fund separation we need to show that $\alpha^{P_1} = \alpha^{P_2}$. Suppose this is not the case. Then let $\hat{\alpha}^1$ be a linear combination of $\alpha^{P_2}$ and the risk-free asset such that $E(\hat{\alpha}^1 \cdot \tilde{z}) = \mu^1$. Similarly, let $\hat{\alpha}^2$ be a portfolio of $\alpha^{P_1}$ and the risk-free asset such that $E(\hat{\alpha}^2 \cdot \tilde{z}) = \mu^2$. By convexity of $R(\cdot)$, $\alpha^1$ and $\alpha^2$ are the unique solutions to Program (2.7) for $j = 1, 2$. Hence,

$$R(\hat{\alpha}^1 \cdot \tilde{z}) > R(\alpha^1 \cdot \tilde{z}) \quad \text{and} \quad R(\hat{\alpha}^2 \cdot \tilde{z}) > R(\alpha^2 \cdot \tilde{z}).$$

(2.8)

Thus, as illustrated in Figure 2.2, the two curves must cross at least once. We will now show that such crossings are impossible. Indeed, by scaling independence (2.8) implies that for any $\lambda > 0$,

$$R(\lambda \alpha^1 \cdot \tilde{z}) < R(\lambda \hat{\alpha}^1 \cdot \tilde{z}),$$

which together with risk-free property implies

$$R(\lambda \alpha^1 \cdot \tilde{z} + (1 - \lambda) r_f) < R(\lambda \hat{\alpha}^1 \cdot \tilde{z} + (1 - \lambda) r_f).$$

This means that all linear combinations of $\alpha^1$ with the risk-free asset (with positive $\lambda$) lie strictly to the left of all linear combinations of $\hat{\alpha}^1$ with the risk-free asset. In particular, $\hat{\alpha}^2$

\(^{13}\)If only one investor holds non-zero weights in risky assets then two-fund separation is trivial.
can be obtained as a linear combination of \( \alpha^1 \) with the risk-free asset by setting

\[
\lambda = \frac{\mu^2 - r_f}{\mu^1 - r_f} > 0,
\]

where the inequality follows since \( \mu^j > r_f \) for \( j = 1, 2 \). But, using this \( \lambda \) we obtain

\[
R(\hat{\alpha}^2 \cdot \tilde{z}) < R(\alpha^2 \cdot \tilde{z}),
\]

contradicting (2.8). Thus, two-fund separation must hold.

A corollary is that the unique portfolio \( \alpha^P \) is efficient. Indeed, let \( \mu^P = E(\alpha^P \cdot \tilde{z}) \). Since in equilibrium all investors hold a linear combination of the risk-free asset and \( \alpha^P \), and since \( \mu^j = E(\alpha^j \cdot \tilde{z}) \geq r_f \) for all \( j \) with strict inequality for some \( j \), we have two cases:\(^{14}\) (i) all investors hold \( \alpha^P \) with a non-negative weight, and \( \mu^P > r_f \); or (ii) all investors hold \( \alpha^P \) with a non-positive weight, and \( \mu^P < r_f \). But, the second case is impossible since then the market cannot clear for at least one risky asset, which is held in positive weight in \( \alpha^P \). Thus, \( \mu^P > r_f \).

Now, assume that \( \alpha' \neq \alpha^P \) solves Problem (2.7) for \( \mu^j = \mu^P \). Then, \( R(\alpha' \cdot \tilde{z}) < R(\alpha^P \cdot \tilde{z}) \), and so by the same argument as in the proof of Theorem 41, all linear combinations of \( \alpha' \) with the risk-free asset would have strictly lower risk than the corresponding linear combinations of \( \alpha^P \) with the risk-free asset. This contradicts that \( \alpha^P \) and the risk-free asset span all optimal portfolios. We thus have:

**Corollary 42** Under the conditions of Theorem 41, the portfolio \( \alpha^P \) is efficient. In particular, it solves Problem (2.7) for some \( \mu^P > r_f \).

Let \( x_i^M = \sum_{j=1}^\ell x_i^j \) be the total amount invested in asset \( i \) in equilibrium. We call \( x^M = (x_1^M, ..., x_n^M) \) the *market portfolio* (consisting of risky assets only). Let \( \alpha^M \) be the corresponding portfolio weights. By Theorem 41, in equilibrium, the market portfolio is

\(^{14}\)If all investors choose the risk-free asset then the market for risky assets cannot clear.
equal to \( \alpha^P \), the unique portfolio of risky assets that together with the risk-free asset spans all optimal portfolios.\(^{15}\) Moreover, by corollary 42, the market portfolio is efficient, and its expected return is strictly higher than \( r_f \).

**Corollary 43** Under the conditions of Theorem 41, the market portfolio is efficient. In particular, it solves Problem (2.7) for some \( \mu^M > r_f \).

**A Generalized Security Market Line**

In the traditional CAPM framework, the security market line describes the equilibrium relation between the expected returns of individual assets and the market expected return. Specifically, the expected return of any asset in excess of the risk-free rate is proportional to the excess market expected return, with the coefficient of proportionality being equal to the traditional beta. The following theorem provides sufficient conditions under which a similar relation holds for a broad set of risk measures.

\(^{15}\)Note that \( \alpha^P \) is of dimension \( n + 1 \), but its first component is zero. By saying that \( \alpha^P = \alpha^M \) we mean that \( \alpha^P_i = \alpha^M_i \) for \( i = 1, ..., n \).
Theorem 44 Consider an equilibrium with positive prices and let $\alpha^M$ be the market portfolio. Assume that $R(\cdot)$ is scaling independent, convex, smooth, and satisfies the risk-free property. Then, for each asset $i = 1, ..., n$,

$$E(\tilde{z}_i) = r_f + B^R_i \left( E(\alpha^M \cdot \tilde{z}) - r_f \right),$$

where

$$B^R_i = \frac{R_i(\alpha^M)}{\sum_{h=1}^{n} \alpha^M_h R_h(\alpha^M)}.$$ (2.10)

If $R(\cdot)$ is also homogeneous of some degree $k$, then (2.10) takes the form

$$B^R_i = \frac{R_i(\alpha^M)}{k R(\alpha^M)}.$$ (2.11)

Thus, the security market line has the traditional form, with the generalized systematic risk measure $(B^R_i)$ given as the marginal contribution of asset $i$ to the total risk of the market portfolio, scaled by the weighted average of marginal contributions of all assets. If $R$ is furthermore homogeneous, it is simply given by the marginal contribution of asset $i$ scaled by total risk multiplied by the degree of homogeneity.

To see the intuition for this result, start with an efficient portfolio $\alpha^*$ and consider borrowing one dollar at the risk-free rate and investing this dollar in asset $i$. The effect of this exercise on the risk of the portfolio is (up to first-order approximation) $R_i(\alpha^*) - R_0(\alpha^*)$, which by the risk-free property is just $R_i(\alpha^*)$. Since $\alpha^*$ is efficient, the effect of this exercise on risk is equal to the shift in the expected return constraint times the shadow price of the constraint, $\xi$, i.e.,

$$R_i(\alpha^*) = \xi (E(z_i) - r_f).$$ (2.11)

Taking the weighted average using the portfolio weights gives

$$\sum_{i=1}^{n} \alpha^*_i R_i(\alpha^*) = \xi (E(\alpha^* \cdot \tilde{z}) - r_f).$$ (2.12)
Using (2.11) and (2.12) we obtain that for any efficient portfolio $\alpha^*$,

$$\frac{R_i(\alpha^*)}{\sum_{i=1}^n \alpha_i^* R_i(\alpha^*)} = \frac{E(z_i) - rf}{E(\alpha^* \cdot \tilde{z}) - rf}.$$ 

Namely, in equilibrium, $B_i^R$ (as given in (2.10)) equals the ratio of the expected excess return of any asset $i$ to the expected excess return of the efficient portfolio $\alpha^*$. Finally, since $\alpha^M$ has been shown to be efficient (Corollary 43) we obtain the result.

### 2.4.3 Applications and Empirical Implementation

We now provide several applications to illustrate the versatility and power of Theorem 44 and its potential empirical usefulness. We show how to use this theorem to generalize the traditional CAPM to account for high distribution moments, downside risk, rare disasters, as well as combinations thereof. We also discuss the economic interpretation of systematic risk in these cases and explain how these results can be implemented empirically.

**Applications**

**Application I: The standard CAPM.** When the risk measure $R$ is the variance, i.e., $R(\tilde{z}) = \text{Var}(\tilde{z})$, Theorem 44 coincides with the standard CAPM (see Appendix II for the derivation). Namely,

$$B_i^R = \frac{\text{Cov}(\tilde{z}_i, \alpha^M \cdot \tilde{z})}{\text{Var}(\alpha^M \cdot \tilde{z})}. \quad (2.13)$$

Thus, in this case systematic risk is measured as the standard regression coefficient. The same result is obtained when $R(\tilde{z}) = w_2(\tilde{z})$, i.e., the standard deviation of returns.

**Application II: A CAPM reflecting aversion to tail risk.** The simplest generalization of the standard CAPM is to the case in which investors are averse to any moment of an even degree. That is, set $R(\tilde{z}) = m_k(\tilde{z}) = E(\tilde{z} - E(\tilde{z}))^k$, $k$ even. This risk measure satisfies all of the conditions in Theorem 44 (see Example 29). In this case the systematic risk takes the
form (see Appendix II for the derivation)

\[
B_i^R = \frac{\text{Cov} \left( \tilde{z}_i, \left( \alpha^M \cdot \tilde{z} - \alpha^M \cdot \text{E} (\tilde{z}) \right)^{k-1} \right)}{m_k (\alpha^M \cdot \tilde{z})}.
\]  

(2.14)

That is, the systematic risk of asset \( i \) is proportional to the covariance of \( \tilde{z}_i \) with the \((k - 1)^{th}\) power of the demeaned market return. In the special case of \( k = 2 \) (variance), this reduces to (2.13) as expected. Another important special case is \( k = 4 \); in which \( R (\tilde{z}) \) measures the tail risk of \( \tilde{z} \). Then,

\[
B_i^R = \frac{\text{Cov} \left( \tilde{z}_i, \left( \alpha^M \cdot \tilde{z} - \alpha^M \cdot \text{E} (\tilde{z}) \right)^3 \right)}{m_4 (\alpha^M \cdot \tilde{z})}.
\]

Namely, the systematic risk of asset \( i \) is proportional to the co-kurtosis of \( \tilde{z}_i \) with the demeaned market return. Similarly, when \( R (\tilde{z}) = w_k (\tilde{z}) \), the normalized \( k^{th} \) central moment, \( B_i^R \) again takes the form (2.14).

**Application III: A CAPM reflecting aversion to downside risk.** Assume \( R (\tilde{z}) = DR_k (\tilde{z}) \) for \( k \geq 2 \). This risk measure satisfies all of the conditions in Theorem 44 (see Example 32). The systematic risk is then given by (see Appendix II for the derivation)

\[
B_i^R = (-1)^k \frac{\text{Cov} \left( \tilde{z}_i, \left( \alpha^M \cdot \tilde{z} - \alpha^M \cdot \text{E} (\tilde{z}) \right)^{-} \right)^{k-1}}{(DR_k (\alpha^M \cdot \tilde{z}))^k}.
\]

(2.15)

That is, the systematic risk of asset \( i \) is proportional to the covariance of \( \tilde{z}_i \) with the \((k - 1)^{th}\) power of the demeaned market return, censored at zero.

**Application IV: A CAPM reflecting aversion to rare disasters.** To account for rare disasters we can use the demeaned expected shortfall measure, which satisfies all the requirements in Theorem 44 (see Example 35). Assume then that \( R (\tilde{z}) = \text{DES}_\delta (\tilde{z}) \), where \( 0 < \delta < 1 \) is some confidence level. The systematic risk in this case is given by (see Appendix II for the derivation)

\[
B_i^R = -\frac{\text{E} \left[ \tilde{z}_i - \text{E} (\tilde{z}_i) \mid \alpha^M \cdot \tilde{z} \leq -\text{VaR}_\delta (\alpha^M \cdot \tilde{z}) \right]}{\text{DES}_\delta (\alpha^M \cdot \tilde{z})}.
\]

(2.16)
Thus, the systematic risk of asset \( i \) in this case equals (the negative of) the expected demeaned return of asset \( i \) conditional on the market being in a disaster, scaled by the market’s demeaned expected shortfall. An equivalent expression is

\[
B^R_i = -\frac{\text{Cov} \left[ \tilde{z}_i, 1_{\alpha^M \cdot \tilde{z} \leq -\text{VaR}_\delta(\alpha^M \cdot \tilde{z})} \right]}{\delta \cdot \text{DES}_\delta(\alpha^M \cdot \tilde{z})},
\]

showing that systematic risk in this case is proportional to the covariance of the asset return with an indicator equal to one when the market is in a disaster.

So far we have restricted our applications to cases in which investors are averse to just one risk aspect. In reality, it is likely that investors are averse to several risk attributes. Our framework allows for this by constructing risk measures that account for several risk characteristics using Lemma 28. The next application illustrates this point.

**Application V: A CAPM reflecting aversion to variance, downside skewness, tail risk, and rare disasters.** Consider the following family of risk measures

\[
R(\tilde{z}) = \theta_1 w_2(\tilde{z}) + \theta_2 DR_3(\tilde{z}) + \theta_3 w_4(\tilde{z}) + \theta_4 \text{DES}_\delta(\tilde{z})
\]

for some confidence level \( \delta \). Here \( \theta_1, ..., \theta_4 \) are non-negative weights accounting for the degree of aversion to variance, downside skewness, tail risk, and rare disasters, respectively.\(^{16}\) The case \( \theta_1 = 1 \) and \( \theta_2 = \theta_3 = \theta_4 = 0 \) corresponds to the traditional CAPM, whereas different values of the weights allow us to reflect different levels of aversion to the different risk attributes.

By Lemma 28 these risk measures satisfy all the conditions in Theorem 44 and so all the CAPM results above hold. The resulting systematic risk measure accounts for the contribution of asset \( i \) to all four risk attributes. It is simply given by a weighted average

---

\(^{16}\)Note that we are using here \( w_2(\cdot) \) and \( w_4(\cdot) \) (the normalized second and fourth moments) instead of \( m_2(\cdot) \) and \( m_4(\cdot) \). This is done to make sure that all of the components in \( R(\cdot) \) are homogeneous of degree 1, and so \( R(\cdot) \) is homogeneous.
of the systematic risk measures as calculated in the above applications (see Appendix II). Namely,

\[
B_i^R = \frac{R^1(\alpha^M)}{R(\alpha^M)}B_i^{R_1} + \frac{R^2(\alpha^M)}{R(\alpha^M)}B_i^{R_2} + \frac{R^3(\alpha^M)}{R(\alpha^M)}B_i^{R_3} + \frac{R^4(\alpha^M)}{R(\alpha^M)}B_i^{R_4},
\]

(2.17)

where \(R^1(\cdot) = \theta_1w_2(\cdot), \ R^2(\cdot) = \theta_2DR_3(\cdot), \ R^3(\cdot) = \theta_3w_4(\cdot),\) and \(R^4(\cdot) = \theta_4DES_\delta(\cdot),\) and where \(B_i^{R_1}, B_i^{R_2}, B_i^{R_3},\) and \(B_i^{R_4}\) are given by (2.13)–(2.16).

**Empirical Implementation**

Similar to the classic CAPM, Theorem 44 and its applications lend themselves naturally to empirical investigation. The standard approach for testing and applying the CAPM follows Fama and MacBeth (1973) and Fama and French (1992). The first stage in their approach consists of estimating beta through time-series regressions, whereas the second stage consists of cross-sectional regressions of excess asset returns on estimated betas.

To apply this approach in our case, one needs to first take a stand on what the risk measure \(R\) is. Then, using Theorem 44 one can estimate \(B_i^R\) from time-series data. For example, if \(R\) takes the form as in Application V above, then we need time series return data for asset \(i\) and for the market portfolio in order to estimate \(B_i^R\) from (2.17). This will be a weighted average of the betas prescribed in Applications I–IV. Note that unlike in the classic CAPM, \(B_i^R\) is in general not a regression coefficient. Nevertheless, it often takes the form of some scaled covariance of the asset returns and some function of the market returns (see Applications I–IV). Thus, \(B_i^R\) can still be readily estimated from time-series return data. The cross-sectional part is then identical in nature to that in Fama and MacBeth (1973).

It is important to note that the model does not provide us with guidance as to what \(R\) is. Rather, for any given risk measure the model provides an expression for the associated systematic risk. In practice we believe that the data can guide us in finding what the “true” risk is, to which investors are averse. For example, consider Application V, which allows the risk measure to reflect aversion to variance, downside skewness, tail risk, and disaster
risk. One still has a lot of flexibility in choosing the weights \( \theta_1, ..., \theta_4 \), which determine the degree of aversion to each particular aspect of risk. The model can then allow the data to determine which set of weights obtains the most support. This flexibility is tantamount to the freedom provided by the Arbitrage Pricing Theory (Ross (1976b)) in which the model suggests the existence of multiple systematic factors but does not provide guidance as to what these factors are.

### 2.4.4 Further Discussion

Note that Theorem 44 relies on the market portfolio being efficient, and that two-fund separation is a way to achieve this efficiency result. Our assumptions on the risk measure are sufficient for two-fund separation, but they are by no means necessary. Weaker conditions that guarantee two-fund separation may exist. Further, even when two-fund separation fails, it does not necessarily mean that market efficiency is rejected. The literature explores market efficiency from both theoretical (see, for example, Dybvig and Ross (1982)) and empirical (see, for example, Levy and Roll (2010)) views. Our generalized SML remains valid as long as we have evidence that the market portfolio is mean-risk efficient.

We should also mention that the classical notion of beta and its relation to expected returns go beyond the standard CAPM setup. Specifically, as long as there is no arbitrage and so a stochastic discount factor exists, a beta representation of the form

\[
E(\tilde{z}_i) = \gamma + B_i \lambda
\]

exists (see Hansen and Richard (1987) and Cochrane (2001) Ch. 6). This does not stand in conflict to the results in this section. Rather, our results essentially identify a class of stochastic discount factors driven by the mean-risk preferences being assumed.
2.5 Systematic Risk as a Solution to a Risk Allocation Problem

The equilibrium approach presented in the previous section generalizes the classic CAPM, but it has two limitations. First, this approach allows us to calculate the contribution of an asset to the risk of the market portfolio, but not to arbitrary portfolios of risky assets. Second, to obtain the equilibrium results we imposed restrictions on the risk measures (scaling independence, convexity, and the risk-free property). These restrictions allow us to establish existence of equilibrium and efficiency of the market portfolio. However, some risk measures do not satisfy these conditions.

In this section we offer an alternative approach to developing a systematic risk measure. This approach applies to any portfolio of risky assets and to a broader class of risk measures. For example, if a bank would like to use the VaR measure to estimate the contributions of different assets on its balance sheet to the total VaR of the bank, then the results in this section can be applied. Importantly, when the risk measure is homogeneous, the two approaches lead to an identical result, generalizing the traditional beta.

Our approach is to consider this issue as a risk allocation problem, where the total risk of a given portfolio needs to be “fairly” allocated among its components. We offer four axioms that describe reasonable properties of solutions to risk allocation problems. We then show that these axioms determine a unique formula for the systematic risk of an asset, the contribution of the asset to the risk of the portfolio.

2.5.1 Axiomatic Characterization of Systematic Risk

A risk allocation problem of order $n \geq 1$ is a pair $(R, x)$, where $R$ is a risk measure and $x \in \mathbb{R}_+^n$ is a portfolio specifying the dollar amount invested in each of $n$ assets $\tilde{z} = (\tilde{z}_1, ..., \tilde{z}_n)$, and $R(x \cdot \tilde{z}) \neq 0$. Denote the total dollar amount invested by $\bar{x} = \sum_{i=1}^n x_i$. Also, let $\alpha$ be the vector of corresponding portfolio weights, i.e., $\alpha_i = x_i/\bar{x}$. The only two requirements we
impose on $R$ in this section are that $R(0) = 0$ (i.e., zero investment entails no risk) and that $R(\cdot)$ is smooth.

A *systematic risk measure* is a function mapping any risk allocation problem of order $n$ to a vector $B^R(x) = (B^R_1(x), ..., B^R_n(x))$ in $\mathbb{R}^n$. Intuitively, one can think of $B^R_i(x)$ as the contribution of asset $i$ to the total risk of portfolio $x$, which is $R(x \cdot \bar{z})$. Note that a systematic risk measure applies to all possible pairs of risk measures and portfolios, rather than to a given pair.

We now state four axioms specifying desirable economic properties of systematic risk measures. The intuition for why these axioms make sense mostly comes from the traditional beta. Here we simply try to identify properties of beta and ask how these properties could be generalized to arbitrary risk measures. It is important to emphasize that these axioms do not impose any restriction on the risk measure. Rather, they impose structure on what would constitute a solution to the risk allocation problem.

The first axiom postulates that (as for the traditional beta) the weighted average of systematic risk values across all assets is normalized to 1.

**Axiom 45** *Normalization:* $\sum_{i=1}^{n} \alpha_i B^R_i(x) = 1$.

The sum of any two risk measures is itself a risk measure. The next axiom requires that in such a case the systematic risk measure of the sum will be a risk-weighted average of systematic risk based on each of the two risk components.

**Axiom 46** *Linearity:* If $R(\cdot) = R^1(\cdot) + R^2(\cdot)$, then

$$B^R_i(x) = \frac{R^1(x)}{R(x)} B^R_i^1(x) + \frac{R^2(x)}{R(x)} B^R_i^2(x) \text{ for all } i = 1, ..., n.$$
risk measures. However, two features can be easily generalized laying the foundations for the next two axioms.

First, while the concept of “correlation” is not easy to generalize, the idea of “perfect correlation” does lend itself to a natural generalization. The intuition is that if several assets are perfectly correlated, then essentially they can be thought of as the same asset. Thus, a portfolio of perfectly correlated assets can be viewed as one “big” asset. This intuition comes from the standard notion of correlation relating to risk being measured by the variance, but it can easily be generalized to arbitrary risk measures.

Formally, given a risk measure $R$, we say that assets $\mathbf{z} = (z_1, \ldots, z_n)$ are $R$-perfectly correlated if there exists a function $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ and a non-zero vector $\mathbf{q} = (q_1, \ldots, q_n) \in \mathbb{R}_+^n$, such that for any portfolio $\mathbf{\eta} = (\eta_1, \ldots, \eta_n) \in \mathbb{R}_+^n$ we have $R(\mathbf{\eta} \cdot \mathbf{z}) = g(\mathbf{\eta} \cdot \mathbf{q})$. That is, the $n$ assets are $R$-perfectly correlated if the risk of any portfolio of these assets as measured by $R$ only depends on some linear combination of their investment amounts. In essence, this means that the $n$ assets can be aggregated into one “big” asset by assigning each asset a certain weight specified by the vector $\mathbf{q}$.

To see the correspondence to the standard notion of perfect correlation, consider the following example. Assume risk is measured using variance and let $\mathbf{z} = (z_1, \tilde{z}_2, \tilde{z}_3)$ with $\tilde{z}_2 = 2 \tilde{z}_1$ and $\tilde{z}_3 = 5 \tilde{z}_1$. Then, all three assets are perfectly correlated and for any portfolio $(\eta_1, \eta_2, \eta_3)$ we have

$$\text{Var}(\eta_1 \tilde{z}_1 + \eta_2 \tilde{z}_2 + \eta_3 \tilde{z}_3) = (\eta_1 + 2\eta_2 + 5\eta_3)^2 \text{Var}(\tilde{z}_1).$$

Thus, we can set $g(t) = t^2$ and the vector of weights is $\mathbf{q} = \sqrt{\text{Var}(\tilde{z}_1)} (1, 2, 5)$. More generally, it is easy to verify that when risk is measured using variance, the concept of $R$-perfect correlation coincides with the standard definition of perfect correlation.

In the standard notion of perfect correlation, we differentiate between positive and negative perfect correlation. We could do the same here by allowing elements of $\mathbf{q}$ to take negative values. However, this is not needed for our axiomatic characterization.
Axiom 47 \textbf{Proportionality:} If $\mathbf{z} = (\tilde{z}_1, ..., \tilde{z}_n)$ are $R$-perfectly correlated with weights $\mathbf{q} = (q_1, ..., q_n)$, then

$$q_j B^R_i(x) = q_i B^R_j(x) \text{ for all } i, j = 1, ..., n.$$ (2.18)

Next we turn to generalize the idea of “positive correlation.” Assume first that risk is measured using variance. Then, if two assets are positively correlated, adding additional units of an asset to any portfolio of the two always increases total variance. We can then rely on this feature to get a generalized notion of positive correlation. Specifically, given a risk measure $R$, we say that assets $\mathbf{z} = (\tilde{z}_1, ..., \tilde{z}_n)$ are $R$-positively correlated if $R_i(\eta \cdot \mathbf{z}) \geq 0$ for all $\eta \in \mathbb{R}^n_+$ and for all $i = 1, ..., n$. Namely, the assets are $R$-positively correlated if adding one more unit of an asset to any portfolio with non-negative weights can never reduce total risk. The key to this definition is that for the assets to be $R$-perfectly correlated it is not enough that adding one more unit of an asset would increase risk for a particular portfolio. Rather, this property has to hold for all possible portfolios of these assets.\footnote{It is easy to check that when risk is measured using variance, the assets are $R$-positively correlated if and only if the correlation between any two assets is non-negative.} The next axiom requires that when the assets are $R$-positively correlated, the systematic risk of all assets is non-negative.

Axiom 48 \textbf{Monotonicity:} If $\mathbf{z} = (\tilde{z}_1, ..., \tilde{z}_n)$ are $R$-positively correlated, then $B^R_i(x) \geq 0$ for all $i = 1, ..., n$.

Our main result in this section follows. It states that Axioms 1–4 are sufficient to pin down a unique systematic risk measure, which takes on a very simple and intuitive form. Moreover, when the risk measure is homogeneous, the solution coincides with the equilibrium result in Theorem 44.
Theorem 49 There exists a unique systematic risk measure satisfying Axioms 45–48. For each risk allocation problem \((R, x)\) of order \(n\), it is given by

\[
B_i^R(x) = \frac{\bar{x} \int_0^1 R_i(tx_1, \ldots, tx_n) \, dt}{R(x_1, \ldots, x_n)} \quad \text{for } i = 1, \ldots, n. \tag{2.19}
\]

Furthermore, if \(R\) is homogeneous of some degree \(k\), then (2.19) reduces to

\[
B_i^R(x) = \frac{R_i(\alpha)}{\sum_{h=1}^n \alpha_h R_h(\alpha)} \tag{2.20}
\]

\[
= \frac{R_i(\alpha)}{k R(\alpha)} \tag{2.21}
\]

Thus, when \(R\) is homogeneous (which is a common case), the systematic risk of asset \(i\) is measured simply as the marginal contribution of asset \(i\) to the total risk of the portfolio, scaled by the weighted average of marginal contributions of all assets. This is identical to the result in Theorem 44 only with respect to an arbitrary portfolio rather than the market portfolio. When the risk measure is not homogeneous, the expression in (2.19) shows that systematic risk depends not only on marginal contributions at \(x\), but rather on marginal contributions along a diagonal between \((0, \ldots, 0)\) and \(x\). This is a variation of the diagonal formula of Aumann and Shapley (1974). The integral can be interpreted as an average of marginal contributions of asset \(i\) to the risk of portfolios along the diagonal. Then, \(B_i^R(x)\) is simply a scaled version of the integral where the scaling ensures that Axiom 1 is satisfied.

Note that when the risk measure is homogeneous, \(B_i^R(x)\) depends only on the portfolio weights \(\alpha\) (and not on the dollar amounts invested in each asset). Indeed, in the homogeneous case \(R_i(tx_1, \ldots, tx_n)\) is proportional to \(R_i(x_1, \ldots, x_n)\) for all \(t \in [0, 1]\), yielding the simple expression in (2.20). When the risk measure is not homogeneous, the actual investment amounts (not just the weights) are necessary for the calculation of systematic risk.

The uniqueness part of the proof of Theorem 49 is in Appendix I. It relies on the solutions to cost allocation problems established in Billera and Heath (1982).\(^{20}\) In this proof we draw

\(^{20}\)Billera and Heath (1982) define a cost allocation problem of order \(n\) as a pair \((h, x)\) where \(h : \mathbb{R}_+^n \to \mathbb{R}^n\)
a one-to-one mapping between risk allocation problems and cost allocation problems, and from systematic risk measures to solutions of cost allocation problems. Then, we show that given these mappings, our set of axioms is stronger than the set of conditions specified in Billera and Heath (1982). This in turn allows us to apply their result to obtain uniqueness.

Existence is straightforward and we show it below by demonstrating that (2.19) satisfies Axioms 1–4. Suppose that $\mathcal{B}_i^R(x)$ is given by (2.19). Then,

$$\sum_{i=1}^{n} \alpha_i \mathcal{B}_i^R(x) = \sum_{i=1}^{n} \frac{x_i \bar{x} \int_0^1 R_i (tx_1, ..., tx_n) \, dt}{R(x_1, ..., x_n)}$$

$$= \int_0^1 \sum_{i=1}^{n} x_i R_i (tx_1, ..., tx_n) \, dt \frac{1}{R(x_1, ..., x_n)}$$

$$= \frac{\int_0^1 dR(tx_1, ..., tx_n) \, dt}{R(x_1, ..., x_n)} = 1, \quad \text{(since } R(0) = 0)$$

and so Axiom 1 holds. To see Axiom 2, suppose $R(\cdot) = R^1(\cdot) + R^2(\cdot)$. Then,

$$\mathcal{B}_i^R(x) = \frac{\bar{x} \int_0^1 R_i (tx_1, ..., tx_n) \, dt}{R(x_1, ..., x_n)}$$

$$= \frac{\bar{x} \int_0^1 R_i^1 (tx_1, ..., tx_n) \, dt}{R(x_1, ..., x_n)} R^1(x_1, ..., x_n) + \frac{\bar{x} \int_0^1 R_i^2 (tx_1, ..., tx_n) \, dt}{R^2(x_1, ..., x_n)} R^2(x_1, ..., x_n),$$

as required. Next, for Axiom 3, suppose that $\tilde{z} = (\tilde{z}_1, ..., \tilde{z}_n)$ are $R$-perfectly correlated. Then, there exists $g(\cdot) : \mathbb{R} \rightarrow \mathbb{R}$ and a nonzero vector $q \in \mathbb{R}_n^+$ such that for all $\eta = (\eta_1, ..., \eta_n)$ we have $R(\eta) = g(\eta \cdot q)$. It follows that

$$R_i(\eta) = q_i g'(\eta \cdot q) \text{ for all } i = 1, ..., n.$$ 

Hence, for all $i = 1, ..., n$, 

$$\mathcal{B}_i^R(x) = \frac{\bar{x} q_i \int_0^1 g'(tx \cdot q) \, dt}{R(x_1, ..., x_n)},$$

which implies (2.18). Finally, given the definition of $R$-positive correlation, it is immediate that (2.19) satisfies Axiom 4.

is continuously differentiable and $h(0) = 0$. They interpret $x$ as a vector of inputs and $h$ as a cost function. The question they ask is how to allocate total cost among the different inputs. See Appendix I for more details on their model.
2.5.2 Applying the Result

In Section 2.4.3 we have provided several applications and shown how to calculate systematic risk for different risk measures. All of these results apply to the approach presented in this section as well, but now they can be used with respect to arbitrary portfolios rather than just the market portfolio. The next example illustrates a case of risk measures that do not satisfy the conditions in Section 2.4, but for which Theorem 49 applies.

Recall the Aumann-Serrano and Foster-Hart risk measures in Example 37. These measures are homogeneous, convex, and smooth, but they do not satisfy the risk-free property.\footnote{Although $R^{AS}(0)$ and $R^{FH}(0)$ are not defined, they can be approximated using a limiting argument. Specifically, take any random return $\tilde{z}$ satisfying $E(\tilde{z}) > 0$ and $P(\{\tilde{z} < 0\}) > 0$. Then, for both $R(\cdot) = R^{AS}(\cdot)$ and $R(\cdot) = R^{FH}(\cdot)$, we can define $R(0)$ by}

Still, Theorem 49 allows us to calculate the systematic risk associated with these risk measures.

Using Theorem 49 and applying the implicit function theorem to (2.4) and (2.5) yields the systematic risk of individual assets associated with the AS and FH measures relative to any portfolio weights $\alpha$ as follows:

$$B^{R^{AS}}_i(\alpha) = \frac{E\left[\exp\left(-\frac{\alpha \tilde{z}}{R(\alpha)}\right) \tilde{z}_i\right]}{E\left[\exp\left(-\frac{\alpha \tilde{z}}{R(\alpha)}\right) \alpha \cdot \tilde{z}\right]},$$

and

$$B^{R^{FH}}_i(\alpha) = \frac{E\left[\tilde{z}_i \right]}{E\left[\frac{\alpha \tilde{z}}{R(\alpha) + \alpha \tilde{z}}\right]},$$

2.5.3 Discussion

It is interesting to ask what would happen if we used (2.20) to define systematic risk when $R$ is not homogeneous (instead of using (2.19)). In particular, this alternative measure only

$$R(0) \equiv \lim_{t \to 0} R(t \tilde{z}) = 0,$$

where the equality follows since both the AS and the FH measures are homogeneous of degree 1.
relies on the marginal contribution of asset \( i \) at \( \alpha \) and not along the diagonal. In the absence of homogeneity these two alternative definitions yield different results. Thus, given Theorem 49, it must be that (2.20) violates at least one of our axioms. It is straightforward to check that the axiom being violated in this case is Axiom 2 while the other three axioms are satisfied.

Another alternative to measuring systematic risk might be to define

\[
B^R_i(\alpha) = \frac{R_i(\alpha)}{R(\alpha)},
\]

namely, the systematic risk of an asset is the marginal contribution of the asset to total risk, scaled by total risk. This measure satisfies Axioms 2, 3, and 4 but it fails Axiom 1, so it cannot be considered as a generalization of the traditional beta.

Finally, it is worth noting that (2.20) can also be written as

\[
B^R_i(\alpha) = \left. \frac{\frac{d}{dt} R(\alpha + t\varepsilon^i)}{\frac{d}{dt} R(\alpha + t\alpha)} \right|_{t=0},
\]

where \( \varepsilon^i \) is an \( n \)-dimensional vector equal to 1 at the \( i^{th} \) dimension and zero elsewhere. Namely, when the risk measure is homogeneous, systematic risk of asset \( i \) can be thought of as the directional derivative of total risk along the \( i^{th} \) dimension scaled by the derivative along the diagonal in the direction of the portfolio itself.

### 2.6 Conclusion

In this paper we generalize the concept of systematic risk to account for a variety of risk characteristics. Our equilibrium approach shows that results attributed to the classic CAPM hold much more broadly. In particular, aspects of the geometry of efficient portfolios, two-fund separation, and the security market line are derived in a setting where risk can account for a variety of attributes. Our axiomatic approach specifies four economically meaningful conditions that pin down a unique measure of systematic risk. Both approaches lead to similar generalizations of the traditional beta.
When risk is confined to measure the variance of a distribution, our systematic risk measures coincide with the traditional beta, the slope from regressing asset returns on portfolio returns. More generally, systematic risk is not a regression coefficient. Our equilibrium setting leads to the conclusion that systematic risk is simply the marginal contribution of the asset to the risk of the portfolio of interest, scaled by the weighted average of all such marginal contributions. An identical result is obtained in the axiomatic approach for homogeneous risk measures. When the risk measure is not homogeneous, the axiomatic approach gives rise to an expression for systematic risk that involves averaging marginal contributions of the asset along a diagonal from the origin to the portfolio of interest.

Our axiomatic approach applies to a wide variety of risk measures, requiring of them only smoothness and zero risk for zero investment. The equilibrium framework imposes additional conditions in the form of scaling independence, convexity, and the risk-free property. Nevertheless, even in the equilibrium framework we are still left with an extensive class of risk measures. Indeed, this class is sufficiently broad to potentially account for high distribution moments, downside risk, rare disasters, and other aspects of risk. A limitation of our framework is that we restrict all investors to use the same risk measure. Future research may direct at developing weaker conditions on the risk measures and introducing more heterogeneity to investor risk preferences.

Finally, our approach is agnostic regarding the choice of a particular risk measure. Indeed, which risk measures better capture the risk preferences of investors is ultimately an empirical question. Our framework therefore provides foundations for testing the appropriateness of risk measures and consequently selecting those that are supported by the data.
Chapter 3

Recovering Conditional Return Distributions by Regression: Estimation and Applications

This chapter proposes a regression approach to recovering the return distribution of an individual asset conditional on the return of an aggregate index based on their marginal distributions. This approach relies on the identifying assumption that the conditional return distribution of the asset given the index return does not vary over time. I show how to empirically implement this approach using option price data. I then apply this approach to examine the cross-sectional equity risk premium associated with systematic disaster risk, to estimate the exposure of banks to systemic shocks, and to extend the Ross (Journal of Finance, 2015) recovery theorem to individual assets.

3.1 Introduction

The recent financial crisis has witnessed dramatic declines in the prices of most securities, which suggests strong return comovement between various assets. It is desirable to understand how the returns of different securities move along with each other. In this paper, I propose a regression approach to recovering the return distribution of an individual asset conditional on the return of an aggregate index based on their marginal distributions. I show
how this approach can be implemented empirically using option prices. I then demonstrate the usefulness of this approach in testing the cross-sectional equity premium associated with systematic disaster risk, in estimating the exposure of banks to systemic shocks, and in an extension of the Ross (2015) recovery theorem.

The simplest and most widely used approach to describing the joint return behavior between two securities is to run a linear regression based on their historical returns. Indeed, this is what we are used to doing when estimating the CAPM beta by regressing excess returns of an asset on those of the market portfolio. This approach, however, has a number of drawbacks. First, it estimates the mean return of one security conditional on the return of the other, but it fails to capture high-moment properties. For example, Figure 3.1 plots the returns of two pairs of hypothetical securities, both of which predict the same conditional mean return of one security given the return of the other. However, the second-moment patterns of the two pairs are clearly distinct in the sense that the first pair has increasing correlation when the returns become lower, whereas the second pair has symmetric correlation over the entire region of returns. Second, running a linear regression between the two securities focuses on the linear relation only, neglecting other aspects of their joint behavior. To illustrate this point, Figure 3.2 depicts the returns of two hypothetical assets against the return of the market portfolio. The returns of both assets fit the same linear relation with the market return. Nevertheless, the nonlinear patterns show that asset 1 is more sensitive to the market disaster risk than asset 2 in the sense that the former tends to deliver lower returns when the market return becomes disastrously low. Third, estimation based on historical returns is backward-looking, which does not necessarily represent future return distributions. Finally, using historical returns makes it difficult to capture the effects of rare events, especially when the sample size is not large enough.

Alternative approaches used in the literature resolve some of the above issues. One such approach is the quantile regression, which predicts the conditional return quantiles of
Figure 3.1: Second Moment Properties in Joint Return Behavior

Figure 3.2: Nonlinearity in Joint Return Behavior
one security given the return of the other. (See Koenker (2005) for detailed discussions on quantile regression.) This approach generates the entire conditional distributions, thus capturing high-moment properties as well as nonlinear aspects in the joint return behavior. Nevertheless, given that the quantile regression is also implemented using historical returns, it is backward-looking and does not adequately reflect the effects of rare events.

Another alternative relies on option prices. We learn from the results of Ross (1976a) and Breeden and Litzenberger (1978) that one can estimate the risk-neutral probability distribution of security returns using prices of options written on the security under consideration. The advantage of using option prices is that it is forward-looking and accounts for rare events even if such events do not occur within sample. However, the risk-neutral distributions obtained from option prices typically differ from the physical distributions due to the adjustment for risk aversion. In addition, if we are interested in the joint return distribution of two securities, we would need options written on the joint values of these two securities. Given that most traded options are written on a single security, this method generally allows one to estimate the risk-neutral marginal return distribution of each single security, but not their joint distribution.

A question that follows naturally is whether we can recover the joint return distribution of two securities, assuming that the associated marginal distributions are known or can be estimated. This is indeed straightforward in the special cases in which the returns of the two securities are perfectly correlated or independent of each other. For more general cases, a well-known tool for this purpose is the copula, which can be used to map the marginal return distributions of multiple securities to their joint distribution. (See Nelsen (1999) for a general overview of the copula method.) However, a drawback of this approach is that it is parametric in the sense that it typically relies on specifying a particular class of copulas. When the copula class is misspecified, the accuracy of estimation might be affected.

In light of all the problems discussed above, it is desirable to have a better approach to
evaluating the joint return behavior of two different securities. The term “better” includes the following aspects. First, it should capture all moment properties of the joint return distribution. Second, it should capture linear as well as nonlinear relations in the returns of the two securities. Third, it should reflect forward-looking information. Fourth, it should naturally account for rare events, whose ex-ante probabilities of occurrence are extremely small. Finally, it should not depend on any parametric assumptions on the return distributions of the securities.

I propose a novel approach of recovering the conditional return distribution of an individual asset given the return of an aggregate index from their marginal distributions. The index return can be viewed as a factor that determines the state of the economy. Examples of the aggregate index include the market portfolio or a sector portfolio, etc. According to the total probability formula, the marginal return distributions of the two securities are linked to each other through the conditional return distribution of the asset given the index return. I assume that the conditional return distribution of the asset given any particular value of the index return remains fixed over time, meaning that the time variation in the return distribution of the asset is solely driven by that of the index. This allows me to estimate the time-invariant conditional return probabilities of the asset as the coefficients from a constrained linear regression of the marginal return distribution of the asset on that of the index over time. I show that under the standard OLS assumptions, the estimates from this constrained regression are consistent, i.e., they converge to the true conditional probabilities as the sample size becomes large enough.

I further assume that the variation in the index return is the only priced risk (systematic risk) such that any variation left in the asset return is idiosyncratic and does not get priced. This implies that the conditional return distribution of the asset given the index return is the same under the physical and the risk-neutral probability measures. Since risk-neutral marginal distributions of security returns can be extracted from option prices, my approach
can be implemented under the risk-neutral measure using option pricing data. The resulting conditional return distribution of the asset estimated this way coincides with the physical conditional distribution.

The advantages of my approach include the following. First, it generates the entire conditional return distribution of the asset given the index return, thus capturing all moment properties and potential nonlinearity in their joint behavior. In addition, since this approach can be implemented using option prices, it is forward-looking and accounts for the likelihood of rare events perceived by investors even if such events do not truly occur within sample. Finally, this approach does not rely on any parametric assumptions on the return distributions of the two securities.

I then study three important applications of my approach. In the first application, I examine the cross-sectional equity premium associated with the sensitivity of stock returns to the market disaster risk. To capture this sensitivity, I construct a "systematic disaster risk" measure based on the conditional return distribution of a stock given the return of the market proxied by the S&P500 index. For both the market and the individual stocks, I define a "normal" state and a "disaster" state. Then, the systematic disaster risk of each stock is defined as the difference in the conditional disaster probabilities of the stock given that the market is in the disaster versus the normal states, respectively. This measure captures the extent to which an individual stock is more likely to be hit by a rare disaster when the market moves from the normal state to the disaster state.

Intuitively, if a stock is more sensitive to the market disaster risk, then it should be less desirable for investors to hold, especially during time periods when a market crash is considered likely. As such, investors should require higher expected returns for holding stocks with higher systematic disaster risk, and this effect should be more pronounced when the market disaster risk is high. To test this hypothesis, I apply the Fama and MacBeth (1973) methodology. I find that systematic disaster risk is not priced when the option implied
market disaster risk is low. However, when I restrict attention to time periods during which a market crash is perceived likely, then I find strong evidence that stocks with higher systematic disaster risk earn significantly higher expected returns after controlling for well documented risk factors. In fact, increasing the systematic disaster risk by one standard deviation raises expected monthly stock returns by 63 basis points, which is equivalent to over 7% per year.

My second application turns to the banking sector. I ask the question of how to estimate the exposure of banks to systemic shocks and what bank characteristics are related to banks’ systemic exposure. To this end, I construct a “systemic exposure” measure based on the conditional return distribution of a bank given the banking sector return, where the sector portfolio is empirically proxied by the KBW Bank Index. For both the sector portfolio and the individual banks, I define a “normal” state and a “disaster” state. I then estimate the systemic exposure of each bank as the difference in the conditional disaster probabilities of the bank given that the sector is in the disaster versus the normal states, respectively. Intuitively, the systemic exposure measure captures by how much an individual bank is more likely to experience a disaster when the whole banking sector falls from the normal state to the disaster state.

My estimates show that the systemic exposure measure is typically positive, indicating that banks are generally more likely to experience a disaster when the banking sector as a whole is in the disaster state relative to when the sector performs normally. I also find that the systemic exposure of a bank increases with its equity beta and the total return volatility. It is also increasing in the non-interest to interest income ratio, reflecting that a bank’s exposure to systemic shocks is largely driven by its non-traditional businesses. In addition, there is some evidence that systemic exposure decreases with total market capitalization.

Finally, in the third application I explore an extension of the recent Ross (2015) recovery theorem, which is aimed to recover the physical return distribution of the market portfolio from the corresponding risk-neutral distribution. While the recovery theorem deals with
the market portfolio only, I seek to extend it to recover the physical return distribution of an individual asset. I show that this can be achieved through the risk-neutral joint return distribution of the asset with the market portfolio, which is given by the product of the risk-neutral marginal return distribution of the market and the conditional return distribution of the asset given the market return. Since my approach generates an estimate for this conditional return distribution, it lends itself naturally to the extension of the Ross recovery theorem to individual assets.

The rest of the paper proceeds as follows. Section 2 reviews the literature. Section 3 introduces the setup, linking the return distribution of an individual asset to that of an aggregate index. Section 4 discusses the estimation methodology and how it can be implemented using option prices. Section 5 provides some discussions and extensions. Section 6 applies the framework to examine the cross-sectional systematic disaster risk premium. Section 7 studies banks’ exposure to systemic shocks. Section 8 shows how my approach can be used to extend the Ross recovery theorem to individual assets. Section 9 concludes. Proofs of propositions are shown in Appendix A, and other technical discussions are delegated to Appendix B.

3.2 Literature Review

The paper contributes to several strands of the literature. First, it adds to the study of the joint return behavior of different securities. Roll (1988), Jorion (2000), and Longin and Solnik (2001), among others, show that the return correlation between securities is not symmetric under all market conditions, but instead increases during market crashes. Ang and Chen (2002), Hong, Tu and Zhou (2007), and Jiang, Wu and Zhou (2014) develop methods to test this asymmetric dependence between security returns. Skinzi and Refenes (2004) and Driessen, Maenhout, and Vilkov (2013) propose methods of inferring equity return correlations from option prices. In this paper, I provide a novel approach of estimating the
entire conditional return distribution of an asset given the return of an aggregate index, thus accounting for potential asymmetries in their joint behavior. In addition, this approach captures all moment properties of their joint distribution, which is beyond the return correlation alone.

The paper also adds to the extensive literature on estimating the risk-neutral distributions of security returns using option pricing data. Ross (1976a) and Breeden and Litzenberger (1978) first show that one can extract the risk-neutral probability distributions of security returns from option prices. Since option prices are available at discrete strike prices and maturities only, some smoothing techniques are needed to estimate the full risk-neutral distribution. (See Melick and Thomas (1997), Posner and Milevsky (1998), and Rubinstein (1998) for parametric methods and Shimko (1993), Jackwerth and Rubinstein (1996), Malz (1997), and Ait-Sahalia and Lo (1998) for non-parametric methods.) Jackwerth (1999) provides a comprehensive review on various methods used to extract the risk-neutral return distributions from option prices. Figlewski (2010) provides an empirical demonstration based on the U.S. market portfolio. Overall, the literature has restricted attention to the return distributions of single securities. My paper extends this literature by introducing an approach of estimating the conditional return distribution of an asset given the return of an aggregate index using option prices. Under the assumption that the variation in the index return is the only priced risk, the risk-neutral conditional distribution estimated from option prices coincides with the physical conditional distribution.

This paper is also related to research on equity premium associated with disaster risk. A large body of theoretical research shows that investors are averse to rare disasters (e.g., Barro (2006, 2009), Gabaix (2008, 2012), Gourio (2012), Chen, Joslin, and Tran (2012), and Wachter (2013)). Consistent with this, a number of empirical papers have documented a positive relation between disaster risk and expected market returns (e.g., Bali, Demirtas, and Levy (2009), Bollerslev and Todorov (2011), and Kelly and Jiang (2014)). Cross-
sectionally, Siriwardane (2013) studies the relation between expected asset returns and the option-implied disaster risk of assets, and finds a positive premium. Van Oordt and Zhou (2012), Kelly and Jiang (2014), Ruenzi and Weigert (2013), on the other hand, focus on the systematic portion of disaster risk by looking at the disaster risk of an asset in relation to that of the market. Given the challenge of estimating the joint disaster risk due to the rare occurrence of disastrous events, all three papers use historical equity returns and resort to either the power law distribution or parametric copulas to model the tails, which unfortunately does not necessarily represent the true probability distributions. My paper contributes to this literature by suggesting a measure of systematic disaster risk based on the conditional return distribution of an asset given the market return. Since the measure is estimated using option prices, it is forward-looking and naturally captures rare disasters. This measure reflects investors’ perceived sensitivity of asset returns to the market disaster risk, which can be conveniently used to test the cross-sectional disaster risk premium.

This paper also adds to the literature on bank systemic risk. By definition, systemic risk focuses on risk associated with the collapse of the entire banking system. Hence, the main challenge of estimating systemic risk comes from the rare occurrence of disastrous events. Different methods have been proposed to tackle this problem. For example, Huang, Zhou, and Zhu (2009) measure systemic risk by the price of insurance against financial distress, in which the default correlation between banks is proxied by the equity return correlation. Acharya, Pedersen, Philippon and Richardson (2010) propose the systemic expected shortfall (SES) measure, which estimates the propensity of a bank to be undercapitalized when the system as a whole is undercapitalized. In particular, their method relies on the power law distribution to model the tails. Adrian and Brunnermeier (2011) propose the $\Delta CoVaR$ measure as the difference between the value-at-risk of the banking system conditional on an individual bank being in distress and the value-at-risk of the banking system conditional on the bank being solvent. Empirical estimation of $\Delta CoVaR$ uses quantile regression to
capture the tail distributions. The contribution of my paper is that it provides a measure of banks’ exposure to systemic shocks based on the conditional return distribution of a bank given the banking sector return. Since this measure is estimated using option prices, it is forward-looking and naturally captures investors’ perceived exposure of a bank to sector-wide disastrous shocks even if such shocks do not occur within sample.

Finally, the paper also contributes to the literature on Ross (2015) recovery theorem, which recovers the physical probability distribution of the market return from the associated risk-neutral distribution. Subsequent research has been done to further explore this problem. Carr and Yu (2012) provide alternative assumptions that allow for recovery for diffusions on a bounded state space. Huang and Shaliastovich (2013) develop a recursive-utility framework to separately identify physical probabilities and risk adjustments. Martin and Ross (2013) show that recovery can indeed be effected by observing the behavior of the long end of the yield curve. Walden (2013) extends the Ross recovery result to unbounded diffusion processes. See also Dubynskiy and Goldstein (2013) and Borovička, Hansen and Scheinkman (2014) for criticism of the Ross recovery theorem. This literature primarily focuses on the market portfolio. My approach contributes to this literature by extending the recovery results to any individual asset through its risk-neutral joint return distribution with the market.

3.3 Setup

In this section, I introduce a simple setup that links the return distribution of an individual asset with that of an aggregate index through the total probability formula. The return of the aggregate index can be viewed as a factor that determines the state of the economy. Examples of the index include the market portfolio or a sector portfolio, etc. The next section will discuss how this setup allows me to estimate the conditional return distribution of the asset given the index return by a regression approach.
There are $T$ discrete time points $t \in \{1, 2, \ldots, T\}$. At any time $t$, I consider security returns over one period ahead, that is, from $t$ to $t + 1$.

Consider an aggregate index $I$, whose return over any one period can take $N$ values

$$(r^I (1), r^I (2), \ldots, r^I (N)).$$

Denote by $\tilde{r}^I_{t,t+1}$ the random return of the index over the period from $t$ to $t + 1$. Evaluated at time $t$, the probability distribution of $\tilde{r}^I_{t,t+1}$ is given by the vector

$$p^I_{t,t+1} = (p^I_{t,t+1} (1), p^I_{t,t+1} (2), \ldots, p^I_{t,t+1} (N)),$$

where $p^I_{t,t+1} (n)$ represents the probability of $\tilde{r}^I_{t,t+1} = r^I (n)$ for any $n \in \{1, 2, \ldots, N\}$.

Consider an asset, whose return over any one period can take $K$ distinct values

$$(r (1), r (2), \ldots, r (K)).$$

Denote by $\tilde{r}_{t,t+1}$ the random return of the asset over the period from $t$ to $t + 1$. Evaluated at $t$, the probability distribution of $\tilde{r}_{t,t+1}$ is given by the vector

$$p_{t,t+1} = (p_{t,t+1} (1), p_{t,t+1} (2), \ldots, p_{t,t+1} (K)),$$

where $p_{t,t+1} (k)$ represents the probability of $\tilde{r}_{t,t+1} = r (k)$ for any $k \in \{1, 2, \ldots, K\}$.

**Assumption 50 (Identifying)** The conditional probability distribution of the asset return given any value of the contemporaneous index return does not vary over time.

This assumption may be understood in relation to the one we make when empirically estimating the CAPM beta that the conditional mean return of an asset given any value of the market return is fixed over time. My assumption is stronger in the sense that it requires not only the conditional mean but indeed the entire conditional distribution to be time-invariant. It implies that the time variation in the return distribution of the asset ($p_{t,t+1}$) is solely driven by the time variation in the return distribution of the index ($p^I_{t,t+1}$).
Denote the time-invariant conditional distribution of the asset return given the index return by the matrix
\[
\theta = \begin{pmatrix}
\theta(1|1) & \cdots & \theta(K|1) \\
\vdots & \ddots & \vdots \\
\theta(1|N) & \cdots & \theta(K|N)
\end{pmatrix},
\]
where \(\theta(k|n)\) stands for the conditional probability of \(\tilde{r}_{t+1} = r(k)\) given \(\tilde{r}_{t+1}^I = r^I(n)\) evaluated at the beginning of the period for any \(n \in \{1, 2, \ldots, N\}\) and \(k \in \{1, 2, \ldots, K\}\).

By Assumption 50, \(\theta\) does not depend on time. According to the properties of conditional probabilities, it must be that given any value of the index return, the conditional probabilities of the asset return sum up to one, i.e.,
\[
\sum_{k=1}^{K} \theta(k|n) = 1, \quad \forall n. \tag{3.1}
\]

At any \(t\), the marginal return distribution of the asset is related to that of the index by the total probability formula. Specifically, for any \(k\),
\[
p_{t+1}(k) = \sum_{n=1}^{N} p_{t+1}(n) \theta(k|n) .
\]
That is, the marginal distribution of the asset return is equal to the weighted average of its conditional distribution given the index return, with the weights given by the marginal return distribution of the index. This can be conveniently rewritten in matrix form as
\[
p_{t+1} = p_{t+1}^I \cdot \theta. \tag{3.2}
\]

The discussion so far is based on the physical probability measure. I now make an additional assumption to link the physical measure with the risk-neutral measure.

**Assumption 51** The variation in the index return is the only priced (systematic) risk.

Formally, Assumption 51 is satisfied if there exists a stochastic discount factor, whose value related to future payoffs depends on the future value of the index return only. This
assumption can be understood in relation to the CAPM framework, which implies that the stochastic discount factor is a linear function of the market return. Assumption 51 is weaker in the sense that it requires the stochastic discount factor to be a function of the index return only, but it does not impose any restriction on the functional form of this relation. This assumption implies that conditional on a particular value of the index return, any variation left in the asset return is purely idiosyncratic and hence is risk-neutrally priced. It is then straightforward to show that the conditional distribution of the asset return given any value of the index return is the same under both the physical and the risk-neutral probability measures. This is formally stated in the following proposition.

**Proposition 52** At any time $t$, the risk-neutral conditional probability of $\tilde{r}_{t,t+1} = r(k)$ given $\tilde{r}_{t,t+1} = r^I(n)$ is equal to $\theta(k|n)$ for all $n$ and $k$.

A conclusion of Proposition 52 is that the total probability formula (3.2) holds just as well under the risk-neutral probability measure with respect to the same conditional probability matrix $\theta$. At time $t$, denote the risk-neutral distributions of $\tilde{r}_{t,t+1}$ and $\tilde{r}_{I,t+1}$ by

$$
\mathbf{q}_{I,t+1} = \left(q^I_{I,t+1}(1), q^I_{I,t+1}(2), \ldots, q^I_{I,t+1}(N)\right),
$$

$$
\mathbf{q}_{t,t+1} = \left(q_{t,t+1}(1), q_{t,t+1}(2), \ldots, q_{t,t+1}(K)\right).
$$

Formally,

**Corollary 53** At any time $t$,

$$
\mathbf{q}_{t,t+1} = \mathbf{q}_{I,t+1} \cdot \theta. \tag{3.3}
$$

### 3.4 Estimation Methodology

The setup introduced in Section 3.3 can be used to estimate the conditional distribution matrix $\theta$ from the marginal return distributions of the two securities. Proposition 52 and
Corollary 53 suggest that for this purpose one can work under either the physical measure or the risk-neutral measure, and the conditional probabilities obtained under both measures would be identical. In practice, there can be different ways of estimating the marginal return distributions in either probability measure. In this paper, I choose to do so under the risk-neutral measure using option prices based on the work of Ross (1976a) and Breeden and Litzenberger (1978). My approach would work in the same manner if the marginal return distributions are obtained using other methods.

I assume that both index $I$ and the individual asset of interest are traded in the option market. Examples of aggregate indices with traded options include the S&P500 index and the KBW Bank Index, etc. The procedure of estimating $\theta$ includes two steps. In the first, I extract the risk-neutral marginal return distributions of the index and the asset, $q_{I,t+1}^t$ and $q_{t,t+1}$, from option prices. Then in the second step, I perform a constrained regression of $q_{t,t+1}$ on $q_{I,t+1}^t$ over time to estimate the conditional distribution matrix $\theta$. Below I discuss each of the two steps separately.

### 3.4.1 Extracting Risk-Neutral Marginal Distributions from Option Prices

I first discuss how the risk-neutral return distributions $q_{I,t+\tau}^t$ and $q_{t,t+\tau}$ can be extracted from option prices. The estimation procedures for $q_{I,t+1}^t$ and $q_{t,t+1}$ are parallel, and hence in this section I focus on $q_{t,t+1}$ for brevity.

Ross (1976a) and Breeden and Litzenberger (1978) show that given a continuous range of strike prices covering all possible values of the underlying asset at maturity, the entire risk-neutral probability distribution of the asset’s future value can be estimated from European option prices. Suppose that $t$ is the current time point and consider an European put option that matures at time $t + 1$. Let $S_t$ represent the current price of the underlying asset, and let $\tilde{S}_{t+1}$ be the random price of the asset in one period. Denote the strike price of the option
by $X$ and the risk-free rate by $r_f$. The price of the put option can then be expressed as a function of the strike price:

$$
\text{Put}(X) = e^{-r_f} \int_{0}^{\infty} (X - S_{t+1})^+ dF(S_{t+1})
$$

(3.4)

$$
= e^{-r_f} \int_{0}^{X} (X - S_{t+1}) dF(S_{t+1}),
$$

where $F(\cdot)$ is the risk-neutral cumulative distribution function (CDF) of $\tilde{S}_{t+1}$ evaluated at time $t$. Differentiating (3.4) with respect to $X$ obtains

$$
\frac{\partial \text{Put}(X)}{\partial X} = e^{-r_f} F(X).
$$

Solving for $F(X)$ leads to

$$
F(X) = e^{r_f} \frac{\partial \text{Put}(X)}{\partial X}.
$$

(3.5)

Evaluating $F(X)$ at all possible values of $X$ thus yields the entire risk-neutral distribution of the asset price at maturity.\(^1\)

Since I am interested in the risk-neutral distribution of the asset return from time $t$ to $t + 1$, I need to relate the return to the price at maturity. Assume that the dividend yield paid by the asset from $t$ to $t + 1$ is equal to $d$. Then, the return of the asset $\tilde{r}_{t,t+1}$ is related to the future asset price $\tilde{S}_{t+1}$ according to the following approximation

$$
\tilde{S}_{t+1} = S_t (1 + \tilde{r}_{t,t+1} - d).
$$

(3.6)

Therefore, the risk-neutral CDF of the asset return is given by $\forall r$,

$$
G(r) = F(S_t (1 + r - d)).
$$

Section 3.3 assumed that the asset return took a finite number ($K$) of values. This, however, is a simplification of the real world in which asset returns have continuous ranges.

\(^1\)One can alternatively estimate $F(\cdot)$ based on prices of European call options. By the call-put parity, the results using call and put options are identical. For simplicity, I choose to work with put options.
To be consistent with the setup, I discretize the continuous asset return by dividing its range into $K$ mutually disjoint intervals with $K - 1$ thresholds. At each time $t$, the risk-neutral probabilities that the one-period asset return lies in each of these $K$ intervals are taken as elements of the vector $q_{t,t+1}$. In particular, let $rr(1), rr(2), \ldots, rr(K - 1)$ denote the $K - 1$ thresholds separating the $K$ intervals of the asset return. Then, the vector $q_{t,t+1}$ can be estimated as

$$
q_{t,t+1}(1) = G(rr(1)),
$$

$$
q_{t,t+1}(k) = G(rr(k)) - G(rr(k - 1)), \forall k = 2, 3, \ldots, K - 1,
$$

$$
q_{t,t+1}(K) = 1 - G(rr(K - 1)).
$$

Two additional technical issues need to be dealt with for the empirical estimation of $q_{t,t+1}$. First, the estimation of the risk-neutral CDF (3.5) relies on differentiating the option price with respect to the strike price. Since it is generally very difficult to obtain a close-form expression for this derivative, I estimate it by linear approximation. A second empirical challenge has to do with obtaining European option prices. Nearly all individual stock options are American options. While indices are generally represented by European options, the market option prices are only available at discrete values of the strike price and time to maturity. To obtain the European option price for any security at any arbitrary point, I adopt a simple and commonly used approach of first fitting the implied volatility surface by kernel smoothing and then deriving the Black-Merton-Scholes (BMS) option price (Black and Scholes (1973) and Merton (1973)) using the fitted volatility. I delegate detailed discussions of these issues to Appendix B.

### 3.4.2 Estimating Conditional Distributions by Constrained Regression

This section discusses how to estimate the conditional distribution matrix $\theta$ based on the marginal return distributions $q_{t,t+\tau}^I$ and $q_{t,t+\tau}$. Since $q_{t,t+\tau}^I$ and $q_{t,t+\tau}$ are extracted from
option prices, they are often subject to measurement errors. In particular, since options
written on individual assets are more thinly traded than index options, one would expect
\( q_{t,t+\tau} \) to be much noisier than \( q_{t,t+\tau}^I \). To reflect the different degrees of noisiness in \( q_{t,t+\tau}^I \)
and \( q_{t,t+\tau} \), I assume that \( q_{t,t+\tau}^I \) can be accurately estimated and that \( q_{t,t+\tau} \)
contains noises, which are captured by an error term

\[
\epsilon_{t,t+1} = (\epsilon_{t,t+1}(1), \epsilon_{t,t+1}(2), \ldots, \epsilon_{t,t+1}(K)).
\]

Now the risk-neutral total probability formula (3.3) becomes

\[
q_{t,t+1} = q_{t,t+1}^I \cdot \theta + \epsilon_{t,t+1}.
\] (3.7)

I will estimate \( \theta \) from \( q_{t,t+\tau}^I \) and \( q_{t,t+\tau} \) using a constrained linear regression based on
(3.7). Before discussing the detailed estimation procedures, I need to make some additional
assumptions, which are sufficient to maintain the consistency of my estimates. In particular,
I assume the following.

**Assumption 54** The pair of vectors \( \{q_{t,t+1}^I, q_{t,t+1}\} \) are jointly stationary and weakly de-
pendent over time.\(^2\)

**Assumption 55** At any time \( t \), \( E[\epsilon_{t,t+1}(k) | q_{t,t+1}] = 0 \) for all \( k \in \{1, 2, \ldots, K\} \), where the
expectation is taken under the physical probability measure.

**Assumption 56** The \( T \times N \) matrix

\[
Q^I = \begin{pmatrix}
q_{1,2}^I \\
q_{2,3}^I \\
\vdots \\
q_{T,T+1}^I
\end{pmatrix}
\]

is of rank \( N \).

\(^2\)The pair of vectors \( \{q_{t,t+1}^I, q_{t,t+1}\} \) are weakly dependent over time if for any \( t \), \( \{q_{t,t+1}^I, q_{t,t+1}\} \) and
\( \{q_{t+\Delta t,t+\Delta t+1}^I, q_{t+\Delta t,t+\Delta t+1}\} \) become approximately independent as \( \Delta t \to \infty \).
To see how $\theta$ can be estimated by linear regression, it is useful to rewrite (3.7) as

$$
q_{t,t+1}(k) = \sum_{n=1}^{N} q_{t,t+1}^I(n) \theta(k|n) + \epsilon_{t,t+1}(k), \ \forall k.
$$

This indicates that one can estimate $(\theta(k|1), \theta(k|2), \ldots, \theta(k|N))^t$ (the $k^{th}$ column of the conditional distribution matrix $\theta$) by running an OLS regression of $q_{t,t+1}(k)$ on the vector $p_{t,t+1}^I$ over time. Assumptions 54–56 guarantee that the resulting OLS estimates are consistent, i.e., they converge to the true parameter values when the sample size approaches infinity.

Then, to estimate the entire $\theta$ matrix, a natural idea would be to run a total of $K$ regressions corresponding to each of the $K$ values of the asset return. However, since $\theta$ represents the conditional probabilities, two implicit constraints must be satisfied. First is that the conditional probabilities given any value of the index return must sum up to one (as required by (3.1)), and the second constraint says that all elements of $\theta$ must lie between 0 and 1, i.e., $0 \leq \theta(k|n) \leq 1$ for all $k$ and $n$. Unfortunately, running $K$ OLS regressions independently does not guarantee that these constraints are satisfied.

A solution to this issue is to conduct the $K$ regressions jointly subject to the above two constraints. Formally, define

$$
Q = \begin{pmatrix}
q_{1,2} \\
q_{2,3} \\
\vdots \\
q_{T,T+1}
\end{pmatrix},
$$

$$
\epsilon = \begin{pmatrix}
\epsilon_{1,2}(1) & \cdots & \epsilon_{1,2}(K) \\
\vdots & \ddots & \vdots \\
\epsilon_{T,T+1}(1) & \cdots & \epsilon_{T,T+1}(K)
\end{pmatrix},
$$

and let $1_{a \times b}$ and $0_{a \times b}$ denote the $a \times b$ matrices of ones and zeros for any positive integers $a$ and $b$, respectively. Then, the problem can be represented by the following constrained
linear regression

\[ Q = Q^I \cdot \theta + \epsilon, \quad (3.8) \]

\[ s.t. \]
\[ \theta \cdot 1_{K \times 1} = 1_{N \times 1}, \]
\[ 0_{N \times K} \leq \theta \leq 1_{N \times K}. \]

I denote the resulting estimates from problem (3.8) by \( \hat{\theta}_T \), where the subscript \( T \) reflects dependence of the estimates on the sample size.

A priori, it is not clear whether imposing the constraints would affect the consistency of my estimation. The following proposition establishes that consistency is indeed preserved in the presence of the constraints.

**Proposition 57** Under Assumptions 54-56, \( \hat{\theta}_T \) is a consistent estimator of \( \theta \), i.e.,

\[ \lim_{T \to \infty} \hat{\theta}_T = \theta. \]

### 3.5 Discussions and Extensions

This section provides some discussions and extensions of the estimation framework introduced above.

#### 3.5.1 Elaboration on Key Assumptions

Assumptions 50 and 51 are key to my approach in that they point out two important roles of the index return.

Assumption 50 states that the conditional return distribution of the asset given the index return is time invariant. The intuition is that while the return distribution of the asset can change over time, its variation is solely driven by the time variation in the distribution of the index return. In particular, once the index return is fixed, the conditional distribution
of the asset return is also fixed, regardless of the time point under consideration. This is the identifying assumption of my approach in that it allows me to make use of the time series information on the marginal return distributions of the two securities to determine the time-invariant conditional probabilities. Without this assumption, my estimation model is not identified.

Another important role of the index return is reflected in Assumption 51, which states that the variation in the index return is the only priced (systematic) risk. As such, fixing a certain value of the index return, any variation left of the asset return is purely idiosyncratic and is thus risk-neutrally priced. This assumption implies that the conditional return distribution of the asset given the index return is the same under the physical and the risk-neutral probability measures. Since investors are averse to risk, security return distributions generally differ under the physical versus the risk neutral measures to reflect the adjustment for risk aversion. In fact, there is a recent literature on Ross (2015) recovery theorem that aims to recover the physical return distributions from the associated risk-neutral distributions. The benefit of Assumption 51 is that it aligns the analyses under the two probability measures once I condition on a particular value of the index return. This allows me to perform empirical estimation under the risk-neutral measure using option prices, and the resulting conditional probabilities would be exactly the same as if I work under the physical measure. However, the failure of this assumption does not necessarily invalidate my approach. Even when this assumption is violated (e.g., when the Fama-French three-factor pricing model holds), my approach can still be applied under either the physical or the risk-neutral measure, but the conditional distribution of the asset return would no longer be the same under the two probability measures.
3.5.2 Multi-Factor Framework

The baseline model discussed earlier is a one-factor framework, in which the index return is the only factor that determines the asset return distributions. In practice, asset return distributions can be affected by more than one factor. Macroeconomic variables such as the consumption growth rate, inflation rate, or VIX may serve as additional factors. In this case, I need to extend my model into a multi-factor framework.

Suppose that there exist $M$ factors. Each factor $m \in \{1, 2, \ldots, M\}$ takes $N^m$ values. Evaluated at time $t$, the joint distribution of the $M$ factors at time $t+1$ is given by the joint distribution function $p_{t,t+1}^I (n^1, n^2, \ldots, n^M)$, where $n^m \in \{1, 2, \ldots, N^m\}$ for every $m$. Similar to Assumption 50, I assume that given any set of joint values of these $M$ factors, the conditional return distribution of an asset does not vary over time, which is denoted by the conditional distribution function $\theta (\cdot | n^1, n^2, \ldots, n^M)$. Then, the total probability formula links the marginal return distribution of the asset to the joint distribution of the $M$ factors through $\theta (\cdot | n^1, n^2, \ldots, n^M)$, i.e.,

$$p_{t,t+1} (k) = \sum_{n^1=1}^{N^1} \sum_{n^2=1}^{N^2} \cdots \sum_{n^M=1}^{N^M} p_{t,t+1}^I (n^1, n^2, \ldots, n^M) \theta (k | n^1, n^2, \ldots, n^M).$$

If I further assume that variations in these $M$ factors constitute the only priced (systematic) risk (the multi-factor version of Assumption 51), then $\theta (\cdot | n^1, n^2, \ldots, n^M)$ is the same under the physical and the risk-neutral measures. This allows me to write down the risk-neutral total probability formula as

$$q_{t,t+1} (k) = \sum_{n^1=1}^{N^1} \sum_{n^2=1}^{N^2} \cdots \sum_{n^M=1}^{N^M} q_{t,t+1}^I (n^1, n^2, \ldots, n^M) \theta (k | n^1, n^2, \ldots, n^M),$$

where $q_{t,t+1}^I (n^1, n^2, \ldots, n^M)$ represents the risk-neutral joint distribution of the $M$ factors at $t+1$ evaluated at time $t$.

If I have the marginal return distribution of the asset and the joint distribution of the factors under either the physical or the risk neutral measure, I can estimate the conditional
distribution function $\theta (k \, |n^1, n^2, \ldots, n^M)$ by regressing the former on the latter over time, as in the baseline framework. Unfortunately, the risk-neutral joint distribution of the factors can no longer be directly extracted from option prices, because options written on the joint values of multiple factors are generally not available. As a result, one need to resort to other approaches to obtain the joint distribution of the factors. Once this joint distribution is obtained, I can estimate $\theta (k \, |n^1, n^2, \ldots, n^M)$ in exactly the same manner as in the baseline case.

### 3.5.3 Continuous Security Returns

Up till now, I have assumed that the returns of both the individual asset and the aggregate index take a finite number of discrete values. This section considers the case of continuous security returns. Given the analogy between the physical and risk-neutral analyses (as in the baseline case), in this section I work directly with the risk-neutral measure.

At any time $t$, suppose that the one-period index return $\tilde{r}_{t,t+1}^I$ and the one-period asset return $\tilde{r}_{t,t+1}$ take continuous values from the interval $[-1, \infty)$. The marginal probability distributions of $\tilde{r}_{t,t+1}^I$ and $\tilde{r}_{t,t+1}$ are given by the density functions $q_{t,t+1}^I (\cdot)$ and $q_{t,t+1} (\cdot)$, respectively. By Assumption 50, given any value of $\tilde{r}_{t,t+1}^I$ the conditional distribution of $\tilde{r}_{t,t+1}$ does not change over time. I denote this time-invariant conditional distribution by the conditional density function $\theta (\cdot | \cdot)$, which integrates to 1 given any $\tilde{r}_{t,t+1}^I = r^I$, i.e.,

$$\int_{-1}^{\infty} \theta (r | r^I) dr = 1, \forall r^I.$$  

At any $t$, the marginal return distributions of the two securities are linked to each other by the total probability formula

$$q_{t,t+1} (r) = \int_{-1}^{\infty} q_{t,t+1}^I (r^I) \theta (r | r^I) dr^I.$$  

Assume that $q_{t,t+1}^I (\cdot)$ can be accurately measured, whereas $q_{t,t+1} (\cdot)$ is subject to noises,
which are captured by the error term $\epsilon_{t,t+1}(\cdot)$. Then, the total probability formula becomes

$$q_{t,t+1}(r) = \int_{-1}^{\infty} q_{t,t+1}^I(r^I) \theta(r|r^I) \, dr^I + \epsilon_{t,t+1}(r).$$

Since $\theta(r|r^I)$ is infinite-dimensional, its empirical estimation is difficult without further information on the structure of $\theta(r|r^I)$. While there are different ways of reducing dimensionality, one of the simplest methods is to make parametric assumptions on the functional form of $\theta(r|r^I)$. Specifically, let

$$\theta(r|r^I) = g(r,r^I; \lambda),$$

where $g(r,r^I; \lambda)$ is the assumed functional form of $\theta(r|r^I)$ and $\lambda$ represents the parameters of choice. Then, one can estimate $\theta(r|r^I)$ by choosing the values of $\lambda$ to solve the following least square problem:

$$\min_{\lambda} \left( \sum_{t=1}^{T} \int_{-1}^{\infty} \left[ q_{t,t+1}(r) - \int_{-1}^{\infty} q_{t,t+1}^I(r^I) g(r,r^I; \lambda) \, dr^I \right]^2 \, dr \right),$$

s.t.

$$\int_{-1}^{\infty} g(r,r^I; \lambda) \, dr = 1, \forall r^I,$$

$$g(r,r^I; \lambda) \geq 0, \forall r^I, r.$$

### 3.5.4 Alternative Econometric Models

In Section 3.4.2, I used a constrained linear regression to estimate the conditional distribution matrix $\theta$ from the marginal return distributions of the two securities. In fact, there are some alternative econometric models (rather than the constrained linear regression) that may also seem appealing for my purpose. I discuss the potentials and limitations of some alternatives in this section.

**Probit and Logit Models**

The Probit and Logit models both can be used to predict the probability distribution of an outcome variable based on the values of the independent variables. The two models
differ in the assumed distribution of the error term. The benefit of these models is that they automatically guarantee that the estimated probabilities of the outcome variable lie between zero and one. It may seem that the Probit and the Logit models are well suited for my purpose. However, a key difference is that in these two models, the dependent variable is a discrete variable representing the outcome of an event. In contrast, in my case the dependent variable itself is the probability distribution of the asset return. Therefore, the Probit and the Logit models do not apply here. In additional, to obtain reasonable results I impose two constraints on my estimates, requiring that each conditional probability lie between zero and one and that they sum up to one given any particular value of the index return. It is not trivial to incorporate these constraints into the Probit and Logit models. In fact, it is not hard to see that once these constraints are met, the predicted marginal probabilities of the asset return based on the current linear model are guaranteed to lie between zero and one with no need for additional restrictions.

**Maximum Likelihood Estimation**

Another alternative econometric approach worth mentioning is the Maximum Likelihood Estimation (MLE). Suppose that the joint distribution of the error terms \( \epsilon_{t,t+1} = (\epsilon_{t,t+1}(1), \epsilon_{t,t+1}(2), \ldots, \epsilon_{t,t+1}(K)) \) is given by the joint density function \( \Lambda(\cdot) \). If \( \epsilon_{t,t+1} \) is independent and identically distributed over time, then the conditional distribution matrix \( \theta \) can be estimated by the following constrained MLE:

\[
\max_{\theta} \prod_{t=1}^{T} \Lambda \left( \mathbf{q}_{t,t+1} - \mathbf{q}'_{t,t+1} \cdot \theta \right),
\]

s.t.

\[
\theta \cdot \mathbf{1}_{K \times 1} = \mathbf{1}_{N \times 1},
\]

\[
\mathbf{0}_{N \times K} \leq \theta \leq \mathbf{1}_{N \times K}.
\]

The key here is the joint density function \( \Lambda(\cdot) \). It is not clear what the best assumption would be for the joint distribution of \( \epsilon_{t,t+1} \). The normal distribution, for instance, may not be
a good choice. This is because both the true and the estimated marginal return distributions of the asset \( q_{t,t+1} \) have bounded values, and hence the associated measurement errors \( \epsilon_{t,t+1} \) should also be bounded, which is clearly not the case for normally distribution variables. In addition, it is also likely that the different elements of \( \epsilon_{t,t+1} \) are correlated with each other, rendering the assumption on \( \Lambda(\cdot) \) even more complicated.

While the constrained linear regression model adopted in this paper seems simple, I will provide evidence for the out-of-sample validity of my estimation in the applications to be discussed in the following sections.

### 3.6 Application I: Systematic Disaster Risk Premium

Starting from the seminal work of the Capital Asset Pricing Model (CAPM), independently developed by Sharpe (1964), Lintner (1965a,b), and Mossin (1966), researchers have found that the cross-sectional risk-return relation is driven by the comovement of individual asset returns with the market return, which is usually termed “systematic risk.” Rubinstein (1973) and Kraus and Litzenberger (1976) extend the CAPM framework, which focuses on the second moment of security returns, to account for higher moments. More recently, Kadan, Liu, and Liu (2015) propose a general framework of evaluating systematic risk for a broad class of risk measures, potentially accounting for various risk attributes such as high distribution moments, downside risk and rare disasters.

On the other hand, a large body of theoretical research shows that investors are averse to rare disasters (e.g., Barro (2006, 2009), Gabaix (2008, 2012), Gourio (2012), Chen, Joslin, and Tran (2012), and Wachter (2013)). Consistent with this, a number of empirical papers have documented a positive relation between disaster risk and expected market returns (e.g., Bali, Demirtas, and Levy (2009), Bollerslev and Todorov (2011), and Kelly and Jiang (2014)).

Given that investors exhibit aversion to rare disasters and that they are concerned with
the comovement of asset returns with the market, it is then natural to conjecture that investors require higher compensation for holding assets that are more sensitive to the market disaster risk. This idea is empirically tested in the literature by Van Oordt and Zhou (2012), Kelly and Jiang (2014), and Ruenzi and Weigert (2013). All three papers show that stocks with higher sensitivity to the market disaster risk earn higher expected returns, at least during some time period. Given the challenge of estimating disaster risk due to the rare occurrence of disastrous events, all of these papers use historical returns and model the tails by either the power law distribution or parametric copulas.

In this section, I propose a “systematic disaster risk” measure, which captures the sensitivity of asset returns to the market disaster risk. The measure is constructed based on the conditional return distribution of an asset given the market return, which can be readily estimated using my approach by taking the market portfolio as the related index. My measure is estimated using option prices rather than historical security returns. Thus, it is forward-looking and naturally accounts for investors’ beliefs on the likelihood of rare events even if such events do not occur within sample. I empirically examine whether this measure predicts the cross-section of expected stock returns.

3.6.1 Measure of Systematic Disaster Risk

I construct the systematic disaster risk measure of a stock based on its conditional return distribution given the market return. For both the stocks and the market portfolio, I focus on three-month returns, whose associated risk-neutral distributions can be extracted from prices of options maturing in three months. The reason for using three-month returns is because options with a maturity around three months are considered to have the most accurate prices.

For both the market portfolio and the individual stocks, I define two states \( (N = K = 2) \), the normal state \( (H) \) and the disaster state \( (L) \). I choose the disaster thresholds for the
three-month returns of the market and individual stocks to be $-1/3$ and $-1/2$, respectively. In words, the market (a stock) experiences a disaster if it loses more than one third (one half) of its original value within three months. The disaster threshold for individual stocks is chosen to be lower than that for the market portfolio, because individual stock returns are more volatile than the market return. I will later show that the choice of the disaster thresholds is consistent with our notion of rare disasters.

For each stock, I define the systematic disaster risk measure as

$$SysDis = \theta (L|L) - \theta (L|H),$$

where $\theta (L|L)$ and $\theta (L|H)$ stand for the disaster probabilities of the stock conditional on the market being in the disaster and normal states, respectively. Intuitively, this measure captures the extent to which an individual stock is more likely to experience a disaster when the whole market moves from the normal state to the disaster state. In other words, it is a measure of the sensitivity of individual stock returns to the market disaster risk.

Since $\theta (L|L)$ and $\theta (L|H)$ are conditional probabilities and take values from 0 to 1, it is apparent that the systematic disaster risk measure $SysDis$ ranges from -1 to 1. In particular, a positive value means that a stock is more likely to have a disaster when the market as a whole is in the disaster state relative to when the market performs normally. The higher the value of $SysDis$, the more sensitive the stock is to the market disaster risk. On the other hand, a negative $SysDis$ implies that the stock is less likely to experience a disaster when the market falls into the disaster state, and therefore can be viewed as a hedge against the market disaster risk.

### 3.6.2 Data

My sample period is from January 1996 to December 2012. I start from year 1996 because option pricing information is not available before then. I use the S&P500 index as a proxy for the market portfolio. This index has actively traded call and put options that cover a wide
range of moneyness and time to maturity levels. I focus on all firms traded on NYSE, AMEX, and NASDAQ with option prices available. My primary source of data is the OptionMetrics database, which contains multiple data sets and provides me with the following information for both the S&P500 index and the individual stocks.

- The *option prices* data set contains daily records of the implied volatility and option vega for all available options with a variety of strike prices and expiration dates. I use this information to fit the implied volatility surface by kernel smoothing. (See Appendix B for details.)

- The *security prices* data set has daily security prices and returns.

- The *index dividend yield* data set provides me with annualized dividend yield for the S&P500 index.

- The *dividend distribution history* data set has the dates and amounts of dividend payments for individual stocks. On each date $t$, I estimate the dividend yield for each stock over the three-month period ahead as the ratio of the total dividend payments during the period to the stock price on date $t$.

- The *zero coupon yield curve* data set contains continuously compounded zero-coupon interest rates for various numbers of days to maturity. The zero-coupon rate is used as a proxy for the risk-free rate in the calculation of the BMS option prices. Since I focus on prices of options with a maturity of three months, I need the three-month interest rate. When the interest rate with this maturity level is not directly given in the data set, I estimate it by linear approximation using rates for adjacent numbers of days to maturity available.

In addition to the above information from OptionMetrics, I also collect the following data. First, I obtain monthly returns for all stocks during the sample period from the
Center for Research in Security Prices (CRSP) database. These monthly returns are used as the dependent variable in the cross-sectional test. In addition, to control for the CAPM beta in my test, I obtain daily excess returns of the CRSP value-weighted market portfolio from the Kenneth French online data library, which are useful for the estimation of beta. Finally, to control for the “value” effect documented in Fama and French (1993), I include the book-to-market ratio as a risk factor in the test. For this sake, I obtain the book value of equity for all stocks from the Compustat database.

3.6.3 Empirical Strategy

My hypothesis states that stocks with higher systematic disaster risk earn higher expected returns. In addition, I conjecture that this relation should be more pronounced when the market disaster risk becomes higher, i.e., when a market crash is considered to be more likely to happen. The intuition is as follows. Suppose that the market performs well and that the probability of a market crash is very low. Then, even if a stock could be sensitive to the market disaster risk, investors may not be much concerned given that a market disaster is very unlikely to happen. On the other hand, when investors believe that the probability of a market crash is considerably high, they may have strong incentives to avoid investing in stocks with high systematic disaster risk. This is because it now becomes very likely that investments in these stocks could incur large losses upon the occurrence of a market crash.

To test these hypotheses, I follow the standard approach of Fama and MacBeth (1973) and Fama and French (1992). For each month within the sample period, I perform a cross-sectional regression of monthly excess stock returns on the lagged values of various stock characteristics. I then test the risk premium associated with a certain characteristic by conducting a $t$-test on the corresponding coefficient estimates across all months. To see how the systematic disaster risk premium depends on the market condition, I define a market disaster risk variable $Dis^M$, estimated annually as the average option-implied risk-neutral
disaster probability of the market portfolio across all dates of the year. Apparently, a high value of $Dis^M$ indicates that investors’ perceived likelihood of a market crash is high. I then repeat the test using subsamples constructed based on the lagged value of $Dis^M$, and see how the relation between systematic disaster risk and expected stock returns changes across different subsamples.

The main stock characteristic of interest is the systematic disaster risk $SysDis$ defined in Section 3.6.1. I estimate this variable for each stock on an annual basis in three steps. First, on each date, I compute the option-implied risk-neutral probabilities of disaster for both the S&P500 index and each individual stock. For stocks with a low level of option trading around the disaster threshold, the limited availability of option prices can potentially affect the accuracy of estimation. To enhance accuracy, I compute the risk-neutral disaster probability for a stock on a date only when (1) there are at least ten different options written on the stock with pricing information available on that date, (2) the lowest moneyness level (ratio of strike price to stock price) of available options for the stock on that date is no higher than 0.95, and (3) the shortest (longest) time to maturity of available options for the stock on that date is lower (higher) than three months. In the second step, I run the constrained linear regression (3.8) for each stock-year using all dates within the year to obtain the conditional disaster probabilities of the stock given different states of the market. To improve accuracy, I perform this step only for stocks with risk-neutral disaster probability available for no less than 200 days during the year. Finally, I compute $SysDis$ using the definition (3.9).

To disentangle the effect of the systematic disaster risk on the expected stock return from that of the “unconditional” propensity of a stock to experience a disastrous event, I control for each stock’s unconditional disaster risk ($Dis$) in the cross-sectional regressions. The $Dis$ variable is estimated annually for each stock as the average option-implied risk-neutral disaster probability of the stock across all dates during the entire year.

To further control for other risk factors, I include the following stock characteristic vari-
ables in the test:

- CAPM beta ($Beta$), estimated annually using daily equity returns. Following Amihud (2002), I estimate beta by size portfolios. At the end of each year, I rank all stocks into ten size portfolios based on NYSE breakpoints.\(^3\) For each size portfolio, I compute the portfolio return as the equal-weighted average return of all stocks in the portfolio, based on which I then estimate the portfolio beta. Finally, I assign the portfolio beta to all stocks in the portfolio.

- Coskewness ($Coskew$), estimated annually using daily equity returns. Following Harvey and Siddique (2000), I measure coskewness as the covariance of stock excess returns and the square of market excess returns.

- Firm size ($Size$), estimated monthly as the logarithm of the total market capitalization.

- Book-to-market ratio ($B2M$), estimated annually as the ratio of the end-of-year book value of equity to the end-of-year market capitalization.

- Lagged stock return, proxied by the stock return ($Ret$) of the previous month. This is to capture the momentum effect documented in Jegadeesh and Titman (1993).

- Mean-adjusted illiquidity ($IlliqMA$), estimated annually for each stock according to Amihud (2002) as the annual average of the ratio of absolute daily return to daily dollar volume of trading, adjusted by the cross-sectional mean.

Notice that the variables $SysDis$, $Dis$, $Beta$, $Coskew$, $B2M$, and $IlliqMA$ are all estimated on an annual basis. Therefore, in the cross-sectional regressions, I always regress excess stock returns in a particular month on the values of these variables estimated from the previous year. On the other hand, since $Size$ and $Ret$ are available monthly, I use the

\(^3\)As pointed out in Fama and French (1992), the idea of using NYSE breakpoints is to avoid having most size portfolios dominated by small stocks traded on NASDAQ.
one-month lagged values of these two variables in the cross-sectional regressions to capture the size and momentum effects.

### 3.6.4 Results

To make sure that my estimation of $SysDis$ generates reasonable results, I first check the out-of-sample validity by trying to predict future states of stocks using my conditional disaster probability estimation. In particular, for each year I take the first quarter and try to predict the state of a stock over the quarter based on the realized state of the market during the same quarter and the conditional disaster probabilities estimated from the previous year. Year 2008 is a special year in my sample, because towards the end of the year the market indeed experienced a rare disaster with the three-month S&P500 return lower than -1/3. This turns out to be the only occurrence of a market crash throughout the entire sample period according to my definition in Section 3.6.1. Hence, for this year instead of focusing on the first quarter, I try to predict the state of each stock over the 91-day period from August 22, 2008 to November 20, 2008, during which the S&P500 return was -41%.

For each quarter-length period mentioned above, I determine the states of each individual stock and the S&P500 index ($H$ or $L$) over the period based on their returns using the definition of Section 3.6.1. Then, for each stock and each quarter-length period, I compare two quantities: $1_{\text{stock in state } L}$, a dummy variable that takes on a value of 1 if the stock delivers a disastrous return of over the period and 0 otherwise, and $\hat{\theta}(L|\text{realized state of S&P500})$, the conditional disaster probability of the stock estimated from the previous year given the realized state of S&P500 over the period. I first check the correlation between these two quantities and find that they exhibit a strong positive correlation of 0.52. This indicates that the higher the predicted conditional disaster probability, the higher the actual occurrence of a firm disaster.
I then compute the following difference

$$\Delta = 1_{\text{stock in state } L} - \hat{\theta}(L | \text{realized state of S&P500}).$$

Since $1_{\text{stock in state } L}$ is either zero or one depending on whether a firm disaster occurs, whereas $\hat{\theta}(L | \text{realized state of S&P500})$ represents the conditional disaster probability, their difference $\Delta$ for any one occasion does not contain much information. However, when I take the average across different stocks and periods, according to the law of large numbers these two quantities should be close to each other if my estimations are meaningful. Motivated by this idea, I compute the average of $\Delta$ for normal and disaster states of the market separately.

Conditional on a realized disaster market state (August 22, 2008 to November 20, 2008) and a realized normal market state (all other quarter-length periods), the average values of $\Delta$ are -0.01 and 0.13, respectively. These numbers are reasonably small in magnitude, providing evidence for good out-of-sample validity.

Next I compute summary statistics for the disaster risk variables $Dis_M^M$, $Dis$ and $SysDis$. The results are reported in Figure 3.3. The mean value of the market disaster risk $Dis_M^M$ across the sample period is around 0.013, implying that on average a market crash is believed to occur with a risk-neutral probability of 0.013. In addition, the standard deviation, median, minimum, and maximum of $Dis_M^M$ are estimated as 0.010, 0.012, 0.002, and 0.033, respectively. For the unconditional disaster risk ($Dis$) and the systematic disaster risk ($SysDis$) of individual stocks, I compute the mean and standard deviation across all stocks for each single year, and I report summary statistics of these annual cross-sectional means and standard deviations in the figure. In particular, the mean value of the annual cross-sectional mean of $Dis$ is equal to 0.031, indicating that on average individual stocks are believed to experience a disastrous event with a risk-neutral probability of 0.031. The mean value of the annual cross-sectional mean of $SysDis$ is around 0.335, suggesting that the conditional disaster probability of an individual stock given a market disaster is on average higher by
<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean of annual mean</th>
<th>Mean of annual S.D.</th>
<th>Median of annual mean</th>
<th>Min of annual mean</th>
<th>Max of annual mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>SysDis</td>
<td>0.3346</td>
<td>0.3643</td>
<td>0.2886</td>
<td>0.1164</td>
<td>0.6910</td>
</tr>
<tr>
<td>Dis</td>
<td>0.0314</td>
<td>0.0377</td>
<td>0.0295</td>
<td>0.0112</td>
<td>0.0577</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>S.D.</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dis M</td>
<td>0.0131</td>
<td>0.0100</td>
<td>0.0123</td>
<td>0.0015</td>
<td>0.0329</td>
</tr>
</tbody>
</table>

Figure 3.3: Summary Statistics of Disaster Risk Variables

<table>
<thead>
<tr>
<th></th>
<th>SysDis</th>
<th>Dis</th>
<th>Beta</th>
<th>Coskew</th>
<th>B2M</th>
<th>IlliqMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SysDis</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dis</td>
<td>0.4048</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Beta</td>
<td>0.1353</td>
<td>0.1104</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coskew</td>
<td>-0.0604</td>
<td>-0.0323</td>
<td>0.0150</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2M</td>
<td>0.0123</td>
<td>0.0248</td>
<td>0.0050</td>
<td>-0.0112</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>IlliqMA</td>
<td>-0.0236</td>
<td>0.1044</td>
<td>0.0227</td>
<td>-0.0210</td>
<td>0.0251</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 3.4: Pairwise Correlations of Stock Characteristics

0.335 than its conditional disaster probability given normal market conditions.

Figure 3.4 reports the correlation coefficients of $SysDis$ with other stock characteristics.\(^4\) The results show that $SysDis$ has a strong positive correlation of 0.40 with the unconditional disaster risk ($Dis$), meaning that stocks with a higher sensitivity to the market disaster risk are also prone to rare disasters by themselves. In addition, $SysDis$ also has a positive correlation of 0.14 with the CAPM beta ($Beta$), indicating that stocks with a higher sensitivity to the market disaster risk are also more sensitive to the overall market movement. Further, $SysDis$ has a small positive correlation with the book-to-market ratio ($B2M$), and is negatively correlated with coskewness ($Coskew$) and illiquidity ($IlliqMA$).

I next perform the Fama-MacBeth test on the cross-sectional risk premium associated with $SysDis$. The test period extends from January 1997 to December 2012, which covers 16 years (192 months). In each month, the number of stocks included in the cross-sectional regression varies from 553 to 2270. The test results for the entire sample period are reported

\(^4\)Since $SysDis$ is estimated annually, I examine its correlation with other annually estimated stock characteristics only.
in Figure 3.5. I first conduct the test with $SysDis$ as the only regressor (column 1). The result shows that $SysDis$ has a positive coefficient, but it is statistically insignificant. In fact, it remain insignificant after controlling for other stock characteristics (column 2). Consistent with existing evidence, beta has no significant effect on expected stock returns, and its coefficient is even negative. Further, $Coskew$ has a negative coefficient, as expected, but it is also statistically insignificant. Perhaps due to the relatively short sample period and the fact that I focus on the set of stocks with option trading only, some well-documented risk factors including size, book-to-market, and illiquidity do not show a significant effect, either. The one-month lagged return, on the other hand, has a significantly negative effect on future expected stock returns, implying a reversal in the return process.

I then examine how the relation between systematic disaster risk and expected stock returns changes with the market disaster risk. For this purpose, I rank all months in the sample by the market disaster risk ($Dis^M$) of the previous year and divide the entire sample into

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1)</th>
<th>(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SysDis_{t-1}$</td>
<td>0.0040 (1.06)</td>
<td>0.0017 (0.84)</td>
</tr>
<tr>
<td>$Dis_{t-1}$</td>
<td></td>
<td>0.0261 (0.46)</td>
</tr>
<tr>
<td>$Beta_{t-1}$</td>
<td>-0.0107 (-1.41)</td>
<td></td>
</tr>
<tr>
<td>$Coskew_{t-1}$</td>
<td>-2104.67 (-0.81)</td>
<td></td>
</tr>
<tr>
<td>$Size_{t-1}$</td>
<td>-0.0011 (-1.16)</td>
<td></td>
</tr>
<tr>
<td>$B2M_{t-1}$</td>
<td>-0.0001 (-0.80)</td>
<td></td>
</tr>
<tr>
<td>$Ret_{t-1}$</td>
<td>-0.0222 (-2.60)**</td>
<td></td>
</tr>
<tr>
<td>$IllicMA_{t-1}$</td>
<td>0.0308 (1.17)</td>
<td></td>
</tr>
<tr>
<td>$Constant$</td>
<td>0.0063 (1.81)*</td>
<td>0.0333 (1.73)*</td>
</tr>
</tbody>
</table>

Figure 3.5: Systematic Disaster Risk Premium for Entire Sample
four subsamples, each containing 48 months. I repeat the test for each of the four subsamples separately. The results are reported in Figure 3.6 For the first three subsamples for which the lagged market disaster risk is below the first quartile (Columns 1), between the first and the second quartiles (Column 2), and between the second and the third quartiles (Column 3), \( SysDis \) does not carry a positive premium. In fact, when the lagged market disaster risk is between the first and the second quartiles (Column 2), \( SysDis \) even has a negative coefficient significant at the 10% level after controlling for other risk factors. However, if I restrict the test to months for which the previous year market disaster risk is higher than the third quartile (column 4), \( SysDis \) becomes significant at the 1% level. Indeed, in this case increasing \( SysDis \) by one standard deviation raises the expected monthly stock return by 63 basis points, which is equivalent to an increase of over 7% per year. Interestingly, when restricted to this subsample, the unconditional disaster risk \( (Dis) \) also carries a significantly positive premium. This implies that investors also require compensation for holding stocks with high unconditional disaster risk when a market crash is considered likely.

To sum up, the empirical study in this section shows no clear evidence that systematic disaster risk is priced by investors under normal market conditions. However, it has a strongly positive effect on expected future stock returns when the market-wide disaster risk is considerably high.

### 3.7 Application II: Bank Systemic Exposure

This section applies my approach to study the systemic risk of the banking sector. Systemic risk is the risk of collapse of the entire banking system due to the interrelation and interdependence of all banks. The study of systemic risk has recently attracted much attention, especially following the financial crisis of 2007–2009. Two aspects of systemic risk are of particular interest to researchers, one looking at the contribution of individual banks to the risk of the whole banking system, and the other focusing on the exposure of individual banks
Figure 3.6: Systematic Disaster Risk Premium and Market Disaster Risk

to systemic shocks.

Despite the general interests in learning about systemic risk, the empirical estimation is challenging due to the small probability of disastrous events. Different methods have been proposed in the literature to tackle this problem. For example, Huang, Zhou, and Zhu (2009) measure systemic risk by the price of insurance against financial distress, in which the default correlation between banks is proxied by the equity return correlation. Acharya, Pedersen, Philippon and Richardson (2010) propose the systemic expected shortfall ($SES$) measure, which estimates the propensity of a bank to be undercapitalized when the system as a whole is undercapitalized. Their approach relies on the power law distribution to model the tails. In addition, Adrian and Brunnermeier (2011) define the $\Delta CoVaR$ measure as the difference between the value-at-risk of the banking system conditional on an individual bank being in distress and the value-at-risk of the banking system conditional on the bank being solvent. The empirical estimation of $\Delta CoVaR$ uses quantile regression to capture the tails.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Low $Dis_{t-1}^M$</th>
<th>(2)</th>
<th>(3)</th>
<th>High $Dis_{t-1}^M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SysDis_{t-1}$</td>
<td>0.0013</td>
<td>-0.0060</td>
<td>-0.0056</td>
<td>0.0172</td>
</tr>
<tr>
<td></td>
<td>(0.48)</td>
<td>(-1.73)*</td>
<td>(-1.59)</td>
<td>(3.18)**</td>
</tr>
<tr>
<td>$Dis_{t-1}$</td>
<td>-0.0674</td>
<td>0.0852</td>
<td>-0.1657</td>
<td>0.2523</td>
</tr>
<tr>
<td></td>
<td>(-0.76)</td>
<td>(0.68)</td>
<td>(-1.23)</td>
<td>(2.77)**</td>
</tr>
<tr>
<td>$Beta_{t-1}$</td>
<td>-0.0080</td>
<td>-0.0113</td>
<td>-0.0139</td>
<td>-0.0098</td>
</tr>
<tr>
<td></td>
<td>(-1.16)</td>
<td>(-0.61)</td>
<td>(-0.90)</td>
<td>(-0.55)</td>
</tr>
<tr>
<td>$Coskew_{t-1}$</td>
<td>-1.0652.98</td>
<td>1321.01</td>
<td>-26.07</td>
<td>939.38</td>
</tr>
<tr>
<td></td>
<td>(-1.12)</td>
<td>(0.37)</td>
<td>(-0.01)</td>
<td>(0.92)</td>
</tr>
<tr>
<td>$Size_{t-1}$</td>
<td>0.0009</td>
<td>-0.0013</td>
<td>-0.0007</td>
<td>-0.0034</td>
</tr>
<tr>
<td></td>
<td>(0.76)</td>
<td>(-0.67)</td>
<td>(-0.27)</td>
<td>(-1.78)*</td>
</tr>
<tr>
<td>$B2M_{t-1}$</td>
<td>0.0001</td>
<td>-0.0002</td>
<td>-0.0004</td>
<td>0.0001</td>
</tr>
<tr>
<td></td>
<td>(0.59)</td>
<td>(-0.58)</td>
<td>(-2.56)**</td>
<td>(0.55)</td>
</tr>
<tr>
<td>$Ret_{t-1}$</td>
<td>-0.0052</td>
<td>-0.0486</td>
<td>-0.0137</td>
<td>-0.0213</td>
</tr>
<tr>
<td></td>
<td>(-0.38)</td>
<td>(-2.30)**</td>
<td>(-0.72)</td>
<td>(-1.62)</td>
</tr>
<tr>
<td>$IlliqMA_{t-1}$</td>
<td>0.0628</td>
<td>0.0449</td>
<td>0.0367</td>
<td>-0.0210</td>
</tr>
<tr>
<td></td>
<td>(2.43)**</td>
<td>(0.84)</td>
<td>(0.71)</td>
<td>(-0.29)</td>
</tr>
<tr>
<td>$Constant$</td>
<td>0.0030</td>
<td>0.0305</td>
<td>0.0272</td>
<td>0.0728</td>
</tr>
<tr>
<td></td>
<td>(0.13)</td>
<td>(0.77)</td>
<td>(0.65)</td>
<td>(1.57)</td>
</tr>
</tbody>
</table>
One commonality of these papers is that they all measure systemic risk based on historical equity returns.

In this section, I examine the exposure of banks to systemic shocks and ask what bank characteristics are related to banks’ systemic exposure. To this end, I propose a “systemic exposure” measure based on the conditional return distribution of a bank given the return of the banking sector portfolio, which can be estimated using my approach by taking the sector portfolio as the related index. Since the measure is empirically estimated using option prices as opposed to historical equity returns, it is forward-looking and naturally accounts for investors’ beliefs on the likelihood of disastrous events even if such events do not occur within sample.

3.7.1 Measure of Systemic Exposure

The systemic exposure measure resembles the systematic disaster risk measure introduced in Section 3.6.1. The main difference is that I now focus on the joint performance of a bank and the whole banking system, and hence I use the banking sector portfolio (as opposed to the market portfolio) as the related index.

As before, I look at equity returns over periods of three months. For both the banking sector and individual banks, I define two states \( N = K = 2 \), the normal state \( H \) and the disaster state \( L \). I choose the disaster thresholds for the three-month returns of both the sector portfolio and individual banks to be \(-1/2\). Intuitively, the sector (a bank) experiences a disastrous event if it loses more than one half of its original value within three months. Then, for each bank I define the systemic exposure measure as

\[
SysExp = \theta (L|L) - \theta (L|H),
\]

(3.10)

where \( \theta (L|L) \) and \( \theta (L|H) \) stand for the disaster probability of the bank conditional on the sector being in the disaster and the normal states, respectively. This measure captures the extent to which a bank is more likely to experience a disaster when the whole sector falls...
from the normal state to the disaster state. Put differently, it is a measure of the sensitivity of individual bank returns to the sector-wide disaster risk.

The systemic exposure measure takes values from -1 to 1. A positive value means that a bank is more likely to experience a disaster when the sector as a whole is in the disaster state relative to when the sector performs normally. The higher the value of $SysExp$, the more sensitive the bank is to sector-wide disaster risk. On the other hand, a negative value of $SysExp$ implies that the bank becomes safer when the sector is hit by a disastrous event.

### 3.7.2 Data and Estimation

I use the KBW Bank Index as a proxy for the banking sector portfolio, which consists of 24 banking companies and has options actively traded on the Philadelphia Stock Exchange. I focus on computing the systemic exposure measure for each of the 24 current constituent banks of the index.\(^5\) The current composition of the index is shown in Figure 3.6.

My sample period is from January 1996 to December 2012. For both the KBW Bank Index and the 24 banks, I collect daily information on option prices, security prices and dividend distributions from the OptionMetrics database. Notice that some of the 24 banks were formed by mergers and acquisitions at some points during my sample, and some others temporarily stopped trading on the option market for some periods. As a result, not all of the 24 banks have full records throughout the entire sample period.

I estimate $SysExp$ in three steps. First, on each date I compute the option-implied risk-neutral probability of disaster for both the KBW Index and the individual banks. To maintain the accuracy of estimation, I compute the risk-neutral disaster probability for a bank on a date only when (1) there are at least ten different options with pricing information available for the bank on that date, (2) the lowest moneyness level of available options for

---

\(^5\)While I focus on the 24 current constituent banks of the KBW index in this application, there is nothing preventing me from estimating $SysExp$ for other banks. The 24 banks are among the largest banks in the sector and are considered to have the most influences upon the economy. Hence, results based on these banks are of particular economic import.
the bank on that date is no higher than 0.95, and (3) the shortest (longest) time to maturity of available options for the bank on that date is lower (higher) than three months. The average risk-neutral disaster probabilities for the KBW Bank Index and for the individual banks throughout the sample period are estimated as 0.010 and 0.018, respectively. In the second step, I run the constrained linear regression (3.8) for each bank over time to obtain the conditional disaster probabilities of the bank given different states of the sector. Finally, I compute $SysExp$ using the definition (3.10).

I am also interested in what bank characteristics predict banks’ systemic exposure. To examine this issue, I construct the following bank characteristics. The first three are computed using daily security pricing information taken from the OptionMetrics database, whereas the rest are constructed from quarterly bank fundamentals obtained from the Compustat database.

- Market beta ($Beta$), estimated with respect to the market portfolio using daily equity returns over the most recent 90 business days. The market excess returns and the risk-free rates needed for the estimation are taken from the Kenneth French online data library.

- Sector beta ($Beta^S$), estimated with respect to the banking sector portfolio using daily equity returns over the most recent 90 business days. The banking sector portfolio is proxied by the KBW Bank Index.

- Equity return volatility ($Vol$), estimated over the most recent 90 business days as the standard deviation of the daily equity returns.

- Bank size ($Size$), computed as the logarithm of the total market capitalization.

- Leverage ($Lever$), estimated as the ratio of the total book value of assets to the total book value of equity.
• Market-to-book ratio \((M2B)\), computed as the ratio of the total market capitalization to the total book value of equity.

• Non-interest to interest income ratio \((N2I)\), estimated as the ratio of total non-interest income to the total interest income. The inclusion of this variable is motivated by Brunnermeier, Dong and Palia (2012), who show that the \(N2I\) ratio contributes to banks’ systemic risk.

• Maturity mismatch. I construct two proxies for maturity mismatch. First is the ratio of total demand deposits to the total book value of assets \((MatMis1)\), and the second is the ratio of total interest-bearing deposits to the total book value of assets \((MatMis2)\). Since interest-bearing deposits generally have longer maturities and are less often withdrawn than demand deposits, the extent of maturity mismatch of a bank is positively related to the first proxy and negatively related to the second.

3.7.3 Results

I first estimate the systemic exposure of each bank using the entire sample period. That is, I run the constrained linear regression (3.8) based on the risk-neutral disaster probabilities of both the KBW Index and the bank of interest throughout the whole sample. Notice that due to missing option prices for some banks during some periods of the sample, the effective estimation period may differ from bank to bank. Figure 3.7 reports the estimation results. A first observation is that the systemic exposure is positive for all banks, indicating that banks are generally more likely to experience a disastrous event when the banking sector as a whole is hit by a disaster relative to when the sector operates in the normal state. For example, the systemic exposure of JPMorgan Chase & Co. is 0.686, meaning that when the banking sector moves from the normal state to the disaster state, the conditional disaster probability of the bank increases by 0.686. In addition, it can be seen that the systemic exposure varies
<table>
<thead>
<tr>
<th>Bank Name</th>
<th>$SysExp$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB&amp;T Corporation</td>
<td>0.7636</td>
</tr>
<tr>
<td>Bank of America Corp</td>
<td>0.9881</td>
</tr>
<tr>
<td>Capital One Financial Corp</td>
<td>0.9789</td>
</tr>
<tr>
<td>JPMorgan Chase &amp; Co</td>
<td>0.6860</td>
</tr>
<tr>
<td>Citigroup</td>
<td>0.9880</td>
</tr>
<tr>
<td>Comerica Inc</td>
<td>0.9925</td>
</tr>
<tr>
<td>Commerce Bancshares Inc</td>
<td>0.1164</td>
</tr>
<tr>
<td>Cullen/Frost Bankers Inc</td>
<td>0.2079</td>
</tr>
<tr>
<td>Fifth Third Bancorp</td>
<td>0.9865</td>
</tr>
<tr>
<td>First Niagara Financial Group</td>
<td>0.2534</td>
</tr>
<tr>
<td>U.S. Bancorp</td>
<td>0.8038</td>
</tr>
<tr>
<td>Huntington Bancshares Inc</td>
<td>0.9907</td>
</tr>
<tr>
<td>KeyCorp</td>
<td>0.9883</td>
</tr>
<tr>
<td>M&amp;T Bank Corporation</td>
<td>0.7331</td>
</tr>
<tr>
<td>Bank of New York Mellon</td>
<td>0.7632</td>
</tr>
<tr>
<td>New York Community Bank</td>
<td>0.5715</td>
</tr>
<tr>
<td>Northern Trust Corp</td>
<td>0.4948</td>
</tr>
<tr>
<td>PNC Financial Services Group</td>
<td>0.7805</td>
</tr>
<tr>
<td>People’s United Financial Inc</td>
<td>0.1553</td>
</tr>
<tr>
<td>Regions Financial Corporation</td>
<td>0.9852</td>
</tr>
<tr>
<td>State Street Corporation</td>
<td>0.8056</td>
</tr>
<tr>
<td>SunTrust Banks Inc</td>
<td>0.9919</td>
</tr>
<tr>
<td>Wells Fargo &amp; Company</td>
<td>0.9912</td>
</tr>
<tr>
<td>Zions Bancorporation</td>
<td>0.9831</td>
</tr>
</tbody>
</table>

Figure 3.7: Systemic Exposure of Banks

much across different banks, from a minimum of 0.116 for Commerce Bancshares Inc. to a maximum of 0.993 for Comerica Inc.

I next ask the question of what bank characteristics predict systemic exposure. To address this question, I estimate $SysExp$ for each bank at the end of each month based on option prices of the most recent 90 business days. To improve accuracy, I perform estimation only if the risk-neutral disaster probability is available for the bank for at least 60 days in the 90-business-day window. This yields a panel for the $SysExp$ measure across different banks over time.

To make sure that my estimation of $SysExp$ generates reasonable results, I check the out-
of-sample validity by trying to predict future states of banks using my conditional disaster probability estimation. As in the previous application, I take the first quarter of each year and try to predict the state of a bank over the quarter based on the realized state of the sector during the same quarter and the conditional disaster probabilities estimated from the most recent 90-business-day window. From late 2008 to early 2009, the banking sector was hit by a disastrous shock, during which the three-month KBW return dropped below \(-1/2\). This turns out to be the only occurrence of a sector-wide disaster throughout the entire sample period according to my definition in Section 3.7.1. Hence, instead of looking at the first quarter of 2009, I try to predict the state of each bank over the 91-day period from November 5, 2008 to February 3, 2009, during which the return of the KBW Index was \(-56\%\).

For each quarter-length period mentioned above, I determine the states of each bank and the KBW Index (\(H\) or \(L\)) over the period based on their returns using the definition of Section 3.7.1. Then, for each bank and each quarter-length period, I compare two quantities: \(1_{\text{bank in state } L}\), a dummy variable that takes on a value of 1 if the bank delivers a disastrous return over the period and 0 otherwise, and \(\hat{\theta}(L|\text{realized state of KBW})\), the conditional disaster probability of the bank estimated from the most recent 90-business-day window given the realized state of KBW over the period. I first check the correlation between these two quantities and find that they exhibit a strong positive correlation of 0.75. This indicates that the higher the predicted conditional disaster probability, the higher the actual occurrence of a bank disaster.

I then compute the following difference

\[
\Delta = 1_{\text{bank in state } L} - \hat{\theta}(L|\text{realized state of KBW}),
\]

and take the average of \(\Delta\) across all banks and across different periods. My results show that conditional on a realized disaster state of KBW (November 5, 2008 to February 3, 2009) and
a realized normal state of KBW (all other quarter-length periods), the average values of $\Delta$ are -0.01 and 0.0006, respectively. This serves as evidence for good out-of-sample validity of my estimation.

I then regress $SysExp$ on lagged values of the bank characteristics discussed in Section 3.7.2. Since some characteristics have low frequency (quarterly available), I take as regressors the most recent value of each bank characteristic computed prior to the 90-business-day window used to estimate $SysExp$. Figure 3.8 reports the regression results. I first run the regression without any fixed effects (column 1). The results show that the total return volatility, the market-to-book ratio, and the bank size have positive and statistically significant coefficients. On the other hand, the market beta, the leverage ratio, the non-interest to interest income ratio, and the maturity mismatch proxies do not seem to be related to banks’ systemic exposure.

I then rerun the regression controlling for the bank-fixed effects (column 2). The return volatility and the market-to-book ratio remain positive and significant. However, the size variable loses its significance, and now it even switches to a negative sign. This implies that the positive correlation between size and systemic exposure is completely absorbed by the bank-fixed effects. In addition, the non-interest to interest income ratio becomes significant at the 10% level. Its positive coefficient indicates that the systemic exposure of a bank increases with its noncore activities (e.g., investment banking, venture capital, and trading activities) relative to the traditional banking businesses.

Next, I perform the regression with both bank- and year-fixed effects. The market beta now has a significantly positive coefficient, indicating that a bank’s sensitivity to the overall market movement positively predicts its systemic exposure. The return volatility continues to have a positive and significant coefficient, reinforcing the result that systemic exposure is increasing in the total return volatility. The market-to-book ratio loses its significance after controlling for the year-fixed effects. The coefficient of the size variable remains negative, and
<table>
<thead>
<tr>
<th>Variable</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Beta_{t-1}$</td>
<td>-0.0023</td>
<td>0.0112</td>
<td>0.0587</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(-0.07)</td>
<td>(0.32)</td>
<td>(2.54)**</td>
<td></td>
</tr>
<tr>
<td>$Beta_{t-1}$S</td>
<td>6.5414</td>
<td>5.4125</td>
<td>4.1575</td>
<td>3.1228</td>
</tr>
<tr>
<td></td>
<td>(6.20)**</td>
<td>(6.10)**</td>
<td>(2.97)**</td>
<td>(2.29)**</td>
</tr>
<tr>
<td>$Vol_{t-1}$</td>
<td>0.0054</td>
<td>-0.0003</td>
<td>-0.0077</td>
<td>-0.0088</td>
</tr>
<tr>
<td></td>
<td>(0.93)</td>
<td>(-0.04)</td>
<td>(-0.89)</td>
<td>(-1.12)</td>
</tr>
<tr>
<td>$M2B_{t-1}$</td>
<td>0.0395</td>
<td>0.0547</td>
<td>0.0258</td>
<td>0.0270</td>
</tr>
<tr>
<td></td>
<td>(2.99)**</td>
<td>(3.55)**</td>
<td>(0.81)</td>
<td>(0.90)</td>
</tr>
<tr>
<td>$Size_{t-1}$</td>
<td>0.0349</td>
<td>-0.0581</td>
<td>-0.0739</td>
<td>-0.0574</td>
</tr>
<tr>
<td></td>
<td>(2.25)**</td>
<td>(-1.73)</td>
<td>(-1.81)*</td>
<td>(-1.45)</td>
</tr>
<tr>
<td>$N2I_{t-1}$</td>
<td>0.0199</td>
<td>0.0480</td>
<td>0.0974</td>
<td>0.1019</td>
</tr>
<tr>
<td></td>
<td>(0.79)</td>
<td>(1.84)*</td>
<td>(3.02)**</td>
<td>(3.29)**</td>
</tr>
<tr>
<td>$MatMis1_{t-1}$</td>
<td>-0.1616</td>
<td>-0.3353</td>
<td>0.5214</td>
<td>0.4235</td>
</tr>
<tr>
<td></td>
<td>(-0.71)</td>
<td>(-1.10)</td>
<td>(1.25)</td>
<td>(1.13)</td>
</tr>
<tr>
<td>$MatMis2_{t-1}$</td>
<td>-0.2059</td>
<td>0.1767</td>
<td>0.0183</td>
<td>0.0421</td>
</tr>
<tr>
<td></td>
<td>(-0.88)</td>
<td>(0.60)</td>
<td>(0.07)</td>
<td>(0.16)</td>
</tr>
<tr>
<td>$Constant$</td>
<td>-0.4852</td>
<td>0.9277</td>
<td>1.1158</td>
<td>0.7549</td>
</tr>
<tr>
<td></td>
<td>(-1.16)</td>
<td>(1.47)</td>
<td>(1.54)</td>
<td>(1.07)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bank-fixed effects</th>
<th>N</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year-fixed effects</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>3123</th>
<th>3123</th>
<th>3123</th>
<th>3123</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>0.1092</td>
<td>0.1371</td>
<td>0.2141</td>
<td>0.2205</td>
</tr>
</tbody>
</table>

Figure 3.8: Systemic Exposure and Bank Characteristics
it now becomes significant at the 10% level, suggesting weak evidence for decreasing systemic exposure with size. Interestingly, the non-interest to interest income ratio continues to have a positive and even more significant effect on systemic exposure. This confirms the result that a bank’s exposure to systemic shocks is driven by its non-traditional businesses. Finally, notice that after controlling for the bank- and year-fixed effects, the R-squared almost doubles, from 0.109 to 0.214, meaning that a large portion of the variation in the systemic exposure can be explained by bank- and year-specific factors.

Finally, I replace the market beta by the sector beta, controlling for all other bank characteristics and fixed effects. The sector beta has a positive coefficient and is significant at the 1% level, implying that banks that are more sensitive to the overall sector movement also tend to be more sensitive to the sector disaster risk. Results concerning other characteristics remain mostly unchanged, but the size variable now becomes insignificant with a $t$-statistic of -1.45. Overall, the results suggest that the systemic exposure cannot be fully predicted by the linear relation between individual bank returns and the sector returns (as captured by the sector beta) alone. Other bank characteristics such as the total return volatility and the non-interest to interest income ratio are also important predictors.

To sum up, I find that a bank exhibits higher exposure to systemic shocks when it is more sensitive to the overall market and sector movement, when its equity return becomes more volatile, and when it puts a larger weight on non-traditional businesses. I also find weak evidence that bank size negatively predicts the exposure to systemic shocks. In addition, banks’ systemic exposure is also considerably driven by bank- and time-specific effects.

3.8 Application III: Ross Recovery for Individual Assets

The distributions of security returns generally differ under the physical versus the risk-neutral measures to reflect the adjustment for risk aversion. Ross (1976a) and Breeden and Litzen-
berger (1978) show that option prices reveal information on the risk-neutral distributions of security returns (see Section 3.4.1). Unfortunately, it is a priori not clear how the physical return distributions can be estimated. Ross (2015) proposes a novel idea called the recovery theorem that enables us to separate physical return distributions from risk adjustments based on the risk-neutral distributions alone. Subsequent research further explores extensions and alternative approaches to tackle this problem (e.g., Carr and Yu (2012), Huang and Shaliastovich (2013), Martin and Ross (2013), Walden (2013), Dubynskiy and Goldstein (2013), and Borovička, Hansen and Scheinkman (2014)). All of these papers primarily focus on the market portfolio, and the recovery results do not directly apply to individual assets.

In this section, I extend the recovery results from the market portfolio to individual assets. I show that this can be accomplished through the risk-neutral joint distribution of the asset return with the market return, which can be conveniently estimated using my approach by taking the market portfolio as the related index. While I base my discussion on the original Ross (2015) recovery theorem, it is worth noticing that my extension is not restricted to Ross’s original result but rather can be applied to any recovery result developed for the market portfolio. I start by reviewing the Ross recovery theorem.

### 3.8.1 Ross Recovery Theorem

Consider a one-period economy with two time points 1 and 2. Suppose that there is an aggregate market portfolio whose value defines \( N \) states indexed by \( \{1, 2, ..., N\} \).\(^6\) For any \( n_1, n_2 \in \{1, 2, ..., N\} \), let \( \pi (n_1, n_2) \) denote the physical transition probability of moving from state \( n_1 \) at time 1 to state \( n_2 \) at time 2. In addition, let \( \phi (n_1, n_2) \) denote the price at time 1 of an asset that pays one dollar at time 2 if the market is in state \( n_1 \) at time 1 and in state \( n_2 \) at time 2. Notice that \( \phi (n_1, n_2) \) is typically referred to as the state price, and it is equal to the risk-neutral transition probability from state \( n_1 \) to state \( n_2 \) discounted at the

---

\(^6\)In Ross’s model, it is the value of the market portfolio, instead of the return, that determines the state of nature.
risk-free rate. The recovery theorem works directly with the state prices rather than the corresponding risk-neutral probabilities.

The Ross recovery theorem is motivated by the following model. Assume that a representative agent exists, who has an additively time-separable von Neumann-Morgenstern utility function \( U(\cdot) \) with \( U'(\cdot) > 0 \) and \( U''(\cdot) < 0 \). Also assume that the representative agent has initial wealth \( W \) and a subjective time discount factor \( \delta \in (0, 1) \). The agent chooses the optimal consumptions at time 1 and 2 subject to the budget constraint. Assume that the optimal consumption level depends on the state of the market only, and is in particular time independent. Denote the optimal consumption in state \( n \) by \( c(n) \). Then, at time 1 the agent solves the following utility optimization problem:

\[
\max_{\{c(1), \ldots, c(N)\}} U(c(n_1)) + \delta \sum_{n_2=1}^{N} U(c(n_2)) \pi(n_1, n_2)
\]

s.t.

\[
c(n_1) + \sum_{n_2=1}^{N} c(n_2) \phi(n_1, n_2) = W.
\]

The solution to the above optimization problem is given by the first order condition

\[
U'(c(n_1)) \phi(n_1, n_2) = \delta U'(c(n_2)) \pi(n_1, n_2).
\]

Rewriting the first order condition in matrix form, we obtain

\[
D \cdot \Phi = \delta \Pi \cdot D,
\]

where

\[
D = \begin{pmatrix}
U'(c(1)) & 0 & \ldots & 0 \\
0 & U'(c(2)) & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & U'(c(N))
\end{pmatrix},
\]

\[
\Phi = \begin{pmatrix}
\phi(1,1) & \ldots & \phi(1,N) \\
\vdots & \ddots & \vdots \\
\phi(N,1) & \ldots & \phi(N,N)
\end{pmatrix},
\]

\[
\Pi = \begin{pmatrix}
\begin{pmatrix}
(1,1) & \ldots & (1,N) \\
\vdots & \ddots & \vdots \\
(N,1) & \ldots & (N,N)
\end{pmatrix}
\end{pmatrix},
\]

132
and

$$\Pi = \begin{pmatrix}
\pi(1,1) & \ldots & \pi(1,N) \\
\vdots & \ddots & \vdots \\
\pi(N,1) & \ldots & \pi(N,N)
\end{pmatrix}.$$ 

Since $U'(\cdot) > 0$, the matrix $\mathbf{D}$ is invertible. Then, solving for $\Pi$ yields

$$\Pi = \frac{1}{\delta} \mathbf{D} \cdot \Phi \cdot \mathbf{D}^{-1}. \quad (3.11)$$

Since $\Pi$ is a transition probability matrix, we must have that starting from any state, the transition probabilities to all states sum up to 1, i.e.,

$$\Pi \cdot \mathbf{1}_{N \times 1} = \mathbf{1}_{N \times 1}.$$ 

This condition and (3.11) together imply

$$\Pi \cdot \mathbf{1}_{N \times 1} = \frac{1}{\delta} \mathbf{D} \cdot \Phi \cdot \mathbf{D}^{-1} \cdot \mathbf{1}_{N \times 1} = \mathbf{1}_{N \times 1},$$

which can be rewritten as

$$\Phi \cdot \eta = \delta \cdot \eta,$$

where $\eta = \mathbf{D}^{-1} \cdot \mathbf{1}_{N \times 1}$. In particular, $U'(\cdot) > 0$ implies $\eta \gg \mathbf{0}$, where $\gg$ means “strictly greater than” in each and every element. That is, $\eta$ is a strictly positive eigenvector of $\Phi$ with a corresponding positive eigenvalue $\delta$.

The Perron-Frobenius theorem (see Meyer (2000)) shows that all nonnegative irreducible matrices have a unique positive eigenvalue associated with a unique positive eigenvector up to scaling. If we assume no arbitrage, then the matrix of state prices $\Phi$ is non-negative, with zero elements if and only if the corresponding physical probabilities are zero. If $\Phi$ is also irreducible, one can uniquely identify the subjective discount factor $\delta$ and the vector $\eta$ (or equivalently the diagonal matrix $\mathbf{D}$) up to scaling. Finally, plugging $\delta$ and $\mathbf{D}$ into (3.11) allows one to recover the physical transition matrix $\Pi$. Notice that while the matrix $\mathbf{D}$ has an undetermined scaling parameter, the same parameter in $\mathbf{D}$ and $\mathbf{D}^{-1}$ exactly cancels out, rendering the physical transition matrix $\Pi$ uniquely identified.
To apply the recovery theorem empirically, an essential step is to estimate the state price matrix $\Phi$ from data. As discussed in Section 3.4.1, we can extract the risk-neutral probabilities and equivalently the corresponding state prices from option prices. One main challenge, however, is that this method only allows us to obtain the state prices starting from the realized initial state at time 1. Without loss of generality, assume that the realized state of the market is $n_1 = 1$ at time 1. Then, one can estimate from option prices $\phi(1, n_2)$ for all $n_2$ (the first row of $\Phi$) but not $\phi(n_1, n_2)$ for $n_1 \neq 1$.

To resolve this problem, Ross (2015) comes up with a brilliant idea. Let

$$\phi^l = (\phi^l (1, 1), \phi^l (1, 2), \ldots, \phi^l (1, N))$$

denote the vector of the $l$-period state prices starting from state 1. Specifically, $\phi^l (1, n)$ equals the price of an asset that pays one dollar in $l$ periods if the market is in state 1 today and in state $n$ after $l$ periods. Notice that $\phi^1$ coincides with $(\phi (1, 1), \phi (1, 2), \ldots, \phi (1, N))$ from the above one-period setting. Further assume that the state of the market portfolio follows a Markov process. Then for any $l = 1, 2, \ldots, N - 1$, we have the following recursive forward equations

$$\phi^{l+1} = \phi^l \cdot \Phi.$$  \hspace{1cm} (3.12)

This forms a linear system of the state price matrix $\Phi$, where the $\phi^l$’s can be estimated from option prices with different maturities. Then, solving this linear system yields $\Phi$, which can then be used to recover the physical transition matrix $\Pi$.

The recovery theorem deals with the entire transition matrix, which in particular includes transition probabilities starting from the realized initial state as well as from all other hypothetical initial states. Often times, however, we are mostly interested in the transition probabilities starting from the realized initial state $(\pi (1, 1), \pi (1, 2), \ldots, \pi (1, N))$. Having the initial state fixed, we then have a 1–1 mapping between the market return and the future market value (given by (3.6)). Using the notations of my framework in Section 3.3
and focusing on transition probabilities starting from the realized initial state, (3.11) can be readily rewritten as
\[ p^I_{t,t+1}(n) = \frac{1}{\delta} e^{-r^f} q^I_{t,t+1}(n) \frac{U'(c(1))}{U'(c(n))}, \] (3.13)
where index I is chosen as the market portfolio, \( p^I_{t,t+1}(n) \) is the physical probability of moving from the current market state to state \( n \) in one period, \( e^{-r^f} q^I_{t,t+1}(n) \) is the state price, and \( \delta \) as well as \( \frac{U'(c(1))}{U'(c(n))} \) are uniquely determined from the Ross recovery result.

### 3.8.2 Extension to Individual Assets

To see how the recovery result can be applied to individual assets, it is useful to consider (3.13) for the market portfolio. This formula shows that the physical distribution of the market return can be decomposed into three parts. The first part is the reciprocal of the subjective discount factor \( \frac{1}{\delta} \), which is a constant and in particular does not depend on the future market state. The second component is the state price \( e^{-r^f} q^I_{t,t+1}(n) \), which can be estimated from option prices. Finally, the third part is the pricing kernel \( \frac{U'(c(1))}{U'(c(n))} \), representing the adjustment for risk aversion in different states of the market.

One cannot directly apply (3.13) to recover the physical return distribution of an individual asset from the associated risk-neutral distribution. The reason is that the same value of the asset return can appear in different states of the market unless there is perfect correlation between the asset return and the market return. Given that the pricing kernel \( \frac{U'(c(1))}{U'(c(n))} \) is a function of the market return (fixing the initial market value), there is in general no single \( \frac{U'(c(1))}{U'(c(n))} \) that can be used for any particular value of the asset return.

Fortunately, this problem can be resolved through the risk-neutral joint return distribution of the asset with the market. In particular, one can express the risk-neutral marginal distribution of the asset return as the sum of its risk-neutral joint probabilities with the
market return, i.e.,
\[ q_{t,t+1}(k) = \sum_{n=1}^{N} q \left( \tilde{r}_{t,t+1} = r(k), \tilde{r}_{t,t+1}^I = r^I(n) \right), \forall k. \]

Then, each of the risk-neutral joint probabilities \( q \left( \tilde{r}_{t,t+1} = r(k), \tilde{r}_{t,t+1}^I = r^I(n) \right) \) can be adjusted for risk aversion by the appropriate value of \( \frac{U'(c(1))}{U'(c(n))} \), which allows one to recover the physical joint probabilities. Finally, summing up the physical joint probabilities across all states of the market yields the physical marginal return distribution of the asset. This is formalized in the following proposition.

**Proposition 58** At any time \( t \), the physical distribution of the asset return over one period head can be recovered as
\[ p_{t,t+1}(k) = \frac{1}{\delta} e^{-r_f} \sum_{n=1}^{N} q \left( \tilde{r}_{t,t+1} = r(k), \tilde{r}_{t,t+1}^I = r^I(n) \right) \frac{U'(c(1))}{U'(c(n))}, \forall k, \] where \( \delta \) and \( \frac{U'(c(1))}{U'(c(n))} \) are uniquely determined from the Ross recovery theorem.

It is clear from Proposition 58 that the key to recovering the physical return distribution of an individual asset is the risk-neutral joint return distribution of this asset with the market portfolio. We know from the probability theory that this risk-neutral joint distribution equals the product of the risk-neutral marginal return distribution of the market and the conditional return distribution of the asset given the market return, i.e.,
\[ q \left( \tilde{r}_{t,t+1}^I = r^I(n), \tilde{r}_{t,t+1} = r(k) \right) = q_{t,t+1}(n) \theta(k|n), \forall n,k. \] While the former can be obtained from option prices, my approach allows me to estimate the latter by taking the market portfolio as the related index. Hence, my framework lends itself naturally to the extension of the Ross recovery theorem to individual assets.
3.9 Conclusion

In this paper I propose a novel approach to recovering the conditional return distribution of an individual asset given the return of an aggregate index by regressing the marginal return distribution of the asset on that of the index. The identifying assumption that underlies this approach states that the conditional return distribution of the asset given any value of the index return does not vary over time. Intuitively, this means that the time variation in the return distribution of the asset is solely driven by the time variation in the return distribution of the index. This assumption allows me to make use of the time series information on the marginal return distributions of both securities to pin-down the time-invariant conditional distribution of the asset return given the index return. Empirically, I show how this approach can be implemented using option prices. I also discuss a variety of applications of this approach to the cross-sectional test of equity risk premium associated with systematic disaster risk, to the estimation of banks’ exposure to systemic shocks, and to the extension the Ross (2015) recovery theorem to individual assets.

The advantage of my approach is that it generates the entire conditional return distribution of the asset given the index return, thus accounting for high moment properties and potential nonlinear patterns of the joint behavior of the two securities. In addition, since the estimation relies on option prices rather than historical equity returns, it is forward-looking and reflects the effects of rare events even if they do not truly occur within sample. Further, compared to the copula approach, which is widely used to back out the joint distribution of multiple random variables from the associated marginals, my approach does not rely on any parametric assumptions on the security returns, thus allowing for greater flexibility. Yet, unlike some non-parametric procedures which are computationally very expensive, my approach can be easily implemented by a constrained linear regression.

Future work can be conducted in different directions. On the methodology side, research
needs to be done on extending the framework to estimate the joint behavior of potentially more than two arbitrary economic variables (besides the returns of an asset and an aggregate index). In terms of applications, this approach can be used to address other important questions related to risk management and firm cyclicality, etc.
Bibliography


Appendix A: Proofs for Chapter 1

As a preparation for the formal proof of Theorem 9, we now review the concept of Wronskian, which is first introduced by Józef Hoene-Wronski (1812) and named by Thomas Muir (1882).

Let \( \Phi(\lambda) = (\phi_1(\lambda), \phi_2(\lambda), \ldots, \phi_K(\lambda)) \) denote a vector of functions with the \( \phi_k(\lambda) \)'s defined over a real interval \( \Lambda \subset \mathbb{R} \) with at least \( K - 1 \) finite continuous derivatives. Then the Wronskian of \( \Phi \) at \( \lambda \) is defined as

\[
W[\Phi](\lambda) = \begin{vmatrix}
\phi_1(\lambda) & \phi_2(\lambda) & \cdots & \phi_K(\lambda) \\
\phi_1'(\lambda) & \phi_2'(\lambda) & \cdots & \phi_K'(\lambda) \\
\vdots & \vdots & \ddots & \vdots \\
\phi_1^{(K-1)}(\lambda) & \phi_2^{(K-1)}(\lambda) & \cdots & \phi_K^{(K-1)}(\lambda)
\end{vmatrix}.
\]

It is apparent that if \( \Phi \) is linearly dependent on \( \Lambda \), we have \( W[\Phi](\lambda) = 0 \) for all \( \lambda \in \Lambda \). Somewhat surprisingly, however, the reverse does not hold generally. That is, the identical vanishing of the Wronskian does not imply linear dependence. Bôcher (1900) shows that if each of the \( \phi_k(\lambda) \)'s is a globally analytic function on \( \Lambda \), then the vanishing of the Wronskian implies that \( \Phi \) is linearly dependent on \( \Lambda \). We now extend this result and show in the following proposition that local analyticity is actually sufficient for the vanishing of the Wronskian to imply linear dependence.

**Proposition 59** (Slight generalization of Bôcher (1900)) Suppose that each of the \( \phi_k(\lambda) \)'s is locally analytic. If \( W[\Phi](\lambda) = 0 \) holds identically on \( \Lambda \), then \( \Phi(\lambda) = (\phi_1(\lambda), \phi_2(\lambda), \ldots, \phi_K(\lambda)) \) is linear dependent on \( \Lambda \).

**Proof of Proposition 59:** We follow Wolsson (1989) in defining a critical point \( \lambda \) of the set of functions \( \Phi \) as a point at which \( W[\Phi](\lambda) = 0 \), and the order of a critical point \( \lambda \)
as the largest positive integer \( r \) such that some \( r \times r \) sub-Wronskian of \( \Phi \) is not zero at \( \lambda \). If all sub-Wronskians vanish at \( \lambda \) (i.e., \( \phi_k (\lambda) = 0 \ \forall k \)), the order is defined to be zero. Apparently, if the maximum order on \( \Lambda \) is zero, then the \( \phi_k (\lambda) \)'s are identically zero, and it trivially follows that \( \Phi \) is linearly dependent. If the maximum order is \( r' > 0 \), which occurs at \( \lambda' \), then this implies that there is some \( r' \times r' \) sub-Wronskian of \( \Phi \) that is not equal to zero at \( \lambda' \). By continuity, the same \( r' \times r' \) sub-Wronskian must be non-zero in an open neighborhood of \( \lambda' \) as well, which we denote by \( \Lambda' \). Given that \( r' \) is the maximum order, we obtain that every point on \( \Lambda' \) is of order \( r' \). Theorem 3 of Wolsson (1989) shows that on any open interval consisting only of critical points of the same order \( r > 0 \), the space spanned by \( \Phi \) has dimension \( r \) and hence \( \Phi \) is linearly dependent. Thus we know that \( \Phi \) is linearly dependent on \( \Lambda' \). Namely, there exist \( c_1, c_2, \ldots, c_K \) such that \( \forall \lambda \in \Lambda', \)

\[
g(\lambda) \equiv \sum_{k=1}^{K} c_k \phi_k (\lambda) = 0. \tag{3.16}
\]

Suppose that \( \Lambda^* \) is a maximum open interval containing \( \Lambda' \) on which (3.16) holds. If \( \Lambda^* = \text{int}\Lambda \), then we are done. Suppose not. Then \( \Lambda^* \) is a proper subset of \( \text{int}\Lambda \), so there exists \( p \in \text{int}\Lambda \setminus \Lambda^* \) that is a boundary point of \( \Lambda^* \). Since each \( \phi_k \) is locally analytic at \( p \) and there are a finite number of them, we can find an open interval \( \tilde{\Lambda} \) around \( p \) on which all of the \( \phi_k \)'s are analytic. This in turn implies that \( g(\lambda) \) is analytic on \( \tilde{\Lambda} \). We know that a non-zero analytic function can only be zero at isolated points. Since \( g(\lambda) \) equals zero on \( \Lambda^* \cap \tilde{\Lambda} \), \( f(\lambda) \) must be identically zero on \( \tilde{\Lambda} \). This contradicts the condition that \( \Lambda^* \) is the maximum open interval on which (3.16) holds. Therefore, we must have \( \Lambda^* = \text{int}\Lambda \), implying that (3.16) holds on \( \text{int}\Lambda \), and by continuity, on the entire \( \Lambda \) as well. Hence, we establish that \( \Phi \) is linearly dependent on \( \Lambda \). 

\[\blacksquare\]

**Proof of Theorem 9:** (\( K \)-fund separation implies (1.3)) The sketch of the proof in the text shows that when \( M_0 (\lambda) \) given by (1.10) is invertible for some value of \( \lambda \), a utility function exhibiting \( K \)-fund separation must satisfy (1.3). To prove that (1.3) is necessary, we only
need to show that (1.3) can be obtained even when \( M_0 (\lambda) \) is not invertible for any value of \( \lambda \).

For this purpose, consider multiplying (1.2) by \( \lambda^l \) for \( l = 0, 1, \cdots, K \), and obtain

\[
\lambda^l I (\lambda \rho) = \sum_{k=1}^{K} \lambda^l \alpha_k (\lambda) f_k (\rho),
\]

where \( l = 0 \) corresponds to the basic case explored in the sketch of the proof. Taking derivatives with respect to \( \lambda \) yields

\[
\begin{pmatrix}
\frac{\partial [\lambda^l I (\lambda \rho)]}{\partial \lambda} \\
\vdots \\
\frac{\partial^K [\lambda^l I (\lambda \rho)]}{\partial \lambda^K}
\end{pmatrix} = M_l (\lambda) \begin{pmatrix}
f_1 (\rho) \\
\vdots \\
f_K (\rho)
\end{pmatrix},
\]

where \( M_l (\lambda) = \begin{pmatrix}
\frac{\partial [\lambda^l \alpha_1 (\lambda)]}{\partial \lambda} & \cdots & \frac{\partial [\lambda^l \alpha_K (\lambda)]}{\partial \lambda} \\
\vdots & \ddots & \vdots \\
\frac{\partial^K [\lambda^l \alpha_1 (\lambda)]}{\partial \lambda^K} & \cdots & \frac{\partial^K [\lambda^l \alpha_K (\lambda)]}{\partial \lambda^K}
\end{pmatrix} \). If any one of the \( M_l (\lambda) \)'s is invertible, we can follow the same procedures as we did in the sketch of the proof. Specifically, we solve for the \( f_k (\rho) \)'s as functions of \( I (\lambda \rho) \) and its derivatives, and plug them back into (1.2). Rearranging the terms and evaluating at \( \lambda = 1 \) (or any positive value) give us a differential equation of the form (1.13). Then, solving (1.13) yields \( I (\xi) \), which is again given by (1.3).

Suppose that for all \( l = 0, 1, \cdots, K \), \( M_l (\lambda) \) is not invertible. Notice that \( M_l (\lambda) \) is the Wronskian of \( \frac{\partial [\lambda^l \alpha_1 (\lambda)]}{\partial \lambda}, \frac{\partial [\lambda^l \alpha_2 (\lambda)]}{\partial \lambda}, \cdots, \frac{\partial [\lambda^l \alpha_K (\lambda)]}{\partial \lambda} \). Since the \( \alpha_k (\lambda) \)'s are locally analytic functions, we know that the \( \frac{\partial [\lambda^l \alpha_k (\lambda)]}{\partial \lambda} \)'s are also locally analytic. By Proposition 59, we have that the \( \frac{\partial [\lambda^l \alpha_k (\lambda)]}{\partial \lambda} \)'s are linearly dependent. In other words, there exist \( t_1^l, t_2^l, \cdots, t_K^l \) not all equal to zero such that

\[
\sum_{k=1}^{K} t_k^l \frac{\partial [\lambda^l \alpha_k (\lambda)]}{\partial \lambda} = 0.
\]

Integrating once and dividing by \( \lambda^l \) yield

\[
\sum_{k=1}^{K} t_k^l \alpha_k (\lambda) = \frac{t_0^l}{\lambda^l},
\]

(3.17)
for some constant $t_0^l$. This gives us $K + 1$ equations corresponding to $l = 0, 1, \cdots, K$. Since for each of these $K + 1$ equations, we have the same $K$ factors \(\{\alpha_k(\lambda)\}_{k=1}^K\) on the left hand side, there must exist $p^0, p^1, \cdots, p^K$ not all equal to zero such that

\[
\sum_{l=0}^K p^l \sum_{k=1}^K t^l_k \alpha_k(\lambda) = 0.
\]

By (3.17), this implies that

\[
\sum_{l=0}^K p^l t_0^l = 0.
\]

Since $1, \frac{1}{\lambda}, \cdots, \frac{1}{\lambda^K}$ are linearly independent, their linear combination vanishes only if all the coefficients are equal to zero, i.e., $p^l t_0^l = 0$ for all $l$. Since the $p^l$’s are not all equal to zero, there must be some $l$ for which $t_0^l = 0$. We will show that this cannot happen.

If $t_0^l$ were zero, by (3.17) we have

\[
\sum_{k=1}^K t_k^l \alpha_k(\lambda) = 0.
\]

Since the $t_k^l$’s are not all equal to zero, take any $k' \in \{1, 2, \cdots, K\}$ such that $t_0^k \neq 0$, and we have

\[
\alpha_{k'}(\lambda) = -\sum_{k \neq k'} \frac{t_k^l}{t_{k'}^l} \alpha_k(\lambda).
\]

Plugging this into (1.2) produces

\[
I(\lambda \rho) = \sum_{k \neq k'} \alpha_k(\lambda) f_k(\rho) - \sum_{k \neq k'} \frac{t_k^l}{t_{k'}^l} \alpha_k(\lambda) f_{k'}(\rho)
\]

\[= \sum_{k \neq k'} \alpha_k(\lambda) \left[ f_k(\rho) - \frac{t_k^l}{t_{k'}^l} f_{k'}(\rho) \right].\]

Thus, we have (no more than) $(K - 1)$-fund separation, with the separating funds being $f_k(\rho) - \frac{t_k^l}{t_{k'}^l} f_{k'}(\rho)$ for $k \neq k'$. This contradicts $K$-fund separation, therefore contradicting $t_0^l = 0$. Hence, we know that $t_0^l \neq 0$.

Hence, we have proved that we can always find $l \in \{0, 1, \cdots, K\}$ such that $M_l(\lambda)$ is invertible and consequently (1.3) can be obtained.
((1.3) implies $K$-fund separation) To establish sufficiency we need to show that any utility function $u$ satisfying (1.3) indeed exhibits $K$-fund separation. In other words, the optimal consumption portfolio $I(\lambda \rho)$ can be expressed in the form of (1.2). To this end, we have

$$I(\lambda \rho) = \sum_{k=1}^{J} (\lambda \rho)^{\gamma_k} \cos (b_k \log (\lambda \rho)) P_{k,1} (\log (\lambda \rho)) + \sum_{k=1}^{J} (\lambda \rho)^{\gamma_k} \sin (b_k \log (\lambda \rho)) P_{k,2} (\log (\lambda \rho))$$

$$= \sum_{k=1}^{J} (\lambda \rho)^{\gamma_k} \cos (b_k \log (\lambda \rho)) \sum_{j=0}^{d_{k,1}} C_{k,1,j} (\log (\lambda \rho))^j$$

$$+ \sum_{k=1}^{J} (\lambda \rho)^{\gamma_k} \sin (b_k \log (\lambda \rho)) \sum_{j=0}^{d_{k,2}} C_{k,2,j} (\log (\lambda \rho))^j$$

$$= \sum_{k=1}^{J} (\lambda \rho)^{\gamma_k} \left[ \cos (b_k \log \lambda) \cos (b_k \log \rho) - \sin (b_k \log \lambda) \sin (b_k \log \rho) \right] \sum_{l=1}^{d_{k,1}} (\log \rho)^l \sum_{j=0}^{d_{k,1}} C_{k,1,j} \left( \left( \begin{array}{c} j \\ l \end{array} \right) \right) (\log \lambda)^{j-l}$$

$$+ \sum_{k=1}^{J} (\lambda \rho)^{\gamma_k} \left[ \sin (b_k \log \lambda) \cos (b_k \log \rho) + \cos (b_k \log \lambda) \sin (b_k \log \rho) \right] \sum_{l=1}^{d_{k,2}} (\log \rho)^l \sum_{j=0}^{d_{k,2}} C_{k,2,j} \left( \left( \begin{array}{c} j \\ l \end{array} \right) \right) (\log \lambda)^{j-l}$$

$$= \sum_{(k,b_k=0)} \sum_{l=0}^{d_k} \alpha_{k,l} (\lambda) f_{k,l}(\rho) + \sum_{(k,b_k \neq 0)} \sum_{l=0}^{d_k} \left[ \alpha_{k,1,l} (\lambda) f_{k,1,l}(\rho) + \alpha_{k,2,l} (\lambda) f_{k,2,l}(\rho) \right],$$

where $d_k = \max_{i=1,2} (d_{k,i})$ and $f_{k,l}(\rho)$, $f_{k,1,l}(\rho)$, $f_{k,2,l}(\rho)$, $\alpha_{k,l}(\lambda)$, $\alpha_{k,1,l}(\lambda)$, $\alpha_{k,2,l}(\lambda)$ are given by (1.4)–(1.9). Notice that when $b_k = 0$, $d_{k,2}$ is set to be 0, and hence $d_k = d_{k,1}$.

**Proof of Theorem 13:** This proof relies on the strict concavity requirement. See Section 1.6 for a detailed discussion.

(1) When $K = 1$, the class of one-fund separating preferences has

$$I(\xi) = C \xi^{\gamma}.$$

If $\gamma > 0$, then we need $C < 0$ to guarantee strict concavity. In this case, $I(\xi) \in (-\infty, 0).$ Similarly, if $\gamma < 0$, we need $C > 0$, in which case we have $I(\xi) \in (0, +\infty).$

(2) When $K \geq 2$, since $I(\xi)$ is a continuous function and $\xi \in (0, +\infty)$ is open and connected, we know that $\{I(\xi) : \xi > 0\}$ must also be open and connected. This implies that
\{I (\xi) : \xi > 0\} is an open interval.

On the other hand, \{I (\xi) : \xi > 0\} is bounded only if

\[ I (\xi) = C, \]

or

\[ I (\xi) = \sum_k [C_{k,1} \cos (b_k \log \xi) + C_{k,2} \cos (b_k \log \xi)], \]

where \( C, \{C_{k,1}\}_k \) and \( \{C_{k,2}\}_k \) are constants. However, strict concavity implies that neither form is possible (see Proposition 23). Therefore, \{I (\xi) : \xi > 0\} must be unbounded.

Finally, we need to show that \{I (\xi) : \xi > 0\} can be any open unbounded interval. To this end, let us consider a simple case in which \( K = 2 \) and

\[ I (\xi) = C_1 \xi^{\gamma_1} + C_2 \xi^{\gamma_2}, \]

for \( \gamma_1 \neq \gamma_2 \). Without loss of generality, suppose \( \gamma_1 < \gamma_2 \). Then we have the following three cases.

Case 1: If \( \gamma_1 = 0 \) and \( \gamma_2 > 0 \), then strict concavity implies \( C_2 < 0 \), and hence, \( I (\xi) \in (-\infty, C_1) \). We can change the value of \( C_1 \) to obtain any open interval that is bounded from above but unbounded from below.

Case 2: If \( \gamma_1 < 0 \) and \( \gamma_2 = 0 \), then strict concavity implies \( C_1 > 0 \), and hence, \( I (\xi) \in (C_2, +\infty) \). We can change the value of \( C_2 \) to obtain any open interval that is bounded from below but unbounded from above.

Case 3: If \( \gamma_1 < 0 \) and \( \gamma_2 > 0 \), then strict concavity implies that \( C_1 > 0 \) and \( C_2 < 0 \). In this case, \( I (\xi) \in (-\infty, +\infty) \).

Proof of Proposition 22: Differentiating (1.36) yields

\[ I' (\xi) = \sum_{k=1}^K \gamma_k C_k \xi^{\gamma_k-1}. \]  

(3.18)
Parts 1 and 2: If \( \gamma_1 \neq 0 \), we can rewrite (3.18) as

\[
P' (\xi) = \xi^{\gamma_1 - 1} \left( \gamma_1 C_1 + \sum_{k=2}^{K} \gamma_k C_k \xi^{\gamma_k - \gamma_1} \right).
\]

Since \( \xi^{\gamma_1 - 1} > 0 \), strict concavity requires \( \forall \xi > 0 \),

\[
\gamma_1 C_1 + \sum_{k=2}^{K} \gamma_k C_k \xi^{\gamma_k - \gamma_1} < 0.
\]

As \( \xi \) approaches 0, we obtain

\[
\lim_{\xi \to 0} \gamma_1 C_1 + \sum_{k=2}^{K} \gamma_k C_k \xi^{\gamma_k - \gamma_1} = \gamma_1 C_1.
\]

Hence, we must have \( \gamma_1 C_1 < 0 \).

Similarly, if \( \gamma_1 = 0 \), strict concavity implies \( \gamma_2 C_2 < 0 \).

Parts 3 and 4: If \( \gamma_K \neq 0 \), we can rewrite (3.18) as

\[
P' (\xi) = \xi^{\gamma_K - 1} \left( \sum_{k=1}^{K-1} \gamma_k C_k \xi^{\gamma_k - \gamma_K} + \gamma_K C_K \right),
\]

and strict concavity requires \( \forall \xi > 0 \),

\[
\sum_{k=1}^{K-1} \gamma_k C_k \xi^{\gamma_k - \gamma_K} + \gamma_K C_K < 0.
\]

As \( \xi \) approaches infinity, we obtain

\[
\lim_{\xi \to +\infty} \sum_{k=1}^{K-1} \gamma_k C_k \xi^{\gamma_k - \gamma_K} + \gamma_K C_K = \gamma_K C_K,
\]

which indicates \( \gamma_K C_K < 0 \).

Similarly, if \( \gamma_K = 0 \), strict concavity implies \( \gamma_{K-1} C_{K-1} < 0 \).

**Proof of Proposition 23:** Our goal is to show that in each case we can find \( \xi > 0 \) such that \( P' (\xi) \geq 0 \).

Case 1. This case can be rewritten as

\[
I (\xi) = \sum_{k=0}^{d} C_k \left( \log \xi \right)^k,
\]

153
where $C_d \neq 0$ and $d \geq 2$ is an even number. Taking derivative yields

$$I'(\xi) = \sum_{k=1}^{d} kC_k \frac{(\log \xi)^{k-1}}{\xi}.$$ 

For all $\xi \neq 1$, this is equivalent to

$$I'(\xi) = \frac{(\log \xi)^{d-2}}{\xi} \sum_{k=1}^{d} kC_k (\log \xi)^{k-d+1}. $$

Since $d$ is even, we have $\frac{(\log \xi)^{d-2}}{\xi} > 0$. On the other hand, it is easy to verify that if $C_d < 0$,

$$\lim_{\xi \to 0} \sum_{k=1}^{d} kC_k (\log \xi)^{k-d+1} = +\infty,$$

and that if $C_d > 0$,

$$\lim_{\xi \to +\infty} \sum_{k=1}^{d} kC_k (\log \xi)^{k-d+1} = +\infty.$$ 

This suggests that we can always find $\xi$ small or large enough such that $I'(\xi) \geq 0$.

Case 2. This case can be rewritten as

$$I(\xi) = \sum_{k=0}^{d} C_k \xi^\gamma (\log \xi)^k,$$

where $C_d \neq 0$, $\gamma \neq 0$, and $d$ is an odd positive integer. Taking derivative yields

$$I'(\xi) = \xi^{\gamma-1} \left( \sum_{k=0}^{d-1} (C_k \gamma + C_{k+1} (k + 1)) (\log \xi)^k + C_d \gamma (\log \xi)^d \right).$$

For all $\xi \neq 1$, this is equivalent to

$$I'(\xi) = \xi^{\gamma-1} (\log \xi)^{d-1} \left( \sum_{k=0}^{d-1} (C_k \gamma + C_{k+1} (k + 1)) (\log \xi)^{k-d+1} + C_d \gamma \log \xi \right).$$

Since $d$ is odd, we know $\xi^{\gamma-1} (\log \xi)^{d-1} > 0$. On the other hand, it is easy to verify that if $C_d \gamma < 0$,

$$\lim_{\xi \to 0} \sum_{k=0}^{d-1} (C_k \gamma + C_{k+1} (k + 1)) (\log \xi)^{k-d+1} + C_d \gamma \log \xi = +\infty,$$
and that if $C_d \gamma > 0$,

$$
\lim_{\xi \to 0} \xi^{-d-1} \sum_{k=0}^{d-1} (C_k \gamma + C_{k+1} (k + 1)) (\log \xi)^{k-d+1} + C_d \gamma \log \xi = +\infty.
$$

Hence, we can always find $\xi$ small or large enough such that $I'(\xi) \geq 0$.

Case 3. This case can be rewritten as

$$
I(\xi) = \sum_{k=1}^{J} \sum_{l=0}^{d_k} [C_{k,1,l} \cos (b_k \log \xi) + C_{k,2,l} \sin (b_k \log \xi)] \xi^{\gamma_k} (\log \xi)^l,
$$

where for any $k$ we have $b_k \neq 0$ and at least one of $C_{k,1,d_k}$ and $C_{k,2,d_k}$ is not zero.

Without loss of generality, we assume $\gamma_1 < \gamma_2 < \cdots < \gamma_J$. Taking derivative yields

$$
I'(\xi) = \sum_{k=1}^{J} \xi^{\gamma_k-1} \sum_{l=0}^{d_k-1} \left[ (C_{k,1,l} \gamma_k + C_{k,2,l} b_k + C_{k,1,l+1} (l + 1)) \cos (b_k \log \xi) - (C_{k,1,l} b_k - C_{k,2,l} \gamma_k - C_{k,2,l+1} (l + 1)) \sin (b_k \log \xi) \right] (\log \xi)^l.
$$

For all $\xi \neq 1$, this is equivalent to

$$
I'(\xi) = \xi^{\gamma_J-1} \sum_{k=1}^{d_J} \xi^{\gamma_k-\gamma_J} (\log \xi)^{d_k}
$$

$$
\sum_{l=0}^{d_k-1} \left[ (C_{k,1,l} \gamma_k + C_{k,2,l} b_k + C_{k,1,l+1} (l + 1)) \cos (b_k \log \xi) - (C_{k,1,l} b_k - C_{k,2,l} \gamma_k - C_{k,2,l+1} (l + 1)) \sin (b_k \log \xi) \right] (\log \xi)^{l-d_k}.
$$

It is easy to verify that

$$
\lim_{\xi \to +\infty} I'(\xi) = \lim_{\xi \to +\infty} \xi^{\gamma_J-1} (\log \xi)^{d_J}
$$

$$
\left[ (C_{J,1,d_J} \gamma_J + C_{J,2,d_J} b_J) \cos (b_J \log \xi) - (C_{J,1,d_J} b_J - C_{J,2,d_J} \gamma_J) \sin (b_J \log \xi) \right],
$$

which apparently does not converge due to the cyclicity of $\cos (\cdot)$ and $\sin (\cdot)$. In fact, $I'(\xi)$ switches between positive and negative values as $\xi$ approaches infinity. Therefore, we can always find $\xi$ such that $I'(\xi) \geq 0$. ■
Appendix B: Proofs, Derivations and Discussions for Chapter 2

Proofs

Proof of Proposition 30: We need to show that for any random returns \( \hat{z}_1 \) and \( \hat{z}_2 \), and any \( 0 < \lambda < 1 \),

\[
w_k (\lambda \hat{z}_1 + (1 - \lambda) \hat{z}_2) \leq \lambda w_k (\hat{z}_1) + (1 - \lambda) w_k (\hat{z}_2) .
\]  

(3.19)

Letting \( \hat{z}_1 = \hat{z}_1 - \text{E}(\hat{z}_1) \) and \( \hat{z}_2 = \hat{z}_2 - \text{E}(\hat{z}_2) \), (3.19) can be rewritten as

\[
\left( \text{E} \left[ (\lambda \hat{z}_1 + (1 - \lambda) \hat{z}_2)^k \right] \right)^{\frac{1}{k}} \leq \lambda \left( \text{E} \left[ \hat{z}_1^k \right] \right)^{\frac{1}{k}} + (1 - \lambda) \left( \text{E} \left[ \hat{z}_2^k \right] \right)^{\frac{1}{k}} .
\]

(3.20)

Applying the binomial formula to the LHS of (3.20) implies that we need to show

\[
\left( \sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} (1 - \lambda)^i \text{E}(\hat{z}_1^{k-i} \hat{z}_2^i) \right)^{\frac{1}{k}} \leq \lambda \left( \text{E} \left[ \hat{z}_1^k \right] \right)^{\frac{1}{k}} + (1 - \lambda) \left( \text{E} \left[ \hat{z}_2^k \right] \right)^{\frac{1}{k}} .
\]

Since \( k \) is even, replacing each \( \hat{z}_1 \) and \( \hat{z}_2 \) with \( |\hat{z}_1| \) and \( |\hat{z}_2| \) will not affect the RHS, but it might increase the LHS. So, it is sufficient to show that

\[
\left( \sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} (1 - \lambda)^i \text{E}(|\hat{z}_1^{k-i} \hat{z}_2^i|) \right)^{\frac{1}{k}} \leq \lambda \left( \text{E} \left[ |\hat{z}_1|^k \right] \right)^{\frac{1}{k}} + (1 - \lambda) \left( \text{E} \left[ |\hat{z}_2|^k \right] \right)^{\frac{1}{k}} .
\]

Since both sides are positive we can raise both sides to the \( k^{th} \) power, maintaining the inequality. Thus, it would be sufficient to show

\[
\sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} (1 - \lambda)^i \text{E}(|\hat{z}_1^{k-i} \hat{z}_2^i|) \leq \left( \lambda \left( \text{E} \left[ |\hat{z}_1|^k \right] \right)^{\frac{1}{k}} + (1 - \lambda) \left( \text{E} \left[ |\hat{z}_2|^k \right] \right)^{\frac{1}{k}} \right)^{k} .
\]
Applying the binomial formula to the RHS implies that it would be sufficient to show
\[ \sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} (1 - \lambda)^i \mathbb{E}(|\hat{z}_1^{k-i} \hat{z}_2^i|) \leq \sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} (1 - \lambda)^i \left( \mathbb{E}(|\hat{z}_1|^{k}) \right)^{\frac{k-i}{k}} \left( \mathbb{E}(|\hat{z}_2|^{k}) \right)^{\frac{i}{k}}. \]

To establish this inequality we will show that it actually holds term by term. That is, it is sufficient to show that for each \( i = 0, ..., k, \)
\[ \mathbb{E}(|\hat{z}_1^{k-i} \hat{z}_2^i|) \leq \left( \mathbb{E}(|\hat{z}_1|^{k}) \right)^{\frac{k-i}{k}} \left( \mathbb{E}(|\hat{z}_2|^{k}) \right)^{\frac{i}{k}}. \]
To see this, note that it is equivalent to show that
\[ \mathbb{E}(|\hat{z}_1^{k-i} \hat{z}_2^i|) \leq \left( \mathbb{E}(|\hat{z}_1|^{k}) \right)^{\frac{k-i}{k}} \left( \mathbb{E}(|\hat{z}_2|^{k}) \right)^{\frac{i}{k}}. \]
But, this is immediate from Hölder’s inequality, and we are done. \( \blacksquare \)

**Proof of Proposition 33:** For any integer \( k \geq 2, \) we can rewrite the downside risk measure as
\[ \text{DR}_k (\hat{z}) = (-1)^k \left( \mathbb{E}([\hat{z} - \mathbb{E}(\hat{z})]^k) \right)^{\frac{1}{k}} = \left( \mathbb{E}((\hat{z} - \hat{\hat{z}})^+)^k \right)^{\frac{1}{k}}, \]
where \([t]^+ = \max (t, 0)\) for \( t \in \mathbb{R}.\)

Consider any two random returns \( \hat{z}_1 \) and \( \hat{z}_2, \) and let \( \hat{\hat{z}}_1 = \mathbb{E}(\hat{z}_1) - \hat{\hat{z}}_1 \) and \( \hat{\hat{z}}_2 = \mathbb{E}(\hat{z}_2) - \hat{\hat{z}}_2 \). Obviously, we have \( \hat{\hat{z}}_1 \geq 0 \) and \( \hat{\hat{z}}_2 \geq 0. \) What we need to show is that for any \( 0 < \lambda < 1, \)
\[ \left( \mathbb{E}((\lambda \hat{z}_1 + (1 - \lambda) \hat{z}_2) - \lambda \hat{\hat{z}}_1 - (1 - \lambda) \hat{\hat{z}}_2)^+)^k \right)^{\frac{1}{k}} \leq \lambda \left( \mathbb{E}(\hat{z}_1^k) \right)^{\frac{1}{k}} + (1 - \lambda) \left( \mathbb{E}(\hat{z}_2^k) \right)^{\frac{1}{k}}. \]
Now,
\[ \mathbb{E}(\lambda \hat{z}_1 + (1 - \lambda) \hat{z}_2) - \lambda \hat{\hat{z}}_1 - (1 - \lambda) \hat{\hat{z}}_2^+] = \lambda (\mathbb{E}(\hat{z}_1) - \hat{\hat{z}}_1) + (1 - \lambda) (\mathbb{E}(\hat{z}_2) - \hat{\hat{z}}_2)]^+ \]
\[ \leq \lambda [\mathbb{E}(\hat{z}_1) - \hat{\hat{z}}_1]^+ + (1 - \lambda) [\mathbb{E}(\hat{z}_2) - \hat{\hat{z}}_2]^+ \]
\[ = \lambda \hat{\hat{z}}_1 + (1 - \lambda) \hat{\hat{z}}_2, \]
where the inequality follows from Jensen’s inequality using that \([\cdot]^+\) is a convex function.
Therefore, it is sufficient to show that
\[
\left( \mathbb{E} \left( \lambda \tilde{z}_1 + (1 - \lambda) \tilde{z}_2 \right)^k \right)^{\frac{1}{k}} \leq \lambda \left( \mathbb{E} \left( \tilde{z}_1^k \right) \right)^{\frac{1}{k}} + (1 - \lambda) \left( \mathbb{E} \left( \tilde{z}_2^k \right) \right)^{\frac{1}{k}}.
\]

The rest of the proof follows closely the proof of Proposition 30. Indeed, since \( \tilde{z}_1 \) and \( \tilde{z}_2 \) are non-negative here, the arguments in the proof of Proposition 30 apply in this case to any positive \( k \) (odd or even).

**Proof of Proposition 36:** In the definition of expected shortfall we assumed the existence of a cumulative distribution function \( F(\cdot) \) applied to realizations of random variables. For the sake of this proof it will be more useful to work directly with the state space \( \Omega \) and with the underlying probability measure \( P(\cdot) \). We first prove that \( \text{ES}_\delta(\tilde{z}) \) is subadditive. That is, for any two random returns \( \tilde{z}_1 \) and \( \tilde{z}_2 \),
\[
\text{ES}_\delta(\tilde{z}_1 + \tilde{z}_2) \leq \text{ES}_\delta(\tilde{z}_1) + \text{ES}_\delta(\tilde{z}_2).
\]
If either \( \tilde{z}_1 \) or \( \tilde{z}_2 \) is equal to a constant with probability 1, then the result is immediate. We shall thus only consider the case in which both of them are not equal to a constant. By (2.3), for any random return \( \tilde{z} \) (which is not constant), \( \text{ES}_\delta(\tilde{z}) \) can be expressed as
\[
\text{ES}_\delta(\tilde{z}) = \frac{1}{\delta} \int_{\{\omega: \tilde{z} \leq -\text{VaR}_\delta(\tilde{z})\}} \tilde{z} dP(\omega).
\]

Let \( \tilde{z}_1 \) and \( \tilde{z}_2 \) be random returns and define \( \tilde{z}_3 = \tilde{z}_1 + \tilde{z}_2 \). Let \( \Omega_i = \{\omega \in \Omega : \tilde{z}_i \leq -\text{VaR}_\delta(\tilde{z}_i)\} \) for \( i = 1, 2, 3 \). Then, (3.21) is equivalent to
\[
\int_{\Omega_3} \tilde{z}_3 dP(\omega) \geq \int_{\Omega_1} \tilde{z}_1 dP(\omega) + \int_{\Omega_2} \tilde{z}_2 dP(\omega),
\]
which can be rewritten as
\[
\int_{\Omega_3} \tilde{z}_1 dP(\omega) + \int_{\Omega_3} \tilde{z}_2 dP(\omega) \geq \int_{\Omega_1} \tilde{z}_1 dP(\omega) + \int_{\Omega_2} \tilde{z}_2 dP(\omega).
\]

It is sufficient to show that
\[
\int_{\Omega_3} \tilde{z}_1 dP(\omega) \geq \int_{\Omega_1} \tilde{z}_1 dP(\omega),
\]
(3.22)
and
\[ \int_{\Omega_3} \tilde{z}_2 dP(\omega) \geq \int_{\Omega_2} \tilde{z}_2 dP(\omega). \] (3.23)

For brevity, we will only prove (3.22). The proof of (3.23) is parallel.

Define
\[
\begin{align*}
\Omega_4 &= \{ \omega \in \Omega : \tilde{z}_1 \leq -\text{VaR}_\delta (\tilde{z}_1), \tilde{z}_3 \leq -\text{VaR}_\delta (\tilde{z}_3) \}, \\
\Omega_5 &= \{ \omega \in \Omega : \tilde{z}_1 \leq -\text{VaR}_\delta (\tilde{z}_1), \tilde{z}_3 > -\text{VaR}_\delta (\tilde{z}_3) \}, \text{ and} \\
\Omega_6 &= \{ \omega \in \Omega : \tilde{z}_1 > -\text{VaR}_\delta (\tilde{z}_1), \tilde{z}_3 \leq -\text{VaR}_\delta (\tilde{z}_3) \}.
\end{align*}
\]

Clearly, \( \Omega_4 \cap \Omega_5 = \emptyset \), \( \Omega_4 \cup \Omega_5 = \Omega_1 \), \( \Omega_4 \cap \Omega_6 = \emptyset \), and \( \Omega_4 \cup \Omega_6 = \Omega_3 \). Thus,
\[ \int_{\Omega_1} dP(\omega) = \int_{\Omega_4} dP(\omega) + \int_{\Omega_5} dP(\omega), \]
and
\[ \int_{\Omega_3} dP(\omega) = \int_{\Omega_4} dP(\omega) + \int_{\Omega_6} dP(\omega). \]

By the definition of VaR, we know
\[ \int_{\Omega_1} dP(\omega) = \int_{\Omega_3} dP(\omega) = \delta. \]

Thus, we obtain
\[ \int_{\Omega_5} dP(\omega) = \int_{\Omega_6} dP(\omega). \] (3.24)

Similarly, we have
\[ \int_{\Omega_1} \tilde{z}_1 dP(\omega) = \int_{\Omega_4} \tilde{z}_1 dP(\omega) + \int_{\Omega_5} \tilde{z}_1 dP(\omega), \]
and
\[ \int_{\Omega_3} \tilde{z}_1 dP(\omega) = \int_{\Omega_4} \tilde{z}_1 dP(\omega) + \int_{\Omega_6} \tilde{z}_1 dP(\omega). \]
Hence,

\[
\int_{\Omega_5} \tilde{z}_1 dP(\omega) - \int_{\Omega_6} \tilde{z}_1 dP(\omega)
\]

\[
= \int_{\Omega_5} \tilde{z}_1 dP(\omega) - \int_{\Omega_6} \tilde{z}_1 dP(\omega)
\]

\[
\leq \int_{\Omega_5} [-\text{VaR}_\delta(\tilde{z}_1)] dP(\omega) - \int_{\Omega_6} [-\text{VaR}_\delta(\tilde{z}_1)] dP(\omega)
\]

\[
= -\text{VaR}_\delta(\tilde{z}_1) \left[ \int_{\Omega_5} dP(\omega) - \int_{\Omega_6} dP(\omega) \right]
\]

\[
= 0,
\]

where the inequality follows from \( \tilde{z}_1 \leq -\text{VaR}_\delta(\tilde{z}_1) \) when \( \omega \in \Omega_5 \) and \( \tilde{z}_1 > -\text{VaR}_\delta(\tilde{z}_1) \) when \( \omega \in \Omega_6 \), and where the last equality follows from (3.24). Therefore, (3.22) is obtained, and hence \( \text{ES}_\delta(\tilde{z}) \) is subadditive. Since \( \text{DES}_\delta(\tilde{z}) = \text{ES}_\delta(\tilde{z}) + \text{E}(\tilde{z}) \) we have that \( \text{DES} \) is also subadditive.

Convexity now follows immediately from homogeneity of degree 1 and subadditivity.

Proof of Theorem 39: Our setting is a special case of the setting in Nielsen (1989).

To show the existence of equilibrium Nielsen requires that preferences satisfy the following three conditions: (i) each investor’s choice set is closed and convex, and contains her initial endowment; (ii) The set of \( \{ \zeta \in \mathbb{R}^{n+1} : U^j(\zeta) \geq U^j(\zeta') \} \) is closed for all \( \zeta' \in \mathbb{R}^{n+1} \) and for all \( j = 1, \ldots, \ell \); (iii) If \( \zeta, \zeta' \in \mathbb{R}^{n+1} \) and \( U^j(\zeta') > U^j(\zeta) \), then \( U^j(t\zeta' + (1-t)\zeta) > U^j(\zeta) \) for all \( t \) in \( (0,1) \).

Condition (i) is satisfied in our setting since the choice set of each investor is \( \mathbb{R}^{n+1} \), which is closed and convex, and contains \( e^j \) for all \( j \). Condition (ii) holds since \( V \) is assumed continuous and \( R \) is assumed smooth, and so their composition is continuous. Condition (iii) follows since \( V(\cdot) \) is quasi-concave, strictly increasing in its first argument and strictly decreasing in its second argument, and \( R(\cdot) \) is a convex risk measure.

Given these properties of the preferences, Nielsen (1989) establishes two conditions as sufficient for the existence of a quasi-equilibrium: (i) positive semi-independence of directions
of improvement, and (ii) non-satiation at Pareto attainable portfolios. Condition (i) follows in our setting as in Nielsen (1990, Proposition 1) since in our setting all investors agree on all parameters of the problem (in particular on the expected returns), and due to the non-redundancy of risky assets assumption. To see why condition (ii) holds in our setting note that we assume the existence of a risk-free asset paying a non-zero payoff with probability 1. Since $R(\cdot)$ satisfies the risk-free property, we have that $R(\tilde{z}_1 + \tilde{z}_2) \leq R(\tilde{z}_1)$ whenever $\tilde{z}_2$ is risk-free with $P(\{\tilde{z}_2 > 0\}) = 1$. Thus, adding a positive risk-free asset can only (weakly) reduce risk. It follows that we can always add this positive risk-free asset to any bundle $\zeta$, strictly increasing the expected return while weakly decreasing risk. This implies that in our model there is no satiation globally. Thus, a quasi-equilibrium exists in our setting. Moreover, any quasi-equilibrium is, in fact, an equilibrium in our setting. This follows from the conditions in Nielsen (1989 p. 469). Indeed, in our setting each investor’s choice set is convex and unbounded, and the set $\{\zeta \in \mathbb{R}^{n+1} : U^j(\zeta) > U^j(\zeta')\}$ is open for all $j$ and $\zeta' \in \mathbb{R}^{n+1}$.

**Proof of Theorem 40:** Suppose that the equilibrium bundle of investor $j$ is $\zeta^j$. Let $\bar{x}^j = \sum_{i=0}^{n} x^j_i = p \cdot \zeta^j$ be the total dollar amount of investment of investor $j$. Then,

$$U^j(\zeta^j) = V^j \left( E \left( \sum_{i=0}^{n} \zeta^j_i \bar{y}_i \right) , R \left( \sum_{i=0}^{n} \zeta^j_i \bar{y}_i \right) \right)$$

$$= V^j \left( \bar{x}^j E \left( \sum_{i=0}^{n} \frac{\zeta^j_i p_i \bar{y}_i}{\bar{x}^j} \frac{\bar{y}_i}{p_i} \right) , R \left( \bar{x}^j \sum_{i=0}^{n} \frac{\zeta^j_i p_i \bar{y}_i}{\bar{x}^j} \frac{\bar{y}_i}{p_i} \right) \right)$$

$$= V^j \left( \bar{x}^j \sum_{i=0}^{n} \frac{x^j_i \bar{z}_i}{\bar{x}^j} \frac{\bar{z}_i}{\bar{x}^j} \right) , R \left( \bar{x}^j \sum_{i=0}^{n} \frac{x^j_i \bar{z}_i}{\bar{x}^j} \frac{\bar{z}_i}{\bar{x}^j} \right)$$

$$= V^j \left( \bar{x}^j \sum_{i=0}^{n} \frac{\alpha^j_i \bar{z}_i}{\bar{x}^j} \frac{\bar{z}_i}{\bar{x}^j} \right) , R \left( \bar{x}^j \sum_{i=0}^{n} \frac{\alpha^j_i \bar{z}_i}{\bar{x}^j} \frac{\bar{z}_i}{\bar{x}^j} \right)$$

$$= V^j \left( \bar{x}^j E \left( \alpha^j \cdot \bar{z} \right) , R \left( \bar{x}^j \left( \alpha^j \cdot \bar{z} \right) \right) \right).$$

161
From the definition of equilibrium, each investor chooses $\zeta^j$ to maximize $U^j (\zeta^j)$ subject to $\bar{x}^j \leq p \cdot e^j$, where by the positivity of prices $\bar{x}^j = p \cdot e^j > 0$ (using that $e^j \in \mathbb{R}^{n+1}_+$ is not zero by assumption). From (3.25) and since $V^j$ is strictly increasing in the first argument and strictly decreasing in the second argument, we have that for any positive $\bar{x}^j$, $U^j (\zeta^j)$ is strictly increasing in $E (\alpha^j \cdot \tilde{z})$ and strictly decreasing in $R (\bar{x}^j (\alpha^j \cdot \tilde{z}))$. Therefore, in equilibrium, $\alpha^j$ must minimize $R (\bar{x}^j (\alpha \cdot \tilde{z}))$ for a given level of expected return $E (\alpha^j \cdot \tilde{z})$. By scaling independence, this is equivalent to minimizing $R (\alpha \cdot \tilde{z})$ for a given level of expected return, and thus, to solving Problem (2.7). The solution is unique since we assumed that $R (\cdot)$ is a convex risk measure, and so $R (\alpha \cdot \tilde{z})$ is convex as a function of $\alpha$. ■

**Proof of Theorem 44:** By the smoothness of $R (\cdot)$ and by Theorem 39, the solution to Problem (2.7) for some $\mu^j = \mu$ is determined by the first order conditions. To solve this program, form the Lagrangian

$$
L (\alpha) = R (\alpha) - \xi \left( \sum_{i=1}^{n} \alpha_i E (\tilde{z}_i) + \left(1 - \sum_{i=1}^{n} \alpha_i\right) r_f - \mu \right),
$$

where $\xi$ is a Lagrange multiplier. Equivalently,

$$
L (\alpha) = R \left(1 - \sum_{i=1}^{n} \alpha_i, \alpha_1, \ldots, \alpha_n\right) - \xi \left( \sum_{i=1}^{n} \alpha_i E (\tilde{z}_i) + \left(1 - \sum_{i=1}^{n} \alpha_i\right) r_f - \mu \right).
$$

The first order condition states that for all $i = 1, \ldots, n$,

$$
-R_0 (\alpha^*) + R_i (\alpha^*) - \xi (E (\tilde{z}_i) - r_f) = 0,
$$

where $\alpha^*$ is any efficient portfolio (the market portfolio being a special case). By the risk-free property, $R_0 (\alpha^*) = 0$. Hence,

$$
R_i (\alpha^*) = \xi (E (\tilde{z}_i) - r_f).
$$

It follows that

$$
\sum_{i=1}^{n} \alpha_i^* R_i (\alpha^*) = \xi \sum_{i=1}^{n} \alpha_i^* (E (\tilde{z}_i) - r_f) = \xi (E (\tilde{z} \cdot \alpha^*) - r_f).
$$
From (3.27) and (3.28) we obtain

\[
\frac{R_i(\alpha^*)}{\sum_{h=1}^{n} \alpha_h^* R_h(\alpha^*)} = \frac{E(\tilde{z}_i) - r_f}{E(\tilde{z} \cdot \alpha^*) - r_f},
\]

as required. If \( R \) is homogeneous of degree \( k \), then by Euler’s homogeneous function theorem and using the risk-free property this is also equivalent to

\[
\frac{R_i(\alpha^*)}{k R(\alpha^*)} = \frac{E(\tilde{z}_i) - r_f}{E(\tilde{z} \cdot \alpha^*) - r_f}.
\]

\[\blacksquare\]

**Proof of Theorem 49:** The proof relies on a mapping between risk allocation problems as defined in Section 2.5.1 and cost allocation problems as defined in Billera and Heath (1982, hereafter BH). Specifically, BH define a cost allocation problem of order \( n \) as a pair \((h, x)\) where \( h : \mathbb{R}_{+}^n \to \mathbb{R} \) is continuously differentiable and \( h(0) = 0 \). Since \( R \) is smooth and satisfies \( R(0) = 0 \) we can view any risk allocation problem, \((R, x)\), of order \( n \), as a cost allocation problem as defined in BH by setting \( h(x) = R(x \cdot \tilde{z}) \). Given this mapping we will use \((R, x)\) to denote both the risk allocation problem and its corresponding cost allocation problem. BH define a cost allocation procedure as a function assigning each cost allocation problem \((R, x)\) of order \( n \) a vector \( c(R, x) \in \mathbb{R}^n \). That is, \( c(R, x) \) should be interpreted as the cost allocated to each of the \( n \) goods or services.

We can then consider a natural mapping between systematic risk measures as defined in Section 2.5.1 and the BH cost allocation procedures as follows. If \( B^R(x) \) is a systematic risk measure of the risk allocation problem \((R, x)\), then

\[
c(R, x) = \frac{B^R(x) R(x \cdot \tilde{z})}{x}
\]

is a cost allocation procedure for the corresponding cost allocation problem \((R, x)\). Namely, risk allocation measures can be viewed as scaled versions of cost allocation procedures for the corresponding problems.
Lemma 60  If a systematic risk measure $B^R(x)$ satisfies Axioms 1-4, then the corresponding cost allocation procedure $c(R, x)$ satisfies Conditions (2.1)-(2.4) in BH.

It is important to note that Axioms 1-4 and Conditions (2.1)-(2.4) in BH are not equivalent to each other either as a group or individually. Rather, our four axioms as a set are stronger than their four conditions as a set. The proof of this lemma follows from the next four steps.

Step 1. Axiom 1 is satisfied if and only if Condition (2.1) in BH holds. Indeed, $\sum_{i=1}^n \alpha_i B^R_i(x) = 1$ is equivalent to $\sum_{i=1}^n x_i R(x) B^R_i(x) / \bar{x} = R(x)$, which using (3.29) is equivalent to $\sum_{i=1}^n x_i c_i(R, x) = R(x)$. This is Condition (2.1).

Step 2. Axiom 2 is satisfied if and only if Condition (2.2) in BH holds. Indeed, suppose $R(\cdot) = R^1(\cdot) + R^2(\cdot)$ and

$$B^R_i(x) = \frac{R^1(x) B^{R^1}_i(x)}{R(x)} + \frac{R^2(x) B^{R^2}_i(x)}{R(x)}.$$

Then

$$\frac{B^R_i(x) R(x)}{\bar{x}} = \frac{R^1(x) B^{R^1}_i(x)}{\bar{x}} + \frac{R^2(x) B^{R^2}_i(x)}{\bar{x}}.$$

That is,

$$c_i(R, x) = c_i(R^1, x) + c_i(R^2, x),$$

which is Condition (2.2).

Step 3. Axioms 1 and 3 jointly imply Condition (2.3).

Assume that both Axioms 1 and 3 are satisfied and assume that for all $\eta \in \mathbb{R}^n_+$,

$$R(\eta \cdot \tilde{z}) = g(\eta \cdot q)$$

(3.30)

for some function $g(\cdot)$ and a non-zero vector $q \in \mathbb{R}^n_+$. Then, $(\tilde{z}_1, ..., \tilde{z}_n)$ are $R$-perfectly correlated.
By Axiom 3 for all \( i, j = 1, \ldots, n \),

\[
q_j B^R_i(x) = q_i B^R_j(x),  \tag{3.31}
\]

and hence

\[
\alpha_i q_j B^R_i(x) = \alpha_i q_i B^R_j(x).
\]

Summing over \( i = 1, \ldots, n \) gives

\[
q_j \sum_{i=1}^{n} \alpha_i B^R_i(x) = (\alpha \cdot \mathbf{q}) B^R_j(x). \tag{3.32}
\]

By Axiom 1 we know that \( \sum_{i=1}^{n} \alpha_i B^R_i(x) = 1 \). Plugging this into (3.32) we have

\[
q_j = (\alpha \cdot \mathbf{q}) B^R_j(x) \text{ for } j = 1, \ldots, n.
\]

By (3.29), and recalling that \( R(x) \neq 0 \),

\[
q_j = (\alpha \cdot \mathbf{q}) \frac{c_j(R, x)}{R(x)} = (x \cdot \mathbf{q}) \frac{c_j(R, x)}{R(x)} \text{ for } j = 1, \ldots, n. \tag{3.33}
\]

If \( x \cdot \mathbf{q} = 0 \) this implies that \( q_j = 0 \) for all \( j \), contradicting that \( \mathbf{q} \) is a non-zero vector. Hence, \( x \cdot \mathbf{q} \) is not zero. We then have

\[
c_j(R, x) = \frac{q_j R(x)}{(x \cdot \mathbf{q})} \text{ for all } j = 1, \ldots, n. \tag{3.34}
\]

Consider an asset with return \( \tilde{w} = \frac{x \cdot \tilde{x}}{x \cdot \mathbf{q}} \). Namely, investing \( x \cdot \mathbf{q} \) dollars in this asset yields the same return as of the portfolio \( x \). Then,

\[
R((x \cdot \mathbf{q}) \tilde{w}) = R(x \cdot \tilde{x}) = g(x \cdot \mathbf{q}).
\]

Consider now the risk allocation problem of order 1 with the single asset \( \tilde{w} \) held at the amount \( x \cdot \mathbf{q} \). By Axiom 1 the systematic risk measure of this asset must satisfy

\[
B^R(x \cdot \mathbf{q}) = 1,
\]
or equivalently using (3.29),

\[
c (g, x \cdot q) = \frac{R((x \cdot q) \bar{w})}{x \cdot q} = \frac{g(x \cdot q)}{x \cdot q}.
\]

Plugging back into (3.34) and using that \( R(x) = g(x \cdot q) \) we have

\[
c_j (R, x) = c(g, x \cdot q) q_j.
\]

This is exactly what Condition (2.3) in BH requires, restricting attention to the case that \( q \) is a non-zero vector of non-negative integers.

Step 4. Axiom 4 is satisfied if and only if Condition (2.4) holds. This follows directly from (3.29) and the definition of \( R \)-positive correlation.

Having established Lemma 60 we now turn to completing the proof of the theorem. First, existence has been proved in the text by showing that (2.19) satisfies Axioms 1-4. To show uniqueness note that Lemma 60 implies that Axioms 1-4 are jointly stronger than Conditions (2.1)-(2.4) in BH. From BH’s main result we know that there is a unique cost allocation procedure \( c(R, x) \) satisfying Conditions (2.1)-(2.4). It follows (using the mapping (3.29)) that there is a unique systematic risk measure satisfying Axioms 1-4. Thus, the unique systematic risk measure is given by (2.19).

Finally, to see that (2.19) and (2.20) are equivalent when \( R \) is homogeneous of degree \( k \), note first that in this case

\[
\int_0^1 R_i(t x_1, ..., t x_n) dt = R_i(x_1, ..., x_n) \int_0^1 t^{k-1} dt = \frac{R_i(x_1, ..., x_n)}{k},
\]
where the first equality follows since $R_i$ is homogeneous of degree $k - 1$. It follows that 

$$B^R_i(x) = \frac{\bar{x} \int_0^1 R_i(tx_1, \ldots, tx_n) \, dt}{R(x_1, \ldots, x_n)}$$

$$= \frac{\bar{x} R_i(x_1, \ldots, x_n)}{k R(x_1, \ldots, x_n)}$$

$$= \frac{\bar{x} R_i(\bar{x} \alpha_1, \ldots, \bar{x} \alpha_n)}{k R(\bar{x} \alpha_1, \ldots, \bar{x} \alpha_n)}$$

$$= \frac{R_i(\alpha_1, \ldots, \alpha_n)}{k R(\alpha_1, \ldots, \alpha_n)}$$

$$= \frac{R_i(\alpha_1, \ldots, \alpha_n)}{\sum_{h=1}^n \alpha_h R_h(\alpha_1, \ldots, \alpha_n)},$$

where the penultimate equality follows from the homogeneity of degrees $k$ and $k - 1$ of $R$ and $R_i$ respectively, and the last equality follows from Euler’s homogeneous function theorem.

This completes the proof of Theorem 49. ■

**Derivations of Systematic Risk for Applications I–V**

Here we provide derivations of the systematic risk associated with different risk measures discussed in Section 2.4.3.

**Application I:** This is a special case of Application II.

**Application II:** Consider the risk measure $R(\tilde{z}) = m_k(\tilde{z})$ for even $k \geq 2$. The risk of the market portfolio is

$$R(\alpha^M \cdot \tilde{z}) = m_k(\alpha^M \cdot \tilde{z}) = E(\alpha^M \cdot \tilde{z} - E(\alpha^M \cdot \tilde{z}))^k.$$

Differentiating with respect to the weight of asset $i$ yields

$$\frac{\partial m_k(\alpha^M \cdot \tilde{z})}{\partial \alpha^M_i} = k E \left[ (\tilde{z}_i - E(\tilde{z}_i)) \left( \alpha^M \cdot \tilde{z} - E(\alpha^M \cdot \tilde{z}) \right)^{k-1} \right]$$

$$= k \text{Cov} \left( \tilde{z}_i, \left( \alpha^M \cdot \tilde{z} - \alpha^M \cdot E(\tilde{z}) \right)^{k-1} \right).$$

167
By Theorem 44, and since \( m_k (\cdot) \) is homogeneous of degree \( k \), the systematic risk is then given by

\[
B_i^R = \frac{\partial m_k (\alpha^M \cdot \tilde{z})}{\partial \alpha_i^M} = \frac{\Cov (\tilde{z}_i, (\alpha^M \cdot \tilde{z} - \alpha^M \cdot \E (\tilde{z}))^{k-1})}{m_k (\alpha^M \cdot \tilde{z})},
\] (3.35)
as required.

Now suppose alternatively that \( R (\tilde{z}) = w_k (\tilde{z}) \). The market portfolio risk is

\[
R (\alpha^M \cdot \tilde{z}) = w_k (\alpha^M \cdot \tilde{z}) = \left( m_k (\alpha^M \cdot \tilde{z}) \right)^{1/k}.
\]

Differentiating with respect to the weight of asset \( i \) gives

\[
\frac{\partial w_k (\alpha^M \cdot \tilde{z})}{\partial \alpha_i^M} = \frac{1}{k} \left( m_k (\alpha^M \cdot \tilde{z}) \right)^{(1/k)-1} \frac{\partial m_k (\alpha^M \cdot \tilde{z})}{\partial \alpha_i^M}.
\]

By Theorem 44, and since \( w_k (\cdot) \) is homogeneous of degree 1, the systematic risk is

\[
B_i^R = \frac{1}{k} \left( m_k (\alpha^M \cdot \tilde{z}) \right)^{(1/k)-1} \frac{\partial m_k (\alpha^M \cdot \tilde{z})}{\partial \alpha_i^M} = \frac{\partial m_k (\alpha^M \cdot \tilde{z})}{\partial \alpha_i^M} = \frac{\Cov (\tilde{z}_i, (\alpha^M \cdot \tilde{z} - \alpha^M \cdot \E (\tilde{z}))^{k-1})}{k m_k (\alpha^M \cdot \tilde{z})},
\]

which is identical to (3.35).

**Application III**: Assume \( R (\tilde{z}) = DR_k (\tilde{z}) \) for \( k \geq 2 \). The risk of the market portfolio \( \alpha^M \) is given by

\[
DR_k (\alpha^M \cdot \tilde{z}) = (-1)^k \left( \E \left( [\alpha^M \cdot \tilde{z} - \E (\alpha^M \cdot \tilde{z})]^{-1} \right) \right)^{1/k}.
\]

Differentiating with respect to \( \alpha_i^M \) gives\(^7\)

\[
\frac{\partial DR_k (\alpha^M \cdot \tilde{z})}{\partial \alpha_i^M} = (-1)^k \left( \E \left( [\alpha^M \cdot \tilde{z} - \E (\alpha^M \cdot \tilde{z})]^{-1} \right) \right)^{(1/k)-1} \E \left[ (\tilde{z}_i - \E (\tilde{z}_i)) \left( [\alpha^M \cdot \tilde{z} - \E (\alpha^M \cdot \tilde{z})]^{-1} \right)^{k-1} \right]
\]

\[
= (-1)^k \left( \E \left( [\alpha^M \cdot \tilde{z} - \E (\alpha^M \cdot \tilde{z})]^{-1} \right) \right)^{(1/k)-1} \Cov \left( \tilde{z}_i, \left( [\alpha^M \cdot \tilde{z} - \E (\alpha^M \cdot \tilde{z})]^{-1} \right)^{k-1} \right).
\]

\(^7\)Note that we are essentially relying here on Leibniz’s rule for differentiation under the integral. While \( \left( [\alpha^M \cdot \tilde{z} - \E (\alpha^M \cdot \tilde{z})]^{-1} \right)^{k} \) is not everywhere differentiable, it is continuous and differentiable almost everywhere. This guarantees that Leibniz’s rule applies.
By Theorem 44, and since \( DR_k(\cdot) \) is homogeneous of degree 1, the systematic risk is given by

\[
B_i^R = \frac{\partial DR_k(\alpha^M \bar{z})}{\partial \alpha^M_i} \cdot \frac{\text{Cov} \left[ \tilde{z}_i, \left( \mathbf{\alpha}^M \cdot \bar{z} - E(\mathbf{\alpha}^M \cdot \bar{z}) \right)^{-1} \right]}{E \left( \left( \mathbf{\alpha}^M \cdot \bar{z} - E(\mathbf{\alpha}^M \cdot \bar{z}) \right)^k \right)} \cdot \frac{\text{Cov} \left[ \tilde{z}_i, \left( \mathbf{\alpha}^M \cdot \bar{z} - E(\mathbf{\alpha}^M \cdot \bar{z}) \right)^{-1} \right]}{(DR_k(\alpha^M \cdot \bar{z}))^k}.
\]

**Application IV**: Assume \( R(\bar{z}) = \text{DES}_\delta(\bar{z}) \) for some confidence level \( 0 < \delta < 1 \). Let \( f(z_1, ..., z_n) \) denote the joint density function of \( \bar{z} \). Since all risky assets have positive net supply and since asset prices are positive, we have \( \alpha_1^M > 0 \). Hence, the risk of the market portfolio \( \alpha^M \) can be written as follows

\[
\text{DES}_\delta(\mathbf{\alpha}^M \cdot \bar{z}) = \frac{1}{\delta} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \frac{-\text{VAR}_\delta(\mathbf{\alpha}^M \cdot \bar{z}) - \sum_{j=2}^{n} \alpha_j^M \tilde{z}_j}{\alpha_1^M} \left( \sum_{j=1}^{n} \alpha_j^M z_j - E(\mathbf{\alpha}^M \cdot \bar{z}) \right) f(z_1, ..., z_n) dz_1 ... dz_n.
\]

Differentiating \( \text{DES}_\delta(\mathbf{\alpha}^M \cdot \bar{z}) \) using Leibniz’s rule with respect to \( \alpha_i^M \) yields

\[
\frac{\partial \text{DES}_\delta(\mathbf{\alpha}^M \cdot \bar{z})}{\partial \alpha_i^M} = \frac{-1}{\delta} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \frac{-\text{VAR}_\delta(\mathbf{\alpha}^M \cdot \bar{z}) - \sum_{j=2}^{n} \alpha_j^M \tilde{z}_j}{\alpha_1^M} \left( z_i - E(\tilde{z}_i) \right) f(z_1, ..., z_n) dz_1 ... dz_n
\]

\[+ \frac{\text{VAR}_\delta(\mathbf{\alpha}^M \cdot \bar{z}) + E(\mathbf{\alpha}^M \cdot \bar{z})}{\delta} \frac{\partial}{\partial \alpha_i^M} \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \frac{-\text{VAR}_\delta(\mathbf{\alpha}^M \cdot \bar{z}) - \sum_{j=2}^{n} \alpha_j^M \tilde{z}_j}{\alpha_1^M} f(z_1, ..., z_n) dz_1 ... dz_n \right).
\]
Notice that by the definition of VaR\(_M\cdot \vec{z}\),

\[
\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \int_{-\infty}^{-\infty} \frac{-\operatorname{VaR}_\delta(M \cdot \vec{z}) - \sum_{j=2}^{n} \alpha_j^M \cdot z_j}{\alpha_i^M} f(z_1, \ldots, z_n) dz_1 \ldots dz_n = \delta,
\]

which is a constant, implying that the second term in (3.36) is zero. Thus,

\[
\frac{\partial \operatorname{DES}_\delta(M \cdot \vec{z})}{\partial \alpha_i^M} = -\frac{1}{\delta} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \int_{-\infty}^{-\infty} \frac{-\operatorname{VaR}_\delta(M \cdot \vec{z}) - \sum_{j=2}^{n} \alpha_j^M \cdot z_j}{\alpha_i^M} (z_i - \mathbb{E}(\tilde{z}_i)) f(z_1, \ldots, z_n) dz_1 \ldots dz_n
\]

\[
= -\frac{1}{\delta} \mathbb{E} \left[ 1_{\alpha^M \cdot \vec{z} \leq -\operatorname{VaR}_\delta(M \cdot \vec{z})} (z_i - \mathbb{E}(\tilde{z}_i)) \right]
\]

\[
= -\mathbb{E} \left[ \tilde{z}_i - \mathbb{E}(\tilde{z}_i) \mid \alpha^M \cdot \vec{z} \leq -\operatorname{VaR}_\delta(M \cdot \vec{z}) \right].
\]

By Theorem 44, and since DES\(_\delta(M \cdot \vec{z})\) is homogeneous of degree 1, the systematic risk is given by

\[
B_i^R = \frac{\frac{\partial \operatorname{DES}_\delta(M \cdot \vec{z})}{\partial \alpha_i^M}}{\operatorname{DES}_\delta(M \cdot \vec{z})} = -\frac{\mathbb{E} \left[ \tilde{z}_i - \mathbb{E}(\tilde{z}_i) \mid \alpha^M \cdot \vec{z} \leq -\operatorname{VaR}_\delta(M \cdot \vec{z}) \right]}{\operatorname{DES}_\delta(M \cdot \vec{z})}.
\]

**Application V:** Consider the following family of risk measures

\[
R(\vec{z}) = \theta_1 w_2(\vec{z}) + \theta_2 \operatorname{DR}_3(\vec{z}) + \theta_3 w_4(\vec{z}) + \theta_4 \operatorname{DES}_\delta(\vec{z})
\]

for some confidence level \(\delta\) and non-negative weights \(\theta_1, \ldots, \theta_4\). From Lemma 28, this family of risk measures satisfies all of the conditions in Theorem 44. Moreover, it is easy to verify that when

\[
R(\vec{z}) = \sum_{j=1}^{s} R^j(\vec{z})
\]

the expression for \(B_i^R\) given in (2.10) implies

\[
B_i^R = \sum_{j=1}^{s} \frac{R^j(\alpha^M \cdot \vec{z})}{R(\alpha^M \cdot \vec{z})} B_i^{R^j}.
\]

That is, the systematic risk takes the form of the risk-weighted average of the systematic risk associated with each of the risk components. (See also Section 2.5.1 for further discussion of this issue as it relates to Axiom 2 in that section.)
Discussions

Mean-Risk Preferences and Expected Utility

Background

One would wonder how the mean-risk preferences considered in Section 2.4 are related to the commonly assumed von Neumann-Morgenstern utility. It is widely known that a von Neumann-Morgenstern investor with a quadratic utility function only cares about the mean and the variance of his investments in the sense that he prefers a high expected wealth and a low variance. In this sense, the mean-risk preference is consistent with the von Neumann-Morgenstern utility when variance is used as the risk measure. Alternatively, when returns are distributed according to a two-parameter elliptical distribution (normal being a special case), mean-variance preferences can also be supported by expected utility. These instances, however, are quite restrictive. First, the quadratic utility is not very intuitive since it implies increasing absolute risk aversion. Second, elliptical distributions, being determined by the first two moments only, limit our ability to describe the dependence of risk on high distribution moments and other risk characteristics. Thus, in general, mean-variance preferences are not consistent with expect utility. The approach taken in this paper is much more general, allowing for a variety of risk measures. Whether a particular risk measure is consistent with expected utility depends on the actual choice of the risk measure. For example, risk measures that are simple linear combinations of raw moments up to the $k^{th}$ degree can be represented by a $k^{th}$ degree polynomial (Müller and Machina (1987)), generalizing the mean-variance result.

While in general the preferences defined in (2.6) cannot be supported by expected utility, they are often consistent with expected utility locally. The idea is based on Machina’s (1982) “Local Utility Function.” To facilitate this approach we first restrict attention to risk measures that depend on the distribution of the random variables only. Thus, we consider
risk measures that are functions from the distribution of realizations to the reals rather than
functions from the random variables themselves. Practically, this does not present a binding
restriction since all the examples in this paper and all standard risk measures only rely on
the distribution of realizations anyway. In this case the preferences in (2.6) can be written
as

\[ U(\zeta) = V(E(F_{\zeta \tilde{y}}), R(F_{\zeta \tilde{y}})), \]

where \( F_{\zeta \tilde{y}} \) is the cumulative distribution of the random variable \( \zeta \cdot \tilde{y} \). When the random
variable of interest is clear, we will omit it from the notation and write the utility as \( U(F) = V(E(F), R(F)). \)

According to Machina (1982), if the realizations of all random variables are contained
in some bounded and closed interval \( I \) and \( U(F) \) is Fréchet differentiable with respect to
the \( L^1 \) norm,\(^8\) then for any two distributions \( F_1, F_2 \) on \( I \) there exists \( u(\cdot; F_1) \) differentiable
almost everywhere on \( I \) such that

\[
U(F_2) - U(F_1) = \int_I u(y; F_1) \, dF_2(y) - \int_I u(y; F_1) \, dF_1(y) + o(\|F_2 - F_1\|), \tag{3.37}
\]

where \( \|\cdot\| \) denotes the \( L^1 \) norm. That is, starting from a wealth distribution \( F_1 \), if an
investor moves to another “close” distribution \( F_2 \), then he compares the utility from these
two distributions as if he is maximizing his expected utility with a local utility function
\( u(\cdot; F_1) \).

The key to applying Machina’s result is to find sufficient conditions on the risk measure
which guarantee that \( U(F) \) is Fréchet differentiable. This can be done in many ways. Next
we provide one simple but effective approach which is sufficient to validate many popular
risk measures as consistent with local expected utility.

\(^8\)Fréchet differentiability is an infinite dimensional version of differentiability. The idea here is that \( U(F) \)
changes smoothly with \( F \), where changes in \( F \) are topologized using the \( L^1 \) norm. See Luenberger (1969, p.
171).
Risk Measures as Functions of Moments

Let $\mu_k^F = \int y^k dF(y)$ be the $k^{th}$ raw moment given distribution $F$, and $m_k^F = \int (y - \mu_1^F)^k dF(y)$ be the $k^{th}$ central moment given distribution $F$. Consider risk measures which are a function of a finite number of (raw or central) moments. We denote such risk measures by $R(\mu_{j_1}^F, ..., \mu_{j_l}^F, m_{k_1}^F, ..., m_{k_n}^F)$. We assume that $R$ is differentiable in all arguments. The utility function in (2.6) then takes the form

$$U(F) = V\left(\mu_1^F, R(\mu_{j_1}^F, ..., \mu_{j_l}^F, m_{k_1}^F, ..., m_{k_n}^F)\right),$$

(3.38)

where $V$ is differentiable in both mean and risk. This class of utility functions is quite general and it allows the risk measure to depend on a large number of high distribution moments. We then have the following proposition.

**Proposition 61** If $U(F)$ takes the form (3.38) then for any two distributions $F_1, F_2$ on $I$ there exists $u(\cdot; F_1)$ differentiable almost everywhere on $I$ such that (3.37) holds.

**Proof:** We need to show that $U(F)$ is Fréchet differentiable. By the chain rule for Fréchet differentiability (Luenberger (1969, p. 176)), we know that if both $\mu_k^F$ and $m_k^F$ are Fréchet differentiable for any $k$, then so is $U(\cdot)$. The Fréchet differentiability of $\mu_k^F$ is obvious, since

$$\mu_k^{F_2} - \mu_k^{F_1} = \int_I y^k dF_2(y) - \int_I y^k dF_1(y) = -k \int_I (F_2(y) - F_1(y)) y^{k-1} dy.$$

Now we show that $m_k^F$ is Fréchet differentiable. We have

$$m_k^F = \int (y - \mu_1^F)^k dF(y)$$

$$= \int \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} y^i (\mu_1^F)^{k-i} dF(y)$$

$$= \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} (\mu_1^F)^{k-i} \int y^i dF(y)$$

$$= \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} (\mu_1^F)^{k-i} \mu_i^F,$$
which is a differentiable function of the $\mu_i^F$'s. By the chain rule, it follows immediately that $m_k^F$ is also Fréchet differentiable. This completes the proof. ■

Sufficient Conditions for Positive Prices

Here we provide a sufficient condition for the positivity of equilibrium prices following the approach of Nielsen (1992). Let $\zeta \in \mathbb{R}^{n+1}$ be a bundle. Denote the gradient of investor $j$’s utility function at $\zeta$ by $\nabla U_j (\zeta) = (U_j^0(\zeta), ..., U_j^n(\zeta))$, where a subscript designates a partial derivative in the direction of the $i^{th}$ asset. Also, let $\gamma^j(\zeta) = -\frac{V_j^1(E(\zeta \bar{y}), R(\zeta \bar{y})))}{V_j^0(E(\zeta \bar{y}), R(\zeta \bar{y})))} > 0$ be the marginal rate of substitution of the expected payoff of the bundle for the risk of the bundle. This is the slope of investor $j$’s indifference curve in the expected payoff-risk space. For brevity we often omit the arguments of this expression and use $\gamma^j(\zeta) = -\frac{V_j^1}{V_j^0}$.

**Proposition 62** Assume that for each asset $i$ there is some investor $j$ such that $E(\bar{y}_i) > \gamma^j(\zeta) R_i (\zeta \cdot \bar{y})$ for all $\zeta$. Then, prices of all assets are positive in all equilibria.

**Proof:** At an equilibrium, all investors’ gradients point in the direction of the price vector. So the price of asset $i$ must be positive in any equilibrium if there is some investor $j$ such that $U_j^i(\zeta) > 0$ for all $\zeta$. Recall that

$$U_j^i(\zeta) = V_j^i (E(\zeta \cdot \bar{y}), R(\zeta \cdot \bar{y})).$$

Thus,

$$U_j^i(\zeta) = V_j^i E(\bar{y}_i) + V_j^i R_i(\zeta \cdot \bar{y}) = V_j^i [E(\bar{y}_i) - \gamma^j(\zeta) R_i(\zeta \cdot \bar{y})],$$

where $R_i(\zeta \cdot \bar{y})$ denotes the partial derivative of $R(\zeta \cdot \bar{y})$ with respect to $\zeta_i$.

Since $V_j^i > 0$, $U_j^i(\zeta) > 0$ corresponds to

$$E(\bar{y}_i) - \gamma^j(\zeta) R_i(\zeta \cdot \bar{y}) > 0,$$

174
as required. ■

Note that $\gamma_j(\cdot)$ can serve as a measure of risk aversion for investor $j$. We can thus interpret this proposition as follows. If each asset’s expected return is sufficiently high relative to some investor’s risk aversion and the marginal contribution of the asset to total risk, then this asset will always be desirable by some investor, and so, its price will be positive in any equilibrium.
Appendix C: Proofs and Discussions for Chapter 3

Proofs

Proof of Proposition 52: At time \( t \), the risk-neutral conditional probability of \( \tilde{r}_{t,t+1} = r(k) \) given \( \tilde{r}_{t,t+1}^I = r^I(n) \) is

\[
q \left( \tilde{r}_{t,t+1} = r(k) \mid \tilde{r}_{t,t+1}^I = r^I(n) \right) = \frac{q \left( \tilde{r}_{t,t+1} = r(k), \tilde{r}_{t,t+1}^I = r^I(n) \right)}{q_{t,t+1}^I(n)}. \tag{3.39}
\]

By Assumption 51, the variation in the index return is the only priced risk. Thus, there exists a stochastic discount factor, whose value related to future payoffs depends on the future value of the index return only. Let \( (\xi_{t,t+1}(1), \xi_{t,t+1}(2), \ldots, \xi_{t,t+1}(N)) \) denote the stochastic discount factor evaluated at time \( t \) related to payoffs to be received in different states of the index return at time \( t + 1 \). This allows one to convert risk-neutral probabilities to physical probabilities. In particular, we have

\[
q \left( \tilde{r}_{t,t+1} = r(k), \tilde{r}_{t,t+1}^I = r^I(n) \right) = e^{r^I \xi_{t,t+1}(n)} p \left( \tilde{r}_{t,t+1} = r(k), \tilde{r}_{t,t+1}^I = r^I(n) \right), \tag{3.40}
\]

and

\[
q_{t,t+1}^I(n) = e^{r^I \xi_{t,t+1}(n)} p_{t,t+1}^I(n). \tag{3.41}
\]

Plugging (3.40) and (3.41) into (3.39) yields

\[
q \left( \tilde{r}_{t,t+1} = r(k) \mid \tilde{r}_{t,t+1}^I = r^I(n) \right) = \frac{p \left( \tilde{r}_{t,t+1} = r(k), \tilde{r}_{t,t+1}^I = r^I(n) \right)}{p_{t,t+1}^I(n)} = \theta(k|n),
\]

as claimed. \( \blacksquare \)
Proof of Proposition 57: This proof relies on Theorem 2 of Liew (1976), which speaks about the estimates from a inequality-constrained least squares regression. The theorem says that when the inequality constraints are not binding with respect to the true parameter values, the estimates from the original inequality-constrained regression are equal to the estimates from the same regression without the inequality constraints when the sample size becomes large enough.

To prove the consistency of $\hat{\theta}_T$, I consider two possible cases.

Case 1: Suppose that $0 < \theta (k|n) < 1$ for all $n$ and $k$.

According to Theorem 2 of Liew (1976), when the sample size $T$ is large enough, $\hat{\theta}_T$ is equal to the coefficient estimates obtained from the following least squares linear regression problem with the equality constraint only

$$Q = Q' \cdot \theta + \epsilon, \quad (3.42)$$

$$s.t.$$ 

$$\theta \cdot 1_{K \times 1} = 1_{N \times 1}.$$

The exact formula for the equality-constrained least squares estimator can be found in Amemiya (1985) page 21. Since the equality constraint holds with respect to the true parameter values, it is standard to prove that under Assumptions 54 to 56, the equality-constrained least squares estimator is consistent. (The proof is similar to the proof for the consistency of the OLS estimator, and hence I omit it here for brevity.) Since $\hat{\theta}_T$ is equal to a consistent estimator when the sample size is large enough, $\hat{\theta}_T$ itself is also consistent.

Case 2: Suppose that the inequality constraint $0 \leq \theta (k|n) \leq 1$ is binding at one end with respect to the true parameter values for some $n$ and $k$. Without loss of generality, consider a simplest case in which $\theta (1|1) = 0$ and $0 < \theta (k|n) < 1$ for all other pairs of $n$ and $k$. The proofs for other cases are parallel.

Let $\hat{\theta}_T (1|1)$ denote the estimate of $\theta (1|1)$ obtained from the least squares regression
problem (3.8) corresponding to a sample with size $T$. The inequality constraint implies
$
\hat{\theta}_T (1|1) \geq 0.
$
Then, for each $T > 0$ I have the following two subcases.

Subcase 2.1: Suppose that $\hat{\theta}_T (1|1) > 0$. In this scenario, $\hat{\theta}_T$ is equal to the coefficient estimates obtained from the following constrained least squares regression

\[
Q = Q^I \cdot \theta + \epsilon,
\]

\[
s.t.
\]
\[
\theta \cdot 1_{K \times 1} = 1_{N \times 1},
\]
\[
\theta \leq 1_{N \times K},
\]
\[
\theta (k|n) \geq 0, \ \forall n, k \text{ unless } n = k = 1.
\]

Notice that the inequality constraint $\theta (1|1) \geq 0$ can be dropped because it is not binding ($\hat{\theta}_T (1|1) > 0$) and because I have a convex problem (convex objective function and compact domain). Theorem 2 of Liew (1976) indicates that when $T$ is large enough, $\hat{\theta}_T$ is equal to the equality-constrained least squares estimator obtained from (3.42), which converges to the true parameter values $\theta$ when the sample size approaches infinity.

Subcase 2.2: Suppose that $\hat{\theta}_T (1,1) = 0$. In this scenario, $\hat{\theta}_T$ is equal to the coefficient estimates obtained from the following constrained least squares regression

\[
Q = Q^I \cdot \theta + \epsilon,
\]

\[
s.t.
\]
\[
\theta \cdot 1_{K \times 1} = 1_{N \times 1},
\]
\[
\theta (1|1) = 0,
\]
\[
\theta \leq 1_{N \times K},
\]
\[
\theta (k|n) \geq 0, \ \forall n, k \text{ unless } n = k = 1.
\]

According to Theorem 2 of Liew (1976), when $T$ is large enough, $\hat{\theta}_T$ is equal to the coefficient estimates obtained from conducting the following regression problem with equality
As in Case 1, since the equality constraints hold with respect to the true parameter values, the estimator from this equality-constrained problem converges to $\theta$ when the sample size approaches infinity.

The analyses of the above two subcases show that in Case 2, the sequence of estimators $\{\hat{\theta}_T\}_T$ can be viewed as consisting of two subsequences (each corresponding to one of the two subcases), both of which converge to the same true parameter values $\theta$ as the sample size approaches infinity. Therefore, the sequence $\{\hat{\theta}_T\}_T$ as a whole must also converge to $\theta$.

Cases 1 and 2 together complete the proof. $\blacksquare$

**Discussions**

As mentioned in Section 3.4.1, two technical issues need to be dealt with for the empirical estimation of the asset return distribution $q_{t,t+1}$.

First, the estimation of the risk-neutral CDF (3.5) relies on differentiating the option price with respect to the strike price. Unfortunately, we are generally not able to obtain a close-form expression for this derivative. To resolve this issue, I resort to a linear approximation based on two points $X^-$ and $X^+$ around the strike price $X$ of interest, i.e.,

$$F(X) \approx e^r \frac{Put(X^+) - Put(X^-)}{X^+ - X^-}.$$  

For all empirical applications in the paper, I choose $X^-$ and $X^+$ to be $0.01$ to the left and right of $X$. 

179
A second empirical challenge has to do with obtaining European option prices. Nearly all individual stock options are American options. While indices are typically represented by European options, the market option prices are only available at discrete values of the strike price and time to maturity. Thus, reasonable estimates are needed for European option prices at any arbitrary point of interest. In the empirical applications, I adopt a simple and commonly used approach in the literature (e.g., Shimko (1993), Malz (1997), and Figlewski (2010)) of first fitting the implied volatility surface and then deriving the option prices using the fitted volatility. Specifically, I start with the implied volatility of all available options written on the asset on the date of estimation. Information on the implied volatility is provided by the OptionMetrics database, which is computed by the BMS model for European options and by the Cox-Ross-Rubinstein (CRR) model (Cox, Ross and Rubinstein (1979)) for American options. I then fit the implied volatility surface for a range of strike prices and maturities of interest based on the implied volatility of available options. Finally, I plug the fitted volatility at any particular point back into the BMS pricing model to obtain an estimation of the European option price.

I fit the implied volatility surface by kernel smoothing using a procedure similar to that adopted by the OptionMetrics database. On each date of estimation, I collect market pricing information of all options written on the same underlying asset indexed by \( h = 1, 2, \ldots, H \). For each option \( h \), let \( \sigma^h \) represent the implied volatility, and let \( V^h \) be the vega of the option (which measures the sensitivity of option price to the volatility). Denote by \( mn^h = X^h / S_t \) the moneyness of the option, by \( mt^h \) the time to maturity in years, and by \( cp^h \) a dummy variable that equals 0 for call options and 1 for put options. Then at any point with moneyness \( mo^* \) (within the range of moneyness of available options), time to maturity \( ma^* \) (within the range of time to maturity of available options), and call-put indicator \( cp^* \), the estimated volatility
\( \hat{\sigma} \) takes the form
\[
\hat{\sigma} (mn^*, mt^*, cp^*) = \frac{\sum_{h=1}^H V^h \sigma^h \Psi (mn^* - mn^h, mt^* - mt^h, cp^* - cp^h)}{\sum_{h=1}^H V^h \Psi (mn^* - mn^h, mt^* - mt^h, cp^* - cp^h)},
\] (3.43)
where the kernel function \( \Psi \) is given by
\[
\Psi (x, y, z) = \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{x^2}{2c_1} - \frac{y^2}{2c_2} - \frac{z^2}{2c_3} \right].
\]
I naively choose \( c_1 = c_2 = c_3 = 0.001 \). These parameter values have been checked to generate reasonable fitting.

The idea of the kernel smoothing procedure is intuitive. At any point of interest, I estimate volatility as the weighted average of the implied volatility of all observed options, where observations close to the point of interest in terms of moneyness, time to maturity, and the call-put indicator are assigned higher weights than observations far away. In addition, since I am eventually interested in estimating the option prices, I would also want to assign higher weights to observed options whose prices have higher sensitivity to the volatility. To this end, I also include the option vega into the weights.

The kernel smoothing formula (3.43) works for moneyness and time to maturity levels within the range of available options. In my empirical applications, I mostly focus on rare disasters. To estimate the risk-neutral probabilities of rare disasters, I need European option prices (and hence the volatility) around the disaster thresholds, which feature very low moneyness levels, sometime even outside the range of moneyness for available options. When the moneyness at the disaster threshold is lower than the lowest observed moneyness of available options, I estimate volatility at the lowest observed moneyness level by (3.43) and use it as a proxy for the volatility around the disaster threshold.

Finally, notice from (3.43) that to perform kernel smoothing, besides the moneyness and time to maturity of interest, I also need to specify the call-put indicator \( cp^* \). Theoretically, the implied volatility should be identical for a call option and a put option with the same moneyness and time to maturity. However, this is rarely the case in reality. A priori, it
is not clear whether I should use the call or the put value for the estimation. In practice, most options trade at-the-money (ATM) or out-of-the-money (OTM), rendering the prices more accurate for ATM and OTM options. I thus adopt the following rule of thumb. When the moneyness of interest is lower than one ($mn^* < 1$), I estimate volatility for the put option ($cp^* = 1$), which trades OTM. When the moneyness of interest is higher than one ($mn^* > 1$), I estimate volatility for the call option ($cp^* = 0$), which again trades OTM. When the moneyness of interest is equal to one ($mn^* = 1$), both the call and the put options are exactly ATM. I then estimate volatilities for both the call and the put options and take the average.