Spring 5-15-2015

Essays in Financial Economics

Suying Liu
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Olin Business School

Dissertation Examination Committee:
Ohad Kadan, Chair
Guofu Zhou, Co-Chair
Thomas Maurer
Todd T. Milbourn
John Nachbar

Essays in Financial Economics

by

Suying 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

List of Tables

Acknowledgments

For Chapter 1, my coauthors and I thank Phil Dybvig, Brett Green, Sergiu Hart, Steve Ross, and the referees as well as seminar participants at UC Berkeley, Cornell University, Hebrew University, Hong Kong Polytechnic University, Indiana University, University of Pennsylvania, and Washington University in St. Louis for helpful comments and suggestions.

For Chapter 2, I thank my dissertation committee for their valuable guidance and comments. I am grateful to Michael Brennan, Henry Cao, Zhi Da, and Phil Dybvig especially, as well as participants at the summer workshop of SWUFE, and the brownbag workshop of Washington University in St. Louis.

For Chapter 3, I thank my dissertation committee for their invaluable guidance and suggestions. Data from a large life insurance industry experience study are instrumental for this study, and are gratefully acknowledged. This research originated during my summer position with the Fixed Income Strategy Group at J.P. Morgan. I am thankful to the Group, especially Matt Jozoff and Praveen Korapaty, for their tremendous help and encouragement. The views expressed are my own and do not necessarily represent that of the Company.

I would also like to thank my wife, Karen, for her love, care, and courage.

Suying Liu

Washington University in St. Louis

May 2015

To my family.

ABSTRACT OF THE DISSERTATION

Essays in Financial Economics
by
Suying Liu
Doctor of Philosophy in Business Administration
Washington University in St. Louis, 2015
Professor Ohad Kadan, Chair
Professor Guofu Zhou, Co-Chair

In Chapter 1, we generalize 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.

For Chapter 2, note that substantial cross-sectional variation in entrepreneurial compensation has been documented in prior literature, although explanation is scarce. The uniqueness of small businesses, in particular the intrinsic difference between entrepreneurs and corporate managers, calls for additional insights aside from that on executive compensation. This study takes an asymmetric-information perspective, where a continuous-time game-theoretic model is developed, incorporating an interesting trade-off between current and future payoffs. A breakdown in the market of entrepreneurial ventures will not occur, but both separating and pooling equilibria are possible, and consequently an equilibrium is not necessarily informationally consistent. Furthermore, when an equilibrium is indeed revealing, the dissipativeness of the signal emitted by entrepreneurs is completely endogenous. These findings

correspond naturally to empirical predictions about entrepreneurial pay, especially on the cross-industry differential in compensation structure.

In terms of Chapter 3, notice that life insurance often embeds a surrender option that gives the policy holder a right to exchange an existing contract for its cash surrender value. Similar to mortgage prepayment option that imposes a cash-flow risk to MBS investors, this surrender option is a source of concern for life insurers. While prior studies have attempted to quantify this surrender risk by pricing the surrender option, a common theoretical assumption imposed is fully rational response of policy holders to only interest rates. However, actual surrender experience indicates that interest rates are just one of the multiple factors that drive the surrender decision and policy holder response is not necessarily optimal. This research therefore integrates an empirical surrender function into the option pricing framework by employing a novel data set from a large life insurance industry experience study. It shows, for the first time in the literature, that policy vintages are a particularly significant and meaningful factor in addition to macroeconomic variables that impact surrender activity. Using these empirics, I find that the experience-based value of the surrender option is substantially less than its fully rational counterpart. In addition, the competitive landscape of the life insurance industry and the interest rate environment both play an important role in assessing the surrender risk exposure of life insurers.

1 Generalized Systematic Risk

1.1 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.

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.[1] 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

[1] 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.

each risky asset i satisfies

$$\mathrm{E}\left(\tilde{z}_i\right) = r_f + \mathcal{B}_i^R \left(\mathrm{E}\left(\tilde{z}^M\right) - r_f\right),$$

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 $\mathcal{B}_i^R$ is the systematic risk of asset i given the risk measure R. Moreover, $\mathcal{B}_i^R$ 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, $\mathcal{B}_i^R$ 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 1.2 discusses the related literature. In Section 1.3 we define the notion of risk measures. Section 1.4 studies the equilibrium setup and offers a generalization of the CAPM. In Section 1.5 we present the axiomatic approach. Section 1.6 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.

1.2 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.

Additionally, 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.

1.3 Risk Measures and Their Properties

Let $(\Omega, \mathcal{F}, P)$ be a probability space, where Ω is the state space, $\mathcal{F}$ is the σ-algebra of events, and $P(\cdot)$ is a probability measure. As usual, a random variable is a measurable function from Ω 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 $\mathrm{E}(\tilde{z})$ and its k^{th} central moment by $m_k(\tilde{z}) = \mathrm{E}(\tilde{z} - \mathrm{E}(\tilde{z}))^k$, 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 1 *A risk measure is a function mapping random variables to the reals.*[2]

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 λ 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.

Next we would like to formalize a property dealing with the type of assets that are risk-free. We say that a risk measure $R(\cdot)$ has the *risk-free property*, if (i) $R(\tilde{z}) \geq 0$

[2]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.

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 $\widetilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_n)$. A *portfolio* is a vector $\mathbf{x} = (x_1, ..., x_n) \in \mathbb{R}^n$, where x_i is the dollar amount invested in $\tilde{z}_i$.[3] Then, $\mathbf{x} \cdot \widetilde{\mathbf{z}} = \sum_{i=1}^{n} x_i \tilde{z}_i$ is itself a random variable. We then say that the risk of portfolio $\mathbf{x}$ is simply $R(\mathbf{x} \cdot \widetilde{\mathbf{z}})$. When the vector of random variables is unambiguous, we often abuse notation and denote $R(\mathbf{x})$ as a shorthand for $R(\mathbf{x} \cdot \widetilde{\mathbf{z}})$. We say that a risk measure is *smooth* if for any vector of random returns $\widetilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_n)$ and for all portfolios $\mathbf{x} = (x_1, ..., x_n)$ we have that $R(\mathbf{x} \cdot \widetilde{\mathbf{z}})$ is continuously differentiable in x_i for $i = 1, ..., n$. We then write $R_i(\mathbf{x})$ (or $R_i(\mathbf{x} \cdot \widetilde{\mathbf{z}})$) for the partial derivative of $R(\cdot)$ with respect to the amount invested in the i^{th} asset evaluated at $\mathbf{x}$.[4]

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 define a new risk measure by

$$R^{\boldsymbol{\theta}}(\tilde{z}) = \sum_{j=1}^{s} \theta_j R^j(\tilde{z}),$$

where θ_j reflects the weight assigned to the risk attribute measured by R^j. We then have the following trivial but useful lemma.

Lemma 1 *Assume that each R^j is homogeneous of degree k, convex, smooth, and satisfies the risk-free property. Then, $R^{\boldsymbol{\theta}}$ also satisfies all of these properties.*

[3] Throughout the paper we denote vectors using bold notation (for both numbers and random variables).

[4] 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.

1.3.1 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 1 ***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 1 *For all $k \geq 2$ even, $R(\tilde{z}) = w_k(\tilde{z})$ is a convex risk measure.*

Example 2 ***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.*[5]

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.

[5]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,

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

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.

Example 3 ***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*

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

where $[t]^- = \min(t, 0)$ *for* $t \in \mathbb{R}$*. Often, this measure is used in the special case of* $k = 2$*. More generally, for any* $k \geq 2$*,* $\mathrm{DR}_k(\tilde{z})$ *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 2 *For any* $k \geq 2$*,* $\mathrm{DR}_k(\tilde{z})$ *is a convex risk measure.*

Example 4 ***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 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* δ*-quantile of* $\tilde{z}$*, i.e.,*

$$\mathrm{VaR}_\delta(\tilde{z}) = -\inf\{z \in \mathbb{R} : F(z) \geq \delta\}, \tag{1}$$

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 (1) is implicitly determined by*

$$\int_{-\infty}^{-\mathrm{VaR}_\delta(\tilde{z})} f(z)\, dz = \delta. \tag{2}$$

This risk measure is homogeneous of degree 1 and smooth.[6] *For any risk-free return* $\tilde{z}$ *with* $P(\{\tilde{z} = c\}) = 1$*, we have* $\mathrm{VaR}_\delta(\tilde{z}) = -c$*, 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.*

[6] Formally, 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). We omit the proof for brevity.

Example 5 ***Expected shortfall and demeaned expected shortfall.***[7] *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* δ*-quantile. That is,*

$$\mathrm{ES}_\delta(\tilde{z}) = -\frac{1}{\delta}\int_{-\infty}^{-\mathrm{VaR}_\delta(\tilde{z})} z f(z)\, dz. \tag{3}$$

Additionally, when $\tilde{z} = c$ *(a constant) with probability 1 we set* $\mathrm{ES}_\delta(\tilde{z}) = -c$*. Similar to VaR, ES is homogeneous of degree 1 and is smooth, but it does not satisfy the risk-free property. To ensure that the risk-free property is satisfied it is useful to consider the demeaned version of ES defined as*

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

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.[8] *Moreover, unlike VaR, both ES and DES satisfy the convexity property as shown in the next proposition.*

Proposition 3 *For any* $\delta \in (0,1)$*,* $R(\tilde{z}) = \mathrm{ES}_\delta(\tilde{z})$ *and* $R(\tilde{z}) = \mathrm{DES}_\delta(\tilde{z})$ *are convex.*

Example 6 ***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* $R^{AS}(\tilde{z})$ *is given by the unique positive solution to the implicit equation*

$$\mathrm{E}\left[\exp\left(-\frac{\tilde{z}}{R^{AS}(\tilde{z})}\right)\right] = 1. \tag{4}$$

The FH measure $R^{FH}(\tilde{z})$ *is given by the unique positive solution to the implicit equation*

$$\mathrm{E}\left[\log\left(1 + \frac{\tilde{z}}{R^{FH}(\tilde{z})}\right)\right] = 0. \tag{5}$$

[7] Expected shortfall is sometimes termed "conditional VaR."

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

Both these measures are homogeneous of degree 1 and smooth. These two risk measures also satisfy the convexity property.[9] *By contrast, these two measures do not satisfy the risk-free property.*[10]

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.

Example 7 *Let R be a risk measure which is homogeneous of some degree k, convex, and satisfies the risk-free property, and let $h : [0, \infty) \to R$ be a strictly increasing and strictly convex 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, 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 $R = m_k$ 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.

1.4 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.

1.4.1 Model Setup

Investors, Assets, and Timing. Assume a market with $n + 1$ assets $\{0, ..., n\}$. Assets $1, ..., n$ are risky and pay a random amount denoted by $(\tilde{y}_1, ..., \tilde{y}_n)$. Asset 0 is

[9] This follows since these risk measures are subadditive and homogeneous of degree 1.

[10] To see this, note that for any constant $c > 0$, $\tilde{z} + c$ first-order stochastically dominates $\tilde{z}$. Since R^{AS} is consistent with first-order stochastic dominance, we have that $R^{AS}(\tilde{z} + c) < R^{AS}(\tilde{z})$. A similar argument applies to R^{FH}. Also, technically, these two risk measures are not defined for risk-free assets.

risk-free, paying an amount $\tilde{y}_0$ which is equal to some constant $y_0 \neq 0$ with probability 1. Denote $\widetilde{\mathbf{y}} = (\tilde{y}_0, ..., \tilde{y}_n)$. There are ℓ 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 $\boldsymbol{\zeta}^j \in \mathbb{R}^{n+1}$ represents the number of shares investor j chooses in each asset $i = 0, ..., n$, i.e., $\boldsymbol{\zeta}^j$ 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 $\sum_{j=1}^{\ell} e_i^j > 0$ for $i = 1, ..., n$. That is, risky assets are in positive net supply. An *allocation* is an ℓ-tuple $\mathcal{A} = \left(\boldsymbol{\zeta}^1, ..., \boldsymbol{\zeta}^\ell\right)$ consisting of a bundle $\boldsymbol{\zeta}^j \in \mathbb{R}^{n+1}$ for each investor. An allocation $\mathcal{A}$ is *attainable* if $\sum_{j=1}^{\ell} \boldsymbol{\zeta}^j = \sum_{j=1}^{\ell} \mathbf{e}^j$, that is, if it clears the market. A *price system* is a vector $\mathbf{p} = (p_0, ..., p_n)$ 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 $\boldsymbol{\zeta} \in \mathbb{R}^{n+1}$ is given by

$$U^j(\boldsymbol{\zeta}) = V^j\left(\mathrm{E}\left(\boldsymbol{\zeta} \cdot \widetilde{\mathbf{y}}\right), R\left(\boldsymbol{\zeta} \cdot \widetilde{\mathbf{y}}\right)\right), \tag{6}$$

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 $U^j(\boldsymbol{\zeta})$ 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 $U^j(\boldsymbol{\zeta})$ 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

$\boldsymbol{\zeta} = (\zeta_0, \zeta_1, ..., \zeta_n) \in \mathbb{R}^{n+1}$ we have $R(\boldsymbol{\zeta} \cdot \tilde{\mathbf{y}}) \neq 0$ unless $(\zeta_1, ..., \zeta_n) = (0, ..., 0)$.[11]

Equilibrium. An *equilibrium* is a pair $(\mathbf{p}, \mathcal{A})$ where $\mathbf{p} \neq 0$ is a price system and $\mathcal{A} = \left(\boldsymbol{\zeta}^1, ..., \boldsymbol{\zeta}^\ell\right)$ is an attainable allocation, such that for each $j \in \{1, ..., \ell\}$, $\mathbf{p} \cdot \boldsymbol{\zeta}^j = \mathbf{p} \cdot \mathbf{e}^j$, and if $\boldsymbol{\zeta} \in \mathbb{R}^{n+1}$ and $U^j(\boldsymbol{\zeta}) > U^j\left(\boldsymbol{\zeta}^j\right)$ then $\mathbf{p} \cdot \boldsymbol{\zeta} > \mathbf{p} \cdot \mathbf{e}^j$. 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 1 *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.

1.4.2 A Generalized CAPM

Geometry of Efficient Portfolios Let $(\mathbf{p}, \mathcal{A})$ be an equilibrium. The equilibrium allocation $\left(\boldsymbol{\zeta}^1, ..., \boldsymbol{\zeta}^\ell\right)$ naturally induces a portfolio for each investor j given by $\mathbf{x}^j =$

[11] In the standard mean-variance case this condition corresponds to the variance-covariance matrix of risky assets being positive-definite.

$\left(x_0^j, ..., x_n^j\right)$, where $x_i^j = p_i \zeta_i^j$ is the amount invested in asset i, and where the vector of portfolio weights of investor j is denoted by $\boldsymbol{\alpha}^j$ and given by $\alpha_i^j = \frac{x_i^j}{\sum_{h=0}^n x_h^j}$. Let

$$\mu^j = \sum_{i=0}^{n} \alpha_i^j \mathrm{E}\left(\tilde{z}_i\right)$$

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

$$\begin{aligned} &\min_{\boldsymbol{\alpha} \in \mathbb{R}^{n+1}} R\left(\boldsymbol{\alpha} \cdot \widetilde{\mathbf{z}}\right) \\ &s.t. \\ &\sum_{i=0}^{n} \alpha_i \mathrm{E}\left(\tilde{z}_i\right) = \mu^j. \\ &\sum_{i=0}^{n} \alpha_i = 1. \end{aligned} \tag{7}$$

Given this, we can now discuss the geometry of portfolios in the μ-R plane where the horizontal axis is the risk of the return of a portfolio (R) and the vertical axis is the expected return (μ). 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 μ-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 μ-R plane lies to the right of the set of portfolios representing convex combinations of these two portfolios. Figure 1 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 (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 (7) for some $\mu^j \in \mathbb{R}$. Thus, the red curve in Figure 1 corresponds to the set of efficient portfolios.

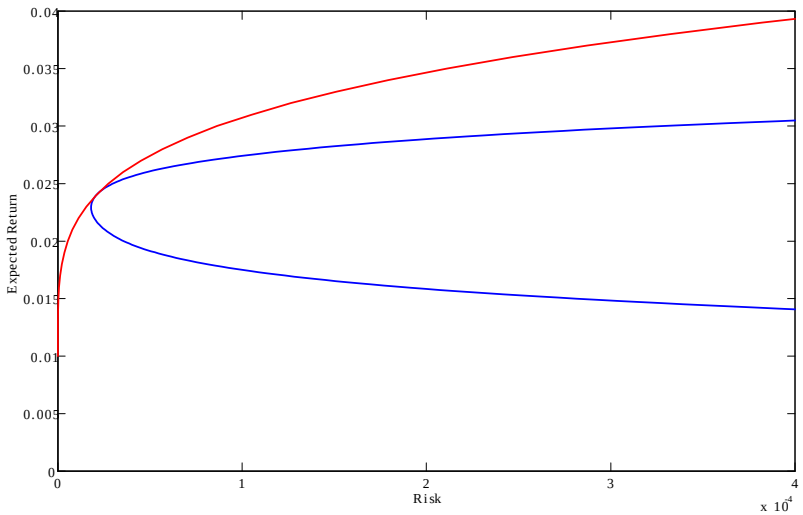

Figure 1: Portfolio Opportunity Set and Efficient Frontier

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 $\boldsymbol{\alpha}^P$ such that $\alpha_0^P = 0$, and for all investors $j \in \{1, ..., \ell\}$, the solution to Problem (7) is a linear combination of $\boldsymbol{\alpha}^P$ and the risk-free asset.

Theorem 3 *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 $\boldsymbol{\alpha}^1$ and $\boldsymbol{\alpha}^2$ be solutions to Problem (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 $\boldsymbol{\alpha}^1$ and $\boldsymbol{\alpha}^2$ have non-zero weights in some risky assets.[12] By the risk-free property and by the non-redundancy assumption, $R(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}) > 0$ for $j = 1, 2$. Hence, $\mu^j = \mathrm{E}(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}) > r_f$ for $j = 1, 2$, since otherwise $\boldsymbol{\alpha}^j$ 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 μ-R

[12] If only one investor holds non-zero weights in risky assets then two-fund separation is trivial.

plane as illustrated in Figure 2. Note that both $\boldsymbol{\alpha}^1$ and $\boldsymbol{\alpha}^2$ can be presented as a linear combination of the risk-free asset and some portfolios $\boldsymbol{\alpha}^{P_1}$ and $\boldsymbol{\alpha}^{P_2}$ 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 $\boldsymbol{\alpha}^{P_1} = \boldsymbol{\alpha}^{P_2}$. Suppose this is not the case. Then let $\hat{\boldsymbol{\alpha}}^1$ be a linear combination of $\boldsymbol{\alpha}^{P_2}$ and the risk-free asset such that $\mathrm{E}\left(\hat{\boldsymbol{\alpha}}^1 \cdot \widetilde{\mathbf{z}}\right) = \mu^1$. Similarly, let $\hat{\boldsymbol{\alpha}}^2$ be a portfolio of $\boldsymbol{\alpha}^{P_1}$ and the risk-free asset such that $\mathrm{E}\left(\hat{\boldsymbol{\alpha}}^2 \cdot \widetilde{\mathbf{z}}\right) = \mu^2$. By convexity of $R(\cdot)$, $\boldsymbol{\alpha}^1$ and $\boldsymbol{\alpha}^2$ are the unique solutions to Program (7) for $j = 1, 2$. Hence,

$$R\left(\hat{\boldsymbol{\alpha}}^1 \cdot \widetilde{\mathbf{z}}\right) > R\left(\boldsymbol{\alpha}^1 \cdot \widetilde{\mathbf{z}}\right) \text{ and } R\left(\hat{\boldsymbol{\alpha}}^2 \cdot \widetilde{\mathbf{z}}\right) > R\left(\boldsymbol{\alpha}^2 \cdot \widetilde{\mathbf{z}}\right). \tag{8}$$

Thus, as illustrated in Figure 2, the two curves must cross at least once. We will now show that such crossings are impossible. Indeed, by scaling independence (8) implies that for any $\lambda > 0$,

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

which together with risk-free property implies

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

This means that all linear combinations of $\boldsymbol{\alpha}^1$ with the risk-free asset (with positive λ) lie strictly to the left of all linear combinations of $\hat{\boldsymbol{\alpha}}^1$ with the risk-free asset. In particular, $\hat{\boldsymbol{\alpha}}^2$ can be obtained as a linear combination of $\boldsymbol{\alpha}^1$ 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 λ we obtain

$$R(\hat{\boldsymbol{\alpha}}^2 \cdot \widetilde{\mathbf{z}}) < R(\boldsymbol{\alpha}^2 \cdot \widetilde{\mathbf{z}}),$$

contradicting (8). Thus, two-fund separation must hold.

A corollary is that the unique portfolio $\boldsymbol{\alpha}^P$ is efficient. Indeed, let $\mu^P = \mathrm{E}\left(\boldsymbol{\alpha}^P \cdot \widetilde{\mathbf{z}}\right)$. Since in equilibrium all investors hold a linear combination of the risk-free asset and $\boldsymbol{\alpha}^P$, and since $\mu^j = \mathrm{E}\left(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}\right) \geq r_f$ for all j with strict inequality for some j, we have two cases:[13] (i) all investors hold $\boldsymbol{\alpha}^P$ with a non-negative weight, and $\mu^P > r_f$; or (ii) all investors hold $\boldsymbol{\alpha}^P$ 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 $\boldsymbol{\alpha}^P$. Thus, $\mu^P > r_f$.

[13] If all investors choose the risk-free asset then the market for risky assets cannot clear.

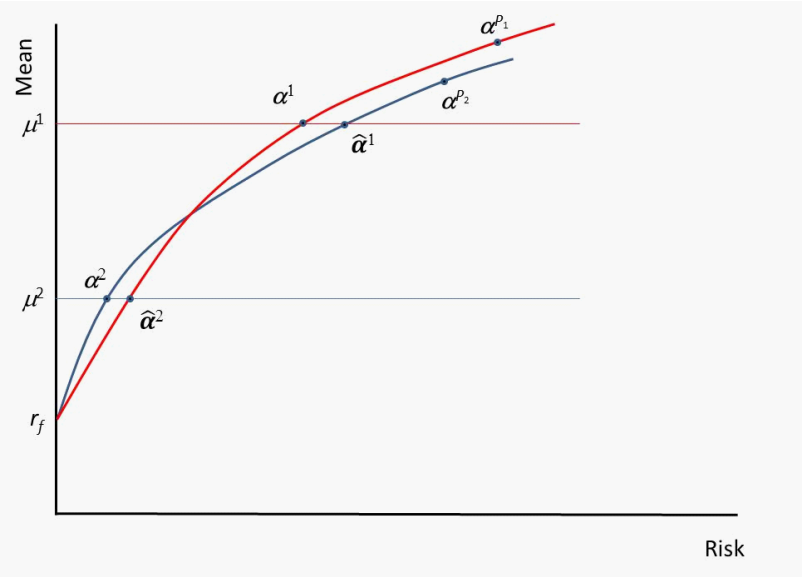

Figure 2: Graphical Illustration of the Proof of Theorem 3

Now, assume that $\boldsymbol{\alpha}' \neq \boldsymbol{\alpha}^P$ solves Problem (7) for $\mu^j = \mu^P$. Then, $R\left(\boldsymbol{\alpha}' \cdot \widetilde{\mathbf{z}}\right) < R\left(\boldsymbol{\alpha}^P \cdot \widetilde{\mathbf{z}}\right)$, and so by the same argument as in the proof of Theorem 3, all linear combinations of $\boldsymbol{\alpha}'$ with the risk-free asset would have strictly lower risk than the corresponding linear combinations of $\boldsymbol{\alpha}^P$ with the risk-free asset. This contradicts that $\boldsymbol{\alpha}^P$ and the risk-free asset span all optimal portfolios. We thus have:

Corollary 1 *Under the conditions of Theorem 3, the portfolio $\boldsymbol{\alpha}^P$ is efficient. In particular, it solves Problem (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 $\mathbf{x}^M = \left(x_1^M, ..., x_n^M\right)$ the *market portfolio* (consisting of risky assets only). Let $\boldsymbol{\alpha}^M$ be the corresponding portfolio weights. By Theorem 3, in equilibrium, the market portfolio is equal to $\boldsymbol{\alpha}^P$, the unique portfolio of risky assets that together with the risk-free asset spans all optimal portfolios.[14] Moreover, by corollary 1, the market portfolio is efficient, and its expected return is strictly higher than r_f.

Corollary 2 *Under the conditions of Theorem 3, the market portfolio is efficient. In particular, it solves Problem (7) for some $\mu^M > r_f$.*

[14]Note that $\boldsymbol{\alpha}^P$ is of dimension $n+1$, but its first component is zero. By saying that $\boldsymbol{\alpha}^P = \boldsymbol{\alpha}^M$ we mean that the $\alpha_i^P = \alpha_i^M$ for $i = 1, ..., n$.

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.

Theorem 4 *Consider an equilibrium with positive prices and let* $\boldsymbol{\alpha}^M$ *be the market portfolio. Assume that* $R(\cdot)$ *is scaling independent, convex, smooth, and satisfies the risk-free property. Then, for each asset* $i = 1, ..., n$,

$$\mathrm{E}(\tilde{z}_i) = r_f + \mathcal{B}_i^R \left(\mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right) - r_f\right), \tag{9}$$

where

$$\mathcal{B}_i^R = \frac{R_i\left(\boldsymbol{\alpha}^M\right)}{\sum_{h=1}^n \alpha_h^M R_h\left(\boldsymbol{\alpha}^M\right)}. \tag{10}$$

If $R(\cdot)$ *is also homogeneous of some degree* k*, then (10) takes the form*

$$\mathcal{B}_i^R = \frac{R_i\left(\boldsymbol{\alpha}^M\right)}{kR\left(\boldsymbol{\alpha}^M\right)}.$$

Thus, the security market line has the traditional form, with the generalized systematic risk measure $\left(\mathcal{B}_i^R\right)$ 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 $\boldsymbol{\alpha}^*$ 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) $R_i(\boldsymbol{\alpha}^*) - R_0(\boldsymbol{\alpha}^*)$, which by the risk-free property is just $R_i(\boldsymbol{\alpha}^*)$. Since $\boldsymbol{\alpha}^*$ 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, ξ, i.e.,

$$R_i(\boldsymbol{\alpha}^*) = \xi\left(\mathrm{E}(z_i) - r_f\right). \tag{11}$$

Taking the weighted average using the portfolio weights gives

$$\sum_{i=1}^n \alpha_i^* R_i(\boldsymbol{\alpha}^*) = \xi\left(\mathrm{E}\left(\boldsymbol{\alpha}^* \cdot \widetilde{\mathbf{z}}\right) - r_f\right). \tag{12}$$

Using (11) and (12) we obtain that for any efficient portfolio $\boldsymbol{\alpha}^*$,

$$\frac{R_i(\boldsymbol{\alpha}^*)}{\sum_{i=1}^n \alpha_i^* R_i(\boldsymbol{\alpha}^*)} = \frac{\mathrm{E}(z_i) - r_f}{\mathrm{E}(\boldsymbol{\alpha}^* \cdot \widetilde{\mathbf{z}}) - r_f}.$$

Namely, in equilibrium, $\mathcal{B}_i^R$ (as given in (10)) equals the ratio of the excess return of any asset i to the excess return of the efficient portfolio $\boldsymbol{\alpha}^*$. Finally, since $\boldsymbol{\alpha}^M$ has been shown to be efficient (Corollary 2) we obtain the result.

1.4.3 Applications and Empirical Implementation

We now provide several applications to illustrate the versatility and power of Theorem 4 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., $R(\tilde{z}) = \mathrm{Var}(\tilde{z})$, Theorem 4 coincides with the standard CAPM (see Appendix II for the derivation). Namely,

$$\mathcal{B}_i^R = \frac{\mathrm{Cov}\left(\tilde{z}_i, \boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\mathrm{Var}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}. \tag{13}$$

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 $R(\tilde{z}) = m_k(\tilde{z}) = \mathrm{E}(\tilde{z} - \mathrm{E}(\tilde{z}))^k$, k even. This risk measure satisfies all of the conditions in Theorem 4 (see Example 1). In this case the systematic risk takes the form (see Appendix II for the derivation)

$$\mathcal{B}_i^R = \frac{\mathrm{Cov}\left(\tilde{z}_i, \left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \boldsymbol{\alpha}^M \cdot \mathrm{E}(\widetilde{\mathbf{z}})\right)^{k-1}\right)}{m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}. \tag{14}$$

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 (13) as expected. Another important special case is $k=4$, in which $R(\tilde{z})$ measures the tail risk of $\tilde{z}$. Then,

$$\mathcal{B}_i^R = \frac{\text{Cov}\left(\tilde{z}_i, \left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \boldsymbol{\alpha}^M \cdot \text{E}\left(\widetilde{\mathbf{z}}\right)\right)^3\right)}{m_4\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}.$$

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, $\mathcal{B}_i^R$ again takes the form (14).

Application III: A CAPM reflecting aversion to downside risk. Assume $R(\tilde{z}) = \text{DR}_k(\tilde{z})$ for $k \geq 2$. This risk measure satisfies all of the conditions in Theorem 4 (see Example 3). The systematic risk is then given by (see Appendix II for the derivation)

$$\mathcal{B}_i^R = (-1)^k \frac{\text{Cov}\left[\tilde{z}_i, \left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \text{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^{k-1}\right]}{\left(\text{DR}_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^k}. \tag{15}$$

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 4 (see Example 5). 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)

$$\mathcal{B}_i^R = -\frac{\text{E}\left[\tilde{z}_i - \text{E}(\tilde{z}_i) \,|\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} \leq -\text{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})\right]}{\text{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})}. \tag{16}$$

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

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

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 1. 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

$$R\left(\tilde{z}\right)=\theta_1 w_2\left(\tilde{z}\right)+\theta_2 \mathrm{DR}_3\left(\tilde{z}\right)+\theta_3 w_4\left(\tilde{z}\right)+\theta_4 \mathrm{DES}_\delta\left(\tilde{z}\right)$$

for some confidence level δ. 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.[15] 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 1 these risk measures satisfy all the conditions in Theorem 4 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 of the systematic risk measures as calculated in the above applications (see Appendix II). Namely,

$$\mathcal{B}_i^R=\frac{R^1\left(\boldsymbol{\alpha}^M\right)}{R\left(\boldsymbol{\alpha}^M\right)}\mathcal{B}_i^{R^1}+\frac{R^2\left(\boldsymbol{\alpha}^M\right)}{R\left(\boldsymbol{\alpha}^M\right)}\mathcal{B}_i^{R^2}+\frac{R^3\left(\boldsymbol{\alpha}^M\right)}{R\left(\boldsymbol{\alpha}^M\right)}\mathcal{B}_i^{R^3}+\frac{R^4\left(\boldsymbol{\alpha}^M\right)}{R\left(\boldsymbol{\alpha}^M\right)}\mathcal{B}_i^{R^4}, \qquad (17)$$

where $R^1\left(\cdot\right)=\theta_1 w_2\left(\cdot\right)$, $R^2\left(\cdot\right)=\theta_2 \mathrm{DR}_3\left(\cdot\right)$, $R^3\left(\cdot\right)=\theta_3 w_4\left(\cdot\right)$, and $R^4\left(\cdot\right)=\theta_4 \mathrm{DES}_\delta\left(\cdot\right)$, and where $\mathcal{B}_i^{R^1}$, $\mathcal{B}_i^{R^2}$, $\mathcal{B}_i^{R^3}$, and $\mathcal{B}_i^{R^4}$ are given by (13)–(16).

Empirical Implementation

Similar to the classic CAPM, Theorem 4 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.

[15] Note that we are using here $w_2\left(\cdot\right)$ and $w_4\left(\cdot\right)$ (the normalized second and fourth moments) instead of $m_2\left(\cdot\right)$ and $m_4\left(\cdot\right)$. This is done to make sure that all of the components in $R\left(\cdot\right)$ are homogeneous of degree 1, and so $R\left(\cdot\right)$ is homogeneous.

To apply this approach in our case, one needs to first take a stand on what the risk measure R is. Then, using Theorem 4 one can estimate $\mathcal{B}_i^R$ 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 $\mathcal{B}_i^R$ from (17). This will be a weighted average of the betas prescribed in Applications I–IV. Note that unlike in the classic CAPM, $\mathcal{B}_i^R$ 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, $\mathcal{B}_i^R$ 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 (1976)) in which the model suggests the existence of multiple systematic factors but does not provide guidance as to what these factors are.

1.4.4 Further Discussion

Note that Theorem 4 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 ex-

pected 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

$$\mathrm{E}\left(\tilde{z}_i\right) = \gamma + \mathcal{B}_i\lambda$$

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.

1.5 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.

1.5.1 Axiomatic Characterization of Systematic Risk

A *risk allocation problem* of order $n \geq 1$ is a pair $(R, \mathbf{x})$, where R is a risk measure and $\mathbf{x} \in \mathbb{R}^n_{++}$ is a portfolio specifying the dollar amount invested in each of n assets

$\widetilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_n)$, and $R(\mathbf{x} \cdot \widetilde{\mathbf{z}}) \neq 0$. Denote the total dollar amount invested by $\bar{x} = \sum_{i=1}^{n} x_i$. Also, let $\boldsymbol{\alpha}$ 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 $\mathbf{B}^R(\mathbf{x}) = \left(\mathcal{B}_1^R(\mathbf{x}), ..., \mathcal{B}_n^R(\mathbf{x})\right)$ in $\mathbb{R}^n$. Intuitively, one can think of $\mathcal{B}_i^R(\mathbf{x})$ as the contribution of asset i to the total risk of portfolio $\mathbf{x}$, which is $R(\mathbf{x} \cdot \widetilde{\mathbf{z}})$. 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 in the traditional beta) the weighted average of systematic risk values across all assets is normalized to 1. Assets that contribute strongly to the risk of the portfolio (aggressive assets) have a beta greater than 1, whereas assets that have little contribution to total risk (defensive assets) have a beta less than 1. The weighted average of all asset betas is 1. We ask that a generalized systematic risk measure have the same property.

Axiom 1 *Normalization:* $\sum_{i=1}^{n} \alpha_i \mathcal{B}_i^R(\mathbf{x}) = 1$.

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

$$\mathcal{B}_i^R(\mathbf{x}) = \frac{R^1(\mathbf{x})}{R(\mathbf{x})}\mathcal{B}_i^{R^1}(\mathbf{x}) + \frac{R^2(\mathbf{x})}{R(\mathbf{x})}\mathcal{B}_i^{R^2}(\mathbf{x}) \quad \textit{for all } i = 1, ..., n.$$

When risk is measured using variance, the notion of systematic risk is closely tied to the concepts of correlation and covariance. It is not easy to generalize these concepts to arbitrary 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 $\widetilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_n)$ are R-*perfectly correlated* if there exists a function $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ and a non-zero vector $\mathbf{q} = (q_1, ..., q_n) \in \mathbb{R}^n_+$, such that for any portfolio $\boldsymbol{\eta} = (\eta_1, ..., \eta_n) \in \mathbb{R}^n_+$ we have $R(\boldsymbol{\eta} \cdot \widetilde{\mathbf{z}}) = g(\boldsymbol{\eta} \cdot \mathbf{q})$. 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 $\mathbf{q}$.[16] Note that different risk measures correspond to different concepts of R-perfect correlation, which typically would not coincide with the standard notion of perfect correlation associated with the variance.[17]

The next axiom imposes that if the n assets are R-perfectly correlated, then their systematic risk measures are proportional to each other.

Axiom 3 *Proportionality: If* $\widetilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_n)$ *are* R*-perfectly correlated with weights* $\mathbf{q} = (q_1, ..., q_n)$*, then*

$$q_j \mathcal{B}_i^R(\mathbf{x}) = q_i \mathcal{B}_j^R(\mathbf{x}) \text{ for all } i, j = 1, ..., n. \tag{18}$$

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

[16] To see the correspondence to the standard notion of perfect correlation, consider the following example. Assume risk is measured using variance and let $\widetilde{\mathbf{z}} = (\tilde{z}_1, \tilde{z}_2, \tilde{z}_3)$ 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 (η_1, η_2, η_3) we have

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

Thus, we can set $g(t) = t^2$ and the vector of weights is $\mathbf{q} = \sqrt{Var(\tilde{z}_1)}(1, 2, 5)$. 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.

[17] 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.

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 $\widetilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_n)$ are *R-positively correlated* if $R_i(\boldsymbol{\eta} \cdot \widetilde{\mathbf{z}}) \geq 0$ for all $\boldsymbol{\eta} \in \mathbb{R}^n_+$ 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.[18] The next axiom requires that when the assets are R-positively correlated, the systematic risk of all assets is non-negative.

Axiom 4 *Monotonicity: If $\widetilde{\mathbf{z}} = (\tilde{z}_1, ..., \tilde{z}_n)$ are R-positively correlated, then $\mathcal{B}_i^R(\mathbf{x}) \geq 0$ 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 4.

Theorem 5 *There exists a unique systematic risk measure satisfying Axioms 1–4. For each risk allocation problem $(R, \mathbf{x})$ of order n, it is given by*

$$\mathcal{B}_i^R(\mathbf{x}) = \frac{\bar{x} \int_0^1 R_i(tx_1, ..., tx_n)\, dt}{R(x_1, ..., x_n)} \text{ for } i = 1, ..., n. \tag{19}$$

Furthermore, if R is homogeneous of some degree k, then (19) reduces to

$$\mathcal{B}_i^R(\mathbf{x}) = \frac{R_i(\boldsymbol{\alpha})}{\sum_{h=1}^n \alpha_h R_h(\boldsymbol{\alpha})} \tag{20}$$

$$= \frac{R_i(\boldsymbol{\alpha})}{kR(\boldsymbol{\alpha})}. \tag{21}$$

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 4 only with respect to an arbitrary portfolio rather

[18] 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.

than the market portfolio. When the risk measure is not homogeneous, the expression in (19) shows that systematic risk depends not only on marginal contributions at $\mathbf{x}$, but rather on marginal contributions along a diagonal between $(0, ..., 0)$ and $\mathbf{x}$. 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, $\mathcal{B}_i^R(\mathbf{x})$ 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, $\mathcal{B}_i^R(\mathbf{x})$ depends only on the portfolio weights $\boldsymbol{\alpha}$ (and not on the dollar amounts invested in each asset). Indeed, in the homogeneous case $R_i(tx_1, ..., tx_n)$ is proportional to $R_i(x_1, ..., x_n)$ for all $t \in [0, 1]$, yielding the simple expression in (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 5 is in Appendix I. It relies on the solutions to cost allocation problems established in Billera and Heath (1982).[19] In this proof we draw 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 (19) satisfies Axioms 1–4. Suppose that $\mathcal{B}_i^R(\mathbf{x})$ is given by (19). Then,

$$\begin{aligned}
\sum_{i=1}^{n} \alpha_i \mathcal{B}_i^R(\mathbf{x}) &= \sum_{i=1}^{n} \frac{x_i}{\bar{x}} \frac{\bar{x} \int_0^1 R_i(tx_1, ..., tx_n)\, dt}{R(x_1, ..., x_n)} \\
&= \frac{\int_0^1 \sum_{i=1}^{n} x_i R_i(tx_1, ..., tx_n)\, dt}{R(x_1, ..., x_n)} \\
&= \frac{\int_0^1 \frac{dR(tx_1, ..., x_n)}{dt} dt}{R(x_1, ..., x_n)} = 1, \quad (\text{since } R(0) = 0)
\end{aligned}$$

[19] Billera and Heath (1982) define a cost allocation problem of order n as a pair $(h, \mathbf{x})$ where $h : \mathbb{R}_+^n \to \mathbb{R}$ is continuously differentiable and $h(\mathbf{0}) = 0$. They interpret $\mathbf{x}$ 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.

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

$$\begin{aligned}\mathcal{B}_i^R(\mathbf{x}) &= \frac{\bar{x}\int_0^1 R_i(tx_1,...,tx_n)\,dt}{R(x_1,...,x_n)} \\ &= \frac{\frac{\bar{x}\int_0^1 R_i^1(tx_1,...,tx_n)dt}{R^1(x_1,...,x_n)}R^1(x_1,...,x_n) + \frac{\bar{x}\int_0^1 R_i^2(tx_1,...,tx_n)dt}{R^2(x_1,...,x_n)}R^2(x_1,...,x_n)}{R(x_1,...,x_n)},\end{aligned}$$

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

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

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

$$\mathcal{B}_i^R(\mathbf{x}) = \frac{\bar{x}q_i\int_0^1 g'(t\mathbf{x}\cdot\mathbf{q})\,dt}{R(x_1,...,x_n)},$$

which implies (18). Finally, given the definition of R-positive correlation, it is immediate that (19) satisfies Axiom 4.

1.5.2 Applying the Result

In Section 1.4.3 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 1.4, but for which Theorem 5 applies.

Recall the Aumann-Serrano and Foster-Hart risk measures in Example 6. These measures are homogeneous, convex, and smooth, but they do not satisfy the risk-free property.[20] Still, Theorem 5 allows us to calculate the systematic risk associated with these risk measures.

[20] 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 $\mathrm{E}(\tilde{z}) > 0$ 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

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

where the equality follows since both the AS and the FH measures are homogeneous of degree 1.

Using Theorem 5 and applying the implicit function theorem to (4) and (5) yields the systematic risk of individual assets associated with the AS and FH measures relative to any portfolio weights $\boldsymbol{\alpha}$ as follows:

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

and

$$\mathcal{B}_i^{R^{FH}}(\boldsymbol{\alpha}) = \frac{\mathrm{E}\left[\frac{\widetilde{z}_i}{R(\boldsymbol{\alpha})+\boldsymbol{\alpha}\cdot\widetilde{\mathbf{z}}}\right]}{\mathrm{E}\left[\frac{\boldsymbol{\alpha}\cdot\widetilde{\mathbf{z}}}{R(\boldsymbol{\alpha})+\boldsymbol{\alpha}\cdot\widetilde{\mathbf{z}}}\right]}.$$

1.5.3 Discussion

When the risk measure is homogeneous both the equilibrium approach and the axiomatic approach yield the same result:

$$\mathcal{B}_i^{R}(\boldsymbol{\alpha}) = \frac{R_i(\boldsymbol{\alpha})}{\sum_{h=1}^{n}\alpha_h R_h(\boldsymbol{\alpha})}. \tag{22}$$

It is interesting to ask what would happen if we used (22) to define systematic risk when R is not homogeneous (instead of using (19)). In particular, this alternative measure only relies on the marginal contribution of asset i at $\boldsymbol{\alpha}$ and not along the diagonal. In the absence of homogeneity these two alternative definitions yield different results. Thus, given Theorem 5, it must be that (22) 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. We are not able to provide an axiomatization that leads to (22) as a solution when the risk measure is not homogeneous. Notably, however, many commonly used risk measures *are* homogeneous, and thus the two approaches often coincide.

Another approach to measuring systematic risk might be to define

$$\mathcal{B}_i^{R}(\boldsymbol{\alpha}) = \frac{R_i(\boldsymbol{\alpha})}{R(\boldsymbol{\alpha})},$$

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 (22) can also be written as

$$\mathcal{B}_i^R(\boldsymbol{\alpha}) = \frac{\frac{d}{dt}\big|_{t=0} R(\boldsymbol{\alpha}+t\boldsymbol{\varepsilon}^i)}{\frac{d}{dt}\big|_{t=0} R(\boldsymbol{\alpha}+t\boldsymbol{\alpha})},$$

where $\boldsymbol{\varepsilon}^i$ 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.

1.6 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 measure coincides 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.

References

[1] Acharya, Viral V., Lasse H. Pedersen, Thomas Philippon, and Mathew Richardson, 2010. 'Measuring systemic risk.' Working Paper, New York University.

[2] Adrian, Tobias, and Markus K. Brunnermeier, 2011. 'CoVaR.' Fed Reserve Bank of New York Staff Reports.

[3] Ang, Andrew, Joseph Chen, and Yuhang Xing, 2006. 'Downside risk.' *Review of Financial Studies*, 19, 1191–1239.

[4] Aumann, Robert J., and Roberto Serrano, 2008. 'An economic index of riskiness.' *Journal of Political Economy*, 116, 810–836.

[5] Aumann, Robert J., and Lloyd S. Shapley, 1974. 'Values of non-atomic games.' *Princeton University Press.*

[6] Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath, 1999. 'Coherent measures of risk.' *Mathematical Finance*, 9, 203–228.

[7] Bansal, Ravi, and Amir Yaron, 2004. 'Risks for the long run: A potential resolution of asset pricing puzzles.' *Journal of Finance*, 59, 1481–1509.

[8] Barro, Robert J., 2006. 'Rare disasters and asset markets in the twentieth century.' *Quarterly Journal of Economics*, 121, 823–866.

[9] Barro, Robert J., 2009. 'Rare disasters, asset prices, and welfare costs.' *American Economic Review*, 99, 243–264.

[10] Billera, Louis J., and David C. Heath, 1982. 'Allocation of shared costs: A set of axioms yielding a unique procedure.' *Mathematics of Operations Research*, 7, 32–39.

[11] Billera, Louis J., David C. Heath, and Joseph Raanan, 1978. 'Internal telephone billing rates - A novel application of non-atomic game theory.' *Operations Research*, 26, 956–965.

[12] Billera, Louis J., David C. Heath, and Robert E. Verrecchia, 1981. 'A unique procedure for allocating common costs from a production process.' *Journal of Accounting Research*, 19, 185–196.

[13] Campbell, John Y., and John H. Cochrane, 1999. 'By force of habit: A consumption-based explanation of aggregate stock market behavior.' *Journal of Political Economy,* 107, 205–251.

[14] Cass, David, and Joseph E. Stiglitz, 1970. 'The structure of investor preferences and asset returns, and separability in portfolio allocation: A contribution to the pure theory of mutual funds.' *Journal of Economic Theory*, 2, 122–160.

[15] Chen, Hui, Scott Joslin, and Ngoc-Khanh Tran, 2012. 'Rare disasters and risk sharing with heterogeneous beliefs.' *Review of Financial Studies*, 25, 2189–2224.

[16] Cochrane, John H., 2001. 'Asset Pricing.' *Princeton University Press.*

[17] Dybvig, Philip, and Fang Liu, 2015. 'On investor preferences and mutual fund separation.' Working paper, Washington University in St. Louis.

[18] Dybvig, Philip H., and Stephen A. Ross, 1982. 'Portfolio efficient set.' *Econometrica*, 50, 1525–1546.

[19] Epstein, Larry G., and Stanley E. Zin, 1989. 'Substitution, risk aversion, and the temporal behavior of consumption and asset returns: A theoretical framework.' *Econometrica*, 57, 937–969.

[20] Fama, Eugene F., and Kenneth R. French, 1992. 'The cross-section of expected stock returns.' *Journal of Finance*, 47, 427–465.

[21] Fama, Eugene F., and James D. MacBeth, 1973. 'Risk, return, and equilibrium: empirical tests.' *Journal of Political Economy,* 81, 607–636.

[22] Foster, Dean P., and Sergiu Hart, 2009. 'An operational measure of riskiness.' *Journal of Political Economy*, 117, 785–814.

[23] Foster, Dean P., and Sergiu Hart, 2013. 'A wealth-requirement axiomatization of riskiness.' *Theoretical Economics,* 8, 591–620.

[24] Gabaix, Xavier, 2008. 'Variable rare disasters: A tractable theory of ten puzzles in macro-finance.' *American Economic Review*, 98, 64–67.

[25] Gabaix, Xavier, 2012. 'Variable rare disasters: An exactly solved framework for ten puzzles in macro-finance.' *Quarterly Journal of Economics*, 127, 645–700.

[26] Gourio, François, 2012. 'Disaster risk and business cycles.' *American Economic Review*, 102, 2734–2766.

[27] Hadar, Josef, and William R. Russell, 1969. 'Rules for ordering uncertain prospects.' *American Economic Review*, 59, 25–34.

[28] Hanoch, Giora, and Haim Levy, 1969. 'The efficiency analysis of choices involving risk.' *Review of Economic Studies*, 36, 335–346.

[29] Hansen, Lars Peter, and Scott F. Richard, 1987. 'The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models.' *Econometrica*, 55, 587–613.

[30] Hart, Sergiu, 2011. 'Comparing risks by acceptance and rejection.' *Journal of Political Economy*, 119, 617–638.

[31] Harvey, Campbell R., and Akhtar Siddique, 2000. 'Conditional skewness in asset pricing tests.' *Journal of Finance*, 55, 1263–1295.

[32] Huang, Xin, Hao Zhou, and Haibin Zhu, 2009. 'A framework for assessing the systemic risk of major financial institutions.' *Journal of Banking and Finance,* 33, 2036–2049.

[33] Jean, William H., 1971. 'The extension of portfolio analysis to three or more parameters.' *Journal of Financial and Quantitative Analysis*, 6, 505–515.

[34] Kadan, Ohad, and Fang Liu, 2014. 'Performance evaluation with high moments and disaster risk.' *Journal of Financial Economics,* 113, 131–155.

[35] Kane, Alex, 1982. 'Skewness preference and portfolio choice.' *Journal of Financial and Quantitative Analysis*, 17, 15–25.

[36] Kraus, Alan, and Robert H. Litzenberger, 1976. 'Skewness preference and the valuation of risk assets.' *Journal of Finance*, 31, 1085–1100.

[37] Lettau, Martin, Matteo Maggiori, and Michael Weber, 2013. 'Conditional risk premia in currency markets and other asset classes.' *Journal of Financial Economics,* forthcoming.

[38] Levy, Moshe, and Richard Roll, 2010. 'The market portfolio may be mean/variance efficient after all.' *Review of Financial Studies*, 23, 2464–2491.

[39] Lintner, John, 1965a. 'The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets.' *Review of Economics and Statistics*, 47, 13–37.

[40] Lintner, John, 1965b. 'Security prices, risk, and maximal gains from diversification.' *Journal of Finance*, 20, 587–615.

[41] Luenberger, David G., 1969. 'Optimization by vector space methods.' *John Wiley & Sons, Inc.*

[42] Machina, Mark J., 1982. '"Expected utility" analysis without the independence axiom.' *Econometrica*, 50, 277–323.

[43] Mossin, Jan, 1966. 'Equilibrium in a capital asset market.' *Econometrica*, 34, 768–783.

[44] Müller, Sigrid M., and Mark J. Machina, 1987. 'Moment preferences and polynomial utility.' *Economics Letters,* 23, 349–353.

[45] Nielsen, Lars T., 1989. 'Asset market equilibrium with short-selling.' *Review of Economic Studies*, 56, 467–473.

[46] Nielsen, Lars T., 1990. 'Existence of equilibrium in CAPM.' *Journal of Economic Theory*, 52, 223–231.

[47] Nielsen, Lars T., 1992. 'Positive prices in CAPM.' *Journal of Finance*, 47, 791–808.

[48] Powers, Michael R., 2007. 'Using Aumann-Shapley values to allocate insurance risk: The case of inhomogeneous losses.' *North American Actuarial Journal*, 11, 113–127.

[49] Rockafellar, R. Tyrrell, Stan Uryasev, and Michael Zabarankin, 2006a. 'Generalized deviations in risk analysis.' *Finance and Stochastics,* 10, 51–74.

[50] Rockafellar, R. Tyrrell, Stan Uryasev, and Michael Zabarankin, 2006b. 'Master funds in portfolio analysis with general deviation measures.' *Journal of Banking and Finance*, 30, 743–778.

[51] Ross, Stephen A., 1976b. 'The arbitrage theory of capital asset pricing.' *Journal of Economic Theory,* 13, 341–360.

[52] Ross, Stephen A., 1978. 'Mutual fund separation in financial theory: The separating distributions.' *Journal of Economic Theory*, 17, 254–286.

[53] Rothschild, Michael, and Joseph E. Stiglitz, 1970. 'Increasing risk I: A definition.' *Journal of Economic Theory*, 2, 225–243.

[54] Rubinstein, Mark, 1973. 'The fundamental theorem of parameter-preference security valuation.' *Journal of Financial and Quantitative Analysis*, 8, 61–69.

[55] Samet, Dov, Yair Tauman, and Israel Zang, 1984. 'An application of the Aumann-Shapley prices for cost allocation in transportation problems.' *Mathematics of Operations Research*, 9, 25–42.

[56] Shapley, Lloyd S., 1953. 'A value for n-person games.' *Annals of Mathematical Studies*, 28, 307–317.

[57] Sharpe, William F., 1964. 'Capital asset prices: A theory of market equilibrium under conditions of risk.' *Journal of Finance*, 19, 425–442.

[58] Tarashev, Nikola A., Claudio E. V. Borio, and Kostas Tsatsaronis, 2010. 'Attributing systemic risk to individual institutions.' BIS Working Paper No. 308.

[59] Tobin, James, 1958. 'Liquidity preferences as behavior towards risk.' *Review of Economic Studies*, 25, 65–86.

[60] Wachter, Jessica A., 2013. 'Can time-varying risk of rare disasters explain aggregate stock market volatility?' *Journal of Finance*, 68, 987–1035.

Appendix

Appendix I: Proofs of Main Theorems

Proof of Theorem 1: 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 $\{\boldsymbol{\zeta} \in \mathbb{R}^{n+1} : U^j(\boldsymbol{\zeta}) \geq U^j(\boldsymbol{\zeta}')\}$ is closed for all $\boldsymbol{\zeta}' \in \mathbb{R}^{n+1}$ and for all $j = \{1, ..., \ell\}$; (iii) If $\boldsymbol{\zeta}, \boldsymbol{\zeta}' \in \mathbb{R}^{n+1}$ and $U^j(\boldsymbol{\zeta}') > U^j(\boldsymbol{\zeta})$, then $U^j(t\boldsymbol{\zeta}' + (1-t)\boldsymbol{\zeta}) > U^j(\boldsymbol{\zeta})$ 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 $\boldsymbol{\zeta}$, 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 $\{\boldsymbol{\zeta} \in \mathbb{R}^{n+1} : U^j(\boldsymbol{\zeta}) > U^j(\boldsymbol{\zeta}')\}$ is open for all j and $\boldsymbol{\zeta}' \in \mathbb{R}^{n+1}$. ■

Proof of Theorem 2: Suppose that the equilibrium bundle of investor j is $\boldsymbol{\zeta}^j$. Let

$\bar{x}^j = \sum_{i=0}^n x_i^j = \mathbf{p} \cdot \boldsymbol{\zeta}^j$ be the total dollar amount of investment of investor j. Then,

$$\begin{aligned} U^j\left(\boldsymbol{\zeta}^j\right) &= V^j\left(\mathrm{E}\left(\sum_{i=0}^n \zeta_i^j \tilde{y}_i\right), R\left(\sum_{i=0}^n \zeta_i^j \tilde{y}_i\right)\right) \\ &= V^j\left(\bar{x}^j \mathrm{E}\left(\sum_{i=0}^n \frac{\zeta_i^j}{\bar{x}^j} \tilde{y}_i\right), R\left(\bar{x}^j \sum_{i=0}^n \frac{\zeta_i^j}{\bar{x}^j} \tilde{y}_i\right)\right) \\ &= V^j\left(\bar{x}^j \mathrm{E}\left(\sum_{i=0}^n \frac{\zeta_i^j p_i}{\bar{x}^j} \frac{\tilde{y}_i}{p_i}\right), R\left(\bar{x}^j \sum_{i=0}^n \frac{\zeta_i^j p_i}{\bar{x}^j} \frac{\tilde{y}_i}{p_i}\right)\right) \\ &= V^j\left(\bar{x}^j \mathrm{E}\left(\sum_{i=0}^n \frac{x_i^j}{\bar{x}^j} \tilde{z}_i\right), R\left(\bar{x}^j \sum_{i=0}^n \frac{x_i^j}{\bar{x}^j} \tilde{z}_i\right)\right) \\ &= V^j\left(\bar{x}^j \mathrm{E}\left(\sum_{i=0}^n \alpha_i^j \tilde{z}_i\right), R\left(\bar{x}^j \sum_{i=0}^n \alpha_i^j \tilde{z}_i\right)\right) \\ &= V^j\left(\bar{x}^j \mathrm{E}\left(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}\right), R\left(\bar{x}^j \left(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}\right)\right)\right). \end{aligned} \tag{23}$$

From the definition of equilibrium, each investor chooses $\boldsymbol{\zeta}^j$ to maximize $U^j\left(\boldsymbol{\zeta}^j\right)$ subject to $\bar{x}^j \leq \mathbf{p} \cdot \mathbf{e}^j$, where by the positivity of prices $\bar{x}^j = \mathbf{p} \cdot \mathbf{e}^j > 0$ (using that $\mathbf{e}^j \in \mathbb{R}_+^{n+1}$ is not zero by assumption). From (23) and since V^j is strictly increasing in the first argument and strictly decreasing in the second argument, we have that for any positive $\bar{x}^j$, $U^j\left(\boldsymbol{\zeta}^j\right)$ is strictly increasing in $\mathrm{E}\left(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}\right)$ and strictly decreasing in $R\left(\bar{x}^j \left(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}\right)\right)$. Therefore, in equilibrium, $\boldsymbol{\alpha}^j$ must minimize $R\left(\bar{x}^j \left(\boldsymbol{\alpha} \cdot \widetilde{\mathbf{z}}\right)\right)$ for a given level of expected return $\mathrm{E}\left(\boldsymbol{\alpha}^j \cdot \widetilde{\mathbf{z}}\right)$. By scaling independence, this is equivalent to minimizing $R\left(\boldsymbol{\alpha} \cdot \widetilde{\mathbf{z}}\right)$ for a given level of expected return, and thus, to solving Problem (7). The solution is unique since we assumed that $R(\cdot)$ is a convex risk measure, and so $R\left(\boldsymbol{\alpha} \cdot \widetilde{\mathbf{z}}\right)$ is convex as a function of $\boldsymbol{\alpha}$. ■

Proof of Theorem 4: By the smoothness of $R(\cdot)$ and by Theorem 1, the solution to Problem (7) for some $\mu^j = \mu$ is determined by the first order conditions. To solve this program, form the Lagrangian

$$\mathcal{L}(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha}) - \xi\left(\sum_{i=1}^n \alpha_i \mathrm{E}\left(\tilde{z}_i\right) + \left(1 - \sum_{i=1}^n \alpha_i\right) r_f - \mu\right),$$

where ξ is a Lagrange multiplier. Equivalently,

$$\mathcal{L}(\boldsymbol{\alpha}) = R\left(1 - \sum_{i=1}^n \alpha_i, \alpha_1, ..., \alpha_n\right) - \xi\left(\sum_{i=1}^n \alpha_i \mathrm{E}\left(\tilde{z}_i\right) + \left(1 - \sum_{i=1}^n \alpha_i\right) r_f - \mu\right).$$

The first order condition states that for all $i = 1, ..., n$,

$$-R_0(\boldsymbol{\alpha}^*) + R_i(\boldsymbol{\alpha}^*) - \xi(\mathrm{E}(\tilde{z}_i) - r_f) = 0, \tag{24}$$

where $\boldsymbol{\alpha}^*$ is any efficient portfolio (the market portfolio being a special case). By the risk-free property, $R_0(\boldsymbol{\alpha}^*) = 0$. Hence,

$$R_i(\boldsymbol{\alpha}^*) = \xi(\mathrm{E}(\tilde{z}_i) - r_f). \tag{25}$$

It follows that

$$\begin{aligned} \sum_{i=1}^{n} \alpha_i^* R_i(\boldsymbol{\alpha}^*) &= \xi \sum_{i=1}^{n} \alpha_i^* (\mathrm{E}(\tilde{z}_i) - r_f) \\ &= \xi(\mathrm{E}(\widetilde{\mathbf{z}} \cdot \boldsymbol{\alpha}^*) - r_f). \end{aligned} \tag{26}$$

From (25) and (26) we obtain

$$\frac{R_i(\boldsymbol{\alpha}^*)}{\sum_{h=1}^{n} \alpha_h^* R_h(\boldsymbol{\alpha}^*)} = \frac{\mathrm{E}(\tilde{z}_i) - r_f}{\mathrm{E}(\widetilde{\mathbf{z}} \cdot \boldsymbol{\alpha}^*) - r_f},$$

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

$$\frac{R_i(\boldsymbol{\alpha}^*)}{kR(\boldsymbol{\alpha}^*)} = \frac{\mathrm{E}(\tilde{z}_i) - r_f}{\mathrm{E}(\widetilde{\mathbf{z}} \cdot \boldsymbol{\alpha}^*) - r_f}.$$

■

Proof of Theorem 5: The proof relies on a mapping between risk allocation problems as defined in Section 1.5.1 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 $(h, \mathbf{x})$ where $h : \mathbb{R}_+^n \to \mathbb{R}$ is continuously differentiable and $h(\mathbf{0}) = 0$. Since R is smooth and satisfies $R(0) = 0$ we can view any risk allocation problem, $(R, \mathbf{x})$, of order n, as a cost allocation problem as defined in BH by setting $h(\mathbf{x}) = R(\mathbf{x} \cdot \widetilde{\mathbf{z}})$. Given this mapping we will use $(R, \mathbf{x})$ 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 $(R, \mathbf{x})$ of order n a vector $\mathbf{c}(R, \mathbf{x}) \in \mathbb{R}^n$. That is, $\mathbf{c}(R, \mathbf{x})$ 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 1.5.1 and the BH cost allocation procedures as follows. If $\mathbf{B}^R(\mathbf{x})$ is a systematic risk measure of the risk allocation problem $(R, \mathbf{x})$, then

$$\mathbf{c}(R, \mathbf{x}) = \frac{\mathbf{B}^R(\mathbf{x}) R(\mathbf{x} \cdot \widetilde{\mathbf{z}})}{\bar{x}} \tag{27}$$

is a cost allocation procedure for the corresponding cost allocation problem $(R, \mathbf{x})$. Namely, risk allocation measures can be viewed as scaled versions of cost allocation procedures for the corresponding problems.

Lemma 2 *If a systematic risk measure* $\mathbf{B}^R(\mathbf{x})$ *satisfies Axioms 1-4, then the corresponding cost allocation procedure* $\mathbf{c}(R, \mathbf{x})$ *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, $\sum_{i=1}^n \alpha_i \mathcal{B}_i^R(\mathbf{x}) = 1$ is equivalent to $\sum_{i=1}^n \frac{x_i R(\mathbf{x}) \mathcal{B}_i^R(\mathbf{x})}{\bar{x}} = R(\mathbf{x})$, which using (27) is equivalent to $\sum_{i=1}^n x_i c_i(R, \mathbf{x}) = R(\mathbf{x})$. 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

$$\mathcal{B}_i^R(\mathbf{x}) = \frac{R^1(\mathbf{x})}{R(\mathbf{x})} \mathcal{B}_i^{R^1}(\mathbf{x}) + \frac{R^2(\mathbf{x})}{R(\mathbf{x})} \mathcal{B}_i^{R^2}(\mathbf{x}).$$

Then

$$\frac{\mathcal{B}_i^R(\mathbf{x}) R(\mathbf{x})}{\bar{x}} = \frac{R^1(\mathbf{x}) \mathcal{B}_i^{R^1}(\mathbf{x})}{\bar{x}} + \frac{R^2(\mathbf{x}) \mathcal{B}_i^{R^2}(\mathbf{x})}{\bar{x}}.$$

That is,

$$c_i(R, \mathbf{x}) = c_i\left(R^1, \mathbf{x}\right) + c_i\left(R^2, \mathbf{x}\right),$$

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 $\boldsymbol{\eta} \in \mathbb{R}_+^n$,

$$R(\boldsymbol{\eta} \cdot \widetilde{\mathbf{z}}) = g(\boldsymbol{\eta} \cdot \mathbf{q}) \tag{28}$$

for some function $g(\cdot)$ and a non-zero vector $\mathbf{q} \in \mathbb{R}^n_+$. Then, $(\tilde{z}_1, ..., \tilde{z}_n)$ are R-perfectly correlated.

By Axiom 3 for all $i, j = 1, ..., n$,

$$q_j \mathcal{B}_i^R(\mathbf{x}) = q_i \mathcal{B}_j^R(\mathbf{x}), \tag{29}$$

and hence

$$\alpha_i q_j \mathcal{B}_i^R(\mathbf{x}) = \alpha_i q_i \mathcal{B}_j^R(\mathbf{x}).$$

Summing over $i = 1, ..., n$ gives

$$q_j \sum_{i=1}^n \alpha_i \mathcal{B}_i^R(\mathbf{x}) = (\boldsymbol{\alpha} \cdot \mathbf{q}) \mathcal{B}_j^R(\mathbf{x}). \tag{30}$$

By Axiom 1 we know that $\sum_{i=1}^n \alpha_i \mathcal{B}_i^R(\mathbf{x}) = 1$. Plugging this into (30) we have

$$q_j = (\boldsymbol{\alpha} \cdot \mathbf{q}) \mathcal{B}_j^R(\mathbf{x}) \text{ for } j = 1, ..., n.$$

By (27), and recalling that $R(\mathbf{x}) \neq 0$,

$$q_j = (\boldsymbol{\alpha} \cdot \mathbf{q}) \frac{c_j(R, \mathbf{x}) \bar{x}}{R(\mathbf{x})} = (\mathbf{x} \cdot \mathbf{q}) \frac{c_j(R, \mathbf{x})}{R(\mathbf{x})} \text{ for } j = 1, ..., n. \tag{31}$$

If $\mathbf{x} \cdot \mathbf{q} = \mathbf{0}$ this implies that $q_j = 0$ for all j, contradicting that $\mathbf{q}$ is a non-zero vector. Hence, $\mathbf{x} \cdot \mathbf{q}$ is not zero. We then have

$$c_j(R, \mathbf{x}) = \frac{q_j R(\mathbf{x})}{(\mathbf{x} \cdot \mathbf{q})} \text{ for all } j = 1, .., n. \tag{32}$$

Consider an asset with return $\tilde{w} = \frac{\mathbf{x} \cdot \tilde{\mathbf{z}}}{\mathbf{x} \cdot \mathbf{q}}$. Namely, investing $\mathbf{x} \cdot \mathbf{q}$ dollars in this asset yields the same return as of the portfolio $\mathbf{x}$. Then,

$$R((\mathbf{x} \cdot \mathbf{q}) \tilde{w}) = R(\mathbf{x} \cdot \tilde{\mathbf{z}}) = g(\mathbf{x} \cdot \mathbf{q}).$$

Consider now the risk allocation problem of order 1 with the single asset $\tilde{w}$ held at the amount $\mathbf{x} \cdot \mathbf{q}$. By Axiom 1 the systematic risk measure of this asset must satisfy

$$\mathcal{B}^R(\mathbf{x} \cdot \mathbf{q}) = 1,$$

or equivalently using (27),

$$c(g, \mathbf{x} \cdot \mathbf{q}) = \frac{R((\mathbf{x} \cdot \mathbf{q}) \tilde{w})}{\mathbf{x} \cdot \mathbf{q}} = \frac{g(\mathbf{x} \cdot \mathbf{q})}{\mathbf{x} \cdot \mathbf{q}}.$$

Plugging back into (32) and using that $R(\mathbf{x}) = g(\mathbf{x} \cdot \mathbf{q})$ we have

$$c_j(R, \mathbf{x}) = c(g, \mathbf{x} \cdot \mathbf{q}) q_j.$$

This is exactly what Condition (2.3) in BH requires, restricting attention to the case that $\mathbf{q}$ 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 (27) and the definition of R-positive correlation.

Having established Lemma 2 we now turn to completing the proof of the theorem. First, existence has been proved in the text by showing that (19) satisfies Axioms 1-4. To show uniqueness note that Lemma 2 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 $\mathbf{c}(R, \mathbf{x})$ satisfying Conditions (2.1)-(2.4). It follows (using the mapping (27)) that there is a unique systematic risk measure satisfying Axioms 1-4. Thus, the unique systematic risk measure is given by (19).

Finally, to see that (19) and (20) are equivalent when R is homogeneous of degree k, note first that in this case

$$\begin{aligned} \int_0^1 R_i(tx_1, ..., tx_n)\, dt &= R_i(x_1, ..., x_n) \int_0^1 t^{k-1} dt \\ &= \frac{R_i(x_1, ..., x_n)}{k}, \end{aligned}$$

where the first equality follows since R_i is homogeneous of degree $k-1$. It follows that

$$\begin{aligned} \mathcal{B}_i^R(\mathbf{x}) &= \frac{\bar{x} \int_0^1 R_i(tx_1, ..., tx_n)\, dt}{R(x_1, ..., x_n)} \\ &= \frac{\bar{x} R_i(x_1, ..., x_n)}{k R(x_1, ..., x_n)} \\ &= \frac{\bar{x} R_i(\bar{x}\alpha_1, ..., \bar{x}\alpha_n)}{k R(\bar{x}\alpha_1, ..., \bar{x}\alpha_n)} \\ &= \frac{R_i(\alpha_1, ..., \alpha_n)}{k R(\alpha_1, ..., \alpha_n)} \\ &= \frac{R_i(\alpha_1, ..., \alpha_n)}{\sum_{h=1}^n \alpha_h R_h(\alpha_1, ..., \alpha_n)}, \end{aligned}$$

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 5. ■

Appendix II: Other Proofs and Derivations

Proofs of Propositions

Proof of Proposition 1: We need to show that for any random returns $\tilde{z}_1$ and $\tilde{z}_2$, and any $0 < \lambda < 1$,

$$w_k\left(\lambda\tilde{z}_1 + (1-\lambda)\,\tilde{z}_2\right) \leq \lambda w_k\left(\tilde{z}_1\right) + (1-\lambda)\,w_k\left(\tilde{z}_2\right). \tag{33}$$

Letting $\hat{z}_1 = \tilde{z}_1 - \mathrm{E}\left(\tilde{z}_1\right)$ and $\hat{z}_2 = \tilde{z}_2 - \mathrm{E}\left(\tilde{z}_2\right)$, (33) can be rewritten as

$$\left(\mathrm{E}\left[\left(\lambda\hat{z}_1 + (1-\lambda)\,\hat{z}_2\right)^k\right]\right)^{\frac{1}{k}} \leq \lambda\left(\mathrm{E}\left[\hat{z}_1^k\right]\right)^{\frac{1}{k}} + (1-\lambda)\left(\mathrm{E}\left[\hat{z}_2^k\right]\right)^{\frac{1}{k}}. \tag{34}$$

Applying the binomial formula to the LHS of (34) implies that we need to show

$$\left(\sum_{i=0}^{k}\binom{k}{i}\lambda^{k-i}\left(1-\lambda\right)^i\mathrm{E}\left(\hat{z}_1^{k-i}\hat{z}_2^i\right)\right)^{\frac{1}{k}} \leq \lambda\left(\mathrm{E}\left[\hat{z}_1^k\right]\right)^{\frac{1}{k}} + (1-\lambda)\left(\mathrm{E}\left[\hat{z}_2^k\right]\right)^{\frac{1}{k}}.$$

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

$$\left(\sum_{i=0}^{k}\binom{k}{i}\lambda^{k-i}\left(1-\lambda\right)^i\mathrm{E}\left(\left|\hat{z}_1^{k-i}\hat{z}_2^i\right|\right)\right)^{\frac{1}{k}} \leq \lambda\left(\mathrm{E}\left[\left|\hat{z}_1\right|^k\right]\right)^{\frac{1}{k}} + (1-\lambda)\left(\mathrm{E}\left[\left|\hat{z}_2\right|^k\right]\right)^{\frac{1}{k}}.$$

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

$$\sum_{i=0}^{k}\binom{k}{i}\lambda^{k-i}\left(1-\lambda\right)^i\mathrm{E}\left(\left|\hat{z}_1^{k-i}\hat{z}_2^i\right|\right) \leq \left(\lambda\left(\mathrm{E}\left[\left|\hat{z}_1\right|^k\right]\right)^{\frac{1}{k}} + (1-\lambda)\left(\mathrm{E}\left[\left|\hat{z}_2\right|^k\right]\right)^{\frac{1}{k}}\right)^k.$$

Applying the binomial formula to the RHS implies that it would be sufficient to show

$$\sum_{i=0}^{k}\binom{k}{i}\lambda^{k-i}\left(1-\lambda\right)^i\mathrm{E}\left(\left|\hat{z}_1^{k-i}\hat{z}_2^i\right|\right) \leq \sum_{i=0}^{k}\binom{k}{i}\lambda^{k-i}\left(1-\lambda\right)^i\left(\mathrm{E}\left[\left|\hat{z}_1\right|^k\right]\right)^{\frac{k-i}{k}}\left(\mathrm{E}\left[\left|\hat{z}_2\right|^k\right]\right)^{\frac{i}{k}}.$$

To establish this inequality we will show that it actually holds term by term. That is, it is sufficient to show that for each $i = 0, ..., k$,

$$\mathrm{E}\left(\left|\hat{z}_1^{k-i}\hat{z}_2^i\right|\right) \leq \left(\mathrm{E}\left[\left|\hat{z}_1\right|^k\right]\right)^{\frac{k-i}{k}}\left(\mathrm{E}\left[\left|\hat{z}_2\right|^k\right]\right)^{\frac{i}{k}}.$$

To see this, note that it is equivalent to show that

$$\mathrm{E}\left(\left|\hat{z}_1^{k-i}\hat{z}_2^i\right|\right) \leq \left(\mathrm{E}\left[\left|\hat{z}_1^{k-i}\right|^{\frac{k}{k-i}}\right]\right)^{\frac{k-i}{k}}\left(\mathrm{E}\left[\left|\hat{z}_2^i\right|^{\frac{k}{i}}\right]\right)^{\frac{i}{k}}.$$

But, this is immediate from Hölder's inequality, and we are done. ■

Proof of Proposition 2: For any integer $k \geq 2$, we can rewrite the downside risk measure as

$$\mathrm{DR}_k(\tilde{z}) = (-1)^k \left(\mathrm{E}\left([\tilde{z} - \mathrm{E}(\tilde{z})]^- \right)^k \right)^{\frac{1}{k}} = \left(\mathrm{E}\left([\mathrm{E}(\tilde{z}) - \tilde{z}]^+ \right)^k \right)^{\frac{1}{k}},$$

where $[t]^+ = \max(t, 0)$ for $t \in \mathbb{R}$.

Consider any two random returns $\tilde{z}_1$ and $\tilde{z}_2$, and let $\hat{z}_1 = [\mathrm{E}(\tilde{z}_1) - \tilde{z}_1]^+$ and $\hat{z}_2 = [\mathrm{E}(\tilde{z}_2) - \tilde{z}_2]^+$. Obviously, we have $\hat{z}_1 \geq 0$ and $\hat{z}_2 \geq 0$. What we need to show is that for any $0 < \lambda < 1$,

$$\left(\mathrm{E}\left([\mathrm{E}(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2) - \lambda\tilde{z}_1 - (1-\lambda)\tilde{z}_2]^+ \right)^k \right)^{\frac{1}{k}} \leq \lambda \left(\mathrm{E}\left(\hat{z}_1^k\right) \right)^{\frac{1}{k}} + (1-\lambda) \left(\mathrm{E}\left(\hat{z}_2^k\right) \right)^{\frac{1}{k}}.$$

Now,

$$\begin{aligned} [\mathrm{E}(\lambda\tilde{z}_1 + (1-\lambda)\tilde{z}_2) - \lambda\tilde{z}_1 - (1-\lambda)\tilde{z}_2]^+ &= [\lambda(\mathrm{E}(\tilde{z}_1) - \tilde{z}_1) + (1-\lambda)(\mathrm{E}(\tilde{z}_2) - \tilde{z}_2)]^+ \\ &\leq \lambda[\mathrm{E}(\tilde{z}_1) - \tilde{z}_1]^+ + (1-\lambda)[\mathrm{E}(\tilde{z}_2) - \tilde{z}_2]^+ \\ &= \lambda\hat{z}_1 + (1-\lambda)\hat{z}_2, \end{aligned}$$

where the inequality follows from Jensen's inequality using that $[\cdot]^+$ is a convex function.

Therefore, it is sufficient to show that

$$\left(\mathrm{E}(\lambda\hat{z}_1 + (1-\lambda)\hat{z}_2)^k \right)^{\frac{1}{k}} \leq \lambda \left(\mathrm{E}\left(\hat{z}_1^k\right) \right)^{\frac{1}{k}} + (1-\lambda) \left(\mathrm{E}\left(\hat{z}_2^k\right) \right)^{\frac{1}{k}}.$$

The rest of the proof follows closely the proof of Proposition 1. Indeed, since $\hat{z}_1$ and $\hat{z}_2$ are non-negative here, the arguments in the proof of Proposition 1 apply in this case to any positive k (odd or even). ■

Proof of Proposition 3: 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 Ω and with the underlying probability measure $P(\cdot)$. We first prove that $\mathrm{ES}_\delta(\tilde{z})$ is subadditive. That is, for any two random returns $\tilde{z}_1$ and $\tilde{z}_2$,

$$\mathrm{ES}_\delta(\tilde{z}_1 + \tilde{z}_2) \leq \mathrm{ES}_\delta(\tilde{z}_1) + \mathrm{ES}_\delta(\tilde{z}_2). \tag{35}$$

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 (3), for any random return $\tilde{z}$ (which is not constant), $\mathrm{ES}_\delta(\tilde{z})$ can be expressed as

$$\mathrm{ES}_\delta(\tilde{z}) = -\frac{1}{\delta}\int_{\{\omega:\tilde{z}\leq -\mathrm{VaR}_\delta(\tilde{z})\}} \tilde{z}dP(\omega).$$

Let $\tilde{z}_1$ and $\tilde{z}_2$ be random returns and define $\tilde{z}_3 = \tilde{z}_1 + \tilde{z}_2$. Let

$$\Omega_i = \{\omega \in \Omega : \tilde{z}_i \leq -\mathrm{VaR}_\delta(\tilde{z}_i)\},$$

for $i = 1, 2, 3$. Then, (35) 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

$$\int_{\Omega_3} \tilde{z}_1 dP(\omega) \geq \int_{\Omega_1} \tilde{z}_1 dP(\omega), \tag{36}$$

and

$$\int_{\Omega_3} \tilde{z}_2 dP(\omega) \geq \int_{\Omega_2} \tilde{z}_2 dP(\omega). \tag{37}$$

For brevity, we will only prove (36). The proof of (37) is parallel.

Define

$$\begin{aligned}
\Omega_4 &= \{\omega \in \Omega : \tilde{z}_1 \leq -\mathrm{VaR}_\delta(\tilde{z}_1), \tilde{z}_3 \leq -\mathrm{VaR}_\delta(\tilde{z}_3)\}, \\
\Omega_5 &= \{\omega \in \Omega : \tilde{z}_1 \leq -\mathrm{VaR}_\delta(\tilde{z}_1), \tilde{z}_3 > -\mathrm{VaR}_\delta(\tilde{z}_3)\}, \text{ and} \\
\Omega_6 &= \{\omega \in \Omega : \tilde{z}_1 > -\mathrm{VaR}_\delta(\tilde{z}_1), \tilde{z}_3 \leq -\mathrm{VaR}_\delta(\tilde{z}_3)\}.
\end{aligned}$$

Clearly, $\Omega_4 \cap \Omega_5 = \varnothing$, $\Omega_4 \cup \Omega_5 = \Omega_1$, $\Omega_4 \cap \Omega_6 = \varnothing$, 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

$$\int_{\Omega_5} dP(\omega) = \int_{\Omega_6} dP(\omega). \tag{38}$$

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,

$$\begin{aligned}
& \int_{\Omega_1} \tilde{z}_1 dP(\omega) - \int_{\Omega_3} \tilde{z}_1 dP(\omega) \\
= & \int_{\Omega_5} \tilde{z}_1 dP(\omega) - \int_{\Omega_6} \tilde{z}_1 dP(\omega) \\
\leq & \int_{\Omega_5} \left[-\mathrm{VaR}_\delta(\tilde{z}_1)\right] dP(\omega) - \int_{\Omega_6} \left[-\mathrm{VaR}_\delta(\tilde{z}_1)\right] dP(\omega) \\
= & -\mathrm{VaR}_\delta(\tilde{z}_1) \left[\int_{\Omega_5} dP(\omega) - \int_{\Omega_6} dP(\omega)\right] \\
= & 0,
\end{aligned}$$

where the inequality follows from $\tilde{z}_1 \leq -\mathrm{VaR}_\delta(\tilde{z}_1)$ when $\omega \in \Omega_5$ and $\tilde{z}_1 > -\mathrm{VaR}_\delta(\tilde{z}_1)$ when $\omega \in \Omega_6$, and where the last equality follows from (38). Therefore, (36) is obtained, and hence $\mathrm{ES}_\delta(\tilde{z})$ is subadditive. Since $\mathrm{DES}_\delta(\tilde{z}) = \mathrm{ES}_\delta(\tilde{z}) + \mathrm{E}(\tilde{z})$ we have that DES is also subadditive.

Convexity now follows immediately from homogeneity of degree 1 and subadditivity. ■

Derivations of Systematic Risk for Applications I–V

Here we provide derivations of the systematic risk associated with different risk measures discussed in Section 1.4.3.

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

$$R\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right) = m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right) = \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^k.$$

Differentiating with respect to the weight of asset i yields

$$\begin{aligned}\frac{\partial m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M} &= k\mathrm{E}\left[\left(\tilde{z}_i - \mathrm{E}\left(\tilde{z}_i\right)\right)\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^{k-1}\right] \\ &= k\mathrm{Cov}\left(\tilde{z}_i, \left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \boldsymbol{\alpha}^M \cdot \mathrm{E}\left(\widetilde{\mathbf{z}}\right)\right)^{k-1}\right).\end{aligned}$$

By Theorem 4, and since $m_k(\cdot)$ is homogeneous of degree k, the systematic risk is then given by

$$\mathcal{B}_i^R = \frac{\frac{\partial m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M}}{k m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)} = \frac{\mathrm{Cov}\left(\tilde{z}_i, \left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \boldsymbol{\alpha}^M \cdot \mathrm{E}\left(\widetilde{\mathbf{z}}\right)\right)^{k-1}\right)}{m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}, \tag{39}$$

as required.

Now suppose alternatively that $R(\tilde{z}) = w_k(\tilde{z})$. The market portfolio risk is

$$R\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right) = w_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right) = \left(m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^{\frac{1}{k}}.$$

Differentiating with respect to the weight of asset i gives

$$\frac{\partial w_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M} = \frac{1}{k}\left(m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^{\frac{1}{k}-1}\frac{\partial m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M}.$$

By Theorem 4, and since $w_k(\cdot)$ is homogeneous of degree 1, the systematic risk is

$$\mathcal{B}_i^R = \frac{\frac{1}{k}\left(m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^{\frac{1}{k}-1}\frac{\partial m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M}}{\left(m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^{\frac{1}{k}}} = \frac{\frac{\partial m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M}}{k m_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)},$$

which is identical to (39).

Application III: Assume $R(\tilde{z}) = \mathrm{DR}_k(\tilde{z})$ for $k \geq 2$. The risk of the market portfolio $\boldsymbol{\alpha}^M$ is given by

$$\mathrm{DR}_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right) = (-1)^k\left(\mathrm{E}\left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^k\right)^{\frac{1}{k}}.$$

Differentiating with respect to α_i^M gives[21]

$$\begin{aligned}
&\frac{\partial \mathrm{DR}_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M} \\
= \;& (-1)^k \left(\mathrm{E}\left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^k\right)^{\frac{1}{k}-1} \\
& * \mathrm{E}\left[\left(\widetilde{z}_i - \mathrm{E}\left(\widetilde{z}_i\right)\right)\left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^{k-1}\right] \\
= \;& (-1)^k \left(\mathrm{E}\left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^k\right)^{\frac{1}{k}-1} \\
& * \mathrm{Cov}\left[\widetilde{z}_i, \left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^{k-1}\right].
\end{aligned}$$

By Theorem 4, and since $\mathrm{DR}_k(\cdot)$ is homogeneous of degree 1, the systematic risk is given by

$$\begin{aligned}
\mathcal{B}_i^R &= \frac{\frac{\partial \mathrm{DR}_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\partial \alpha_i^M}}{\mathrm{DR}_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)} \\
&= \frac{\mathrm{Cov}\left[\widetilde{z}_i, \left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^{k-1}\right]}{\mathrm{E}\left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^k} \\
&= (-1)^k \frac{\mathrm{Cov}\left[\widetilde{z}_i, \left(\left[\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^{k-1}\right]}{\left(\mathrm{DR}_k\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right)^k}.
\end{aligned}$$

Application IV: Assume $R(\tilde{z}) = \mathrm{DES}_\delta(\tilde{z})$ for some confidence level $0 < \delta < 1$. Let $f(z_1, ..., z_n)$ denote the joint density function of $\widetilde{\mathbf{z}}$. 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

[21] Note that we are essentially relying here on Leibniz's rule for differentiation under the integral. While $\left(\left[\boldsymbol{\alpha}^M \cdot \mathbf{z} - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right]^-\right)^k$ is not everywhere differentiable, it is continuous and differentiable almost everywhere. This guarantees that Leibniz's rule applies.

market portfolio $\boldsymbol{\alpha}^M$ can be written as follows

$$\begin{aligned}
& \mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) \\
= \; & \mathrm{ES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) + \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right) \\
= \; & -\frac{1}{\delta}\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{-\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) - \sum_{j=2}^n \alpha_j^M z_j}{\alpha_1^M}} \\
& \left(\sum_{j=1}^n \alpha_j^M z_j - \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)\right) f(z_1, ..., z_n) dz_1 ... dz_n.
\end{aligned}$$

Differentiating $\mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})$ using Leibniz's rule with respect to α_i^M yields

$$\begin{aligned}
& \frac{\partial \mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})}{\partial \alpha_i^M} \qquad (40) \\
= \; & -\frac{1}{\delta}\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{-\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) - \sum_{j=2}^n \alpha_j^M z_j}{\alpha_1^M}} (z_i - \mathrm{E}(\tilde{z}_i)) f(z_1, ..., z_n) dz_1 ... dz_n \\
& + \frac{\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) + \mathrm{E}\left(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}\right)}{\delta} \frac{\partial}{\partial \alpha_i^M} \\
& * \left(\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{-\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) - \sum_{j=2}^n \alpha_j^M z_j}{\alpha_1^M}} f(z_1, ..., z_n) dz_1 ... dz_n\right).
\end{aligned}$$

Notice that by the definition of $\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})$,

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{-\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) - \sum_{j=2}^n \alpha_j^M z_j}{\alpha_1^M}} f(z_1, ..., z_n) dz_1 ... dz_n = \delta,$$

which is a constant, implying that the second term in (40) is zero. Thus,

$$\begin{aligned}
\frac{\partial \mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})}{\partial \alpha_i^M} &= -\frac{1}{\delta}\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{-\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}}) - \sum_{j=2}^n \alpha_j^M z_j}{\alpha_1^M}} \\
& \quad (z_i - \mathrm{E}(\tilde{z}_i)) f(z_1, ..., z_n) dz_1 ... dz_n \\
&= -\frac{1}{\delta}\mathrm{E}\left[1_{\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} \leq -\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})} (z_i - \mathrm{E}(\tilde{z}_i))\right] \\
&= -\mathrm{E}\left[\tilde{z}_i - \mathrm{E}(\tilde{z}_i) \,|\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} \leq -\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})\right].
\end{aligned}$$

By Theorem 4, and since $\mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})$ is homogeneous of degree 1, the systematic risk is given by

$$\mathcal{B}_i^R = \frac{\frac{\partial \mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})}{\partial \alpha_i^M}}{\mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})} = -\frac{\mathrm{E}\left[\tilde{z}_i - \mathrm{E}(\tilde{z}_i) \,|\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}} \leq -\mathrm{VaR}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})\right]}{\mathrm{DES}_\delta(\boldsymbol{\alpha}^M \cdot \widetilde{\mathbf{z}})}.$$

Application V: Consider the following family of risk measures

$$R\left(\tilde{z}\right)=\theta_1 w_2\left(\tilde{z}\right)+\theta_2 \mathrm{DR}_3\left(\tilde{z}\right)+\theta_3 w_4\left(\tilde{z}\right)+\theta_4 \mathrm{DES}_\delta\left(\tilde{z}\right)$$

for some confidence level δ and non-negative weights $\theta_1,...,\theta_4$. From Lemma 1, this family of risk measures satisfies all of the conditions in Theorem 4. Moreover, it is easy to verify that when

$$R\left(\tilde{z}\right)=\sum_{j=1}^{s} R^j\left(\tilde{z}\right),$$

the expression for $\mathcal{B}_i^R$ given in (10) implies

$$\mathcal{B}_i^R=\sum_{j=1}^{s}\frac{R^j\left(\boldsymbol{\alpha}^M\cdot\widetilde{\mathbf{z}}\right)}{R\left(\boldsymbol{\alpha}^M\cdot\widetilde{\mathbf{z}}\right)}\mathcal{B}_i^{R^j}.$$

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 1.5.1 for further discussion of this issue as it relates to Axiom 2 in that section.)

Internet Appendix

Mean-Risk Preferences and Expected Utility

Background One would wonder how the mean-risk preferences considered in Section 1.4 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 (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 (6) can be written as

$$U(\boldsymbol{\zeta}) = V\left(\mathrm{E}\left(F_{\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}}\right), R\left(F_{\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}}\right)\right),$$

where $F_{\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}}$ is the cumulative distribution of the random variable $\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}$. When the

random variable of interest is clear, we will omit it from the notation and write the utility as $U(F) = V(\mathrm{E}(F), R(F))$.

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,[22] then for any two distributions F_1, F_2 on I there exists $u(\cdot; F_1)$ differentiable almost everywhere on I such that

$$U(F_2) - U(F_1) = \int_I u(y; F_1)\, dF_2(y) - \int_I u(y; F_1)\, dF_1(y) + o(\|F_2 - F_1\|), \quad (41)$$

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.

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 $m_k^F = \int \left(y - \mu_1^F\right)^k dF(y)$ 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 $R\left(\mu_{j_1}^F, ..., \mu_{j_l}^F, m_{k_1}^F, ..., m_{k_n}^F\right)$. We assume that R is differentiable in all arguments. The utility function in (6) then takes the form

$$U(F) = V\left(\mu_1^F, R\left(\mu_{j_1}^F, ..., \mu_{j_l}^F, m_{k_1}^F, ..., m_{k_n}^F\right)\right), \quad (42)$$

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 4 *If $U(F)$ takes the form (42) then for any two distributions F_1, F_2 on I there exists $u(\cdot; F_1)$ differentiable almost everywhere on I such that (41) holds.*

[22] 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).

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 μ_k^F and m_k^F are Fréchet differentiable for any k, then so is $U(\cdot)$. The Fréchet differentiability of μ_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

$$\begin{aligned} m_k^F &= \int \left(y - \mu_1^F\right)^k dF(y) \\ &= \int \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} y^i \left(\mu_1^F\right)^{k-i} dF(y) \\ &= \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \left(\mu_1^F\right)^{k-i} \int y^i dF(y) \\ &= \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} \left(\mu_1^F\right)^{k-i} \mu_i^F, \end{aligned}$$

which is a differentiable function of the μ_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 $\boldsymbol{\zeta} \in \mathbb{R}^{n+1}$ be a bundle. Denote the gradient of investor j's utility function at $\boldsymbol{\zeta}$ by $\nabla \mathbf{U}^j(\boldsymbol{\zeta}) = (U_0^j(\boldsymbol{\zeta}), ..., U_n^j(\boldsymbol{\zeta}))$, where a subscript designates a partial derivative in the direction of the i^{th} asset. Also, let $\gamma^j(\boldsymbol{\zeta}) = -\frac{V_2^j(\mathrm{E}(\boldsymbol{\zeta}\cdot\tilde{\mathbf{y}}), R(\boldsymbol{\zeta}\cdot\tilde{\mathbf{y}}))}{V_1^j(\mathrm{E}(\boldsymbol{\zeta}\cdot\tilde{\mathbf{y}}), R(\boldsymbol{\zeta}\cdot\tilde{\mathbf{y}}))} > 0$ 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 $\gamma^j(\boldsymbol{\zeta}) = -\frac{V_2^j}{V_1^j}$.

Proposition 5 *Assume that for each asset i there is some investor j such that $\mathrm{E}(\tilde{y}_i) > \gamma^j(\boldsymbol{\zeta}) R_i(\boldsymbol{\zeta}\cdot\tilde{\mathbf{y}})$ for all $\boldsymbol{\zeta}$. 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 $U_i^j(\boldsymbol{\zeta}) > 0$ for all $\boldsymbol{\zeta}$. Recall that

$$U^j(\boldsymbol{\zeta}) = V^j\left(\mathrm{E}\left(\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}\right), R\left(\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}\right)\right).$$

Thus,

$$\begin{aligned} U_i^j(\boldsymbol{\zeta}) &= V_1^j \mathrm{E}(\tilde{y}_i) + V_2^j R_i\left(\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}\right) \\ &= V_1^j[\mathrm{E}(\tilde{y}_i) - \gamma^j(\boldsymbol{\zeta}) R_i\left(\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}\right)], \end{aligned}$$

where $R_i\left(\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}\right)$ denotes the partial derivative of $R\left(\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}\right)$ with respect to ζ_i.

Since $V_1^j > 0$, $U_i^j(\boldsymbol{\zeta}) > 0$ corresponds to

$$\mathrm{E}(\tilde{y}_i) - \gamma^j(\boldsymbol{\zeta}) R_i\left(\boldsymbol{\zeta}\cdot\widetilde{\mathbf{y}}\right) > 0,$$

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.

2 Signaling with Dynamic Payoffs and Entrepreneurial Compensation

2.1 Introduction

Entrepreneurial compensation is a topic that has drawn academic interest but not been much studied. A number of stylized facts were discovered while unexplained. Indeed, as Wasserman (2004) point out, there is tremendous cross-sectional variation in entrepreneurial compensation, particularly in terms of the relative scale between equity-based and cash-based pay, although explanation for such variation is yet to be provided.

In addition, an emerging area of entrepreneurial study is the "Search Fund" model (Stanford Graduate School of Business, 2012), where a group of investors financially back a searcher-entrepreneur, typically straightly out of a MBA program, to search for an existing small business to invest, operate, and grow. During this search stage, the entrepreneur draws a salary from the capital provided by the investors as well as shares an equity stake in the target business. Historically, targets have resided in a wide range of industries, from burger diners, to nursing homes, and to software development shops. Interestingly, this cash-versus-equity-pay also differs substantially across these industries. A natural question is therefore how to explain this cross-sectional variation.

One may be tempted to borrow the results from the executive compensation literature. In particular, the seminal work of Jensen and Meckling (1976) and Holmstrom (1979) provided an agency-theoretic perspective, addressing the moral hazard problem as to why it is important to provide incentive for effort exertion. Prendergast (1999, 2002) further adds to the discussion by pointing out the trade-off between risk and incentives. However, these insights do not seem to translate directly to entrepreneurial compensation, and the reason is small business uniqueness. Indeed, as Ang (1991) argues, the issue of misaligned incentive is less serious in small firms than in large firms because of the high concentration, both financially and operationally, that the entrepreneur is involved with his or her firm. Moreover, the issue of risk aversion is essentially inapplicable since entrepreneurs and corporate executives intrinsically have distinct risk preferences, where the former are much less risk-averse (if they are not even risk-seeking) than the latter. Indeed, a number of studies have supported such risk preferences, such as Kanbur (1979), Blanchflower and Oswald (1998), and

Cramer et. al (2002). A more serious issue is asymmetric information, as Ang (1991) also suggests, where entrepreneurs know significantly more than investors and other sorts of monitoring such as analyst coverage are absent.

Therefore, this paper provides an asymmetric-information perspective to study entrepreneurial compensation. Essentially, as anecdotal evidence suggests,[23] uninformed investors tend to rely on compensation structure as a signal to distinguish among different types of informed entrepreneurs. Furthermore, an interesting trade-off between current and future payoffs is modeled, where it manifests itself in the balance of cash-based pay (current) versus equity-based pay (deferred). Intuitively, a "high-type" entrepreneur, *ceteris paribus*, is likely more willing to accept equity pay due to better ability in growing the firm and consequently a larger pie to share from in the future.

The signal literature originates from Akerlof's (1970) seminal work, which suggests that in a market with asymmetric information, if the informed party cannot communicate the private information to the uninformed party, then a breakdown can occur due to adverse selection. Since then, various signaling models have been proposed. For example, Spence (1974) and subsequently Rothschild and Stiglitz (1976), Leland and Pyle (1977), Riley (1979), Wilson (1980), and Ofer and Thakor (1987), among others, address this issue by showing that a fully revealing signaling equilibrium can exist, where the uninformed know in equilibrium the true types of the informed. However, the signal is dissipative in that there has to be deadweight loss relative to the "first-best" situation in order for the signal to be effective. The reason is that, by their very being, the low types cause an externality such that the high types are worse off than they would be in the absence of the low types while the low types are no better off than they would be in the absence of the high types. In contrast, Bhattacharya (1980), Heinkel (1982), Brennan and Kraus (1987), and Franke (1987) obtain nondissipative revealing signaling equilibrium. Nevertheless, they impose exogenous conditions on the behavior of the uninformed, such as being able to penalize the informed ex post.

The signaling model in this paper, while built on existing studies, has some interesting features. In particular, a market breakdown will not happen, but both separating and pooling equilibria are possible, and consequently, an equilibrium is not necessarily informationally consistent (i.e., fully revealing). Furthermore, when

[23] Venture Capital Panel Discussion, Skandalaris Center for Entrepreneurial Studies, 2009.

an equilibrium is indeed revealing, whether the signal is dissipative or nondissipative is completely endogenous.

Essentially, it is this dynamic payoff structure embedding the trade-off between present pay and future pay that results in these features. If we are concerned with only one single payoff, such as that in Spence (1974), then a market will break down unless the informed can signal the private information. In addition, the signal needs to be costly to be effective.[24] However, if two payoffs (trading off each other, such as today versus tomorrow) are in the picture, then there may exist self-selections where one type prefers one payoff while the other type prefers the other, as indicated in Salop and Salop (1976). Then the choice of the payoff structure of the informed serves as a natural signal, which is therefore nondissipative, i.e., mimicking the good type may simply be outside the bad type's optimal course of actions. Nevertheless, I do not impose such preferences directly; rather, the case with nondissipative signal arises endogenously depending on the extent of this trade-off.

I employ a continuous-time framework where the informed agent can affect the capital stock process associated with an entrepreneurial project through only its drift. The capital accumulation at each point in time is thus uninformative about the informed agent's private information. The agent chooses a compensation policy which sets her salary rate as well as shapes the course of reinvestment into the project. This reinvestment in turn affects the drift of capital accumulation. The agent's private information relates to an entrepreneurial technology which determines its obsolescence rate. Put differently, this private information determines how relevant the agent is to the project over time. Indeed, this rate impacts the valuation of the project when the entrepreneur retires, which determines her terminal payoff. The agent therefore is explicitly concerned with this intertemporal trade-off between current pay and deferred pay.

The game-theoretic feature of the model manifests itself in a signaling game where the uninformed agent moves first by precommitting to a valuation for each compensation policy. This is appropriate in the entrepreneurial setting given that the two parties sign a contract upon the initiation of the entrepreneur's engagement, which states the salary rates as well as the terminal valuation of the technology. In other words, the uninformed investors offer a menu of salary-valuation contracts, and then the informed entrepreneur chooses among these contracts to signal her type. As for

[24] Unless the uninformed can somehow penalize dishonesty upon seeing the result ex post.

equilibrium concepts, the study primarily employs the Riley Reactive equilibrium (Riley (1979)).

The rest of the paper is organized as follows. Section 2.2 describes the formulation of the model and solves for the case with symmetric information. Section 2.3 defines an equilibrium with asymmetric information and details the equilibrium analysis. Model extensions and empirical predictions of the analysis are in Section 2.4. Section 2.5 concludes.

2.2 Model Formulation

2.2.1 Agents and Technological Possibilities

The economy carries a continuum of investors. There are also some entrepreneurs, each having a project that requires an initial capital of $K_0 > 0$. Without the entrepreneur, the project is valueless to the investors. The interpretation is that it is a piece of knowledge or innovation that is potentially commercializable. So although it is valuable, it does not generate revenue without the know-hows of an entrepreneur.

There are two types of entrepreneurs, H (for high) and L (for low). After the project is initiated, it generates cash flows at any time t via a constant rate $h > 0$ that is common to both types.[25] The two types differ, however, by an entrepreneurial technology that determines the obsolescence rate $\delta^i > 0$ of the project's capital, where $i \in \{H, L\}$ and $\delta^H < \delta^L \leq h$.[26] This information on types is private with the entrepreneur, although investors know the values of δ^H and δ^L as well as the proportion of the high type, denoted by $\xi \in (0, 1)$.

We can interpret this obsolescence rate in primarily two ways. First, in terms of intangible assets such as patents and goodwill, a better technology leads to better management of them and thus adds to the livelihood of the project. Second, in terms of operation strategies, a better entrepreneur can operate the project more efficiently in such a way it is not easily replicated.

For the cash flows generated at each point in time, the entrepreneur has a decision to make. She can either compensate herself or reinvest into the project or both. Let γ_t denote her salary rate at time t. Then the project's capital stock evolves according

[25] Since only with the entrepreneur's evolvement are these cash flows possible and the investors are perfectly competitive, any surplus goes to the entrepreneur.

[26] If $\delta^i > h$, then it is not a viable technology and the entrepreneur simply should not undertake the project to begin with.

to

$$dK_t = (I_t - \delta^i K_t)dt + \sigma K_t dZ_t, \tag{43}$$

where $I_t = (h- \gamma_t)K_t$ is (re)investment amount, $i \in \{H, L\}$, σ is a volatility parameter, and Z_t is a standard Brownian motion. Furthermore, I assume that there is no debt market, so the entrepreneur cannot borrow to reinvest, and thus, $0 \leq \gamma_t \leq h$.

Substituting I_t into (43), we have

$$dK_t = (h - \gamma_t - \delta^i)K_t dt + \sigma K_t dZ_t. \tag{44}$$

The capital accumulation specification (44) is similar to that in Cox, Ingersoll, and Ross (1985) and Sundaresan (1984), where uncertainty of capital accumulation is proportional to the level of capital stock. It is in contrast with that in Albuquerue and Wang (2008), which is a continuous-time version of Greenwood, Hercowitz, and Huffman (1988) and Greenwood, Hercowitz, and Krusell (1997, 2000), where volatility depends on I instead of K. But then the entrepreneur has the option to avoid uncertainty by always investing zero amount. My specification reflects the fact that the entrepreneur cannot (perfectly) control the volatility of capital accumulation. Consequently, even though the investors can observe K_t at each time point t, they cannot infer δ^i based on this information.

An entrepreneur cannot develop the project forever. I assume that with a Poisson arrival rate, λ, her involvement is terminated. We can think of this as her unforeseeable retirement need due to circumstances like health condition. Also, one simple way to illustrate this is that a fixed termination date cannot be written into the contract. In any case, the essence of this assumption is that the termination is not motivated by the entrepreneur's private knowledge of the quality of her entrepreneurial technology. Then, since there is a continuum of investors, when the time for termination comes, the investors are only concerned with breaking even,[27] in which case the terminal value of the project is paid to the entrepreneur. This "fair" terminal value of the project takes into account both the capital stock at the time of termination and the project's growth potential. The latter depends on the project's technology, through the obsolescence rate δ^i, as the lower this rate is, the higher the effective growth rate of the project.

[27] Alternatively, we can assume that the investors and the entrepreneurs engage in a Nash bargainning game. The results will clearly be more complicated although qualitatively similar.

Moreover, in line with the assumption that the investors cannot develop the project unless they hire an entrepreneur, I assume that only the entrepreneur possesses the specific know-hows on how to reinvest in the project. So any reinvestment is productive only if the entrepreneur is operating the project. Thus, when the entrepreneur retires, the investors will not reinvest. Effectively, investors are valuing the project for its future stream of cash flows paid out for example in the form of dividends.[28] The terminal value of the project can then be calculated using the Gordon model (Gordon (1959)). Formally, let τ denote the time of the termination. The capital stock of the project after this termination will evolve as

$$dK_t = (-\delta^i)K_t dt + \sigma K_t dZ_t, \tag{45}$$

for $t \geq \tau$. The expected value of the stream of future cash flows at τ is

$$E_\tau[\int_{t=\tau}^{\infty} hK_t dt], \tag{46}$$

which can be solved in closed-form, as in the following lemma. Intuitively, the higher is the rate of cash flows, and the lower is the obsolescence rate, the higher the terminal value is.

Lemma 3 *The terminal value of the project at time τ is $\frac{h}{\delta^i}K_\tau$.*

Proof. See Appendix. ■

2.2.2 The Entrepreneur's Problem

Let u_1 and u_2 denote the entrepreneur's subutility functions for salary streams and proceeds from the terminal value. Then given the initial capital $K_0 > 0$, the entrepreneur's problem is to choose adapted salary rates $\{\gamma_t\}$ to

$$\begin{aligned} &\max E[\int_{t=0}^{\tau} e^{-\rho t}u_1(\gamma_t K_t)dt + e^{-\rho\tau}u_2(\frac{h}{\delta^i}K_\tau)] \\ &s.t. \\ &dK_t = (h - \gamma_t - \delta^i)K_t dt + \sigma K_t dZ_t \\ &0 \leq \gamma_t \leq h \\ &K_t \geq 0, \end{aligned} \tag{47}$$

[28] Of course, the benefits associated with some projects may be non-pecuniary, so we assume a monetary value for each of them.

where ρ is the natural rate of discount. As noted above, the terminal value recognizes the impact of the entrepreneurial technology. *Ceteris paribus*, in particular for the same amount of capital accumulation, a high type merits a higher terminal payoff than that of a low type. The last constraint, i.e., the nonnegativity of capital stock, reflects the fact that the entrepreneur's problem is null if the project is no longer viable.

Moreover, since termination occurs with Poisson arrival rate $\lambda > 0$, we have $\Pr(\tau \leq s) = 1 - e^{-\lambda s}$, which implies that (47) can be written equivalently as

$$E[\int_{s=0}^{\infty} e^{-\rho s - \lambda s} u_1(\gamma_s K_s) ds + \int_{s=0}^{\infty} \lambda e^{-\rho s - \lambda s} u_2(\frac{h}{\delta^i} K_s) ds],$$

which upon simplifying becomes

$$E\{\int_{s=0}^{\infty} e^{-\rho s - \lambda s} [u_1(\gamma_s K_s) + \lambda u_2(\frac{h}{\delta^i} K_s)] ds\}. \tag{48}$$

Assume further that the entrepreneur possesses linear utility over nonnegative wealth,[29] that is, $u_j(c) = c$, for $c \in [0, \infty)$, $j \in \{1, 2\}$. Then, (48) becomes

$$E[\int_{s=0}^{\infty} e^{-\rho s - \lambda s} (\gamma_s K_s + \lambda \frac{h}{\delta^i} K_s) ds].$$

2.2.3 Optimal Strategies with Symmetric Information

Let us first solve the entrepreneur's problem assuming that there is symmetric information between the entrepreneurs and the investors. We can specify the entrepreneur's problem above as Problem 1.

Problem 1. Given $K_0 > 0$, choose $\{\gamma_t\}$ to

$$\begin{aligned}
&\max E[\int_{s=0}^{\infty} e^{-\rho s - \lambda s} (\gamma_s K_s + \lambda \frac{h}{\delta^i} K_s) ds] \\
&s.t. \\
&dK_t = (h - \gamma_t - \delta^i) K_t dt + \sigma K_t dZ_t \\
&0 \leq \gamma_t \leq h \\
&K_t \geq 0.
\end{aligned}$$

Note that we can apply homogeneity and thus alternatively characterize this problem as Problem 2.

[29] This is similar to risk neutrality but not exactly the same, since risk neutrality requires linear utility over the entire real line. But the essence of the linear utility here is to rid risk aversion.

Problem 2. Given $K_0 > 0$, choose $\{\gamma_t\}$ to

$$\begin{aligned}
&\max K_0 E[\int_{s=0}^{\infty} e^{-\rho s-\lambda s}(\gamma_s \frac{K_s}{K_0} + \lambda \frac{h}{\delta^i}\frac{K_s}{K_0})ds] \\
&s.t. \\
&d(\frac{K_t}{K_0}) = (h - \gamma_t - \delta^i)\frac{K_t}{K_0}dt + \sigma \frac{K_t}{K_0}dZ_t \\
&0 \leq \gamma_t \leq h \\
&\frac{K_t}{K_0} \geq 0.
\end{aligned}$$

Thus, given that K is the only state variable and letting $V(K)$ denote the value function, we have $V(K_0) = K_0 V(1)$. That is, the value function, as a function of K_0, is equal to K_0 multiplied by the value function evaluated at $K_0 = 1$. Thus, since the problem is stationary, the value function takes the form

$$V(K) = CK \tag{49}$$

for some constant C. That is, the value function is linear in K.

We then obtain the closed-form solution to the entrepreneur's problem, summarized in the following proposition.

Proposition 6 *If* $\delta^i \leq \frac{\sqrt{(\rho+\lambda-h)^2+4\lambda h}-(\rho+\lambda-h)}{2}$, *then* $C = \frac{\lambda\frac{h}{\delta^i}}{\rho+\lambda+\delta^i-h}$, *i.e.,* $V(K) = \frac{\lambda\frac{h}{\delta^i}}{\rho+\lambda+\delta^i-h}K$. *In this case, the optimal strategy is* $\gamma_t^* = 0$, $\forall t$. *If* $\delta^i > \frac{\sqrt{(\rho+\lambda-h)^2+4\lambda h}-(\rho+\lambda-h)}{2}$, *then* $C = \frac{\lambda\frac{h}{\delta^i}+h}{\rho+\lambda+\delta^i}$, *i.e.,* $V(K) = \frac{\lambda\frac{h}{\delta^i}+h}{\rho+\lambda+\delta^i}K$. *In this case, the optimal strategy is* $\gamma_t^* = h$, $\forall t$.

Proof. See Appendix. ■

Intuitively, under symmetric information, when the entrepreneurial technology is highly valuable so that it becomes obsolete at a rather slow rate, it effectively scales up the terminal payoff as the investors value the promising growth prospects. This incentivizes the entrepreneur to accumulate capital stock, and as a result she chooses to compensate herself nothing until termination. However, when the technology is not very valuable, it becomes obsolete rather quickly, and the benefit of accumulating capital stock to scale up terminal payoff is limited. Yet doing so is costly, since she has to give up compensation which after all carries time value. Therefore, when the cost dominates the benefit, she chooses to consume all cash flows at each point in

time. As it is standard in the literature, let us call these strategies the *first best strategies*, denoted by γ^i_{FB}, for $i \in \{H, L\}$. In addition, we call this threshold the *first best threshold* and denote it by δ^*, that is,

$$\delta^* = \frac{\sqrt{(\rho + \lambda - h)^2 + 4\lambda h} - (\rho + \lambda - h)}{2}.$$

2.3 Equilibrium with Asymmetric Information

Let us now turn to the case where there is informational asymmetry between the entrepreneurs and the investors. Due to this asymmetric information, the investors are concerned with being "fooled" by the entrepreneurs. Moreover, since the investors cannot infer the value of the entrepreneur's technology based on the capital stock, the only available signal is her compensation structure. Thus, at the initial undertaking of the project, the two parties sign a contract stating the salary rate as well as the terminal valuation of the technology. From now on, therefore, we consider a signaling model between the entrepreneurs and the investors.

The literature has witnessed extensive studies on signaling models. Indeed, ever since Akerlof's (1970) suggestion that in a market with asymmetric information a breakdown can occur due to adverse selection if the informed party cannot communicate the private information to the uninformed party, various signaling models have been proposed. For example, Spence (1974) and subsequently Rothschild and Stiglitz (1976), Riley (1979), Wilson (1980), and Ofer and Thakor (1987), among others, address this issue by showing that a fully revealing signaling equilibrium can exist, where the uninformed know in equilibrium the true types of the informed. However, the signal is dissipative in that there has to be deadweight loss relative to the first best in order for the signal to be effective. In contrast, Bhattacharya (1980), Heinkel (1982), Brennan and Kraus (1987), and Franke (1987) obtain nondissipative revealing signaling equilibrium, although they impose exogenous conditions on the behavior of the uninformed, such as being able to penalize the informed ex post.

Intuitively, if the informed is concerned with only one payoff, then a market will break down unless the informed can signal the private information. In addition, the signal will be costly unless the uninformed can somehow penalize dishonesty upon seeing the result ex post. However, if there are dynamic payoffs, as it is the case in this study, then the market may not break down since in equilibrium either information will be revealed or it is the best course of actions for both types to behave similarly.

Furthermore, there can exist self-selections that naturally signal and separate types, making the signal nondissipative, since it may simply be not worthwhile for one to mimic the other's strategy given the trade-off between current and future payoffs.[30]

Indeed, in what follows, I will show that the current paper resolves all these issues. In particular, a market breakdown will not happen, even though the equilibrium is not necessarily revealing. Furthermore, if a signaling equilibrium is indeed revealing, whether it is dissipative is completely endogenous.

2.3.1 Definition of Equilibrium

The literature, broadly speaking, has seen two branches of signaling models. The first involves the uninformed party moving first, such as Spence (1974), Rothschild and Stiglitz (1976), Riley (1979), and Ofer and Thakor (1987). The second relates to the informed party moving first, such as Grossman and Perry (1986), Banks and Sobel (1987), Cho (1987), and Cho and Kreps (1987). As hinted above, my model belongs to the first branch, as I assume that the uninformed investors offer the informed entrepreneur a contract at the initiation of her project. In fact, an explicit contract is not necessary; all that matters is that the investors can precommit to a certain valuation of the project's technology for each particular pattern of the entrepreneur's salary policy. One particular benefit of this modeling choice is that it facilitates the equilibrium analysis by excluding the need for off-equilibrium beliefs.[31]

We can now define an equilibrium in our setting, which is based on Riley (1979).

Definition 2 (*Riley Reactive Equilibrium*) *An equilibrium is a set of compensation contracts such that for any additional contract which generates an expected gain to the deviating investor i, there exists another contract that can be made by investor j that produces positive profits for j and negative profits for i. Moreover, there does not exist a further contract k such that j can be made to suffer losses.*

As it is well known that the Rothschild and Stiglitz equilibrium (1976), if existent, is the Riley reactive equilibrium. So in what follows, I will characterize the equilibria using the Rothschild and Stiglitz notion, which is often more succinct, with strengthening by the Riley reactive equilibrium whenever necessary. Before proceeding with that, let us highlight the key elements in the Rothschild and Stiglitz equilibrium.

[30] This is similar in spirit to the "Two-Part Wage" structure in Salop and Salop (1976).
[31] There is no off-equilibrium path.

An equilibrium is a *Rothschild and Stiglitz equilibrium* if it is a pair $(\boldsymbol{\eta}, \delta)$ for each type of entrepreneurs, corresponding to a compensation contract where $\boldsymbol{\eta}$ is the vector of salary rates as a fraction of the cash flow rate and δ is the valuation of technology at the time of termination, such that

(i) the entrepreneurs maximize expected utility;

(ii) no contract in the equilibrium set makes negative profits for the investors;

(iii) no contract outside the equilibrium set, if offered, will make a positive profit for the investors.

I now have three remarks. First, although termination can occur at different points in time, recall that the terminal value is the product of a scaling factor (which is the constant cash flow rate h divided by the valuation of technology δ) and the capital stock at the time of the termination, so the contract needs to specify only δ.[32] Second, since the values of δ^H and δ^L are public knowledge, $\delta \in [\delta^H, \delta^L]$. Third, note that the vector of salary rates $\boldsymbol{\eta}$ lies in a continuum, $[0,1]^\infty$. This is an infinite space, but we can focus on one subset of it in which the salary policies are fixed-rate over time.[33] That is, let $\gamma_t = \eta h$ for every t, where $\eta \in [0,1]$ is constant. In other words, $\boldsymbol{\eta} = (\eta, \eta, ...)$. We can then denote the salary-valuation contract as (η, δ) in place of the vector notation. Note that since the value function is continuous in the policy variable, the value of any vector of salary rates can be replicated using a fixed-rate policy. So the investors can provide the entrepreneurs with just a fixed-rate contract rather than one with a "wild" pattern. I do acknowledge that this is still one class of compensation policies, and I will provide suggestions on generalization in Section 4. But for now let us turn to studying these fixed-rate policies.

2.3.2 Fixed-Rate Policies

For a type i entrepreneur choosing (η, δ), the project's capital stock will evolve as

$$dK_t = ((1-\eta)h - \delta^i)K_t dt + \sigma K_t dZ_t, \tag{50}$$

and the value for doing so is

$$V^i = E[\int_{s=0}^{\infty} e^{-\rho s - \lambda s}(\eta h K_s + \lambda \frac{h}{\delta} K_s) ds] \tag{51}$$

where K_t follows (50). In fact, we have the following lemma.

[32] Think about valuation using multiples.

[33] While similar analysis can be applied to other subsets, the advantage of this one is that it is fully tractable.

Lemma 4 *For a type i entrepreneur choosing (η, δ), the value is*

$$V^i = \frac{\eta h + \lambda \frac{h}{\delta}}{\rho + \lambda + \delta^i - (1-\eta)h} K_0. \tag{52}$$

Proof. See Appendix. ■

Differentiating V^i with respect to η and δ, respectively, we observe two properties of the value, summarized in Proposition 7.

Proposition 7 *V^i is strictly decreasing in η if $\delta < \frac{\lambda h}{\rho+\lambda+\delta^i-h}$, constant in η if $\delta = \frac{\lambda h}{\rho+\lambda+\delta^i-h}$, and strictly increasing in η if $\delta > \frac{\lambda h}{\rho+\lambda+\delta^i-h}$. V^i is strictly decreasing in δ.*

Proof. See Appendix. ■

Intuitively, when the offered valuation is high enough, that is, when δ is low enough, the opportunity cost of consuming now and thus lowering the amount of capital accumulation at termination is high. So it incentivizes the entrepreneur to substitute future payoff for current payoff by reducing the salary rate η. However, when the offered valuation is low enough, that is, when δ is high enough, the opportunity cost of consuming now is low, making current compensation effectively more valuable. The entrepreneur consequently would like to raise η. On the other hand, for any level of salary rate, the entrepreneur strictly prefers a lower δ, since it simply gives her a higher future payoff than a higher δ does.

Even more interestingly, the interaction between η and δ reflects the trade-off between current payoff and future payoff, as indicated by the slopes of the indifference curves in the δ-η plane, denoted by $(\frac{d\delta}{d\eta})_i$, where the subscript designates types, i.e., $i \in \{H, L\}$.[34] Proposition 8 formalizes this result.

Proposition 8 *$(\frac{d\delta}{d\eta})_i > 0$ if $\delta > \frac{\lambda h}{\rho+\lambda+\delta^i-h}$, $(\frac{d\delta}{d\eta})_i = 0$ if $\delta = \frac{\lambda h}{\rho+\lambda+\delta^i-h}$, and $(\frac{d\delta}{d\eta})_i < 0$ if $\delta < \frac{\lambda h}{\rho+\lambda+\delta^i-h}$.*

Proof. See Appendix. ■

The intuition is straightforward. The trade-off between current payoff and future payoff ultimately manifests itself in the interaction of two effects: the substitution effect and the income effect. When δ is high enough, lowering δ incentivizes the entrepreneur to reduce salary rate as she substitutes future consumption for current

[34] Note that we suppress the arguments of these functions in this notation.

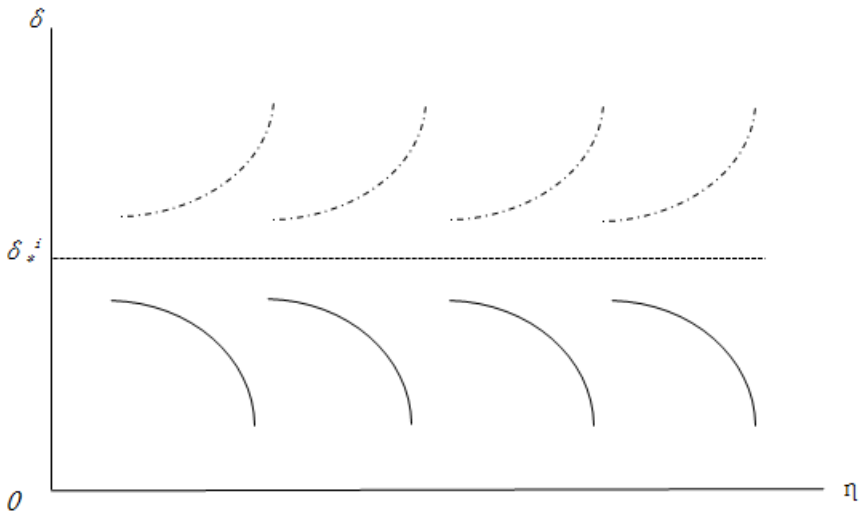

Figure 3: Indifference curves in the δ-η plane. $\delta_*^i = \frac{\lambda h}{\rho+\lambda+\delta^i-h}$, where $i \in \{H, L\}$.

consumption. That is, the substitution effect dominates. When δ is low enough, however, the entrepreneur's future wealth is substantial to the extent that a further reduction in δ disincentivizes her to accumulate capital. In fact, she would like to smooth consumption by raising salary rate. In this case, the income effect dominates.

Denote this critical threshold by δ_*^i, i.e., let $\delta_*^i = \frac{\lambda h}{\rho+\lambda+\delta^i-h}$. Figure 3 depicts the indifference curves in the δ-η plane. Let us highlight four observations. First, note that $\delta = \delta_*^i$ (the dash line) is one indifference curve, as Proposition 8 shows. Second, $\delta = \delta_*^i$ divides the δ-η plane into two regions. It is easy to verify that the lower region (solid curves) corresponds to utility value that is strictly larger than that of $\delta = \delta_*^i$, which in turn is strictly larger than that of the upper region (the dash-dot curves). Third, the direction of increasing utility in the lower region is southwest, whereas it is southeast in the upper region. Fourth, as discussed above, the indifference curves slope downward in the lower region and upward in the upper region.

Moreover, the fact that $\delta^H < \delta^L$ results in crucial relations between the high type's indifference curves and the low type's, which will become useful later. On the one hand, since δ_*^i is decreasing in δ^i, we have $\delta_*^H > \delta_*^L$. On the other hand, at any point, the slope of the low type's indifference curve is larger than that of the high type's, which is summarized in the following lemma.

Lemma 5 *Fix any point in the δ-η plane.*

$$(\frac{d\delta}{d\eta})_{i=H} < (\frac{d\delta}{d\eta})_{i=L}. \tag{53}$$

Proof. See Appendix. ■

Intuitively, the "marginal rate of substitution" of η for reduction in δ is higher for the high type than that for the low type. The reason is that the low type is less efficient at accumulating capital, so even if she gives up some salary and gets a better valuation in return, her capital stock cannot scale up the terminal value as effectively as the high type. In other words, the high type is more willing to give up salary (to accumulate capital stock) for a better valuation.

2.3.3 Equilibrium Analysis

Let us first note the relation among δ^i, δ^*, and δ^i_*, which is summarized in Lemma 6.

Lemma 6 *For $i \in \{H, L\}$, $\delta^i > \delta^i_*$ if and only if $\delta^i > \delta^*$, and $\delta^i \leq \delta^i_*$ if and only if $\delta^i \leq \delta^*$.*

Proof. See Appendix. ■

Indeed, δ^i_* as a function of δ^i has a fixed point at δ^*. That is, $\delta^i_*(\delta^*) = \delta^*$. Thus, the relative magnitude among δ^H, δ^L, and δ^* results in three main cases, and the corresponding relative magnitude among δ^H, δ^L, δ^H_* and δ^L_* leads to additional subcases. Let us now study them.

$\delta^* < \delta^H < \delta^L$

In this case, neither type possesses a technology that is sufficiently valuable to favor terminal over current compensation, as Proposition 6 indicates. Indeed, the first best strategies for both types are to consume all cash flows at each point in time. That is, $\gamma^i_t = h$, $\forall t$, where $i \in \{H, L\}$. Equivalently, $\eta = 1$ is preferable for both types. Now by Lemma 6, we know that $\delta^H > \delta^H_*$, and $\delta^L > \delta^L_*$. Thus, $\delta^L > \delta^H > \delta^H_* > \delta^L_*$. Figure 4 depicts this case.

However, let us show that an equilibrium in this case cannot be pooling when there is asymmetric information between the entrepreneurs and the investors, which is summarized in the following proposition.

Proposition 9 *If $\delta^* < \delta^H < \delta^L$, an equilibrium is never pooling.*

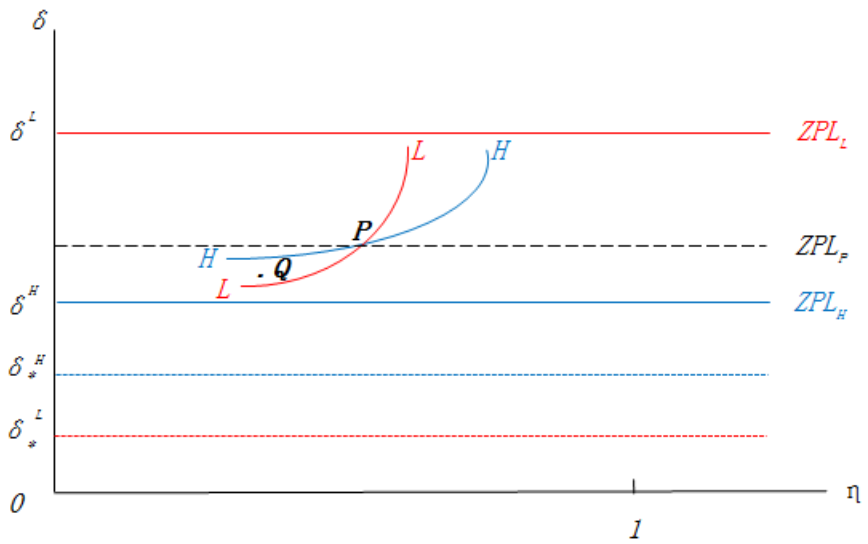

Figure 4: An equilibrium is never pooling if $\delta^* < \delta^H < \delta^L$.

Proof. The proof is similar to that in Rothschild and Stiglitz (1976). Figure 4 depicts in the δ-η plane the indifference curves of the entrepreneurs, including the horizontal $\delta = \delta^i_*$ for $i \in \{H, L\}$, and the zero profit lines of the investors. Consider a pooling contract, P. First note that P must lie on the pooling zero profit line of the investors, ZPL_P. This is due to the fact that if P is below ZPL_P, then the investors offering P lose money, contracting the definition of equilibrium, and if P is above ZPL_P, then there is a contract that offers slightly higher η and slightly lower δ which still makes a profit when all entrepreneurs select it. All will prefer this contract to P, so P cannot be an equilibrium.

Since $(\frac{d\delta}{d\eta})_{i=H} < (\frac{d\delta}{d\eta})_{i=L}$ at P, consider the contract Q, which lies to the southwest of P, between the indifference curves through P and above the zero-profit line for the high type. Observe that the direction of increasing utility is southeast. So if Q is offered, the high type will prefer it over P, while the low type will prefer P over it. Thus, Q attracts away only the high type, and hence makes a positive profit, upsetting the equilibrium. ■

Intuitively, if a pooling contract is offered here, some investors can always "skim the cream" by offering a contract in the neighborhood of the pooling contract. By doing so, they attract away just the high type while making a positive profit. I now

characterize an equilibrium of this case, which is a separating equilibrium, by the following proposition.

Proposition 10 *The contract that corresponds to the low type's first best and the contract that maximizes the utility of the high type subject to that it gives the low type a utility at most equal to that of her first best constitute an equilibrium, if* $\delta^* < \delta^H < \delta^L$.

Proof. Figure 5 depicts the low type's indifference curve L through her first best contract, A, as well as the zero profit lines of the investors in the δ-η plane. As Proposition 9 shows, if there is an equilibrium, each type must select a separate contract. Moreover, as it is clear from the proof of Proposition 9, we see that each contract in the equilibrium set makes zero profits. So an equilibrium contract for the low type must be on the line ZPL_L. A is most preferable (as well as feasible) along ZPL_L. Thus, A must be part of any equilibrium.

An equilibrium contract for the high type must not be more attractive to the low type than A; it must lie on the northwest of L (including L). Clearly, of all such contracts, the one that the high type most prefers is B, the contract at the intersection of ZPL_H and H, since otherwise one can always offer another contract in the neighborhood that attracts away only the high type and thus makes a positive profit. This establishes that the set (A, B) is the only possible equilibrium for a market with high- and low-type entrepreneurs.[35] ■

In essence, we maximize the high type's utility while ensuring that the low type has no incentive to deviate from her first best. This in turn pins down the contract that the investors should offer to the high type.

A question of existence arises naturally. Indeed, Rothschild and Stiglitz (1976) provide conditions under which (A, B) may not be an equilibrium, so that an equilibrium does not exist (see also Hahn (1974) for a suggestive explanation for this nonexistence). However, one can now "strengthen" the equilibrium notion using a Riley reactive equilibrium (RRE), and it is easy to show that (A, B) is a RRE. Thus, if $\delta^* < \delta^H < \delta^L$, an equilibrium exists and is unique. We can therefore refer to (A, B) as *the* equilibrium.

[35] This somewhat heuristic argument can be made completely rigorous. See Wilson (1980).

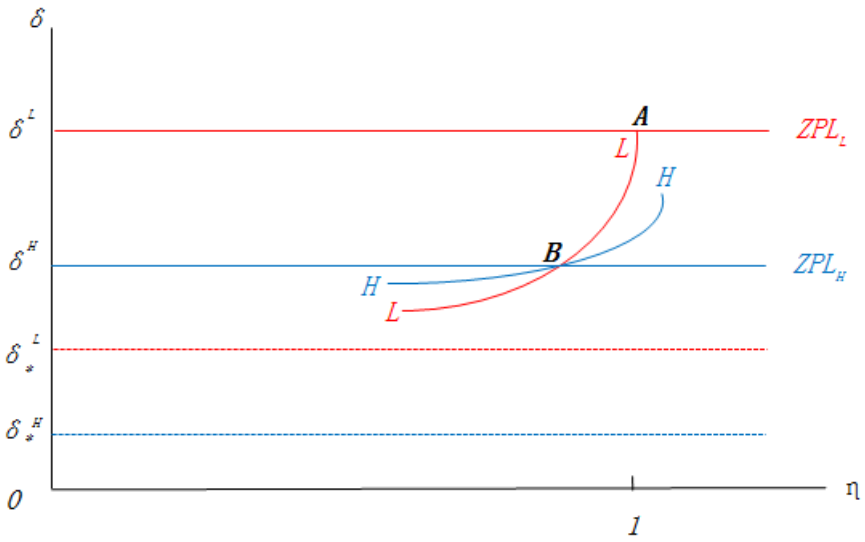

Figure 5: The equilibrium in the case of $\delta^* < \delta^H < \delta^L$

$\delta^H \leq \delta^* < \delta^L$

In this case, the high-type entrepreneur has a sufficiently valuable technology and she would like to compensate herself nothing until termination in absence of her low-type counterpart. In contrast, the low-type entrepreneur's technology makes full salary attractive if there is no high-type entrepreneur. What makes this case particularly interesting and slightly complicated is that the relative magnitude between δ^H and δ^L results in two subcases, each of which includes two situations.

(1) $\delta^H \geq \delta_*^L$

This corresponds to the subcase where although the high type's technology is valuable, it is still not attractive yet for the low type to discard her first best strategy. This is shown in Figure 6 and Figure 7, as even the δ^H valuation is still in the region of upward-sloping indifference curves for the low type, where higher salary rate is more desirable. These two figures correspond to the two situations, which we now discuss in turn.

First, we have $\delta_*^L \leq \delta^H < \delta^L < \delta_*^H$. In this situation, since the investors know the values of δ^H and δ^L, the relevant valuation δ belongs to $[\delta^H, \delta^L]$. The indifference curves of the high type and the low type have opposite signs in this region. Then

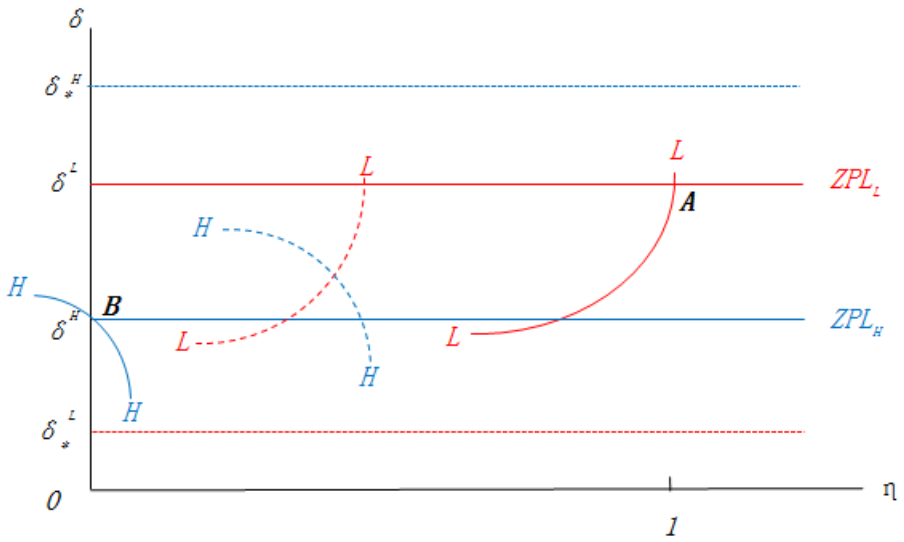

Figure 6: The equilibrium in the case of $\delta_*^L \leq \delta^H < \delta^L < \delta_*^H$

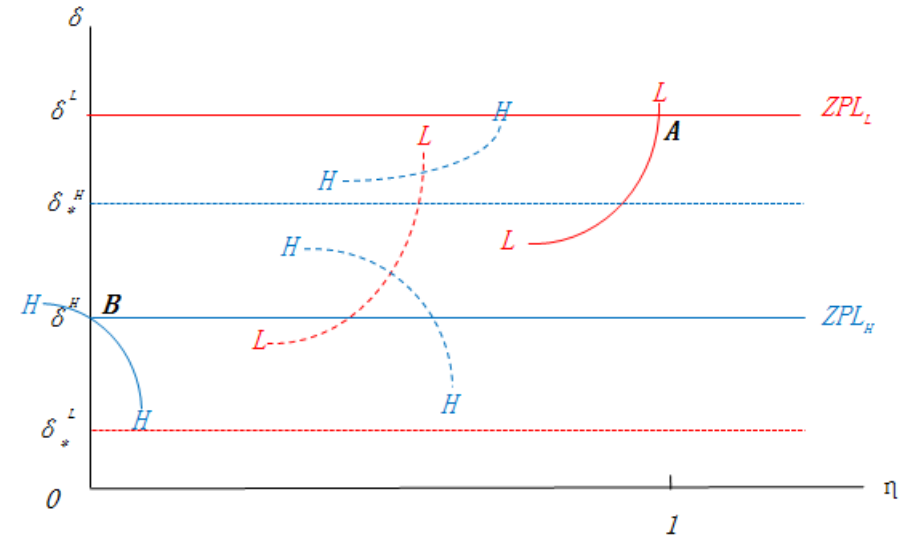

Figure 7: The equilibrium in the case of $\delta_*^L \leq \delta^H < \delta_*^H \leq \delta^L$

it is clear that an equilibrium cannot be pooling, since some investors can always "cherry-pick" the high type in the neighborhood of the pooling contract, making a positive profit.

We now show the unique equilibrium is the set (A, B) in Figure 6, by the next proposition.

Proposition 11 *The contracts that correspond to the low type's first best and the high type's first best constitute the unique equilibrium, if* $\delta_*^L \leq \delta^H < \delta^L < \delta_*^H$.

Proof. Figure 6 depicts both types' indifference curves, L and H, through their respective first best contracts, A and B, as well as the zero profit lines of the investors in the δ-η plane. We know that if there is an equilibrium, it is separating, and since each contract in the equilibrium set makes zero profits, the equilibrium contracts must be on the zero profit lines, ZPL_L and ZPL_H, respectively.

Now, along ZPL_L and ZPL_H, A and B maximize respectively the low type's utility and the high type's utility. Thus, (A, B) satisfies the first two requirements in the definition of an equilibrium. Moreover, any contract that is preferable to A for the low type is to the southeast of A, which is either unfeasible or makes a negative profit (since it attracts only the low type); any contract that is preferable to B for the high type is to the southwest of B, which simply is not feasible. So the third requirement in the equilibrium definition is also fulfilled. Thus, (A, B) is an equilibrium.

Finally, Proposition 6 shows that the solution to the optimization problem for either type is unique. Therefore, (A, B) is the unique equilibrium. ■

The second situation is when $\delta_*^L \leq \delta^H < \delta_*^H \leq \delta^L$. Similar to the first situation, an equilibrium in this situation is never pooling. One slight variation is that although the indifference curves of the low type in the relevant range is everywhere upward-sloping, the indifference curves of the high type can be upward-sloping, flat, as well as downward-sloping. However, recall that a downward-sloping indifference curve corresponds to a higher utility level than that of the flat one, which in turn is higher than that of a upward-sloping one. So no matter where the pooling contract is,[36] some investors can always offer a contract in the region of downward-sloping indifference curves that is attractive only to the high type.

In addition, applying the same argument in the proof of Proposition 11, one can establish that the set (A, B) is the unique equilibrium in Figure 7. Therefore, as long

[36] It has to belong to (δ^H, δ^L) though, since $\xi \in (0, 1)$.

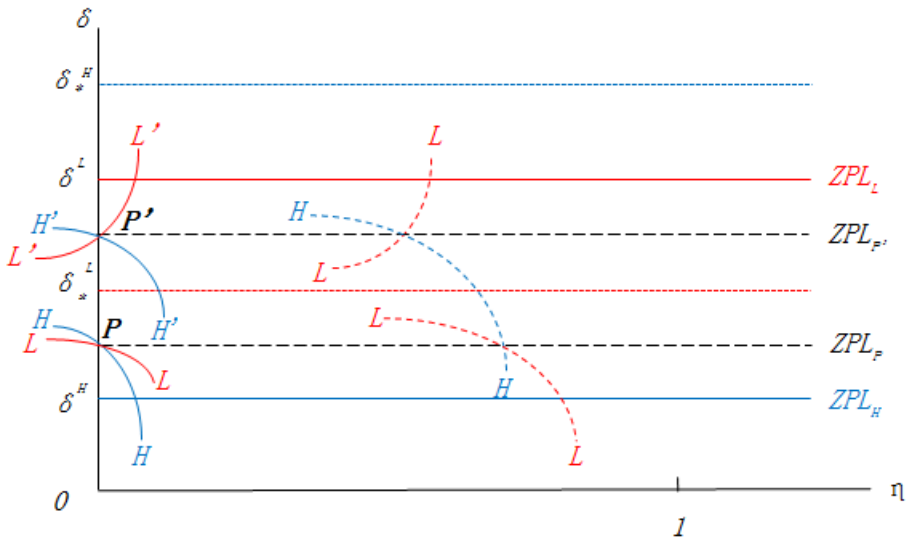

Figure 8: The equilibrium in the case of $\delta^H < \delta_*^L < \delta^L < \delta_*^H$

as $\delta^H \leq \delta^* < \delta^L$ and $\delta^H \geq \delta_*^L$, the contracts that correspond to the "first bests" of the two types constitute the unique separating equilibrium.

(2) $\delta^H < \delta_*^L$

This corresponds to the subcase where the high type's technology is so valuable that it is attractive for the low type to mimic even if the low type has to forgo her first best strategy. As Figure 8 and Figure 9 show, the δ^H valuation belongs to the region of downward-sloping indifference curves for the low type. That is, if the low type can obtain a valuation close to δ^H, she is incentivized to postpone payoff until the time of termination. We again have two situations.

First, we have $\delta^H < \delta_*^L < \delta^L < \delta_*^H$. This is shown in Figure 8. Note that while the indifference curves of the high type in the relevant range $[\delta^H, \delta^L]$ is everywhere downward-sloping, the indifference curves of the low type can be upward-sloping, flat, as well as downward-sloping. In contrast to the cases above where an equilibrium cannot be pooling, we show in the current situation that an equilibrium cannot be separating, which is formalized in the next proposition.

Proposition 12 *If $\delta^H < \delta_*^L < \delta^L < \delta_*^H$, an equilibrium is never separating.*

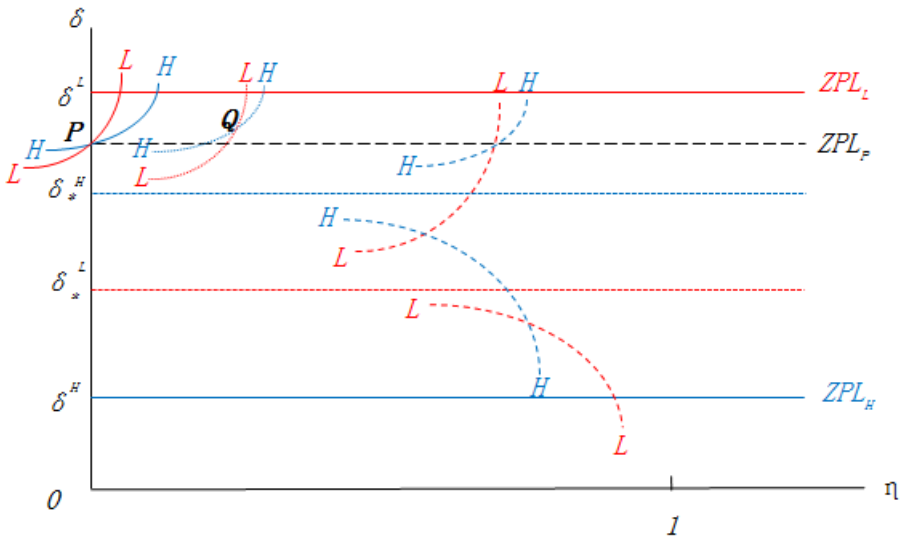

Figure 9: The equilibrium in the case of $\delta^H < \delta_*^L < \delta_*^H \leq \delta^L$

Proof. The proof is almost immediate. If an equilibrium is separating, then since each contract in the equilibrium set makes zero profits, the equilibrium contract for the low type must be on ZPL_L and that for the high type must be on ZPL_H. But we know any indifference curve of the low type crossing ZPL_L corresponds to a utility level that is lower than that of any indifference curve of the low type crossing ZPL_H, as in Figure 8. Thus, if any contract along ZPL_H is offered, despite that it is intended for the high type, the low type will select it as well. But then the investors will make a negative profit, upsetting the equilibrium. ■

In essence, when the high type's technology is substantially valuable, the low type does not mind altering her optimal course of actions since the terminal payoff at a valuation near the high type's will more than compensate for the reduction in current pay. Indeed, mimicking the high type becomes her top priority. We now show that there exists a unique pooling equilibrium, by the following proposition.

Proposition 13 *The pooling contract that corresponds to the weighted average pooling valuation using ξ and the high type's first best choice of salary rate is the unique equilibrium if $\delta^H < \delta_*^L < \delta^L < \delta_*^H$.*

Proof. Depending on the magnitude of ξ, the proportion of the high type, the weighted average pooling valuation can fall into either of the two ranges, $(\delta^H, \delta_*^L]$ or

(δ_*^L, δ^L). The pooling zero profit line ZPL_P in Figure 8 depicts the former range. First, it is obvious that if the contract P is offered and selected by both types, the investors makes zero profits. Second, any contract that is preferable than P to the low type and the high type lies to the southwest of the solid curves L and H, respectively. Given feasibility, these correspond to the regions enclosed by the vertical line $\eta = 0$, ZPL_H, and L or H, respectively. Then, if a contract between L and H is offered, it attracts only the low type and clearly makes a negative profit. If a contract to the southwest of H is offered, it attracts both types but it still makes a negative profit since it is below ZPL_P. Finally, among the pooling contracts on ZPL_P, P maximizes both types' utility.

The pooling zero profit line $ZPL_{P'}$ illustrates the latter case, which is slightly more complicated. Here a pooling valuation does not motivate the low type to choose zero salary rate (indeed, she prefers full salary). However, we claim that the equilibrium is still a pooling contract where both types compensate themselves nothing until the time of termination, which corresponds to the point P'. First, since the equilibrium cannot be separating, it must lie on $ZPL_{P'}$ by the argument in the proof of Proposition 9. Along $ZPL_{P'}$, P' is the unique contract where the investors cannot "cherry-pick" by offering a contract in the neighborhood that attracts only the high type while making a positive profit. Indeed, any contract that is preferable to P' lies in the region between the solid curve H' and the line $\eta = 0$. But any such offer attracts both types and since it is below $ZPL_{P'}$, who offers it will make a negative profit. Thus, no investor will make such an offer. Moreover, any other offer that is to the southeast of L' as well as to the northeast of H' attracts only the low type and thus makes a negative profit. So no investor will make an offer in this region either. Therefore, P' is the only equilibrium when the weighted average pooling valuation is in $(\delta_*^L, \delta^L]$. ■

It may seem counter-intuitive at first that the low type does worse than her first best when the weighted average pooling valuation is in the range $(\delta_*^L, \delta^L]$. After all, why does not she just reveal the truth to the investors that she is the low type and then require the full salary rate? The problem is that this revelation is not credible. Suppose the investors offer two contracts $(\eta = 1, \delta^L)$ and $(\eta = 0, \delta^H)$, intended respectively, to the low type and the high type. But once these two contracts are offered, all entrepreneurs will select $(\eta = 0, \delta^H)$. That is, the low type cannot credibly commit to $(\eta = 1, \delta^L)$, and the investors know this. Therefore, they will only offer

the pooling contract.

The second situation is when $\delta^H < \delta_*^L < \delta_*^H \leq \delta^L$. Similar to the first situation, an equilibrium in this situation is never separating. Also, it is almost parallel to show that the pooling contract that corresponds to the weighted average pooling valuation and the high type's first best choice of salary rate is the unique equilibrium. The only difference is that there are now three ranges that the weighted average pooling valuation can fall into, $(\delta^H, \delta_*^L]$, $(\delta_*^L, \delta_*^H]$, and (δ_*^H, δ^L). We have established the equilibrium corresponding to the first two ranges and we now show the equilibrium associated with the third.

As Figure 9 indicates, the contract Q attracts both types and since it is above the pooling zero profit line ZPL_P, it also makes a positive profit for those investors who offer this. We now apply the stronger Riley Reactive Equilibrium concept and it is easy to show that P is a RRE. Indeed, if Q is offered by a group of investors, then another group of investors can offer a contract that is to the southwest of Q, between the dotted H and L, as well as above ZPL_H. This newly offered contract attracts the high type but not the low type. Moreover, the investors who offer this contract cannot be made to suffer losses, since they already have only the high type (the worst for them is to break even). Applying this argument iteratively to any contract that is a profitable deviation from P and noticing that this "cherry-picking" just described is not available at P since any contract that is to the southwest of P and between the solid H and L is not feasible, we establish that P is the unique equilibrium.

$\delta^H < \delta^L \leq \delta^*$

In this case, by Proposition 6, either type possesses a technology that is sufficiently valuable to motivate the entrepreneur to compensate herself nothing until termination, when there is symmetric information between the entrepreneurs and the investors. Now Lemma 6 implies that $\delta^H < \delta_*^H$, and $\delta^L \leq \delta_*^L$. Thus, $\delta_*^H > \delta_*^L \geq \delta^L > \delta^H$. Figure 10 depicts this case.

It is simple to show that P is the unique equilibrium since any deviating contract that is preferable to either type will make a negative profit for the investor offering it. This is due to the fact that "cherry-picking" requires a contract that is outside the feasible region. Intuitively, the technology of either type is so valuable that following the optimal course of actions of full investment dominates the issue of mimicking.

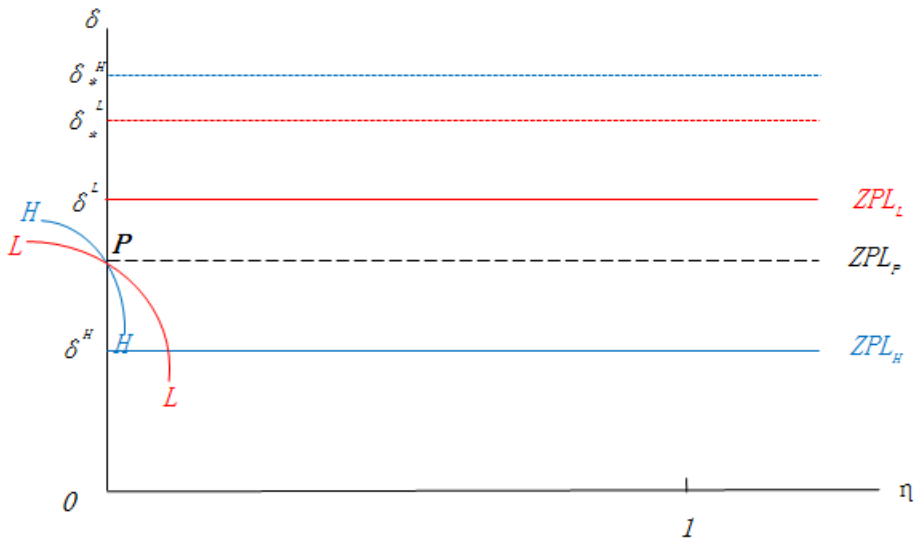

Figure 10: The equilibrium in the case of $\delta^H < \delta^L \leq \delta^*$

Summary Summarizing the results above, we obtain the main theorem of the paper.

Theorem 6 *Depending on the parameter values, we have one of the following:*

(i) If $\delta^* < \delta^H < \delta^L$, a unique separating equilibrium exists where the low type obtains her first best of full salary rate while the high type does not.

(ii) If $\delta^H \leq \delta^* < \delta^L$, a unique separating equilibrium exists where both types obtain their respective first bests when $\delta^H \geq \delta^L_*$; a unique pooling equilibrium exists where both types follow the high type's first best of full investment rate when $\delta^H < \delta^L_*$.

(iii) If $\delta^H < \delta^L \leq \delta^*$, a unique pooling equilibrium exists where both types follow full investment.

When both types' technologies are less valuable than the first best threshold, we have a separating equilibrium that is fully revealing. The low type achieves her first best by paying herself all the cash flows at each point in time. The high type, however, takes a cut in salary rate in return for being valued at her technology's true worth. She is willing to do so because she is more efficient at accumulating capital than the low type, so reducing salary rate has the added benefit of increasing (more effectively than the low type) the capital stock which scales up the terminal payoff.

Nevertheless, having a salary rate that is lower than her first best is still costly to the high type. So the signal is dissipative.

When the low type's technology is less valuable than the first best threshold while the high type's is more valuable than it and yet it is not too much more valuable than the low type's, we again have a fully revealing separating equilibrium. Interestingly, however, both types follow their respective first best compensation policies. In this case, it is better for the low type to stick to her optimal course of actions under symmetric information given her specific trade-off between current payoff and future payoff. As a result, the signal serves to self-select and is therefore nondissipative.

When the low type's technology is less valuable than the first best threshold while the high type's is more valuable than it and it is also too more valuable than the low type's, we have a pooling equilibrium. The two types pool at the high type's first best compensation policy since the high type's valuation is so high that the low type would like to mimic no matter what. The low type may or may not do better overall than her first best total payoff under symmetric information. When she does do worse, it is due to her lack of credible commitment to her first best compensation policy and the natural boundedness of her feasible actions. The investors being aware of this can thus force a pooling result, where neither market breakdown nor information revelation occurs.

Finally, when both types' technologies are more valuable than the first best threshold, we also have a pooling equilibrium. Both types opt for full investment. This is due to the fact that the technologies are so valuable that following the optimal course of actions for capital accumulation is one's top priority.

2.4 Model Extensions and Empirical Predictions

2.4.1 Model Extensions

We have considered one class (although a continuum) of contracts where the salary policies are fixed-rate over time. More generally, we can solve for optimal contract by maximizing the high type's utility subject to the low type's incentive compatibility constraint as well as the capital stock processes and nonnegativity. That is, we can

$$
\begin{aligned}
& \max_{\{\gamma_t\}} E[\int_{s=0}^{\infty} e^{-\rho s-\lambda s}(\gamma_s K_s^H + \lambda \frac{h}{\delta^H} K_s^H)ds] \\
s.t. \quad & E[\int_{s=0}^{\infty} e^{-\rho s-\lambda s}(\gamma_s K_s^L + \lambda \frac{h}{\delta^H} K_s^L)ds] \leq V^L(K_0) \\
& dK_t^H = (h-\gamma_t-\delta^H)K_t^H dt + \sigma K_t^H dZ_t \\
& dK_t^L = (h-\gamma_t-\delta^L)K_t^L dt + \sigma K_t^L dZ_t \\
& 0 \leq \gamma_t \leq h \\
& K_t^H \geq 0 \\
& K_t^L \geq 0,
\end{aligned}
$$

where $V^L(K_0) = \frac{\lambda \frac{h}{\delta^L}}{\rho+\lambda+\delta^L-h} K_0$, if $\delta^L \leq \frac{\sqrt{(\rho+\lambda-h)^2+4\lambda h}-(\rho+\lambda-h)}{2}$, and $V^L(K_0) = \frac{\lambda \frac{h}{\delta^L}+h}{\rho+\lambda+\delta^L} K_0$, if $\delta^L > \frac{\sqrt{(\rho+\lambda-h)^2+4\lambda h}-(\rho+\lambda-h)}{2}$.

Obviously, this problem is more complicated to solve than the setting where the salary policies are fixed-rate. However, it may be tractable given that we can still utilize homogeneity.[37]

2.4.2 Empirical Predictions

The main empirical predictions of the model are summarized here. The focus is on the obsolescence rate δ, which captures the growth prospects of the project as a result of the entrepreneur. In other words, it reflects how relevant the entrepreneur is to the production process of the project.

A key proxy for entrepreneur relevance is the nature of the industry. Specifically, industries such as the fast food industry do not appear to require much hands-on entrepreneur expertise, and it is fairly standardized across firms. On the other hand, industries such as the software development industry typically require a lot of human capital input from the entrepreneurs, and this is likely the case for every single firm. In between perhaps these two extremes are industries such as the service industry, for instance nursing homes, where the relevance of the entrepreneurial input is drastically heterogeneous. It is sometimes the entrepreneur's insight that determines the fate of the venture, whereas in other cases it is the general business environment (industry trend, etc.) that really matters.

[37] That is, dividing both the objective and the constraints by K_0, we can again apply the solution strategy in Section 2.

Therefore, cross-sectionally, we predict that

- In industries that the relevance of entrepreneurial inputs is homogeneously low, entrepreneurs in better performing firms take lower cash-based pay.
- In industries that the relevance of entrepreneurial inputs is homogeneously high, entrepreneurs in all firms take low levels cash-based pay.
- In industries that the relevance of entrepreneurial inputs is heterogeneous, entrepreneurs in better performing firms take lower cash-based pay.

It is often the case that the relevance of entrepreneurial inputs is high among high-tech industries. Therefore, all in all, we expect the across-firm variation in cash-versus-equity pay to be more pronounced in low-tech industries than in high-tech industries.

2.5 Conclusion

Prior literature has documented yet not explained the tremendous cross-sectional variation in entrepreneurial compensation. The uniqueness of small businesses, in particular the difference between entrepreneurs and corporate managers, makes the insights from the executive compensation studies on risk and incentives remote for this issue. Instead, an asymmetric-information perspective seems more relevant given the significant information advantage that entrepreneurs possess over the investors.

This paper employs such a perspective considering the reliance of the uninformed investors on compensation structure as a signal to distinguish among different types of informed entrepreneurs. Furthermore, an interesting trade-off between current and future payoffs is modeled, where an entrepreneur balances her salary drawn from developing a project with its terminal payoff. This terminal payoff is effectively her deferred compensation as she makes reinvestment into the project.

The signaling model with this embedded intertemporal trade-off has some interesting features. In particular, a market breakdown will not occur, with both separating and pooling equilibria possible, and consequently, an equilibrium is not necessarily informationally consistent. Furthermore, when an equilibrium is indeed revealing, whether the signal is dissipative is completely endogenous.

These features correspond naturally to empirical predictions about entrepreneurial compensation. The model focuses on how relevant entrepreneurial inputs are to

the production process, which reflects the intrinsic interest of the investors on the growth prospects. A key proxy for entrepreneurial relevance is the nature of the industry. Therefore, the study may explain some of the cross-industry differential in compensation structure, in particular between equity-based and cash-based pay. Indeed, future research can be designed to directly test these implications.

References

[1] Ang, J. S., 1991. Small Business Uniqueness and the Theory of Financial Management. *Journal of Entrepreneurial Finance* 1: 1-13.

[2] Akerlof, G. A., 1970. The market for "lemons": quality uncertainty and the market mechanism. *Quarterly Journal of Economics* 84: 488-500.

[3] Albuquerue, R. and N. Wang, 2008. Agency conflicts, investment, and asset pricing. *Journal of Finance* 63: 1–40.

[4] Banks, J. and J. Sobel, 1987. Equilibrium selection in signaling games. *Econometrica* 55: 647-661.

[5] Bhattacharya, S, 1980. Nondissipative Signaling Structures and Dividend Policy. *Quarterly Journal of Economics* 95: 1-24.

[6] Blanchflower, D. G. and A. J. Oswald, 1998. What makes an entrepreneur? *Journal of Labor Economics* 16: 26–60.

[7] Brennan M., and A. Kraus, 1987. Efficient Financing Under Asymmetric Information. *Journal of Finance* 42: 1225-1243.

[8] Cho, I., 1987. A refinement of sequential equilibrium. *Econometrica* 55: 1367-1389.

[9] Cho, I. and D. M. Kreps, 1987. Signaling games and stable equilibria. *Quarterly Journal of Economics* 102: 179-222.

[10] Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross, 1985. An intertemporal general equilibrium model of asset prices. *Econometrica* 53: 363–384.

[11] Cramer, J. S., J. Hartog, N. Jonker, and C. M. Van Praag, 2002. Low risk aversion encourages the choice for entrepreneurship: an empirical test of a truism. *Journal of Economic Behavior and Organization* 48: 29–36.

[12] Franke, G., 1987. Costless Signalling in Financial Markets. *Journal of Finance* 42: 809-822.

[13] Gordon, M. J., 1959. Dividends, earnings and stock prices. *Review of Economics and Statistics* 41: 99–105.

[14] Greenwood, J., Z. Hercowitz, and G. W. Huffman, 1988. Investment, capacity utilization and the real business cycle. *American Economic Review* 78: 402–417.

[15] Greenwood, J., Z. Hercowitz, and P. Krusell, 1997. The role of investment-specific technological change in the business cycle. *European Economic Review* 44: 91–115.

[16] Greenwood, J., Z. Hercowitz, and P. Krusell, 2000. Long-run implications of investment-specific technological change. *American Economic Review* 87: 342–362.

[17] Grossman, S. J. and M. Perry, 1986. Perfect sequential equilibrium. *Journal of Economic Theory* 39: 97–119.

[18] Hahn, F. H., 1974. Notes on R-S models of insurance markets. Mimeo, Cambridge University.

[19] Heinkel, R., 1982. A Theory of Capital Structure Relevance under Imperfect Information. *Journal of Finance* 37: 1141-1150.

[20] Holmstrom, B., 1979. Moral hazard and observability. *Bell Journal of Economics* 10: 74-91.

[21] Jensen, M. C. and W. H. Meckling, 1976. Theory of the investor: managerial behaviour, agency costs and ownership structure. *Journal of Financial Economics* 3: 305-360.

[22] Kanbur, S. M., 1979. On risk taking and the personal distribution of income. *Journal of Political Economy* 87: 760–797.

[23] Leland, H. E. and D. H. Pyle, 1977. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance* 32: 371–387.

[24] Ofer, A. and A. V. Thakor, 1987. A theory of corporate cash disbursement mechanisms: stock repurchases and dividends. *Journal of Finance* 42: 365-394.

[25] Prendergast, C., 1999. The provision of incentives in investors. *Journal of Economic Literature* 37: 7–63.

[26] Prendergast, C., 2002. The tenuous trade-off between risk and incentives. *Journal of Political Economy* 110: 1071-1102.

[27] Riley, J. G., 1979. Informational equilibrium. *Econometrica* 47: 331-359.

[28] Rothschild, M. and J. Stiglitz, 1976. Equilibrium in competitive insurance markets: an essay on the economics of imperfect information. *Quarterly Journal of Economics* 90: 629-649.

[29] Salop, J. and S. Salop, 1976. Self-selection and turnover in the labor market. *Quarterly Journal of Economics* 90: 619-627.

[30] Spence, M., 1974. Competitive and optimal responses to signals: an analysis of efficiency and distribution. *Journal of Economic Theory* 7: 296-332.

[31] Stanford Graduate School of Business, 2012. Search Funds-2011: Selected Observations.

[32] Sundaresan, S. M., 1984. Consumption and equilibrium interest rates in stochastic production economies. *Journal of Finance* 39: 77–92.

[33] Wasserman, N., 2004. Executive Compensation in Entrepreneurial Teams: The Founder Gap, Board membership, and Pay for Milestones. *Academy of Management Annual Meeting Paper Proceedings.*

[34] Wilson, C. A., 1980. The nature of equilibrium in markets with adverse selection. *Bell Journal of Economics* 11: 108-30.

Appendix

Proof of Lemma 3. From (45), we know the capital stock process follows a geometric Brownian motion, which implies that

$$K_t = K\tau \exp[(-\delta^i - \frac{1}{2}\sigma^2)(t-\tau) + \sigma(Z_t - Z_\tau)],$$

for $t \geq \tau$

Thus, from (46), interchanging the order of integration, and using the property of log-normal distribution, we have

$$\begin{aligned} E_\tau[\int_{t=\tau}^{\infty} hK_t dt] &= \int_{t=\tau}^{\infty} E_\tau[hK_t]dt] \\ &= h\int_{t=\tau}^{\infty} E_\tau\{K\tau \exp[(-\delta^i - \frac{1}{2}\sigma^2)(t-\tau) + \sigma(Z_t - Z_\tau)]\}dt \\ &= hK\tau \int_{t=\tau}^{\infty} \exp[(-\delta^i)(t-\tau)]dt \\ &= \frac{h}{\delta^i} K_\tau \end{aligned}$$

■

Proof of Proposition 6. Define the process

$$M_t \equiv \int_{s=0}^{t} e^{-\rho s - \lambda s}(\gamma_s K_s + \lambda \frac{h}{\delta^i} K_s)ds + e^{-\rho t - \lambda t} V(K_t). \tag{A1}$$

Applying Ito's lemma and basic calculus, we have from (44) and (49) that

$$\frac{E[dM_t]}{e^{-\rho t - \lambda t}dt} = \gamma_t K_t + \lambda \frac{h}{\delta^i} K_t - (\rho + \lambda)CK_t + C(h - \gamma_t - \delta^i)K_t \tag{A2}$$

If we have the correct value function and the optimal strategy, M_t is the conditional expectation of the value and a martingale. If we have a suboptimal strategy, however, then this is a supermartingale, since the value will fall on average reflecting the shortfall from potential. Thus, if have the right value function, the drift of M_t is maximized at 0 by the optimal controls.[38] Thus, from (A2), we have the Bellman equation

$$\lambda \frac{h}{\delta^i} K - (\rho + \lambda)CK + \max_\gamma [\gamma K + C(h - \gamma - \delta^i)K] = 0, \tag{A3}$$

which upon simplifying becomes

$$\lambda \frac{h}{\delta^i} - (\rho + \lambda)C + \max_\gamma [\gamma + C(h - \gamma - \delta^i)] = 0. \tag{A4}$$

Since the objective is linear in γ and $0 \leq \gamma \leq h$, (A4) is equivalent to

$$\lambda \frac{h}{\delta^i} - (\rho + \lambda)C + \max_\gamma \{(h - \delta^i)C, h - \delta^i C\} = 0, \tag{A5}$$

[38] Since a margingale has zero drift and a supermartingale has zero or negative drift.

where the first component in the max function is obtained when $\gamma = 0$ while the second is obtained when $\gamma = h$.

Define the function

$$f(C) \equiv \lambda \frac{h}{\delta^i} - (\rho + \lambda)C + \max_{\gamma}\{(h - \delta^i)C, h - \delta^i C\},$$

and recognizing that

$$(h - \delta^i)C \geq h - \delta^i C \Leftrightarrow C \geq 1,$$

we have

$$f(C) = \begin{Bmatrix} \lambda \frac{h}{\delta^i} - (\rho + \lambda + \delta^i - h)C & if\ C \geq 1 \\ \lambda \frac{h}{\delta^i} - (\rho + \lambda + \delta^i)C + h & if\ C < 1 \end{Bmatrix}.$$

Note that

$$f(1) = \lambda \frac{h}{\delta^i} - (\rho + \lambda + \delta^i - h).$$

Thus, letting C^* be such that $f(C^*) = 0$, we have that if $\lambda \frac{h}{\delta^i} \geq \rho + \lambda + \delta^i - h$, i.e., $\delta^i \leq \frac{\sqrt{(\rho+\lambda-h)^2+4\lambda h}-(\rho+\lambda-h)}{2}$,[39] then $\lambda \frac{h}{\delta^i} - (\rho + \lambda + \delta^i - h)C^* = 0$, which implies that

$$C^* = \frac{\lambda \frac{h}{\delta^i}}{\rho + \lambda + \delta^i - h},$$

which is indeed greater than or equal to 1. So this corresponds to the optimal strategy that $\gamma^* = 0$.

In contrast, if $\lambda \frac{h}{\delta^i} < \rho + \lambda + \delta^i - h$, i.e., $\delta^i > \frac{\sqrt{(\rho+\lambda-h)^2+4\lambda h}-(\rho+\lambda-h)}{2}$, then $\lambda \frac{h}{\delta^i} - (\rho + \lambda + \delta^i)C^* + h = 0$, which implies that

$$C^* = \frac{\lambda \frac{h}{\delta^i} + h}{\rho + \lambda + \delta^i},$$

which is indeed smaller than 1. In this case, the optimal strategy is that $\gamma^* = h$. ■

Proof of Lemma 4. From (50), we know that the capital stock process follows a geometric Brownian motion, which implies that

$$K_t = K_0 \exp\{[(1 - \eta)h - \delta^i - \frac{1}{2}\sigma^2]t + \sigma Z_t\}. \tag{A6}$$

[39] This is because $\rho + \lambda - h > 0$ (otherwise, the capital process will explode if the obsolescence rate is not large enough) and $\delta^i > 0$ (so the negative root is not relevant).

Thus, from (51), interchanging the order of integration, and using the property of log-normal distribution, we have

$$\begin{aligned} V^i &= \int_{s=0}^{\infty} E[e^{-\rho s-\lambda s}(\eta h K_s + \lambda \frac{h}{\delta} K_s)ds] \qquad \text{(A7)} \\ &= \int_{s=0}^{\infty} e^{-\rho s-\lambda s}[\eta h + \lambda \frac{h}{\delta}]E(K_0 \exp\{[(1-\eta)h - \delta^i - \frac{1}{2}\sigma^2]s + \sigma Z_s\})ds \\ &= \int_{s=0}^{\infty} e^{-\rho s-\lambda s}[\eta h + \lambda \frac{h}{\delta}]K_0 \exp\{[(1-\eta)h - \delta^i]s\}ds \\ &= [\eta h + \lambda \frac{h}{\delta}]K_0 \int_{s=0}^{\infty} \exp\{[-(\rho+\lambda+\delta^i) + (1-\eta)h]s\}ds \\ &= K_0[\eta h + \lambda \frac{h}{\delta}] \int_{s=0}^{\infty} \exp(-A^i s)ds, \end{aligned}$$

where $A^i = \rho + \lambda + \delta^i - (1-\eta)h$. Obviously, $A^i > 0$ since $\rho + \lambda - h > 0$. Then (A7) implies that

$$V^i = \frac{\eta h + \lambda \frac{h}{\delta}}{A^i} K_0,$$

that is,

$$V^i = \frac{\eta h + \lambda \frac{h}{\delta}}{\rho + \lambda + \delta^i - (1-\eta)h} K_0.$$

■

Proof of Proposition 7. Differentiating V^i with respect to η, we have

$$\begin{aligned} \frac{\partial V^i}{\partial \eta} &= \frac{hK_0(\rho + \lambda + \delta^i - h + \eta h) - h(\eta h + \lambda \frac{h}{\delta})K_0}{(\rho + \lambda + \delta^i - h + \eta h)^2} \\ &= \frac{K_0 h(\rho + \lambda + \delta^i - h - \lambda \frac{h}{\delta})}{(\rho + \lambda + \delta^i - h + \eta h)^2}. \end{aligned}$$

Since $K_0 > 0$ and $h > 0$, we have $\frac{\partial V^i}{\partial \eta} > 0$ if and only if $\rho + \lambda + \delta^i - h - \lambda \frac{h}{\delta} > 0$, which corresponds to

$$\delta > \frac{\lambda h}{\rho + \lambda + \delta^i - h}.$$

Similarly, $\frac{\partial V^i}{\partial \eta} = 0$ if and only if

$$\delta = \frac{\lambda h}{\rho + \lambda + \delta^i - h},$$

and $\frac{\partial V^i}{\partial \eta} < 0$ if and only if

$$\delta < \frac{\lambda h}{\rho + \lambda + \delta^i - h}.$$

Now differentiating V^i with respect to δ, we have

$$\frac{\partial V^i}{\partial \delta} = -\frac{K_0 \lambda h}{(\rho + \lambda + \delta^i - h + \eta h)\delta^2} < 0,$$

since $K_0 > 0$, $\lambda > 0$, $h > 0$, and $\rho + \lambda + \delta^i - (1 - \eta)h > 0$. ■

Proof of Proposition 8. Calculating the slopes of the indifference curves in the δ-η plane, we have

$$\begin{aligned}
(\frac{d\delta}{d\eta})_i &= -\frac{\frac{\partial V^i}{\partial \eta}}{\frac{\partial V^i}{\partial \delta}} \\
&= -\frac{\frac{K_0 h(\rho+\lambda+\delta^i-h-\lambda\frac{h}{\delta})}{(\rho+\lambda+\delta^i-h+\eta h)^2}}{[-\frac{K_0\lambda h}{(\rho+\lambda+\delta^i-h+\eta h)\delta^2}]} \\
&= \frac{\delta^2(\rho + \lambda + \delta^i - h - \lambda\frac{h}{\delta})}{\lambda(\rho + \lambda + \delta^i - h + \eta h)}.
\end{aligned}$$

Thus, $\frac{d\delta}{d\eta} > 0$ if and only if $\rho + \lambda + \delta^i - h - \lambda\frac{h}{\delta} > 0$, or

$$\delta > \frac{\lambda h}{\rho + \lambda + \delta^i - h}.$$

Similarly, $\frac{d\delta}{d\eta} = 0$ if and only if

$$\delta = \frac{\lambda h}{\rho + \lambda + \delta^i - h},$$

and $\frac{d\delta}{d\eta} < 0$ if and only if

$$\delta < \frac{\lambda h}{\rho + \lambda + \delta^i - h}.$$

■

Proof of Lemma 5. From the proof of Proposition 8, we know that

$$(\frac{d\delta}{d\eta})_i = \frac{\delta^2(\rho + \lambda + \delta^i - h - \lambda\frac{h}{\delta})}{\lambda(\rho + \lambda + \delta^i - h + \eta h)}.$$

Now differentiate $(\frac{d\delta}{d\eta})_i$ with respect to the parameter δ^i,

$$\begin{aligned}
\frac{\partial(\frac{d\delta}{d\eta})_i}{\partial \delta^i} &= \frac{\delta^2\lambda(\rho + \lambda + \delta^i - h + \eta h) - \lambda\delta^2(\rho + \lambda + \delta^i - h - \lambda\frac{h}{\delta})}{[\lambda(\rho + \lambda + \delta^i - h + \eta h)]^2} \\
&= \frac{\lambda\delta^2(\eta h + \lambda\frac{h}{\delta})}{[\lambda(\rho + \lambda + \delta^i - h + \eta h)]^2},
\end{aligned}$$

which is positive, since λ, h, and δ are all positive, and $\eta \in [0,1]$.

Thus, $\delta^H < \delta^L$ implies that

$$(\frac{d\delta}{d\eta})_{i=H} < (\frac{d\delta}{d\eta})_{i=L}.$$

■

Proof of Lemma 6. Consider any $i \in \{H, L\}$.

$$\delta^i > \delta^i_* = \frac{\lambda h}{\rho + \lambda + \delta^i - h}$$

if and only if (recall $\rho + \lambda + \delta^i - h > 0$)

$$\delta^i(\rho + \lambda + \delta^i - h) > \lambda h,$$

which is true if and only if

$$(\delta^i)^2 + (\rho + \lambda - h)\delta^i - \lambda h > 0. \tag{A8}$$

But (A8) holds if and only if[40]

$$\delta^i > \frac{\sqrt{(\rho + \lambda - h)^2 + 4\lambda h} - (\rho + \lambda - h)}{2} = \delta^*.$$

Similarly, $\delta^i \leq \delta^i_*$ if and only if $\delta^i \leq \delta^*$. ■

[40]The negative root is not relevant.

3 Surrender Risk in Life Insurance Policies

3.1 Introduction

Many financial securities in the fixed income market have embedded option features. The classical example is perhaps a callable bond, which allows the issuer of the bond a right to redeem the bond at some point before maturity. Another example is mortgage-backed securities (MBS), which often embed a prepayment option, where homeowners can pay off the mortgage loans early to take advantage of possibly lower interest payments through refinancing. These options, albeit implicit in the contracts, explicitly impact the associated cash flows. As a result, understanding such optionality is fundamentally important in pricing and trading such financial instruments.

Indeed, these embedded option features have attracted enormous attention from academics and practitioners alike. For instance, the seminal work of Brennan and Schwartz (1977) provides a pricing framework to value callable bonds, while subsequently Dunetz and Mahoney (1988) and Longstaff (1992), among others, address many practical and empirical concerns. Similarly, mortgage prepayment option is heavily studied, and a very selective list includes Dunn and McConnell (1981), Green and Shoven (1986), Schwartz and Torous (1989), Kang and Zenios (1992), Stanton (1995), LeRoy (1996), and Kalotay et al. (2004).

Some life insurance related instruments also encapsulate salient option features.[41] In particular, cash-value life insurance, such as a whole life insurance policy, embeds a surrender option that gives the policy holder a right to exchange an existing contract for its cash surrender value any time during the life of the contract. In other words, a surrender option is an American-style put option that entitles its owner (the policy holder) to sell back the contract to the issuer (the insurer) at the cash surrender value. Just as the prepayment option imposes a cash-flow risk to MBS investors, this surrender option is a source of concern for life insurers. Specifically, the cash surrender value provides effectively an interest rate guarantee to policy holders. So in theory, when interest rates go up, policy holders may surrender their existing policies to buy new contracts offering higher yields, negatively impacting the insurers. Therefore, it is critical to appreciate this surrender risk by studying the optionality intrinsic to the surrender right.

[41] Indeed, as early as the 1980s, Smith (1982) and Walden (1985) argue that life insurance contracts can be considered as option packages.

Prior research, albeit limited, has explored this issue mainly from a theoretical perspective. Albizzati and Geman (1994) first price the surrender option by considering it as a package of European options and applying a Black-Scholes (1973) type of technique with a Heath-Jarrow-Morton (1992) model of interest rate dynamics. Grosen and Jorgensen (1997, 2000) first consider the full American feature of the option by valuing the surrender right as an early exercisable interest rate guarantee and utilizing developments on American option pricing theory (e.g., Johnson (1983), Karatzas (1988), Kim (1990), Carr et al. (1992), Jamshidian (1992), and Myneni (1992)). Bacinello (2003) also prices the American-style option by employing a recursive binomial formula patterned after the Cox et al. (1979) discrete option pricing model, and so do Tanskanen and Lukkarinen (2003) by extending Grosen and Jorgensen (2000).

What most of these studies share in common is the theoretical assumption of fully rational response of policy holders to interest rates. In particular, given the American feature, the premise is that exercising the option or surrendering occur at an optimal stopping time. However, in reality, whether to surrender an existing policy is a multi-faceted decision, often characterized by irrationality from policy holders (See e.g., LIMRA International (2011)). Analogously, interest rates are only one of the multiple factors that drive the mortgage prepayment decision (See Veronesi (2010), for instance). Indeed, ever since Schwartz and Torous (1989)'s pioneering work in integrating an empirically estimated prepayment function into the valuation framework, it has become common practice among financial institutions to rely on empirics to price MBS. Nevertheless, such integration appears absent in the arena of life insurance.

This paper fills this gap in the literature by incorporating an empirical surrender function to price the surrender option embedded in life insurance policies. I first model surrendering econometrically, asking what factors drive actual surrender activity. I then construct an interest rate binomial tree and apply a recursive technique for option pricing. This procedure allows me to price the surrender option under both the assumption of full rationality and the consideration of the empirically estimated surrender function, examining how the option prices compare between the two frameworks. Finally, I analyze some comparative statics, exploring how the option prices behave in different interest rate and industry environments.

Although this is to my knowledge the first paper to integrate empirics into the

valuation of surrender options, there are some previous studies that investigate empirically the surrender activity. Two competing hypotheses have been proposed. One is the so called "Emergency Fund Hypothesis" which actually dates all the way back to Linton (1932), suggesting that surrender activity should increase during periods of economic duress. The other is the so called "Interest Rate Hypothesis" emerging from a New York Life (1986) survey of surrendering policy holders which sees a common reason for surrendering from respondents as a better value on another insurance policy due to higher yield.

However, evidence on this debate is mixed. Dar and Dodds (1989) use UK aggregate data from 1952 through 1985 and find a positive relation between surrender activity and unemployment, but no such relationship with interest rates. Outreville (1990) employs North American macro data for the period of 1955 to 1979, and finds that unemployment has a significantly positive effect while income has a significantly negative effect, but again no significant relationship with interest rates. Nevertheless, more recently, Russell et al. (2013) exploits state-level variation from 1995 through 2009 in the US and find that interest rate variables are significantly positively related to policy surrender, while there is a negative relation between real per capita income and surrender activity, yet counter-intuitively, policy surrender seems to decrease with unemployment.

This paper, on the other hand, provides unifying evidence on both hypotheses. Indeed, I find policy surrender is significantly positively associated with interest rates, supporting the Interest Rate Hypothesis. I also find that policy surrender is significantly positively related to unemployment but negatively related to income, supporting the Emergency Fund Hypothesis.

The data that I utilize come from a large industry experience study in the US spanning 2001 to 2010. An improvement of these data compared to prior country- or state-level data is the consideration for policy vintages. In other words, while country- or state-level data have to lump together policies originated in different years to calculate a surrender rate for this year, our industry experience data can track these policies with different vintages over time, making each surrender rate specific to one particular year of origination. This feature of the data allows me to discover another factor greatly impacting surrender activity, policy vintage, that previous studies cannot capture. Indeed, I find, for the first time, that the first policy year has a significantly positive impact on surrender, and its economic magnitude is

even higher than these other factors. Moreover, with this additional factor, the paper is able to explain more than 80% of the variation in the surrender activity.

Using these empirics, I calculate and compare the value of the surrender option under full rationality with that based on actual surrender experience. I find that the experience-based option value is substantially lower than its fully rational counterpart. This reflects suboptimality in actual surrender activity. Moreover, while the fully rational option value is always positive, the experience-based option value can sometimes be negative, in which case it indicates that life insurers can make a profit from the option exercising even though they are the writers of the option.

Finally, the competitive landscape of the life insurance industry and the interest rate environment both have prominent impact on the value of the surrender option. A less competitive market allows the insurers a greater margin on their products, which reduces the cash flows to the policy holders upon surrendering the policy and hence results in a lower option value, which this paper illustrates. Higher interest rates, however, effectively increase the moneyness of the option, raising its value, which is also confirmed in my findings.

The paper proceeds as follows. In Section 3.2, I provide some information about the life insurance market, detailing the different types of insurance policies and pointing out the focus of this paper. Section 3.3 introduces the various components of the policies that are relevant for the surrender option and highlights the interest rate binomial model that is the centerpiece of the pricing apparatus. In Section 3.4, I discuss the data and the key parameters, illustrate the flow of the methodology, and present the results. Section 3.5 concludes. All tables and figures are in the Appendix.

3.2 Life Insurance Market

The primary reason that people buy life insurance is to protect their dependents against financial hardship when the insured person (the policy holder) dies. Many life insurance products also allow policy holders to accumulate savings that can be used at a later time. This economic protection appears attractive to most American families, as 70% own some type of life insurance (LIMRA International 2011).

Indeed, the sheer volume of life insurance policies makes it a vast market. Americans purchased $2.8 trillion of new life insurance coverage in 2013, and by the end of the year, total life insurance coverage in the U.S. was $19.7 trillion (American Council of Life Insurers (2014)). To put this in perspective, issuance of all mortgage-related

securities totaled $1.97 trillion in 2013, with $8.72 trillion outstanding at year end (Securities Industry and Financial Markets Association (2014)), which is already huge compared to mere $55 billion for the same year in the US IPO market (Renaissance Capital (2014)).[42]

There are three major types of life insurance policies. First is individual insurance, which is underwritten separately for each individual who seeks insurance protection. Second, group insurance is underwritten on a group as a whole, such as the employees of a company or the members of an organization. Third, credit insurance guarantees payment of some debt, such as a mortgage, in the event the insured person dies, and can be bought on either an individual or a group basis.

Individual life is by far the most widely used form of life insurance protection, accounting for 58 percent of all life insurance in force in the U.S. at year-end 2013 (American Council of Life Insurers (2014)).[43] Individual life policies offer two basic types of protection: covering a specified term, or permanently covering one's whole life. The term insurance policies provide life insurance coverage for a fixed period, but no further benefits when the term expires or build-up of cash value, and therefore, the issue of surrendering is typically irrelevant for term insurance. Permanent life insurance, however, provides protection for as long as the insured lives. In addition, permanent life policies accumulate cash value that a policy holder can exchange for the policy's death benefit (the next section will illustrate this surrender feature further).

There are four types of permanent life insurance policies: whole life (WL), universal life (UL), variable life (VL), and variable-universal life (VUL). Table 1 presents the main differences. The premium of WL policies is fixed and the insurer guarantees a fixed death benefit since it is the insurer who makes the investment choice using the insurance premium. The insurer also makes the investment choice providing a fixed benefit with UL, but allows a flexible premium payment schedule. In contrast, for VL, the benefit and cash value vary subject to the performance of a portfolio of investments chosen by the policy holder (a stock index, for instance), and finally, VUL combines the flexible premium payment options of UL with the varied investment options of VL.

The surrender risk is more relevant in WL and UL policies, since it is the insurers

[42] 2013 was in fact the best year for US IPO market in over a decade, with a total of 222 companies gone public.

[43] It has also grown steadily, from a total coverage of $9.7 trillion in 2003 to that of $11.4 trillion in 2013, averaging an annual rate of 1.6 percent.

who bear the risk of financial market fluctuation when they guarantee the benefits to the policy holders. Between WL and UL, WL policies are more popular[44] and the fixed-premium structure simplifies the analysis. Therefore, I focus on WL policies in the subsequent sections of this study.

3.3 Pricing Model

3.3.1 Formulation

Whole life insurance policies provide policy holders with death benefits. Denote the face amount of this death benefit by F. Assume that the insurance premium is paid lump-sum at the inception of the contract.

Consider an age-gender profile $c = (a, g)$, where a is the age at which the policy holder first purchases the whole life insurance policy. The most common value is $a = 35$ and $g = male$, which will be the case in this paper. Then let t denote passage of time (in years) relative to a, i.e., since the inception of the contract. Denote by $\mu^c(t)$ the probability that the insured with profile c dies between times $t-1$ and t conditional on that the policy holder has survived until time $t-1$. This probability is provided in a standard mortality table. In Section 3.4, the 2001 Commissioners Standard Ordinary (CSO) Mortality Table is used, which is the current industry standard (American Council of Life Insurers (2014))).

The *cash surrender value* of this policy at time s (again relative to when the policy holder is at age a, similarly henceforth), $U^c(s)$, is the sum of the present expected value of future benefits where the discounting uses f, a fixed annual interest rate set by regulation (Towers Watson (2014)). That is,

$$U^c(s) = F \sum_{t=1}^{T-s+1} \{\mu^c(s+t) \prod_{i=1}^{s+t-1} [1 - \mu^c(i)]\} \exp(-ft), \tag{54}$$

where T is the difference between the terminal age of the standard mortality table, 120 (due to the fact that the standard mortality table has the probability of death as 1 for age 120 and over), and a. In other words, T is the maximal policy years possible. Given that we focus on $c = (35, male)$ in this paper, we use the mortality table of such profile, although the entire procedure can be easily generalized to any other age-gender combination.

[44] See Hoopes (2015).

Upon surrendering, the policy holder with the original profile c can purchase a new whole life policy at the then market rate to obtain the same level of death benefit. To do so, the policy holder pays (again, WLOG, assume that the insurance premium is paid lump-sum) the new premium at time s

$$P^c(s) = \lambda F \sum_{t=1}^{T-s+1} \{\mu^c(s+t) \prod_{i=1}^{s+t-1} [1-\mu^c(i)]\} \exp[-R(s,s+t)t], \tag{55}$$

where $R(s, s+t)$ is the interest rate at time s for the period from s to $s+t$. In other words, the insurance premium is the actuarially fair value of future benefits (the sum of the present expected value of future benefits) multiplied by a "mark-up" or (gross) margin that the insurer charges, $\lambda > 1$, and the discounting now uses the ongoing annual continuously compounded interest rates at time s. λ can be interpreted as an "inverse" of the expense ratio (see for example, Cummins and VanDerhei (1979)), which is a direct consequence of the operational efficiency of the insurer but fundamentally reflects the competition among insurers. In other words, λ is the parameter that represents the competitive landscape of the life insurance industry.

At time s, information about the interest rates is known, that is, we know $R(s, s+t)$. However, at time 0, one does not know such rates with certainty. Consequently, when calculating option values at time 0, we need to take expectation of these future insurance premiums. We do so via an interest rate binomial tree, which we now introduce.

3.3.2 Interest Rate Tree

A risk-neutral interest rate binomial tree is built based on the Ho-Lee model (Ho and Lee (1986)). This choice is based on its popularity, simplicity, and ability to exactly fit the term structure of interest rates on the one hand, and the fact that it allows substantial (risk-neutral) probability mass to low interest rates, likely to perform well in low interest rate environments (better than for instance the Black, Derman and Toy model (Black, Derman, and Toy (1990)); see Veronesi (2010)) on the other hand. Given the low interest rates in recent years, it is desirable to have such feature in an interest rate model.

Let us first represent a point on the tree as a pair (i, j), where i is the time index and j is the node index. Following convention, an upward movement is described by

an increase in the index i, but not in the index j. A downward movement, however, is described by both an increase in the index i and that in the index j.

Let $r_{i,j}$ be the continuously compounded interest rate in node j between steps i and $i+1$. Then, for every (i,j) the Ho-Lee model postulates that

$$r_{i+1,j} = r_{i,j} + \theta_i \Delta + \sigma\sqrt{\Delta} \text{ with RN prob. } p^* = 1/2, \tag{56}$$

as well as

$$r_{i+1,j+1} = r_{i,j} + \theta_i \Delta - \sigma\sqrt{\Delta} \text{ with RN prob. } p^* = 1/2, \tag{57}$$

where θ_i are free parameters that are chosen to fit exactly the current term structure of interest rates (which will be determined in the numeric section), σ is a volatility parameter based on historical interest rate data, and Δ is the time interval (a year in our case). Clearly, the tree is recombining, as an "up and down" movement in interest rates leads to the same level as a "down and up" movement.

3.3.3 Insurance Premium Calculation via Interest Rate Tree

Let $P^c_{i,j}$ denote the value of the insurance premium at point (i,j) on the tree given profile c. We make the standard assumption of independence between mortality and financial markets (See Albizzati and Geman (1994), for example). Then at this point, there is a probability of $\mu^c(i)$ that the policy holder dies, and thus the insurer pays out the face amount F.

If the policy holder lives, the probability of which is $1-\mu^c(i)$, the continuation value of the policy's death benefits is

$$\exp(-r_{i,j}\Delta)\mathrm{E}^*[P^c_{i+1}] = \exp(-r_{i,j}\Delta)(\frac{1}{2}P^c_{i+1,j} + \frac{1}{2}P^c_{i+1,j+1}),$$

where the expectation is with respect to the risk-neutral probability measure. Therefore, we have

$$P^c_{i,j} = \lambda\{\mu^c(i)F + [1-\mu^c(i)]\exp(-r_{i,j}\Delta)(\frac{1}{2}P^c_{i+1,j} + \frac{1}{2}P^c_{i+1,j+1})\}.$$

At the terminal age of 120, even though the policy holder may still live, for actuarial purpose, we treat it as payout of the face amount with certainty. Thus,

$$V_{I,j} = F \text{ for all } j.$$

We can then construct the tree for the insurance premiums following a backward recursive procedure. We defer to Section 3.4.2 for an illustration of such procedure.

Intuitively, the insurance premium, being the scaled actuarially fair value of future benefits, is analogous to a (non-callable) coupon bond with mortality-adjusted coupon payments. Similarly, just as one would need to introduce the optionality to price a callable coupon bond, we next introduce the corresponding optionality to price the surrender option under the condition of fully rational exercising.

3.3.4 Option Pricing - Fully Rational Exercising

To price the surrender option, we first consider the case that the exercise of the option is fully rational in response to changes in interest rates. Similar to a rational homeowner who closes a mortgage only to refinance it at a lower mortgage rate, a rational policy holder surrenders an existing policy only when he/she can find another policy with the same amount coverage but with more favorable terms. However, again just as a homeowner should not necessarily refinance the mortgage the first instant the ongoing mortgage rate is lower than that of the current mortgage, it is not always optimal for the policy holder to surrender the policy just yet when the ongoing interest rates become more attractive.[45] That is, there is value for waiting. Essentially, the surrender option is an American option,[46] and this American feature fits well with the recursive procedure of pricing with an interest rate tree.

Formally, let $V_{i,j}$ denote the value of the surrender option at point (i, j) on the tree. At this point, the policy holder can decide whether to exercise the option or wait. If he/she exercises, the payoff is

$$V_{i,j}^{Exer} = U_i^c - P_{ij}^c.$$

Note that the cash surrender value depends only on i but not on j, since its schedule is fixed since the inception of the policy. If the policy holder waits, the value for doing so is

$$\begin{aligned} V_{i,j}^{Wait} &= \exp(-r_{i,j}\Delta)\mathrm{E}^*[V_{i+1}] \\ &= \exp(-r_{i,j}\Delta)(\frac{1}{2}V_{i+1,j} + \frac{1}{2}V_{i+1,j+1}). \end{aligned}$$

[45] Recall that the cash surrender value is obtained by discounting using the statutory rate, whereas the insurance premium uses the ongoing interest rates. Consequently, an increase in rates incentivizes a policy holder to cash out an old policy for a new policy, potentially pocketing the difference.

[46] Strictly speaking, the surrender option is a Bermudan option, since although the policy holder can exercise this option at any time, the actual payout of the cash surrender value will typically occur at the next policy anniversary.

In order to maximize the value of the surrender option, the policy holder should choose between exercising and waiting such that

$$V_{i,j} = \max(V_{i,j}^{Exer}, V_{i,j}^{Wait}). \tag{58}$$

Moreover, since the option expires worthless at maturity, we have that at the terminal of the tree I,

$$V_{I,j} = 0 \text{ for all } j.$$

Given this final value of the option, we can conduct a backward procedure in equation (58) to obtain the option value at time 0. Such procedure and its results (the tree of option values) are presented in Section 3.4.2.

3.3.5 Option Pricing - Experience-Based Exercising

As Veronesi (2010) points out, the homeowner's prepayment decision depends on a variety of factors additional to the interest rates. Examples abound, one of which is the sale of the house as a result of a change in job location. Analogously, the surrender decision of a policy holder is driven by a number of variables on top of the interest rates. Section 3.4.2 will detail those that are explicitly considered in this paper. Here let us just layout the schematics for how to price the option when we incorporate these factors.

Let $p(\mathbf{X}, i)$ denote the probability that a policy holder will surrender his/her policy at time period i. It is a function of both time and a vector $\mathbf{X}$ of variables, interest rates and beyond. What this probability function entails is the timing of the the exercise of the surrender option, which in turn determines the resultant cash flow upon exercise.

Formally, if the option is exercised at time i, the payoff to the policy holder is

$$U^c(i) - P^c(i), \tag{59}$$

and the date-0 value of the option is therefore

$$\mathrm{E}^*\{e^{-(r_0+r_1+r_2+\ldots+r_{i-1})\Delta}[U^c(i) - P^c(i)]\}.$$

We now state two observations. First, the sign of (59) can be positive or negative. In other words, in contrast to the fully rational case in the previous subsection where the policy holder cannot lose money by exercising the option, here it is quite

possible. Indeed, conversation with major life insurers suggests that during the early years of the policy, it is detrimental to policy holders to surrender their policies, yet substantial amount of exercising does occur, which is confirmed in the data (See Section 3.4.1). Second, we abstract away the dependency of r, U and P on the nodes (the j subscripts) of the risk-neutral tree, although we are still employing the tree to determine such quantities, through Monte Carlo simulation (which is also used to evaluate the expectation). This will become more clear in Section 3.4.2.

3.4 Numerical Analysis

3.4.1 Data and Parameter Values

We first need term structure of zero coupon rates to calibrate our Ho-Lee tree. I obtain such data via the stripped curve on Bloomberg. I consider two start dates, Jan 2, 2014 and Jan 3, 2000, to represent two (indeed drastically different) interest rate environments. For maturities 30 years or less, the data are readily available using the treasuries, where I follow the same intrapolation technique used by Bloomberg for maturities that do not have direct treasury correspondence. For even longer end of the curve (recall that we need as far out as 85 years, the difference between 120 and 35), Bloomberg relies on swap rates and sometimes has coverage up until 60 years of maturity. For further maturities, I employ a moving-average extrapolation technique following Armstrong and Forecasting (1978). As an illustration, Figure 1 presents the term structure on Jan 2, 2014, which shows an upward-sloping yield curve, with rates low around 0.30% at the near end and high around 4.03% at the far end.

The volatility parameter σ is set to 0.0173, as in Veronesi (2010), who in turn estimates such value based on historical interest rate data, and recall that the time interval Δ is one year in our case. Then the calibration involves solving, iteratively, the parameters θ_i to fit exactly the term structure of interest rates. Table 2 presents the calibrated tree along with the solved θ values for the start date of Jan 2, 2014 (for the first ten periods).[47]

For policy parameter values, μ is provided in a standard mortality table, for which I use the 2001 CSO Table as pointed out in Section 3.3.1. f, the statutory rate, is set by regulation, and obtained from Towers Watson (2014). F is arbitrary (it is simply

[47]The tree for the start date of Jan 3, 2000 is contructed in completely the same way and not shown here to conserve space, but it is available upon request.

a scale parameter), and WLOG it is set to \$1,000. λ varies among insurers, on which we will conduct sensitivity analysis.[48]

To empirically estimate the surrender probability p, the most critical information one would need is surrender activity data. Such data should possess at least one key attribute, which is time-series variation in the sense that one can track each particular vintage of policies over the years. Ideally, it is preferable to have policy-level data, which would indeed tackle this issue perfectly. However, policy-level data are not available due to its proprietary nature, not to mention the difficulty in data collection (and often times the lack of).

Indeed, we face the same issue in mortgage-back securities, and the literature has proposed the use of aggregate data, such as Schwartz and Torous (1989). This paper follows such idea. However, we face a new problem: most aggregate data, country- or state-level (for example, see Russell et al. (2013)), do not enable us to track policy vintages. In other words, those data are aggregated in such a way that, for instance, a surrender rate in a particular state of a given year reflects the surrender activities across all vintage years.

How to solve this problem? The answer is industry experience studies, which are aggregate data as they are aggregated across insurers, yet one of their purposes is to keep track of policy vintages. This paper uses one large industry experience study that spans the years 2001 to 2010. In addition to such time-series variation, the data also provide cross-sectional variation across different age groups, allowing us to employ fixed-effect estimation. Figure 2 presents the data graphically. We first note that there is indeed some across-group variation. Second, for all age groups, surrender activities tend to concentrate in the first year after inception of the policy and level off after year 5. We will analyze these patterns in detail later in this section.

For variables that are likely to have an impact on surrender decisions, data are available from the following sources. Historical treasury rates come from the Wall Street Journal. Historical unemployment rates come from the Bureau of Labor Statistics. Historical real per capita income levels come from the Census Bureau. To price the surrender option at a more recent date, one would also need projections of these variables. The Congressional Budget Office provides such projections.

Table 3 provides the summary statistics. We have several observations. First, there is substantial variation in the surrender rates, as even though they average to

[48] Conversation with major life insurers suggests that it ranges between 1.05 and 1.20.

about 5%, some policy years see surrender activity as high as 23%. Second, variation in the macroeconomic factors is fairly high too, albeit not as drastically as policy surrender. Third, statutory rates over this period have been quite stable, residing in the range of 4% to 4.5%.

3.4.2 Methodology

Recursive Procedure

We now illustrate the backward recursive procedure introduced in Section 3.3, using Table 4, which is an excerpt from the whole option price tree assuming full rationality. We start from the terminal period, that is, 85 in our case. Again, since the option expires worthless at maturity, we have the value of the option as 0, the same across all interest rate possibilities in this period, as shown in Panel B.

The interest rates from Panel A in period 84, however, are each used to discount the values in period 85 that correspond to an up movement and a down movement along the interest rate tree from that interest rate. For example, at node $j = 0$ (which is the case that rates have gone up every period), we have

$$V_{84,0}^{Wait} = \exp(-r_{84,0}\Delta)(\frac{1}{2}V_{85,0} + \frac{1}{2}V_{85,1}),$$

which is obviously 0.

The payoff from exercising the option at the same point on the tree, $V_{84,0}^{Exer}$, is 878.61 (determined using a tree of insurance premiums and recall that the schedule of the cash surrender values is fixed) for the case where $\lambda = 1.2$. Then,

$$V_{84,0} = \max(V_{84,0}^{Exer}, V_{84,0}^{Wait}) = 878.61.$$

Similarly, we have $V_{84,1} = 874.40$, reflecting again the cash surrender value and the insurance premium at this point. Following the same procedure backward and recursively, we have $V_{83,0} = 972.87$, $V_{82,0} = 973.63$, and so on, until we arrive at $V_{0,0} = 9.03$.

Empirical Estimation of Surrender Rates

The empirical literature on policy surrender, albeit limited, has proposed several factors that drive surrender activity. In fact, there are two competing hypotheses that draw the focus. One is the so called "Emergency Fund Hypothesis" (EFH). It links life insurance surrender activity to an urgent need for funds during a time of crisis or need. In other words, the household is more likely to surrender its life

insurance policy when the need for funds is high such as during unemployment. The other is the so called "Interest Rate Hypothesis" (IRH). It states that life insurance surrender activity is directly related to the differential return offered by ongoing market interest rates over life savings products. Based on this proposition, one would expect an increase in interest rates to cause an increase in policy surrender. Prior studies contribute to this debate by providing mixed evidence.

New York Life (1986) surveys surrendering policy holders during a period of high interest rates.[49] It finds that the most common reason for surrendering was a better value on another insurance product, and hence supports the IRH. Its evidence on EFH is somewhat vague however, with only one-third of the respondents said they surrendered their policies because "family circumstances had changed."

Dar and Dodds (1989) explore the relationships amongst interest rates, unemployment, and surrender activity in endowment life insurance policies in the United Kingdom.[50] Using data from 1952 through 1985, they find a direct relation between surrender activity and unemployment, but no such relationship was identified between interest rates and surrender activity. Based on this finding, they concluded that emergency cash needs drive surrender activity, supporting the EFH.

Outreville (1990) analyzed the effects of macroeconomic variables on early lapsation using U.S. and Canadian data from the period 1955–79. He finds that unemployment has a significantly positive effect on early lapsation while personal income has a significantly negative effect. However, similar to Dar and Dodds (1989), he finds no such significant relationship in terms of interest rates, and so he also concludes in favor of the EFH.

More recently, Russell et al. (2013) utilize state-level aggregate data from 1995 through 2009 to test these hypotheses. Life insurance surrender activity data for each state are obtained from the National Association of Insurance Commissioners. They find that interest rate variables are significantly positively related to policy

[49]The survey was completed during a timeframe where short-term interest rates were at extraordinarily high levels (e.g., money market interest rates in excess of 12 percent), so the results clearly show the impact of large interest rate changes on consumer behavior, with fifty-one percent of the respondents who lapsed bought another policy with better value.

[50]Endowment policies pay a fixed amount to the policy holder if he/she lives to the maturity date. If the policy holder dies prior to maturity, the beneficiary receives a death benefit. In recent years, endowment products have nearly vanished from the US life insurance scene because changes in US tax laws (the Deficit Reduction Act of 1984) drastically reduced the attractiveness of this policy as a tax shelter.

surrender and show a negative relation between real per capita income and surrender activity. In other words, at the state level, there is evidence for both the IRH and the EFH. However, surprisingly, they also find that policy surrender decreases with unemployment, which is counter-intuitive and in contrast to expectations from the EFH. Therefore, evidence for EFH remains mixed.

As mentioned in the previous subsection, the aggregate data (country- or state-level) in prior studies cannot track policy vintages, as even at the state-level, one data point of surrender rate in a particular state of a given year reflects the surrender activities across all vintage years. However, from Figure 2, the effect of policy vintages, especially that of the first year, seems quite pronounced. The industry experience study data employed in this paper allows us, for the first time, to capture such effect.

I model life insurance surrender activity as a function of the households' liquidity needs for cash, interest rate arbitrage opportunities, and policy vintages:

$$Surrender = f(Liquidity, Arbitrage, Vintage).$$

Proxies for liquidity needs are unemployment rates and real per-capita income levels. Interest rate arbitrage opportunities are proxied by the differential between (short- and long-term) treasury rate and the regulatory rate provided in the insurance policy. Policy vintages are the number in years since policy inception where both an emphasis on the first year and a year trend will be explored.

More specifically, as unemployment increases, a greater proportion of the insured would likely surrender cash-value life insurance policies to gain access to the cash surrender value, and we therefore expect a positive relation between unemployment rate and surrender activity. Prior research provides evidence that life insurance demand increases as income increases, suggesting that the insured would be less likely to surrender policies (See Zietz (2003)). So we expect that as real per capita income decreases, individuals will be more likely to surrender, and thus a negative relation between income and surrender. For interest rates, since the IRH predicts a positive relation between interest rates and policy surrender, we expect the same as intuitively, the higher the ongoing interest rates relative to the regulatory rate are, the more incentivized policy holders are to effectively "re-finance" their policies. Furthermore, based on what we observe in Figure 2, we expect the first policy year to have a strong positive effect on surrender, whereas the impact of time trend does not appear too substantial.

Finally, the panel data enables us to include age-group fixed effects. This has two potential benefits. First, it may allow us to explain more of the variation in the data. Second, we can laser on one particular age group when we predict the surrender rate (for instance the age group closest to the age 35) in order to price the option related to the insurance policy with that age profile. It is likely important, since people in different age groups presumably make decisions differently given their different stages in the consumption/investment life cycle, and so this may capture such demographic effect.

Monte Carlo Simulation

Monte Carlo simulation in finance, since the pioneer work by Boyle (1977), has become a standard procedure for pricing relative complex financial securities. One uses computer programs to simulate several interest rate scenarios in the future, and then obtain the value of the security by averaging an appropriate discounted value of the payoff. I use Monte Carlo simulation to price the surrender options where exercising is experience-based. The rationale is to capture the additional factors above that impact the policy holder's surrendering decisions beyond interest rates. The key is to simulate such surrender decisions over time.

To illustrate the procedure, let us first recall our (risk-neutral) Ho-Lee interest rate model in Equations (56) and (57). Because the interest rate process goes up or down with equal probabilities, one can simulate the paths on the tree by utilizing a random number generator. For instance, consider a random number generator based on the uniform $(0, 1)$ distribution. Then we can simulate a large number of times, and each time the realization is on one side of 0.5 we say that the interest rate moved up the tree (i.e., $+\sigma\sqrt{\Delta}$), and each time the realization is on the other one side of 0.5 we say that the interest rate movement is negative (i.e., $-\sigma\sqrt{\Delta}$). We can record each path, along which we can discount the corresponding cash flows. Then given all of the simulation paths, we average the discounted payoffs across these paths to obtain the security price (in our case, the price of the surrender option).

As an example, in Table 5, I first provide an excerpt of the simulated interest rate paths (10 simulations from year 0 to year 10) in Panel A. Then we can simulate a discount $Z^s(0, T_i)$ for each simulation path s and for each time period $T_i = 1, 2, ...,$

in Panel B, as[51]

$$Z^s(0,T_i) = e^{-(r_0+r_1^s+r_2^s+\ldots+r_{i-1}^s)\Delta},$$

which can be thought of as one realization of the discount factor $Z(0,T_i)$, that is, the risk-neutral expected discounted value of \$1 at time T_i. In other words, from risk-neutral pricing, we have

$$Z(0,T_i) = \mathrm{E}^*[e^{-(r_0+r_1+r_2+\ldots+r_{i-1})\Delta} * \$1]. \tag{60}$$

Given all of the simulation paths, for each time period T_i we can compute the average discount across simulations, that is,

$$\widehat{Z}(0,T_i) = \frac{1}{N}\sum_{s=1}^{N} Z^s(0,T_i),$$

where N is the number of simulated paths. This is an approximation obtained by Monte Carlo simulations of the true discount factor (60).

One can use these simulated discounts to price the zero coupon bonds. Since this binomial tree is such that the prices of zero coupon bonds computed from it exactly match the data, the simulated zero coupon bond prices should be quite close to those in the data. We indeed obtain close results in Panel C, which serves as a sanity check on the calculation procedure. In addition (not reported here for consideration of space), we determine the number of simulations needed based on this closeness between simulated zero coupon bonds and data, and $10,000$ seems sufficient, although with enough computing power, this study uses $30,000$ simulations.

Finally, to simulate surrender decisions over time, we employ another random number generator and compare it with the empirical surrender function. Specifically, for each period, the random number generated can be either less than or not less than the estimated surrender rate of that period. Now the procedure is such that the first instance the former case occurs, the policy is surrendered. This determines the timing of surrender as well as the corresponding cash flow. The rest of the pricing procedure is then similar to the above.

[51] Note that the last interest rate used to discount a payoff at time T_i (i.e., period i) is the one corresponding to the previous period, $i-1$, as there is a lag of one period between the cash flow and the interest rate needed to discount it. For instance, the first interest rate r_0 is used to discount a cash flow at $i = 1$.

3.4.3 Results

Empirics

I consider both short-term (one-year) and long-term (ten-year) treasury rates as our proxies for interest rates.[52] I report only results related to the ten-year rates while those for the one-year rates are available upon request. I focus on the ten-year rates for three reasons. First, the results for the one-year rates are in fact stronger, with higher statistical significance and expected signs on all variables of interest. So we are being more conservative by using the results for the ten-year rates. Second, the Congressional Budget Office provides projections on the ten-year rates but not the one-year rates, and thus we want to stay close to their estimation. Third, life insurers, being long-term participants of the financial market, tend to focus on the longer end of the curve.

Table 6 presents the results with the three macroeconomic variables that are the focal points of previous studies, namely, interest rate, unemployment rate, and real per capita income. Similar to Russell et al. (2013), we find the coefficient associated with interest rate is positive and highly significant, while that associated with real per capita income is negative and highly significant. Unlike Russell et al. (2013), we find a positive coefficient of unemployment rate (when all three variables are present), albeit insignificant.

We further include age-group fixed effects in our analysis given the panel nature of the data, as Table 7 reports. Coefficients and significance remain the same as before, although the fitness of the model does increase substantially, for example, from 43.5% in Specification (7) of Table 6 to 53.0% in Specification (7) of Table 7. Moreover, as mentioned above, incorporating fixed effects allow us to capture the difference among age-groups, which is important since each insurance policy is after all age-specific.

The most striking improvement comes with the inclusion of policy vintage effect, especially the first year. As Table 8 shows, for instance Specification (4), not only do interest rate and real per capita income remain significant with expected signs, unemployment rate also becomes significant (and positive, as the EFH suggests). In addition, *Year1*, the variable representing the first policy year, is highly significant

[52] The 30-year treasury bond offers the longest maturity among treasury securities and was commonly used as a proxy for long-term interest rates. This role has largely been taken over by the 10-year note, however, as the size and frequency of the 30-year bond issues declined significantly in the 1990s and early 2000s.

and positive. The economic magnitude of this variable is also especially prominent, as the first policy year sees on average approximately 8.2 percentage points more likelihood of surrender. This reflects the conspicuous pattern observed in Figure 2. Furthermore, along with the age-group fixed effects, the model with the first policy year is now able to explain 82% of the variation, which gives us confidence in using the empirically estimated surrender rates in our pricing model. Additionally, year trend does not seem to provide extra mileage and in fact adds more noise to the analysis. Indeed, observing Figure 2 again does not seem to give us indication of the presence of time trend. Therefore, we use Specification (4) of Table 8 for pricing consideration in the next subsection.

All in all, the fact that our industry experience data allow us to capture the policy vintage effect improves upon existing literature in at least two aspects. First, it captures an explanatory variable that appears even more important than all the marcoeconomic factors considered in prior studies. Second, with the inclusion of such effect, all these marcoeconomic factors are now statistically significant, providing full support to both the EFH and the IRH.

As for why policy vintage has such a strong impact, anecdotally, it seems attributable to high pressure agents. Their powerful sales techniques appear capable of bringing new customers that are not yet familiar with the nuts and bolts of life insurance. Some of these customers are not sure if they really want life insurance or do not find their current policy favorable. Whether this is indeed the case is outside the scope of this paper, and we leave it to future research to examine more closely the reason behind the importance of the policy vintage effect.

Option Pricing

Experience-Based Option Price and Sensitivity to the Insurer's Mark-Up

With the empirics above, we can estimate the surrender rates for different policy years using the macroeconomic factors and the policy vintage effect. Specifically, with the projections from the Congressional Budget Office, one can predict surrender rates for the next ten policy years (2015 to 2024). Given the flatness in surrender activity in the later years of Figure 2, we assume similar flatness beyond policy year

10.[53] This is similar to the practice of using the Public Securities Association (PSA)[54] experience to price MBS (See for example, Hayre (2002)). The estimated surrender rates are presented in Figure 3 for a policy whose holder at policy inception is within the age group of 35 to 39.

As mentioned before, based on these estimated surrender rates, we can simulate the surrender decision by generating a series of random numbers for each interest rate path. Then along one path, whenever the first random number is below the surrender rate for that period, the policy is surrendered. The payoff of the option for that path is the difference at the time of surrender between the cash surrender value and new insurance premium, discounted by the sequence of interest rates along the path. Then, the date-0 value of the surrender option is obtained by averaging across all simulated paths.

The mark-up factor, λ, in equation (55) turns out to play a crucial role in the pricing. I present the option values as a function of λ, in Figure 4, using the common levels suggested by the industry, 1.05 to 1.20. A direct observation is that the option value decreases with λ. This is intuitive, as what the mark-up does effectively is to absorb the cost of surrendering (from the insurer's perspective) in addition to making a profit. In fact, when λ is sufficiently high, actually just about 1.06, given the recent interest rate environment (the Jan 2, 2014 term structure), the option value can be negative. In other words, with sufficiently high λ, the experience-based option value is completely priced in, and it imposes minimal surrender risk to the insurers.

Comparison with Fully Rational Option Price

To put the above experience-based option values into perspective, let us now compare them with the fully rational option prices. Figure 5 presents the case. The fully rational option value also decreases with λ, from about \$15.84 for $\lambda = 1.05$ to about \$9.03 for $\lambda = 1.20$,[55] although not as dramatically as that for the experience-based option value. Also for each level of λ, the experience-based option value is lower than the fully rational counterpart. Moreover, the fully rational option value is always positive, which is intuitive given the optimality in option exercising. However,

[53] We use again the moving-average extrapolation technique (Armstrong and Forecasting (1978)), although assuming a straight flatness provides similar results.

[54] It was renamed to Bond Market Association and later merged with the Securities Industry Association to form the Securities Industry and Financial Markets Association.

[55] The scale of the axis makes this decrease less discernible, although zooming in, the change is still substantial percentage wise.

this does illustrate the limitation of a fully rational pricing model. It leads to inflated option prices due to lack of consideration for the additional factors beyond interest rates that drive the surrender decision as well as for the prevalence in the data of sub-optimal option exercising.

Consequently, as a practical policy implication, an insurer considering hedging the surrender risk embedded in their insurance contracts should think beyond the theoretical fair valuation of the surrender options proposed in prior studies. In particular, within a low interest rate environment such as recently, there does not appear much hedging need. Of course, the financial market may recognize this by offering really discounted hedging instruments like high-strike interest rate caps, but investing in such instruments can still be a waste of resources.

Now again, how much surrender risk a life insurer exposes to depends closely on its λ. The life insurer certainly would like to set a high λ, but how high is possible is beyond the insurer's control due to the competitiveness of the market. Indeed, as hinted above, what λ represents is the competitive landscape of the whole industry. Therefore, during episodes of market expansion (such as those of the 1920s and the 1960s, see Obersteadt et al. (2013)), it is advisable that life insurers be more alert for hedging.

Changes in Interest Rate Environment

The recent interest rate environment (as represented by the term structure of Jan 2, 2014) is characterized by low rates at both the near end and the far end of the curve. These are likely results of the Federal Reserve's monetary policy as well as the quantitative easing programs, intended to battle the aftermath of the Great Recession. However, historically, regimes of high interest rates are evident (see, for example, Figure 6 of the ten-year treasury rates between 1962 and 2012). How do changes in interest rate environment impact our surrender option prices?

To address this question, we obtain the term structure of another date, Jan 3, 2000, as mentioned in the data section, which contrasts quite substantially with that of Jan 2, 2014. Indeed, one can see from Figure 7 while the general upward-sloping shapes of the yield curves between the two dates are fairly similar, the levels are drastically different. On Jan 3, 2000, even the very near-end of the curve saw an interest rate that is higher than that at the very far-end of the curve on Jan 2, 2014.

Intuitively, higher rates effectively make it more likely that the option is "in-the-money." So we expect higher option values associated with the term structure of Jan

3, 2000 than that of Jan 2, 2014. Figure 8 compares the fully rational option prices between the two rate environments across different levels of λ. Clearly, option values based on the 2000 term structure are strictly higher than those based on the 2014 term structure.

We expect the same for experience-based option values. To calculate those associated with the 2000 term structure, we need surrender rates across different policy years. Recall that with the 2014 term structure, we relied on projections from the Congressional Budget Office. After all, we do not have data on interest rates, real per capita income, etc. for future years yet. For the 2000 term structure, we have those data readily available, and so we do not need to use projections. Instead, we use actual data to estimate the surrender rates. Then using these estimated surrender rates, we can price the experience-based option prices for the 2000 term structure in the same way as those for the 2014 term structure, and Figure 9 illustrates the comparison. Similar to the fully rational case, option values based on the 2000 term structure are strictly higher than those based on the 2014 term structure. Moreover, for those of the 2000 term structure, it now takes a λ of 1.18 to have a negative option price. In other words, in the high interest rate environment during the early 2000s, it was quite unlikely that the surrender risk was fully priced in, and it was during times such as this that life insurers should really consider hedging the surrender risk embedded in their insurance policies.

3.5 Conclusion

Life insurance often embeds a surrender option that gives the policy holder a right to exchange an existing contract for its cash surrender value. Similar to mortgage prepayment option that imposes a cash-flow risk to MBS investors, this surrender option is a source of concern for life insurers. Understanding the adverse impact of the surrender risk is therefore critical for the risk management of life insurers.

While prior literature has attempted to quantify this surrender risk by pricing the surrender options, what existing studies share in common is the theoretical assumption of fully rational response of policy holders to interest rates. However, realistically, whether to surrender an existing policy is a multi-faceted decision, just as interest rates are only one of the many factors that drive the mortgage prepayment decision. Indeed, MBS valuation in light of empirical prepayment experience has become more favorable after being heavily studied by academics and practitioners alike. Never-

theless, such transition is still at its infancy for surrender options in life insurance policies primarily due to limitation of data.

This paper fills this gap in the literature by integrating an empirical surrender function into the option pricing framework. It relies on a novel data set from a large life insurance industry experience study, and accounts, for the first time, policy vintage years, whose analogy in the MBS community has been shown repeatedly as a prominent factor for prepayment decision. In fact, I show that the first policy year has a significantly positive impact on policy surrender, in addition to a positive effect related to interest rates, a positive effect from unemployment and a negative effect associated with income, the factors that have previously been proposed. Moreover, while prior studies provide mixed evidence on these latter factors, creating a debate between the so called Interest Rate Hypothesis and the Emergency Fund Hypothesis, the current study presents more consistent evidence that unifies both.

With these empirics, the option value based on actual surrender experience is calculated and compared to that based on the assumption of full rationality. I find that the experience-based option value is substantially less than its fully rational counterpart, which suggests suboptimality of surrender activity in practice. Moreover, while the fully rational option value is always positive, the experience-based option value can sometimes be negative, indicating that life insurers can make a profit from the exercising of the option even though they are the writers of the option.

In addition, I find that the value of the surrender option is especially sensitive to the insurer's mark-up, which reflects the competitive landscape of the life insurance industry. A less competitive market allows the insurers a greater margin on their products, which reduces the cash flows to the policy holders upon surrendering the policies. On the other hand, the interest rate environment also plays a critical role in the option value, with higher interest rates effectively increasing the moneyness of the option and hence subject the insurer to more cash-flow risk.

Finally, it is conceivable that factors beyond those in this paper and the existing literature may impact the surrender decision. Future studies can therefore direct at analyzing those factors. Such efforts will likely require more granular data incorporating policy holder demographics, but will also be rewarded with finer estimation of policy surrender that further improves the pricing accuracy.

References

[1] Albizzati, M. O., & Geman, H. (1994). Interest rate risk management and valuation of the surrender option in life insurance policies. *Journal of Risk and Insurance*, 61(4): 616-637.

[2] American Council of Life Insurers (2014). *Life Insurers Fact Book*, 2014.

[3] Armstrong, J. S., & Forecasting, L. R. (1985). From crystal ball to computer. *New York ua*.

[4] Bacinello, A. R. (2003). Fair valuation of a guaranteed life insurance participating contract embedding a surrender option. *Journal of Risk and Insurance*, 70(3): 461-487.

[5] Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy*, 81(3): 637-654.

[6] Black, F., Derman, E., & Toy, W. (1990). A one-factor model of interest rates and its application to treasury bond options. *Financial Analysts Journal*, 46(1): 33-39.

[7] Brennan, M. J., & Schwartz, E. S. (1977). Savings bonds, retractable bonds and callable bonds. *Journal of Financial Economics*, 5(1): 67-88.

[8] Boyle, P. P. (1977). Options: A monte carlo approach. *Journal of Financial Economics*, 4(3): 323-338.

[9] Carr, P., Jarrow, R., & Myneni, R. (1992). Alternative characterizations of American put options. *Mathematical Finance*, 2(2): 87-106.

[10] Cox, J. C., Ross, S. A., & Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics*, 7(3): 229-263.

[11] Cummins, J. D., & VanDerhei, J. (1979). A note on the relative efficiency of property-liability insurance distribution systems. *Bell Journal of Economics*, 10(2): 709-719.

[12] Dar, A., & Dodds, C. (1989). Interest rates, the emergency fund hypothesis and saving through endowment policies: some empirical evidence for the UK. *Journal of Risk and Insurance*, 56(3): 415-433.

[13] Dunetz, M. L., & Mahoney, J. M. (1988). Using Duration and Convexity in the Analysis of Callable Bonds. *Financial Analyst Journal*, 44(3): 53-72.

[14] Dunn, K. B., & McConnell, J. (1981). Valuation of GNMA Mortgage-Backed Securities. *Journal of Finance*, 36(3): 599-616.

[15] Green, J. & Shoven, J. (1986). The effects of interest rates on mortgage prepayments. *Journal of Money, Credit and Banking*, 18(1): 41-59.

[16] Grosen, A., & Jogensen, P. L. (1997). Valuation of early exercisable interest rate guarantees. *Journal of Risk and Insurance*, 64(3): 481-503.

[17] Grosen, A., & Jogensen, P. L. (2000). Fair valuation of life insurance liabilities: the impact of interest rate guarantees, surrender options, and bonus policies. *Insurance: Mathematics and Economics*, 26(1): 37-57.

[18] Hayre, L. (Ed.). (2002). Salomon Smith Barney guide to mortgage-backed and asset-backed securities (Vol. 105). *John Wiley & Sons*.

[19] Heath, D., Jarrow, R., & Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica*, 60(1): 77-105.

[20] Ho, T. S., & Lee, S. B. (1986). Term structure movements and pricing interest rate contingent claims. *Journal of Finance*, 41(5): 1011-1029.

[21] Hoopes, S. (2015). Life Insurance & Annuities in the US. *IBISWorld Industry Report*, January 2015.

[22] Jamshidian, F. (1992). An Analysis of American Options. *Review of Futures Markets*, 11: 72-80.

[23] Johnson, H. E. (1983). An analytic approximation for the American put price. *Journal of Financial and Quantitative Analysis*, 18(1): 141-148.

[24] Kalotay, A., Yang, D., & Fabozzi, F. J. (2004). An option-theoretic prepayment model for mortgages and mortgage-backed securities. *International Journal of Theoretical and Applied Finance*, 7(8): 949-978.

[25] Kang, P., & Zenios, S. A. (1992). Complete prepayment models for mortgage-backed securities. *Management Science*, 38(11): 1665-1685.

[26] Karatzas, I. (1988). On the Pricing of the American Option. *Applied Mathematics and Optimization*, 17: 37-60.

[27] Kim, I. N. (1990). The analytic valuation of American options. *Review of financial studies*, 3(4): 547-572.

[28] LeRoy, S. F. (1996). Mortgage valuation under optimal prepayment. *Review of Financial Studies*, 9(3): 817-844.

[29] LIMRA International (2011). *LIMRA Buyer/Nonbuyer Study*, 2011.

[30] Linton, N. A. (1932). Panics and Cash Values, *Transactions of the Actuarial Society of America*, 33: 265-394.

[31] Longstaff, F. A. (1992). Are negative option prices possible? The callable US Treasury-Bond puzzle. *Journal of Business*, 65(4): 571-592.

[32] Myneni, R. (1992). The pricing of the American option. *The Annals of Applied Probability*, 2(1): 1-23.

[33] New York Life Insurance Company Review (1986).

[34] Obersteadt, A., Bruning, L., Cude, B., DeFrain, K., Fechtel, B., Hall, S., ... & Mazyck, R. (2013). State of the Life Insurance Industry: Implications of Industry Trends. *National Association of Insurance Commissioners & Center for Insurance Policy and Research*

[35] Outreville, J. F. (1990). Whole-life insurance lapse rates and the emergency fund hypothesis. *Insurance: Mathematics and Economics*, 9(4): 249-255.

[36] Renaissance Capital (2014). *US IPO Market 2013 Annual Review.* January 2, 2014.

[37] Russell, D. T., Fier, S. G., Carson, J. M., & Dumm, R. E. (2013). An Empirical Analysis of Life Insurance Policy Surrender Activity. *Journal of Insurance Issues*, 36(1): 35-57.

[38] Schwartz, E. S., & Torous, W. N. (1989). Prepayment and the Valuation of Mortgage-Backed Securities. *Journal of Finance*, 44(2): 375-392.

[39] Securities Industry and Financial Markets Association (2014). *US Research Quarterly*. August 26, 2014.

[40] Smith, M. L. (1982). The life insurance policy as an options package. *Journal of Risk and Insurance*, 49(4): 583-601.

[41] Stanton, R. (1995). Rational prepayment and the valuation of mortgage-backed securities. *Review of Financial Studies*, 8(3): 677-708.

[42] Towers Watson (2014). *Prescribed U.S. Statutory and Tax Interest Rates for the Valuation of Life Insurance and Annuity Products*, October 2014.

[43] Tanskanen, A. J., & Lukkarinen, J. (2003). Fair valuation of path-dependent participating life insurance contracts. *Insurance: Mathematics and Economics*, 33(3): 595-609.

[44] Veronesi, P. (2010). Fixed Income Securities: Valuation, Risk, and Risk Management. *Wiley & Sons*.

[45] Walden, M. L. (1985). The whole life insurance policy as an options package: an empirical investigation. *Journal of Risk and Insurance*, 44-58.

[46] Zietz, E. N. (2003). An examination of the demand for life insurance. *Risk Management and Insurance Review*, 6(2): 159-191.

Appendix

Table 1. Four different types of permanent life insurance policies. What characterize the differences are the premium structure and the party who makes the investment choice, which in turn determines the structure of the death benefit. Whole life policies are the focus of this paper.

	Premium	Investment Choice	Death Benefit
Whole Life	Fixed	Insurer	Fixed
Universal Life	Flexible	Insurer	Fixed
Variable Life	Fixed	Insured	Flexible
Variable-Universal Life	Flexible	Insured	Flexible

Table 2. The Ho-Lee interest rate tree calibrated to the term structure of Jan 2, 2014. The volatility parameter σ is set to 0.0173. The first ten periods are shown.

Period i	0	1	2	3	4	5	6	7	8	9	10
θ_i	0.34%	1.01%	1.33%	0.82%	0.69%	0.59%	0.22%	0.50%	0.23%	0.33%	0.49%
j											
0	0.30%	2.38%	5.12%	8.18%	10.74%	13.16%	15.48%	17.43%	19.66%	21.62%	23.68%
1		-1.08%	1.66%	4.72%	7.28%	9.70%	12.02%	13.97%	16.20%	18.16%	20.22%
2			-1.80%	1.26%	3.82%	6.24%	8.56%	10.51%	12.74%	14.70%	16.76%
3				-2.20%	0.36%	2.78%	5.10%	7.05%	9.28%	11.24%	13.30%
4					-3.10%	-0.68%	1.64%	3.59%	5.82%	7.78%	9.84%
5						-4.14%	-1.82%	0.13%	2.36%	4.32%	6.38%
6							-5.28%	-3.33%	-1.10%	0.86%	2.92%
7								-6.79%	-4.56%	-2.60%	-0.54%
8									-8.02%	-6.06%	-4.00%
9										-9.52%	-7.46%
10											-10.92%

Table 3. Summary statistics of the main variables. Surrender data are from a large industry experience study. Treasury rates come from the Wall Street Journal. Unemployment rates are from the Bureau of Labor Statistics. Real per capita income levels are obtained from the Census Bureau and in 2013 dollars. Statutory rates are from Towers Watson (2014).

	mean	sd	min	max
Surrender Rate (%)	4.82	4.09	0.00	23.02
1-Year Rate (%)	2.63	1.68	0.40	5.11
10-Year Rate (%)	4.18	0.73	2.52	5.16
Unemployment Rate (%)	5.85	1.60	4.20	9.70
Real Income ($)	25102.20	1659.21	22794.00	26964.00
Statutory Rate (%)	4.25	0.25	4.00	4.50
Observations	170			

Table 4. Illustration of the backward recursive procedure. Assume fully rational response from policy holders to interest rates. $\lambda = 1.2$.

Panel A. Interet rate tree.

76	77	78	79	80	81	82	83	84	85
204.64%	207.87%	211.11%	214.35%	217.60%	220.86%	224.13%	227.40%	230.68%	232.41%
	204.41%	207.65%	210.89%	214.14%	217.40%	220.67%	223.94%	227.22%	228.95%
		204.19%	207.43%	210.68%	213.94%	217.21%	220.48%	223.76%	225.49%
			203.97%	207.22%	210.48%	213.75%	217.02%	220.30%	222.03%
				203.76%	207.02%	210.29%	213.56%	216.84%	218.57%
					203.56%	206.83%	210.10%	213.38%	215.11%
						203.37%	206.64%	209.92%	211.65%
							203.18%	206.46%	208.19%
								203.00%	204.73%
									201.27%

Panel B. Option price tree assuming full rationality.

76	77	78	79	80	81	82	83	84	85
917.07	927.10	936.95	946.63	956.13	965.38	973.63	972.87	878.61	0
	925.01	935.15	945.10	954.87	964.38	972.80	971.64	874.40	0
		933.27	943.51	953.56	963.33	971.92	970.34	870.04	0
			941.85	952.19	962.24	971.01	968.96	865.54	0
				950.77	961.11	970.04	967.50	860.87	0
					959.92	969.03	965.95	856.04	0
						967.96	964.31	851.03	0
							962.56	845.86	0
								840.50	0
									0

Table 5. Illustration of the Monte Carlo procedure.

Panel A. Ten simulated paths of Ho-Lee tree.

	Period i										
Simulation	0	1	2	3	4	5	6	7	8	9	10
1	0.31%	-1.07%	-1.79%	1.27%	3.83%	2.79%	5.11%	3.60%	2.37%	0.87%	2.93%
2	0.31%	-1.07%	1.67%	1.27%	0.37%	-0.67%	1.65%	3.60%	5.83%	7.79%	9.85%
3	0.31%	-1.07%	-1.79%	1.27%	0.37%	2.79%	1.65%	3.60%	2.37%	4.33%	2.93%
4	0.31%	2.39%	5.13%	8.19%	10.75%	9.71%	8.57%	7.06%	9.29%	7.79%	9.85%
5	0.31%	-1.07%	-1.79%	1.27%	0.37%	2.79%	5.11%	7.06%	9.29%	11.25%	9.85%
6	0.31%	-1.07%	-1.79%	1.27%	3.83%	6.25%	5.11%	3.60%	2.37%	4.33%	6.39%
7	0.31%	-1.07%	-1.79%	-2.19%	-3.09%	-0.67%	1.65%	3.60%	5.83%	7.79%	9.85%
8	0.31%	-1.07%	-1.79%	-2.19%	0.37%	-0.67%	-1.81%	-3.32%	-1.09%	0.87%	-0.53%
9	0.31%	-1.07%	1.67%	4.73%	7.29%	9.71%	8.57%	10.52%	12.75%	14.71%	16.77%
10	0.31%	2.39%	1.67%	4.73%	3.83%	6.25%	5.11%	3.60%	5.83%	4.33%	2.93%

Panel B. Ten simulated discounts.

	Period i									
Simulation	1	2	3	4	5	6	7	8	9	10
1	0.9969	1.0076	1.0258	1.0128	0.9748	0.9480	0.9008	0.8690	0.8486	0.8413
2	0.9969	1.0076	0.9909	0.9784	0.9748	0.9814	0.9654	0.9312	0.8785	0.8127
3	0.9969	1.0076	1.0258	1.0128	1.0091	0.9814	0.9654	0.9312	0.9094	0.8709
4	0.9969	0.9734	0.9247	0.8519	0.7651	0.6943	0.6373	0.5939	0.5412	0.5007
5	0.9969	1.0076	1.0258	1.0128	1.0091	0.9814	0.9325	0.8690	0.7919	0.7077
6	0.9969	1.0076	1.0258	1.0128	0.9748	0.9158	0.8702	0.8394	0.8198	0.7851
7	0.9969	1.0076	1.0258	1.0485	1.0814	1.0887	1.0709	1.0331	0.9746	0.9016
8	0.9969	1.0076	1.0258	1.0485	1.0447	1.0517	1.0709	1.1071	1.1193	1.1096
9	0.9969	1.0076	0.9909	0.9451	0.8787	0.7974	0.7319	0.6588	0.5800	0.5007
10	0.9969	0.9734	0.9572	0.9130	0.8787	0.8255	0.7844	0.7566	0.7138	0.6836

Panel C. Simulated prices of zero coupon bonds versus data (first ten periods).

Maturity	1	2	3	4	5	6	7	8	9	10
Actual Price	99.70	99.07	97.50	94.75	91.42	87.72	83.81	80.05	76.25	72.65
Simulated Price	99.69	99.05	97.48	94.74	91.43	87.73	83.83	80.08	76.29	72.73

Table 6. Baseline results. The dependent variable for each specification is the ratio of policy surrenders to the total number of policies for a given year fixing the policy origination time. *Interest* is the difference between the ten-year treasury rate and the statutory rate of that year. *Unemploy* is the unemployment rate. *Income* is the real per capita income, in terms of 2013 dollars. Standard errors are heteroskedasticity-robust.

	(1) Surrender	(2) Surrender	(3) Surrender	(4) Surrender	(5) Surrender	(6) Surrender	(7) Surrender
Interest	2.742***			2.357***	1.929***		1.975***
	(5.58)			(5.23)	(5.33)		(5.11)
Unemploy		-0.830***		-0.264*		-0.422***	0.0336
		(-4.86)		(-2.25)		(-3.64)	(0.35)
Income			-0.00147***		-0.00129***	-0.00135***	-0.00129***
			(-7.79)		(-8.13)	(-7.99)	(-8.17)
Constant	5.018***	9.677***	41.78***	6.533***	37.27***	41.30***	37.20***
	(16.15)	(7.81)	(8.41)	(7.55)	(8.94)	(8.69)	(8.84)
Observations	170	170	170	170	170	170	170
Adjusted R^2	0.184	0.100	0.353	0.186	0.438	0.374	0.435

t statistics in parentheses

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 7. Results with age-group fixed effects. The dependent variable for each specification is the ratio of policy surrenders to the total number of policies for a given year fixing the policy origination time. *Interest* is the difference between the ten-year treasury rate and the statutory rate of that year. *Unemploy* is the unemployment rate. *Income* is the real per capita income, in terms of 2013 dollars. Standard errors are heteroskedasticity-robust and clustered at the age-group level.

	(1) Surrender	(2) Surrender	(3) Surrender	(4) Surrender	(5) Surrender	(6) Surrender	(7) Surrender
Interest	2.742***			2.357***	1.929***		1.975***
	(9.32)			(8.72)	(8.93)		(8.48)
Unemploy		-0.830***		-0.264***		-0.422***	0.0336
		(-9.99)		(-10.55)		(-9.19)	(1.31)
Income			-0.00147***		-0.00129***	-0.00135***	-0.00129***
			(-9.16)		(-8.96)	(-8.88)	(-8.81)
Constant	3.432***	8.092***	40.20***	4.947***	35.68***	39.72***	35.61***
	(164.25)	(16.65)	(9.97)	(31.04)	(9.86)	(9.90)	(9.90)
Observations	170	170	170	170	170	170	170
Adjusted R^2	0.252	0.159	0.439	0.254	0.533	0.463	0.530
FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes

t statistics in parentheses

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 8. Results with age-group fixed effects and policy vintage effect. The dependent variable for each specification is the ratio of policy surrenders to the total number of policies for a given year fixing the policy origination time. *Interest* is the difference between the ten-year treasury rate and the statutory rate of that year. *Unemploy* is the unemployment rate. *Income* is the real per capita income, in terms of 2013 dollars. *Year1* is an indicator variable representing the first policy year. *Trend* is the year trend, denoting the number of years the policy has been in effect. Standard errors are heteroskedasticity-robust and clustered at the age-group level for specifications with age-group fixed effects.

	(1)	(2)	(3)	(4)	(5)	(6)
	Surrender	Surrender	Surrender	Surrender	Surrender	Surrender
Interest	1.123***	1.280**	0.977**	1.123***	1.280**	0.977***
	(3.79)	(3.02)	(2.95)	(6.26)	(3.22)	(3.52)
Unemploy	0.199*	1.341**	0.485	0.199***	1.341**	0.485
	(2.23)	(2.81)	(1.69)	(5.66)	(2.85)	(1.90)
Income	−0.0007***	0.00172	-0.0001	−0.0007***	0.00172	-0.0001
	(-5.78)	(1.56)	(-0.14)	(-6.71)	(1.59)	(-0.15)
Year1	8.188***		8.126***	8.188***		8.126***
	(6.65)		(6.58)	(12.57)		(8.84)
Trend		−2.142**	-0.472		−2.142**	-0.472
		(-2.69)	(-1.07)		(-2.74)	(-1.19)
Constant	21.65***	-34.45	5.985	20.07***	-36.04	4.399
	(6.77)	(-1.32)	(0.40)	(7.73)	(-1.41)	(0.33)
Observations	170	170	170	170	170	170
Adjusted R^2	0.696	0.443	0.695	0.820	0.540	0.820
FE	No	No	No	Yes	Yes	Yes

t statistics in parentheses

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

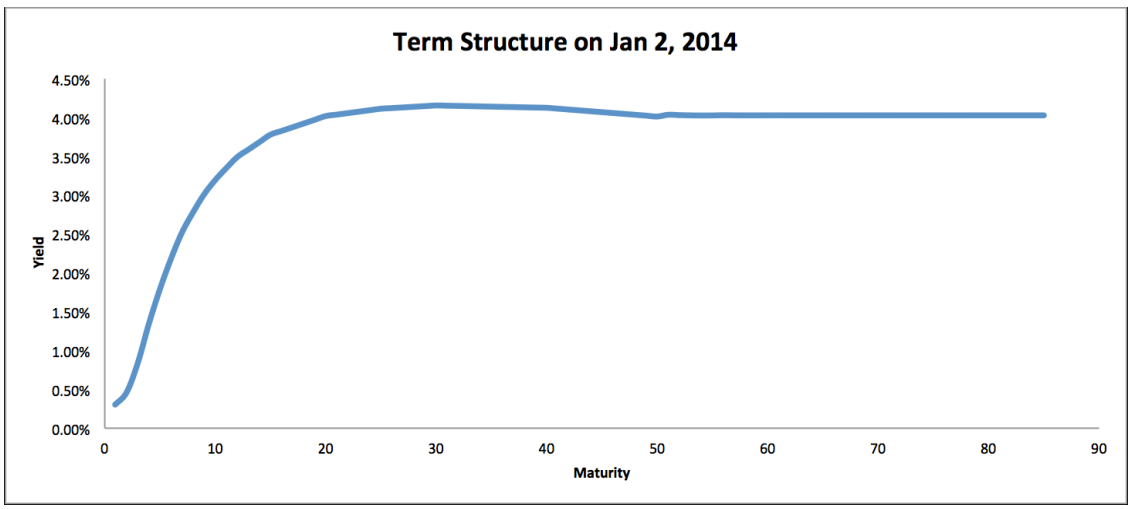

Figure 11: Term structure of interest rates on Jan 2, 2014

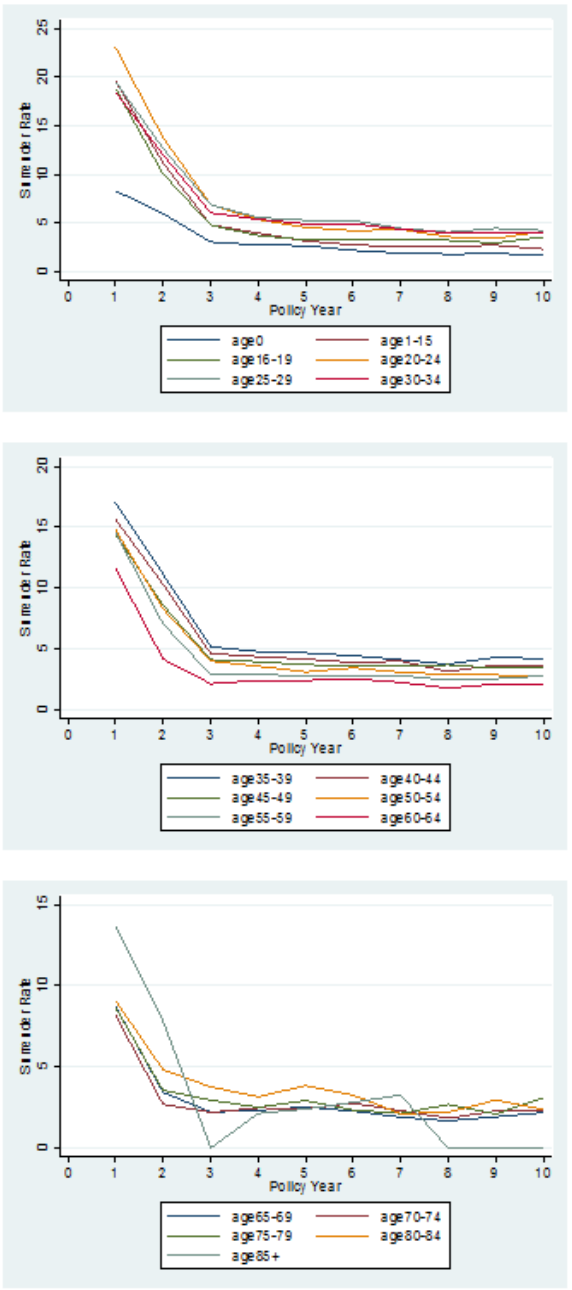

Figure 12: Graphical presentation of the surrender activity over policy years across different age groups. The surrender data are obtained from a large industry experience study.

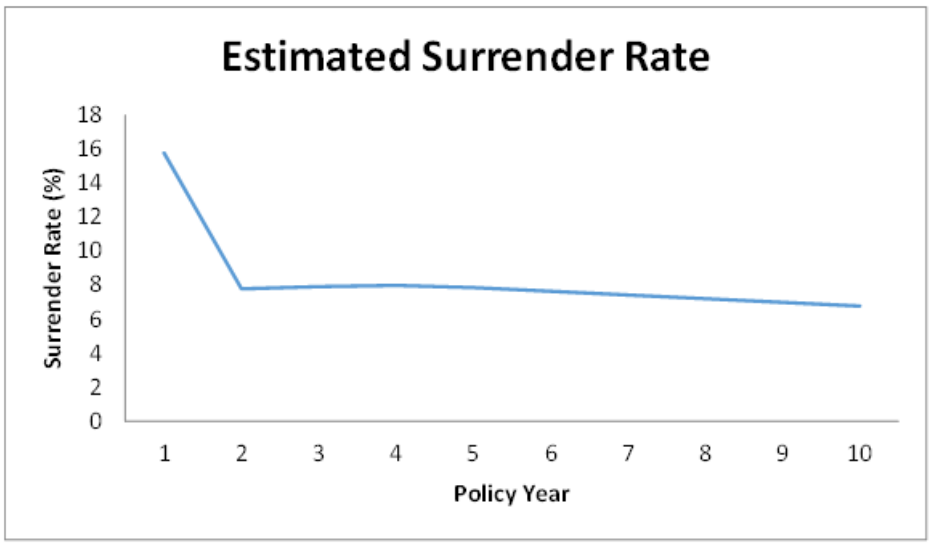

Figure 13: Estimated surrender rate. Specification (4) of Table 8 for the age group of 35 to 39 provides the basis of this estimation. The inputs related to macroeconomic factors are the projections from the Congressional Budget Office. First ten policy years are shown.

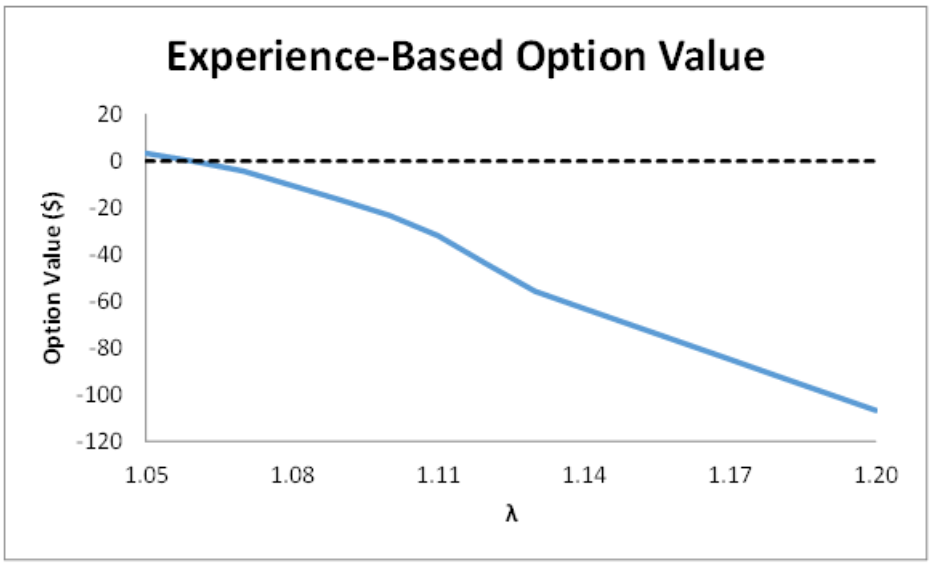

Figure 14: Experience-based option value. Specification (4) of Table 8 for the age group of 35 to 39 provides the basis of the estimation of the surrender rates. The inputs related to macroeconomic factors are the projections from the Congressional Budget Office. Option values are in dollars per $1,000 of policy face amount.

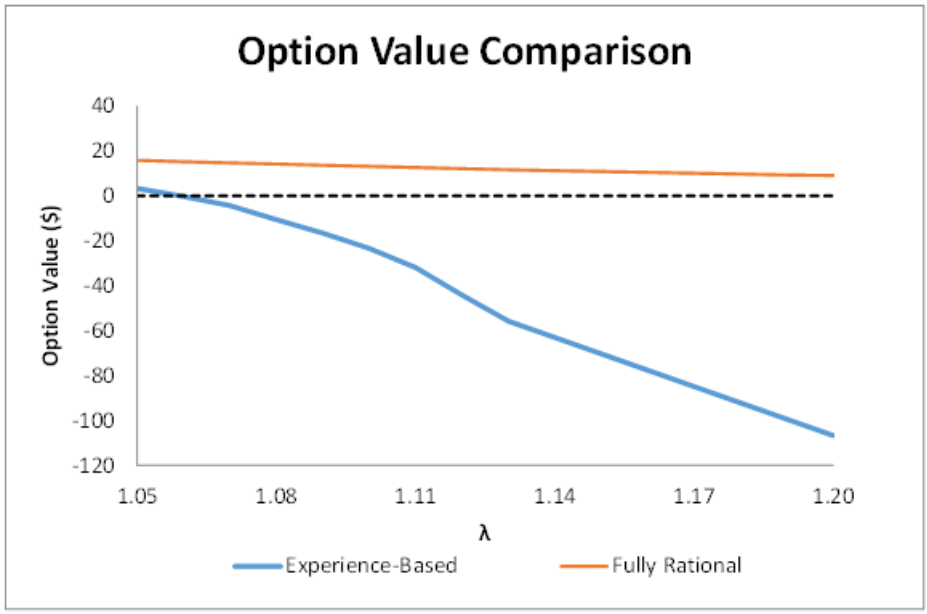

Figure 15: Comparison between fully rational and experience-based option values. Specification (4) of Table 8 for the age group of 35 to 39 provides the basis of the estimation of the surrender rates. The inputs related to macroeconomic factors are the projections from the Congressional Budget Office. Option values are in dollars per $1,000 of policy face amount.

Figure 16: Historical ten-year treasury rates. The time period is 1962 to 2012. Sources: Federal Reserve, Treasury, and Bloomberg.

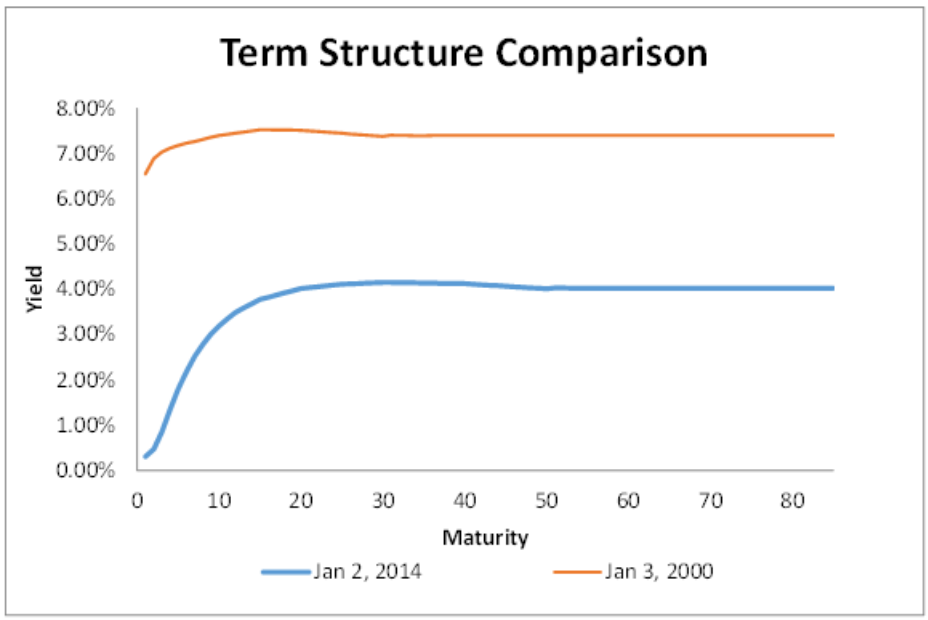

Figure 17: Term structure comparison between Jan 2, 2014 and Jan 3, 2010

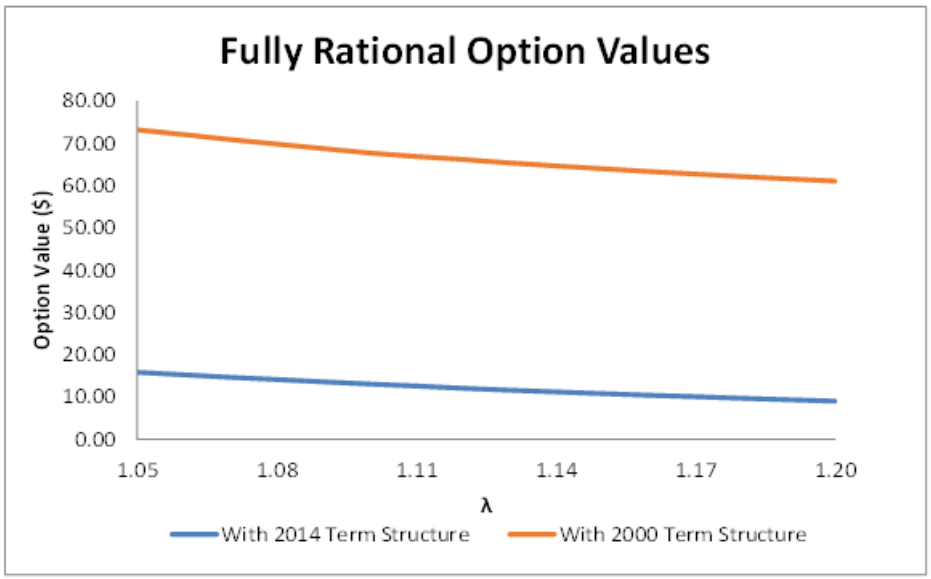

Figure 18: Sensitivity to interest rate environment - fully rational option values. Option values are in dollars per $1,000 of policy face amount.

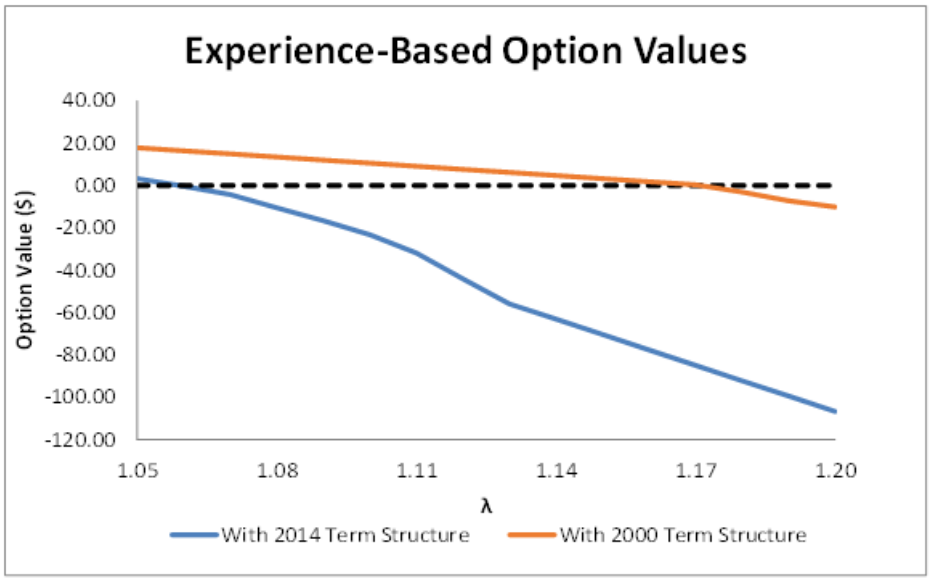

Figure 19: Sensitivity to interest rate environment - experience-based option values. Specification (4) of Table 8 for the age group of 35 to 39 provides the basis of the estimation of the surrender rates. The inputs related to macroeconomic factors are actual historical data. Option values are in dollars per $1,000 of policy face amount.