# Essays in Asset Pricing: Extensions and Applications of the Recovery Theorem

Anthony Sanford

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Mu-Jeung Yang, Chair

Eric Zivot, Chair

Yu-Chin Chen

Program Authorized to Offer Degree:
Economics

University of Washington

## Abstract

Essays in Asset Pricing: Extensions and Applications of the Recovery Theorem

Anthony Sanford

Co-Chairs of the Supervisory Committee:
Professor Mu-Jeung Yang
Economics

Professor Eric Zivot
Economics

This thesis has three separate goals: to provide a methodological framework for extracting risk-neutral densities from options prices, to extend the Recovery Theorem (RT) theoretically, and to apply the RT to firm decision making practices.

The first chapter introduces a new model for estimating the risk-neutral density. Current estimation techniques use a single mathematical model to interpolate option prices on two dimensions: strike price and time-to-maturity (TTM). I demonstrate that, when we vary the interpolating methodology based on which dimension we are interpolating, it allows us to better extract market information. I use B-splines with at-the-money knots for the strike price interpolation and a function that depends on the option expiration horizon for the TTM interpolation. The results of this "hybrid" interpolation technique are particularly striking when compared to the common Aït-Sahalia and Lo benchmark in an application to the Recovery Theorem. My contribution is significant because it illustrates that different risk neutral density estimation techniques will reveal different market information and risk preferences. Hence, the accuracy of the density estimation is critical.

In the second chapter, I redefine the prices derived in Ross's Recovery Theorem (Ross,

2015) using a multivariate Markov chain rather than a univariate one. I employ a mixture transition distribution where the proposed states depend on the level of the S&P 500 index and its options' implied volatilities. I include volatility because the transition path between states depends on the propensity of an underlying asset to vary. An asset that is highly volatile is more likely to transition to a far-away state. These higher transition probabilities should lead to higher state prices. The multivariate method improves upon the univariate RT because the latter does not include the volatility inherent in the state transition, which makes its derived prices less precise. The multivariate RT produces forecast results far superior to the univariate RT. Using quarterly forecasts for the 1996-2015 period, the out-of-sample R-square of the RT increases from around 12% to 30%.

Finally, in the third chapter, I answer the question: what effect does uncertainty about the aggregate economy have on investment, holding news shocks constant? Recent empirical studies have struggled to answer this question, as times of high economic uncertainty are typically also times of bad news. This chapter proposes a new methodology to measure and separate uncertainty and news shocks in stock return data. By using option prices to adjust abnormal returns for the time-varying risk premia, it is possible to estimate the impact of uncertainty shocks on firm investment while controlling for news shocks. Using quarterly data on public firms from 1996 to 2015, we find that uncertainty shocks systematically depress investment, even after controlling for bad news. Moreover, lumpy investments reinforce the negative effect of uncertainty on investment, while better management systematically attenuates this negative effect.

# TABLE OF CONTENTS

# LIST OF FIGURES

# ACKNOWLEDGMENTS

Fifteen years ago, I met the woman who is now my wife. Fourteen years ago, I made the decision to leave the security of the family business to pursue an undergraduate degree. Twelve years ago, I decided that, upon the completion of my undergraduate degree, I would move to the United States and pursue a Ph.D. These are some of the key decisions that culminated in this dissertation. This is my journey. Along the way, I met and was helped by more people than I could possibly thank in these acknowledgements.

Throughout the doctoral process, Mu-Jeung believed in me. He pushed me to produce my best work and, probably most importantly for a soon-to-be scholar, brought me down to earth. One of the most precious things that a doctoral student can ask from his/her advisor is time. With weekly (sometimes more!) meetings, I had all I could ever ask for. MJ truly was, to me, the best mentor a doctoral student could get. I aspire to one day be as good a mentor to the future generations of students as MJ was to me. Thank you!

I would like to express my deepest gratitude to Eric Zivot. Early on, he saw the potential in my ideas and encouraged me to pursue them. His confidence was critical in my early development as an economist. His expertise in financial econometrics was key to my success. His advice about the job market, and ultimately the job that I took, was tremendously helpful. Being Eric's student made me a better scholar. Thank you!

I would also like to thank my committee members, Yu-Chin Chen and Thomas Gilbert, as well as Matt Lorig, for their thoughtful feedback on my work, their great ideas on how to improve it, and their advice on my career going forward. And thank you Simon, without whom I think most of us would be completely lost!

Financial support for this dissertation was provided through the doctoral fellowship from the Fonds de recherche Société et culture (FRQSC) as well as the Henry T. Buechel Memorial Fellowship. Computing resources were provided by the University of Washington's Center for Studies in Demography and Ecology (CSDE). Many professors, colleagues, and fellow students also gave me essential feedback throughout the years at seminars, workshops, and conferences. Thank you!

I would not be here without the constant support of my family. Maman, dad, Vanes, Nic, and grandma, thank you! Thomas Fillebeen, thank you for listening to me talk incessantly about the Recovery Theorem, and for the memorable nights drinking good wine and playing backgammon. I will cherish those times for the rest of my life. To the folks in the Art building offices, Rory, Agraj, Ibrahim, and Pushpak – you guys are awesome!

Last but not least, to my wife, I am eternally grateful for the seed you planted in my brain that I too had what it took to complete a doctoral degree. I am thankful for your love, your drive, and your support. I wholeheartedly believe that, had it not been for you, this dissertation would not exist. Full disclosure: my work has benefited greatly from your gut-wrenching honesty and ruthless editing.

# DEDICATION

To my wife, Joannie, je t'aime fort coquine!

# Chapter 1

# INTRODUCTION

## *1.1 Asset Pricing Theory*

Asset pricing theory attempts to understand and explain the price of assets in financial markets (Cochrane, 2009). These assets can be stocks, bonds, or options, for example. Regardless of the asset type, asset pricing theory attempts to set a price for an asset today that will provide the holder with uncertain payments in the future. Ultimately, what dictates the value of an asset boils down to two major things: 1) how much time happens between today and the future date when payment is expected, and 2) the riskiness of that payment. These two components are discernible easily in a very simple equation of asset prices, based on the discount factor/generalized method of moments view of asset pricing theory (Cochrane, 2009), as follows:

$$p_t = E(m_{t+1}x_{t+1}) \tag{1.1}$$

where $p_t$ is the price of an asset at time $t$, $m_{t+1}$ is what the literature calls the stochastic discount factor at some future time, and $x_{t+1}$ is the asset payoff at some future time. We can define the stochastic discount factor as being a function of, say, data and parameters. The uncertainty and the risk, as previously mentioned, will depend on the amount of time between today, $t$, and the future period, $t+1$, as well as the amount of uncertainty associated with the future payoff of the asset. The uncertainty is captured in this model by the stochastic discount factor. Hence, if we believe that an asset price can be modeled as in equation 1.1, the question that asset pricers must answer is how to estimate the value of $x_{t+1}$ and the

stochastic discount factor.

In essence, this dissertation provides a blueprint to obtain the stochastic discount factor, while standardizing the future payoff. We do this by extending the Recovery Theorem (Ross, 2015) through a multivariate setup instead of a univariate one. This model, which provides us with an expected return on the market portfolio, can then be used in countless applications in economics and finance. The fourth chapter of this dissertation is one such application: we estimate news and uncertainty shocks and determine firm-level responses to these shocks for certain key variables (such as firm investment decisions).

### *1.2 The Risk-Neutral Density*

The second chapter of this dissertation proposes a new methodology to extract a full non-parametric risk-neutral density using market option prices (both call and put options). Many models in asset pricing theory (Ross, 2015; Kadan & Manela, 2017) assume that an infinite number of option prices are observed in the market. In reality, however, this is not the case. For example, on June 1st, 1996, the following strike prices were reported:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 400.00 | 425.00 | 450.00 | 475.00 | 500.00 | 510.00 | 520.00 | 525.00 | 530.00 | 540.00 | 545.00 |
| 550.00 | 560.00 | 565.00 | 570.00 | 575.00 | 580.00 | 585.00 | 590.00 | 595.00 | 600.00 | 605.00 |
| 610.00 | 615.00 | 620.00 | 625.00 | 630.00 | 635.00 | 640.00 | 645.00 | 650.00 | 655.00 | 660.00 |
| 665.00 | 670.00 | 675.00 | 680.00 | 685.00 | 690.00 | 695.00 | 700.00 | 725.00 | 750.00 | |

Table 1.1: Strike prices on S&P 500 call options for 1 April 1996

We have more than enough values here on which to conduct a nonparametric interpolation. However, in the TTM dimension, on the same day, the following values were reported:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.05 | 0.13 | 0.23 | 0.47 | 0.72 | 0.97 | 1.22 | 1.72 |

Table 1.2: Time-to-maturity on S&P 500 call options for 1 April 1996

Once we remove values that were not traded on that day, this is far from enough observations to construct a full density using a nonparametric (such as a spline) method.

This paper serves two purposes. First, this paper proposes a way to interpolate market option prices on two dimensions: strike price and time-to-maturity (TTM). Most articles provide a method for interpolating option prices only on the strike price dimension (Figlewski, 2008; Jiang & Tian, 2007; Jackwerth & Rubinstein, 1996). In an application like the one in chapter 3 of this dissertation, strike price interpolation is not enough. We also need to interpolate on the TTM dimension. Research has provided us with tools that allow us to extract prices on the TTM dimension. However, these methods are often the same methods used for the strike price dimension. I argue that such an approach is often inadequate because there are too few points to use these methodologies reliably (Aït-Sahalia & Lo, 1998; Fengler, 2009; Chernov & Ghysels, 2000).

The second contribution of this paper is to highlight the importance of transparency in applications that use risk-neutral densities. Countless articles (such as Ross, 2015) never explain how they interpolate the risk-neutral densities to obtain their results. I show that methodologies themselves can have a significant impact on the end result. Since different interpolation techniques appear to recover different information from options, researchers should always explain their interpolation methods in some detail in an effort to be more transparent.

### *1.3 The Natural Probability Distribution*

The natural probability distribution is the risk-adjusted expected probability of the market. Prior to the Recovery Theorem (RT), no methodology allowed us to disentangle state prices into their individual components: the risk-aversion parameters, the pricing kernel, and the natural probability distribution. Using the RT, we can focus on deriving a nonparametric measure of the expected distribution of the market. The distribution can be used in a mul-

titude of applications, two of which are explored in this dissertation. The first application, presented in chapter 3 of this dissertation, uses the result from the RT to obtain an expected return of the market. This expected return is then compared to the realized return. The chapter also shows that, given an increased volatility of the pricing kernel, we are able to increase the upper bound on the acceptable out-of-sample $R^2$ that is consistent with the efficient market hypothesis. The second application, presented in chapter 4, uses the expected return from the RT to examine the effect of uncertainty about the aggregate economy on investment by firms, holding news shocks constant. News shocks are defined as the difference between realized return and expected return and uncertainty shocks are defined as the volatility of these news shocks.

One key contribution of the third chapter of this dissertation is to improve upon the theory of the RT. I redefine prices derived in the RT using a multivariate Markov chain rather than a univariate one. This new multivariate RT controls for implied volatility because the transition path between states depends on the propensity of an underlying asset to vary. Controlling for implied volatility is critical in the estimation of the expected distribution of returns for the market. Using quarterly forecasts for the 1996-2015 period, the out-of-sample $R^2$ of the RT increases from around 12% (original version) to 30% (multivariate version). In sum, by adding the implied volatility as a measure of expected uncertainty in the derivation of the natural probability distribution, we obtain a much more accurate representation of the expected distribution of returns.

The fourth chapter uses the RT in a novel way. It paves the way for the use of nonparametric measures of uncertainty shocks obtained via the stock market to explain the different investment levels of individual firms. We posit that we can estimate the impact of uncertainty shocks on firm investment while controlling for news shocks. We find that uncertainty shocks systematically depress investment, even after controlling for bad news. This makes sense, since we would expect firms to adopt a wait-and-see strategy when deciding whether or

not to make investments when faced with a highly uncertain environment. This chapter also shows that lumpy investments further reinforce the negative effect of uncertainty on firm investment and that better management attenuates this effect. In essence, the chapter creates a link between uncertainty in the economy and the decision making process of the firm. This work is related to recent work by Bloom (2009) which shows that, in a controlled simulated environment, periods of higher uncertainty are linked to periods where firms temporarily pause their investment decisions.

### *1.4 Plan of the Dissertation*

This dissertation is divided into four main sections. Chapter two derives the proposed methodology for a full risk-neutral distribution. An application of the proposed interpolation methodology is tested using the RT. Chapter three compares the univariate and multivariate RTs, and discusses the steps required to implement the theorem. It also walks through a simple numerical example for the original univariate and the proposed multivariate RTs. Chapter four is an application of the RT which shows that uncertainty shocks partly explain depressed investments by firms, even when controlling for news shocks. Finally, chapter five explores possible extensions and concludes.

# Chapter 2

# RISK NEUTRAL DENSITY ESTIMATION

This paper presents a new methodology for interpolating option prices. For most assets, the market is rarely, if ever, complete. Yet, many theoretical frameworks require market completeness. In the case of financial options, we require completeness to estimate a full and well behaved risk-neutral probability distribution (implied volatility surface). Market completeness, in this case, implies that we have traded options for every possible strike price and expiration. Current estimation techniques use a single mathematical model to interpolate option prices on two dimensions: strike price and time-to-maturity (TTM). By using only one interpolation methodology for both dimensions, we are assuming that the function that characterizes the strike price dimension is the same as the function that characterizes the TTM dimension. This paper demonstrates that, by allowing the strike price and TTM dimensions to depend on different functions, we are able to better extract market information.

For example, assuming that the strike price dimension is characterized in the same way as the TTM dimension would imply that the difference of going from a strike price of $1,000 to a strike price of $1,001 would have the same impact as an option going from 30 days-to-expiration to 31 days. Because of the non-linearity of option prices, this is highly unlikely. Hence, I argue that the interpolation methodology for option prices should be modeled differently depending on which dimension we are trying to interpolate (strike prices or TTM). This paper shows theoretically and empirically that interpolation of option prices using two different models is more accurate and more efficient.[1] I use B-splines with at-the-money knots

[1]By accurate, I mean that the interpolation better reflects market information. By efficient, I mean that the proposed methodology is faster to compute than alternative methods.

for strike price-based interpolation and a function that depends on the option expiration horizon for TTM-based interpolation. The current literature assumes that interpolating both dimensions using splines is adequate (Figlewski, 2008; Aït-Sahalia & Lo, 1998). However, this paper shows that 1) the TTM dimension does not have enough observations to produce a reliable result from splines, and 2) using two different methods represents the available option information more accurately.

I compare the methodology developed here to several other density estimation techniques: the Aït-Sahalia and Lo (Aït-Sahalia & Lo, 1998), lognormal density (Jondeau et al., 2007), generalized beta (Jondeau et al., 2007), mixed lognormal (Jondeau et al., 2007), Edgeworth approximation (Jarrow & Rudd, 1982), and Shimko methods (Shimko, 1993). Once the extracted densities are compared, the methodologies are applied to the Recovery Theorem (RT). I use the RT because there is no accepted benchmark by which we can judge empirical applications of the derived densities.[2] The method proposed here was originally developed in my work extending the RT, so it was natural to provide results for that theorem as an empirical exercise. Furthermore, applying these methodologies illustrates how important the interpolation method is for empirical work. This piece seeks to promote transparency by showing the radical effects different interpolation methods have on empirical results. Most researchers assume market completeness, but do not mention whether completeness is achieved. As will be shown below, the interpolation technique alone can sometimes account for significant differences in empirical model performance. The results applied to the RT reflect the fact that the model uses market information to obtain an out-of-sample natural probability distribution of returns. As such, if we use a methodology that more accurately reflects the available information in the market, we would expect a better natural probability distribution of returns from the RT.

The paper proceeds as follows: section 2.1 introduces and derives my proposed interpo-

[2]In other words, we do not know what the actual complete density looks like empirically.

lation methodology; section 2.2 defines the benchmark interpolation methodologies; section 2.3 presents the model used to test the efficacy of the interpolation methodologies; section 3.3 presents the data used in this paper and the empirical results; and finally section 3.4 concludes the paper.

### *2.1 Theory and derivation*

In this section, I introduce my proposed implied volatility interpolation method and show how interpolated prices lead to a complete set of option prices (at least for a density). It is important to note that implied volatility interpolation is not an exact science. Most of the research in this area of finance relies on our general understanding of implied volatility distributions. Researchers generally try various methods to: 1) determine which method(s) avoids arbitrage in option prices, and 2) ensure that the graphical representations of the resulting implied volatilities conform to what we observe in the market (smiles/smirks). These two principles are, at the core, how I determined which methodology (or combination of methodologies) was "best" in this paper.

The interpolation method should not result in arbitrage opportunities in the final interpolated prices. In other words, the interpolation should not result in solutions where the implied volatilities are negative or where there are negative option prices.[3] Figlewski (2008) shows that, at least in the strike price dimension, no arbitrage occurs when smoothed splines are used with appropriate knots. If placed properly, knots ensure that we obtain positive implied volatilities (and, by extension, option prices). Basically, these knots "clamp" down the function and ensure that the function passes through the at-the-money strike price. Following Figlewski (2008), I use smoothed splines in this paper. However, instead of using

[3]Negative volatility should not occur because negative volatility is not mathematically possible. Negative option prices are also not possible because they would imply arbitrage opportunities – an investor could by the option (get paid for the option since it is a negative price) and get a certain profit if the option is in the money at some future date.

smoothed quartic splines, I have opted to use B-splines for two reasons. First, the B-spline provides us with a well-behaved risk-neutral distribution, which implies well-behaved implied volatilities. The B-spline does this better than a quartic spline because it allows for more customization of the basis functions (described below). Second, the smoothness implies that we are not being overly influenced by microstructure noise. In other words, we are trying to keep as much market information as possible while removing some market imperfections.

The purpose of interpolation is to simulate market completeness. As such, we want our interpolation to be as similar as possible to what we might observe in the market. Assuming that we are referring to option prices, we can characterize the information set contained in prices by defining prices as a function of the current log price, $S_T$, and the information set up to time $t$, $\mathscr{F}_t$: $C(S_T|\mathscr{F}_t)$. I am arguing that the interpolation methodology will affect the information set, $\mathscr{F}_t$, that we obtain from option prices. We want that information set to be a true reflection of market information rather than simply a construct of our interpolation methodology (which may or may not be correct). We do not want to interpolate option prices just for the sake of obtaining prices. This would defeat the purpose of the interpolation. The B-spline allows us to estimate complete markets without jeopardizing the integrity of the underlying data.

When attempting to use smoothed splines for the time-to-maturity interpolation, certain problems arise. First, in some instances, there are too few points to conduct a proper spline interpolation. This is not an issue in the strike price dimension because there are generally more points being traded than in the TTM dimension. Second, and perhaps most importantly, negative implied volatilities often result when extrapolating volatilities near or far from expiration. On the one hand, those who trade options that are very close to expiration are speculating. This results in prices that are not necessarily driven by fundamentals (since most fundamentals are known in the very short term). On the other hand, options that expire in the longer term are generally traded by investors trying to hedge. Investors who

hedge are trying to avoid extreme movements in the market. Hence, their valuations are based on insurance needs rather than on market outcomes. Moreover, there is much less liquidity at these extreme TTMs. This is why very-short-term and very-long-term option data have a tendency to be unreliable. I explicitly address this difficulty in the function-based interpolation proposed below (see section 2.1.2).

Ultimately, implied volatility interpolation for options is a balancing act. We want precision and feasibility. Since there is no perfect method, I propose a combination of methodologies that provides the best arbitrage-free results.

### *2.1.1 Strike price interpolation – Proposed method part I*

There are only a certain number of option strike prices being traded on any given day. For example, table 2.1 shows the (unique) strike prices for call options on the S&P 500 for 27 April 2011. Yet, to produce a complete volatility surface for this day, we might need a continuous set of strike prices ranging from, say, 350 to 1,200. Table 2.1 may lead us to believe that we have enough option prices. However, most applications assume that there are an infinite number of prices in the range of interest. Thus, interpolation is necessary.

The interpolation method discussed in this paper is a modification of a method proposed by Figlewski (2008). He shows that one of the more precise ways to extrapolate a volatility surface is to use a smoothed quartic spline regression with a single at-the-money (ATM) knot (Figlewski, 2008). The smooth spline allows us to obtain a full and well-behaved function of call options while giving us a more complete set of prices. In other words, the smoothed spline methodology addresses the technical problem of obtaining a sufficient number of valid option prices using a sparse set of option prices, without removing all of the market noise (or market information).

What I propose to do here is different from what Figlewski (2008) proposes in two ways: 1) I use smoothed B-splines instead of regular quartic splines, and 2) I do not append a

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 50 | 100 | 150 | 200 | 225 | 250 | 275 | 280 | 300 | 325 | 350 |
| 375 | 400 | 425 | 450 | 475 | 500 | 525 | 550 | 575 | 600 | 625 |
| 650 | 675 | 700 | 725 | 750 | 770 | 775 | 800 | 810 | 820 | 825 |
| 830 | 840 | 850 | 860 | 870 | 875 | 880 | 890 | 900 | 905 | 910 |
| 915 | 920 | 925 | 930 | 935 | 940 | 945 | 950 | 955 | 960 | 965 |
| 970 | 975 | 980 | 985 | 990 | 995 | 1000 | 1005 | 1010 | 1015 | 1020 |
| 1025 | 1030 | 1035 | 1040 | 1045 | 1050 | 1055 | 1060 | 1065 | 1070 | 1075 |
| 1080 | 1085 | 1090 | 1095 | 1100 | 1105 | 1110 | 1115 | 1120 | 1125 | 1130 |
| 1135 | 1140 | 1145 | 1150 | 1155 | 1160 | 1165 | 1170 | 1175 | 1180 | 1185 |
| 1190 | 1195 | 1200 | 1205 | 1210 | 1215 | 1220 | 1225 | 1230 | 1235 | 1240 |
| 1245 | 1250 | 1255 | 1260 | 1265 | 1270 | 1275 | 1280 | 1285 | 1290 | 1295 |
| 1300 | 1305 | 1310 | 1315 | 1320 | 1325 | 1330 | 1335 | 1340 | 1345 | 1350 |
| 1355 | 1360 | 1365 | 1370 | 1375 | 1380 | 1385 | 1390 | 1395 | 1400 | 1405 |
| 1410 | 1415 | 1420 | 1425 | 1430 | 1435 | 1440 | 1445 | 1450 | 1455 | 1460 |
| 1465 | 1470 | 1475 | 1480 | 1490 | 1500 | 1510 | 1520 | 1525 | 1530 | 1540 |
| 1550 | 1560 | 1570 | 1575 | 1580 | 1590 | 1600 | 1625 | 1650 | 1700 | 1750 |
| 1800 | 1900 | 2000 | 2250 | 2500 | 3000 | | | | | |

Table 2.1: Strike prices on S&P 500 call options for 27 April 2011

generalized extreme value (GEV) distribution to the tails. I explain why I do not use a GEV later in this section.

We can derive the coefficient estimate for the spline regression by first defining the criterion function to be minimized as follows:

$$\min_{\beta} ||C - G\beta||^2 + \lambda\beta'\Omega\beta \tag{2.1}$$

where

$$G_{i,j} = g_j(\sigma_{IV,i}), \quad i,j = 1,...,n \tag{2.2}$$

$$\Omega_{i,j} = \int g_i''(t)g_j''(t)dt, \quad i,j = 1,...,n \tag{2.3}$$

where $n$ is the number of knots, $g()$ are the B-spline basis functions, $\Omega$ is the penalty matrix

based on the basis functions, and $\lambda$ is a smoothing parameter. Here, the number of knots $n$ is equal to one and the smoothing parameter $\lambda$ is equal to 0.35. The smoothing parameter is typically between zero and one. The closer the parameter is to one, the smoother the function will be. I tested various values, but chose a value closer to zero to preserve the informational content of the data. If the smoothing parameter is too large, it would not portray the data accurately.

Next, we need to define the B-spline basis function. We can define the spline function as:

$$G_{i,j} = \sum_{i=1}^{n+1} B_j(\sigma_{IV,i}) G_i, \qquad \sigma_{IV,min} \leq \sigma_{IV,i} < \sigma_{IV,max} \tag{2.4}$$

where $G_i$ corresponds to the control points and $B()$ is the basis function of order $j$. Let the basis function from the B-spline be defined as follows:

$$B_{i,1}(\sigma_{IV}) = \begin{cases} 1, & \text{if } \sigma_{IV,i} \leq \sigma_{IV} < \sigma_{IV,(i+1)} \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

$$B_{i,j}(\sigma_{IV}) = \frac{\sigma_{IV} - \sigma_{IV,i}}{\sigma_{IV,(i+j-1)} - \sigma_{IV,i}} B_{i,j-1}(\sigma_{IV}) + \frac{\sigma_{IV,(i+j)} - \sigma_{IV}}{\sigma_{IV,(i+j)} - \sigma_{IV,(i+1)}} B_{i+1,j-1}(\sigma_{IV}) \tag{2.6}$$

Finally, we obtain the smoothing spline estimate at the knot $C$:

$$\hat{r}(C) = \sum_{j=1}^{n} \hat{\beta}_j g_j(\sigma_{IV}) \tag{2.7}$$

In contrast with Figlewski (2008), I do not extrapolate the tails using the Generalized Extreme Value (GEV) distribution because it removes all of the information in the tails. When trying to extract a well-behaved risk-neutral density, this may be the appropriate procedure. However, here, we are trying to extract as much information from the market as possible. By imposing a distribution on the data, we are limiting the amount of market

information that is fed into the model. For major market indices, like the S&P 500, there are enough strike prices beyond our point of interest for a GEV distribution not to be necessary. For example, table 2.1 shows that there are strikes as low as $50 for 27 April 2011. This strike price is more than six standard deviations from that date's S&P 500 level. Hence, based on the range of available market data, I opted to not append a GEV distribution to the tails.

### *2.1.2 Time-to-maturity interpolation – Proposed method part II*

Now that we have estimated our distribution on the strike price dimension, we need to interpolate our data based on the TTM dimension. Table 2.2 shows the TTM on S&P 500 call options for 27 April 2011 in number of days. For most applications, there are hardly enough TTM points to obtain a full implied volatility distribution. Therefore, I need to interpolate the data. Most of the derivation in this section is purely mechanical and may not have an intuitive explanation. Bloomberg has found that this method performs best in the TTM dimension.[4]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 2.23 | 24.29 | 52.29 | 64.29 | 80.29 | 115.29 | 143.29 | 156.29 |
| 234.33 | 247.33 | 325.29 | 338.29 | 416.29 | 605.33 | 969.33 | |

Table 2.2: Time-to-maturity on S&P 500 call options for 1 April 1996

Table 2.2 shows quite a few TTM observations, but other days have far fewer observations. For example, on 1 April 1997, there were only eight TTM observations. Furthermore, if we only consider options that have volume (meaning observations that were actually traded on

[4]Please note that Bloomberg also has a similar methodology to extrapolate prices in the strike price dimension. The methodology proposed by Bloomberg for the strike price interpolation is not used in this paper.

any given day), we would have a total of five observations. Five traded observations hardly seems enough to interpolate an entire distribution.

For the TTM interpolation, I use a method devised by mathematicians at Bloomberg (Chen, 2011) extending Heston's (1993) work. As of spring 2016, this is the methodology to extract an implied volatility surface at a typical Bloomberg terminal. First, let us define the extrapolated call price as follows:

$$C(T,K) = \sum_{l=1}^{N} p_l(T) \cdot BSP(\xi_l(T) S_{0,p}, K, r_f, \Sigma_l(T)/\sqrt{T}) \tag{2.8}$$

where BSP corresponds to the traditional Black-Scholes equation (Black & Scholes, 1973) where each variable is a regular Black-Scholes input with certain parameters adjusted for interpolation. I discuss the parameters in equation 2.8 later in this section.

I start by defining two functions, $\alpha(t)$ and $\eta_l(t)$, for notational simplicity:

$$\varphi(t) = \frac{T_{i+1} - t}{T_{i+1} - T_i} \tag{2.9}$$

$$\eta_l(t) = log(\frac{\xi_{l+1}(t)}{\xi_l(t)}) \tag{2.10}$$

where $\eta_l(t)$ uniquely determines $\xi_l(t)$ under the assumption that $\sum_l p_l(t)\xi_l(t) = 1$, $\xi_l(T) \geq 0$ is the time-dependent multiplicative means of the $l$-th lognormal, $0 \leq p_l(T) \leq 1$ is the time-dependent weight of the $l$-th lognormal, $t$ is the market maturity at which we want to extrapolate, and $i$ is the index for each of the observed times-to-maturity.

If we assume a Poisson default process and a survival probability $D(t) = 1 - Q(t)$, we obtain the hazard rate $\Lambda(t)$ that is consistent with the survival probability:

$$D(t) = 1 - Q(t) = \sum_l p_l(t) = e^{-\Lambda(t)t} \tag{2.11}$$

where the initial $\Lambda(t)$ is obtained from the Bloomberg survival probability data. It is important to note that, for an underlying asset like the S&P 500, the survival probability would be very close to one. Because the S&P 500 replaces poorly performing firms in its index, it is extremely unlikely that the whole index would default at any one time. This is true even for the longest horizons available in this sample – even at the almost 1,000-day horizon, we would not expect the S&P 500 index to go bankrupt.

Once we have the benchmark hazard rate and survival probability, we simply need to estimate four equations (the new $\Lambda()$, $p_l()$, $\eta_l()$, and $\Sigma_l()$) and use the values as inputs for equation 2.8. The specific equations depend on whether we are extrapolating between TTMs, we are doing a shorter-term TTM interpolation (less than three months), or we are doing a longer-term TTM interpolation (greater than six months).[5] Each of these is derived and discussed in its own section below.

**Shorter-term interpolation** A shorter-term interpolation is an interpolation that occurs either within three months of an available datapoint, or at a TTM below the lowest available TTM (but still less than six months from the lowest available TTM). First, we need the hazard rate $\lambda(t)$ to obtain $p_l(t)$ as follows:

$$\Lambda_{new} = \Lambda e^{\frac{x_m^2 - x^2}{2T_t}} \tag{2.12}$$

$$\hat{\Lambda}_{new} = \Lambda_{new} e^{\frac{x^2}{2}(\frac{1}{T_0} - \frac{1}{t})} \tag{2.13}$$

where $x_m = K_{min}/F(T_i)$ and the $F$ represents the strike price, $x = K/F(T_i)$, $T_i$ is the closest TTM, $F()$ is obtained from the put-call parity ($C() - P() = \frac{1}{r_f}(F - K)$ (Stoll, 1969)), $T_0$ is the smallest TTM, and $t$ is the TTM of interest. Here, we are effectively dampening the hazard rate estimate.

[5]The longer-term interpolation is used only occasionally since we usually have data within six months of interpolations of interest.

Once we have adjusted this hazard rate, we can easily obtain $p_l(t)$ by ensuring that its weights have the same ratio as the lowest TTM.[6] Then, we can obtain the time-dependent standard deviation of the $l$-th lognormal, $\Sigma_l(t)$, and the means of each lognormal as:

$$\Sigma_l(t) = \frac{\Sigma_l(T_1)t}{T_1} \tag{2.14}$$

$$\eta_l(t) = \eta_l(T_1)\sqrt{\frac{t}{T_1}} \tag{2.15}$$

Now, we have all of the necessary components to solve equation 2.8 (Black & Scholes, 1973).

**Interpolation between times-to-maturity** Here, we need to extrapolate between available TTMs. First, we derive the dampened hazard rate using equation 2.12. The only difference is that we adjust $K_{min}$ by defining it as follows:

$$K_{min} = \varphi(t)K^i_{min} + (1 - \varphi(t))K^{i+1}_{min} \tag{2.16}$$

Once we have estimated the dampened hazard rate, we can proceed to estimate the multiplicative means $\xi_l(T)$, the time-dependent weight $p_l(T)$, and the time-dependent standard deviation $\Sigma_l(T)$ using the following equations:

$$p_l(t) = (\frac{p_l(T_i + 1)}{D(T_{i+1})}\frac{\sqrt{t} - \sqrt{T_i}}{\sqrt{T_{i+1}} - \sqrt{T_i}} + \frac{p_l(T_i)}{D(T_i)}\frac{\sqrt{T_{i+1}} - \sqrt{t}}{\sqrt{T_{i+1}} - \sqrt{T_i}})D(t) \tag{2.17}$$

$$\Sigma^2_l(t) = (1 - \varphi(t))\Sigma^2_l(T_{i+1}) + \varphi(t)\Sigma^2_l(T_i) \tag{2.18}$$

$$\eta^2_l(t) = (1 - \varphi(t))\eta^2_l(T_{i+1}) + \varphi(t)\eta^2_l(T_i) \tag{2.19}$$

**Longer-term interpolation** At longer time horizons, we do not dampen the hazard function. We want the full effects of the potential for default. We obtain the time-dependent

[6] In other words, we are ensuring that the weights at $p_l(t)$ are the same as the ratio of weights $\frac{p_{l+1}}{p_l}$ at $T_1$.

weights as:

$$p_l(t) = p_l(T_n)\frac{D(t)}{D(T_n)} \tag{2.20}$$

where $T_n$ is the largest available datapoint with respect to TTM. Please recall that we defined the survival probability $D(t)$ using equation 2.11. We then obtain the time-dependent volatility as:

$$\Sigma_l^2(t) = \Sigma_l^2(T_n)\frac{t}{T_n} \tag{2.21}$$

Finally, we need to derive the means as follows:

$$\eta_l(t) = \eta_l(T_n)\sqrt{\frac{t}{T_n}} \tag{2.22}$$

In addition to being the industry norm, this method performs significantly better than its alternatives (such as previous Bloomberg algorithms, dealer quotes, or the Merrill Lynch volatilities) (Chen, 2011). Chen (2011) notes that linear interpolation could be used instead, but I have found this method to be superior.[7]

### *2.1.3 Implied volatility surface and option prices*

**Implied volatility surface** As noted above, the purpose of the interpolation is to obtain a richer dataset than that obtained from market data, while also maintaining the characteristics of the market. If we modify the implied volatilities too much, we run the risk of distorting the information that we extracted. One way to ensure that the information has not been distorted is to graph the derived implied volatilities. Implied volatilities obtained using the Black-Scholes-Merton equation (Black & Scholes, 1973) display a volatility smirk, smile, or skew. The volatility skew (which is used to characterize implied volatilities in equity markets) is caused by the inability of the Black-Scholes-Merton equation (Black & Scholes, 1973) to

[7]Results available from the author.

account for varying volatility levels and the log-normal distribution of the underlying assets' returns. Hence, it is a desired property when deriving implied volatility surfaces (Brigo & Mercurio, 2002; Cont et al., 2002; Benaim & Friz, 2009; Dumas et al., 1998).

Figure 2.1 illustrates the skew of the extrapolated implied volatilities on 1 April 1996.[8] The implied volatility increases at low strike prices, decreases as the strike price becomes higher, and finally increases again at higher strike prices, displaying a volatility skew (almost a volatility smirk). Simply by looking at the figure, I can confirm that the interpolation produced the desired characteristics.

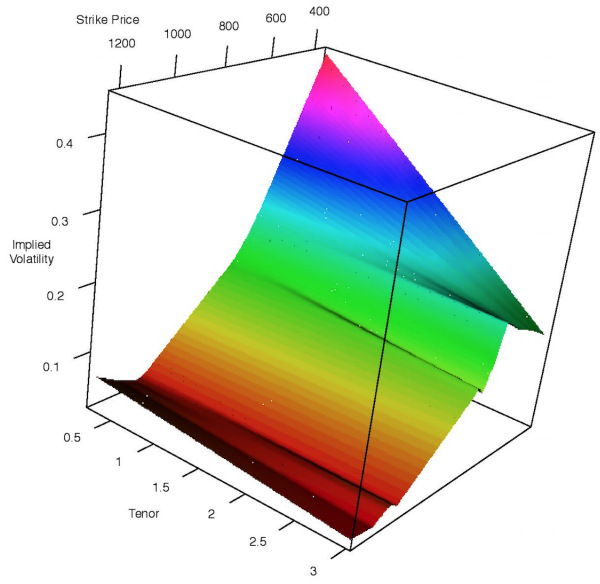


Figure 2.1: Implied volatility surface, 1 April 1996

[8]This date is used, in contrast with the rest of the paper, because there are very few observed data points for this specific day. Hence, this result shows that, even on the most sparsely traded days, we can obtain a complete and well-behaved volatility surface.

**Option prices** Once we have obtained a matrix with implied volatilities at the required strike prices (outlined in section 2.1.1)[9] and TTMs (outlined in section 2.1.2), we can proceed with inputting the data in the Black-Scholes-Merton equation (Black & Scholes, 1973):

$$C(S_{0,p}, t) = N(d_1)S_{0,p} - N(d_2)Ke^{-r_f(T-t)} \tag{2.23}$$

where

$$d_1 = \frac{1}{\sigma\sqrt{T-t}}[ln(\frac{S_{0,p}}{K} + (r_f + \frac{\sigma^2}{2})(T-t)]$$
$$d_2 = \frac{1}{\sigma\sqrt{T-t}}[ln(\frac{S_{0,p}}{K} - (r_f + \frac{\sigma^2}{2})(T-t)]$$

where $N()$ is a value from the normal distribution. Note that this step of transforming implied volatility back into option prices may or may nor be necessary depending on the empirical application. It is necessary in the application presented in this paper.

The above produces a matrix of call prices at our desired strike prices and TTMs. In the empirical analysis below (see section 3.3), the TTMs of interest are from three months to three years in three-month increments ($\Delta = 3$ months). The strike prices of interest are the strike prices up to (plus and minus) six standard deviations from that day's level of the underlying asset in one-dollar increments. But before I move on to the the details of the empirical analysis, let me briefly introduce the benchmark methods used in this paper.

### *2.2 Benchmark models*

This section introduces the models against which my proposed interpolation methodology will be compared. An extensive literature focuses on implied volatility estimation, mak-

[9] In this paper, I use \$1 increments for strike prices.

ing it critical to compare new methodologies to currently accepted one. The benchmark methodologies in this paper are 1) the method proposed by Aït-Sahalia & Lo (1998) and 2) the methodologies proposed by Jondeau et al. (2007) in their book, including the lognormal density, generalized beta, mixed lognormal, Edgeworth approximation, and Shimko's method.

### *2.2.1 Aït-Sahalia and Lo interpolation*

The first benchmark methodology is a non-parametric option pricing/volatility interpolation proposed by Aït-Sahalia & Lo (1998). This method of obtaining the risk-neutral density is one of the most highly cited and most often used in practice because it is nonparametric, computationally fast, and simple to use. The implied volatility interpolation equation is the following:

$$\hat{\sigma}(F_t, K, \tau) = \frac{\sum_{i=1}^{n} k_F\left(\frac{F_t - F_{t_i}}{h_F}\right) k_K\left(\frac{K - K_i}{h_K}\right) k_\tau\left(\frac{\tau - \tau_i}{h_\tau}\right) \sigma_{IV,i}}{\sum_{i=1}^{n} k_F\left(\frac{F_t - F_{t_i}}{h_F}\right) k_K\left(\frac{K - K_i}{h_K}\right) k_\tau\left(\frac{\tau - \tau_i}{h_\tau}\right)} \tag{2.24}$$

where $K$ is the strike price, $\tau$ is the TTM, $\sigma_{IV,i}$ is the implied volatility, and $F_t = S_{t,p} e^{(r_t - \delta_t)\tau}$. For the sake of brevity, I do not derive the method in its entirety here. Interested readers should refer to the original paper (Aït-Sahalia & Lo, 1998).

This interpolation is the one most closely related to the methodology proposed by Figlewski (2008). Both the spline and the kernel methodologies are nonparametric. However, in the kernel regression used by Aït-Sahalia & Lo (1998), the interpolation uses a local polynomial regression. The spline regression, on the other hand, uses a local piecewise polynomial. The piecewise polynomial is what allows us to "clamp" the function at specific places on the curve. In the case of Figlewski (2008), the knot occurs at the money. The two methodologies produce similar results. On average, however, the spline methodology gives results that are less prone to negative option prices or negative implied volatilities. In other words, the B-spline (smoothed spline) methodology adopted in this paper allows for important customizations

of the fitted function such that we do not have arbitrage-based prices.

### *2.2.2 MOE interpolation*

The Jondeau et al. (2007) book discusses most of the derivations and details about the methodologies in the Mother of all Extractions (MOE) interpolations. As such, I will not derive each model in this paper. Instead, I refer the interested reader to the book. The methodologies used in this section are as follows: the lognormal density, generalized beta, mixed lognormal, Edgeworth approximation, and Shimko's method. These methodologies are very fast to compute since most of them are based on parametric models. As such, my benchmarks include methodologies that are both parametric and nonparametric in nature.

## *2.3 Recovery Theorem*

In this paper, I apply the interpolation models to the Recovery Theorem (RT). The RT was not chosen at random. First, I want to demonstrate that, empirically, my model outperforms benchmark models. In the context of this empirical exercise, "outperforming" is defined as producing a more accurate forecast of the underlying asset. Second, I want to highlight the dependence of certain asset pricing models on the underlying interpolation technique. I have found that scholars rarely discuss the underlying methodology they use to model market completeness.[10] As this example will show, the results for the RT vary widely depending on which model is used.

Before discussing the results, I present a brief derivation of the RT. Its purpose is to obtain the natural probability distribution from observed prices in the market. The novelty of this model is that it allows the user to disentangle the natural probability distribution and the pricing kernel from observed prices without imposing a functional form for the risk

[10]Granted, a large quantity of research is conducted based on the results of the Bloomberg terminal's implied volatility surface. In that case, I suggest that we not blindly use black box models of interpolation.

aversion function. Moreover, it does not make any distributional assumptions for the natural probability distribution. Once the distribution is derived, we can use it to obtain a forecast for the return of the underlying asset. The version of the RT used in this paper is based on the multivariate RT proposed in Sanford (2017). Beyond the interpolation of option prices, deriving the RT involves three major steps: state price estimation, contingent state price estimation, and the natural probability distribution recovery. I discuss each in turn below.

**State price estimation** The first step involves estimating state prices using the complete set of option prices. These state prices are butterfly spreads with a strike price differential of one dollar. I define the state prices in the same way as Breeden & Litzenberger (1978). Formally, state prices correspond to the price of a security at some initial time $t_0$ such that, at some future time $T$, the security pays a pre-specified amount (normalized to \$1) if the market is at a pre-specified state of the world, and pays nothing otherwise. For example, assuming that the level of the S&P 500 today is 1,000, a state price would be the price of an asset that pays you 1\$ in, say, three months if the level of the S&P 500 is 1,500 at that time. Mathematically, this corresponds to the following equation:

$$s(K,T) \approx -C_{K_1} + 2C_{K_2} - C_{K_3} \tag{2.25}$$

which gives a guaranteed payoff of \$1 at expiration $T$ if the market ends at $K_2$.

**Contingent state price estimation** Ross (2015) estimates the contingent state price matrix using the following equation:

$$s_{t+1} = s_t P,\ t = 1, ..., m-1 \tag{2.26}$$

where $m$ is the number of states and $P$ is the contingent state price matrix. In this paper, I estimate the contingent state price matrix using a multivariate Markov chain specification

(Ching et al., 2008) as follows:

$$s_{t+1} = s_t P + vol_t \beta, \quad t = 1, ..., m-1 \tag{2.27}$$

where $vol_t$ is the implied volatility state at time $t$. The reasoning for including the implied volatility state variable in the derivation of the contingent state price matrix is explained in Sanford (2017).

**Natural probability distribution** Once we have obtained the contingent state price matrix, we can proceed to apply the RT. The complete proof is available in Ross (2015).

I start from the discrete time specification presented in Ross (2015). Like Ross, I also assume the representative agent formulation:

$$U_i' p_{ij} = \delta U_j' f_{ij}, \tag{2.28}$$

where

$$U_i' \equiv U'(c(\theta_i)) \tag{2.29}$$

which can then be written in terms of the kernel:

$$\phi_j \equiv \phi(\theta_1, \theta_j) = \delta(\frac{U_j'}{U_1'}) \tag{2.30}$$

where $\theta_1$ is the current state. In continuous time, Ross defines the kernel as:

$$\phi(\theta_i, \theta_j) = \delta \frac{h(\theta_j)}{h(\theta_i)} \tag{2.31}$$

using equation 3.16 and assuming transition independence, we have:

$$p(\theta_i, \theta_j) = \phi(\theta_i, \theta_j) f(\theta_i, \theta_j) = \delta \frac{h(\theta_j)}{h(\theta_i)} f(\theta_i, \theta_j) \tag{2.32}$$

where $h(\theta) = U'(c(\theta))$, and $p(\theta_i, \theta_j)$ is the state price transition function that we observe. From there, the objective is to solve for the unknowns: the natural probability transition function $f(\theta_i, \theta_j)$, the kernel $\phi(\theta_i, \theta_j) = \delta \frac{h(\theta_j)}{h(\theta_i)}$, and the discount rate $\delta$. Back to the discrete time specification, we can rewrite equation 3.17 in matrix form as:

$$DP = \delta FD \tag{2.33}$$

where $P$ is the $m$ by $m$ state price matrix derived in equation 2.27, $F$ is the $m$ by $m$ matrix that we are calling the natural probabilities (and is the matrix of interest for this section), and $D$ is the diagonal matrix of undiscounted kernel or the marginal rate of substitution. Rearranging equation 3.18, we get:

$$F = \frac{1}{\delta} DPD^{-1} \tag{2.34}$$

We obtained $P$ in equation 2.27, so now $D$ must be estimated. Up to this point, the RT has not provided us with additional insight into disentangling the risk aversion, pricing kernel, and natural probability distribution because there were not enough equations to solve our system of equations. We can make some assumptions about $P$ that will allow us to use the Perron-Frobenius Theorem (Meyer, 2000). Namely, we can assume that the option prices have no arbitrage opportunities. No arbitrage implies that the contingent state price matrix is going to be nonnegative. The second necessary assumption is that the matrix $P$ be irreducible. A matrix is said to be irreducible if we can reach any state in $k$ steps. As Ross (2015) argues, even if some of the transition probabilities in $P$ are zero, it should

still be possible to reach the desired state via an intermediary state (or states). As such, since $P$ is nonnegative and irreducible, we can apply the Perron-Frobenius Theorem (Meyer, 2000), which states that all nonnegative and irreducible matrices have a unique positive characteristic root (eigenvector) $z$, and a Perron root $\delta$. This allows us to solve for $D$, which we can introduce in the true distribution equation:

$$F = \frac{1}{\delta} DPD^{-1} \tag{2.35}$$

Once we have obtained the natural probability matrix, it is trivial to obtain the market forecast. The question, however, still remains: why does this paper apply its interpolation methodology to the RT? Recall that, in equations 3.4, 2.26, and 2.27, we derived state prices and contingent state prices. Both of these are going to be derived as the second derivative of a call price (Breeden & Litzenberger, 1978) in continuous time or as in equation 3.4 in discrete time. Since we cannot do continuous time in empirical applications, I focus on discrete time specifications here. In order to derive these state prices, we need a complete risk neutral density interpolated on both the strike price and the TTM dimensions. This is the stepping stone, and the first step, of the RT. The following is an example of the state price matrix, $S$, needed in the first step of the RT:

| | $t_1$ | $t_2$ | $t_3$ |
|---|---|---|---|
| $K_1$ | $s_{1,1}$ | $s_{1,2}$ | $s_{1,3}$ |
| $K_2$ | $s_{2,1}$ | $s_{2,2}$ | $s_{2,3}$ |
| $K_3$ | $s_{3,1}$ | $s_{3,2}$ | $s_{3,3}$ |

where $t$ is the TTM and $K$ is the strike price which corresponds to the future expected level of the underlying asset (in this case, the S&P 500). In order to estimate the above table, we will need state prices that correspond to the specific strike prices and TTMs of interest.

However, in most case, these strike prices and TTMs are not directly observed in the market. For example, $t_1$ may be equal to to a quarter, or 0.25 of a year, but there may not be an option that expires in exactly three months. Hence, we will need to extrapolate both in the strike price and the TTM dimensions in order to be able to estimate our full matrix of state prices $S$. This is why the RT lends itself particularly well as an empirical application of the methodology proposed in this paper.

## *2.4 Data and results*

### *2.4.1 Overview of the data*

I collected the data for this paper from two sources: 1) the Wharton Research Data Services (WRDS) database, and 2) the Bloomberg terminal. I use daily option prices on the S&P 500, the S&P 500's closing price, and the risk-free rate. The risk-free rate is the one-month Treasury Bill rate, which can be found in the Fama & French factors data. S&P 500[11] prices are from the CRSP dataset. The S&P 500 is generally thought to be the best proxy for the market portfolio. All of the option data are from OptionMetrics. This paper covers the time period from January 1996 to July 2015, the entire timeframe included in the OptionMetrics database. I use this sample for two major reasons. First, the forecast horizon in the empirical analysis in this paper is quarterly. A quarterly forecast requires a large enough sample size to test the efficacy of the proposed interpolation methodology and this twenty-year sample provides me with approximately 80 data points. Second, it allows me to divide the sample into subsamples and test my model in periods that experience various shocks (such as the tech bubble and the recent financial crisis).

Strike prices on the options obtained from OptionMetrics are quoted for lots of 1,000 securities. The Black-Scholes-Merton equation requires strike prices that are on a per-stock basis, so I divided the strike price by 1,000. Time-to-maturity is converted from a date to a

[11]SECID 108105

fraction of years to expiration, also a required input for the Black-Scholes-Merton equation. Option price is replaced with the midpoint of the bid-ask spread. This is consistent with Figlewski (2008), who argues that bid and ask prices are continuously quoted for almost all strikes regardless of whether a trade takes place. The alternative, transaction prices, occurs irregularly (Figlewski, 2008) and would make it more difficult to extract a proper implied volatility curve. Option data with no volume is also removed from the sample. We are interested in obtaining a density that reflects options being traded. Table 2.3 shows descriptive statistics.

| **Variable** | **Mean** | **Standard Deviation** | **Minimum** | **Maximum** |
|---|---|---|---|---|
| Stock Price | 1249.24 | 326.16 | 598.48 | 2130.82 |
| Risk-free Rate | 0.00000298 | 0.000124 | -0.000895 | 0.001135 |
| Call Price | 285.37 | 289.68 | 0.025 | 1288.35 |
| Call Option Volume | 236.89 | 1265.47 | 0 | 14953 |
| Put Price | 79.98 | 197.03 | 0.025 | 1661.35 |
| Put Option Volume | 286.37 | 1518.04 | 0 | 28886 |

Table 2.3: Descriptive Statistics

The stock price and risk-free rate descriptive statistics represent the values for the entire sample – from January 1996 to July 2015. The summary statistics for the option data represent the descriptive statistics for a single day of option prices – April 27th, 2011. I use a single day for the option data because that day will be the focus of the interpolation in the upcoming section. Nothing in table 2.3 is particularly surprising. Perhaps the only item that warrants attention is the fact that trading volume for some of these options is sometimes zero.[12]

[12]No transaction for a particular option with a specific strike price and a specific TTM combination was traded on that day.

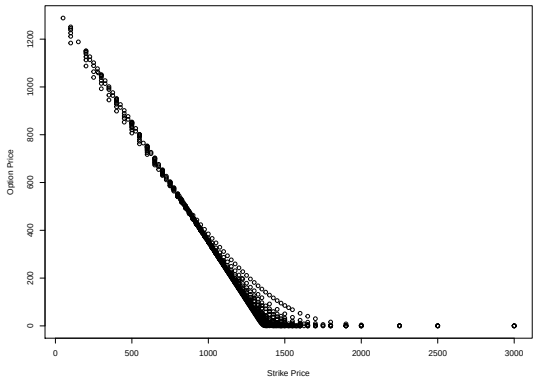


Figure 2.2: Call Option Prices

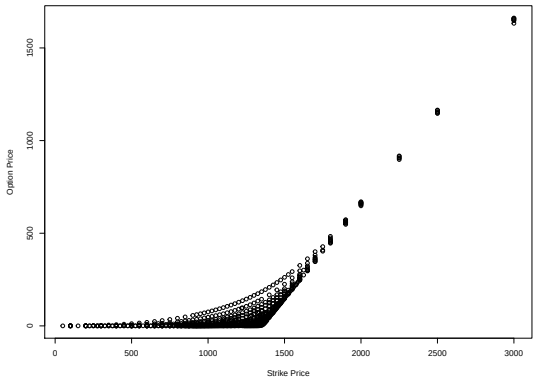


Figure 2.3: Put Option Prices

Figures 2.2 and 2.3 are plots of the option prices on April 27th, 2011 with the x-axis as the strike price and the y-axis as the option price. The prices reflect the kink behavior that we would expect. Option prices become more and more positive as the option becomes more and more in the money. For example, if the current stock price is \$1,200 and we are pricing a call option price with a strike price of \$1,300, we would expect the price of the option to be larger than if the strike price was \$1,200 (which would be almost zero), all else equal. The same can be said for the put option (for more information on the shapes of option prices, please refer to Hull & Basu (2016) or McDonald (2006)).

### *2.4.2 Density Results*

This section provides visual representations of the densities estimated on the strike price dimension based on my interpolation methodology (the B-spline methodology) and the benchmark methodologies. The benchmark results were obtained using the MOE (mother of all extractions) function from the RND package in R (Jondeau et al., 2007). All interpolations use data from April 27, 2011 with a TTM of June 30, 2011. As expected, all of the means of the distributions appear to be around the same level – that of the current S&P 500 level.

This is the be expected because we have a lot of datapoints around the mean. Hence, we know, fairly well and regardless of the assumptions being made, where the mean of the distribution should be. The more interesting results are the standard deviation (the spread of the distribution), the skewness (is the majority of the distribution towards the left or the right?), and the kurtosis (the density in the tails).

One important observation is the wide range of the density around the value close to the current level of the market. For example, figure 2.6, the single lognormal interpolation, has a maximum density around 0.003. Figure 2.5, the B-Spline (my proposed), has a maximum density of around 0.006, which is quite a bit larger than 0.003. Furthermore, the spread of the data when using the lognormal method is quite a bit larger than the spread for the B-spline (my proposed) method.

The tails of the distributions all seem to be quite similar. In contrast, there appears to be quite a bit of variation in the skewness. For example, the ASL method produces a substantially more left-skewed distribution than the B-spline (my proposed) method. All of this being said, it is unclear what the appropriate distribution *should* look like.

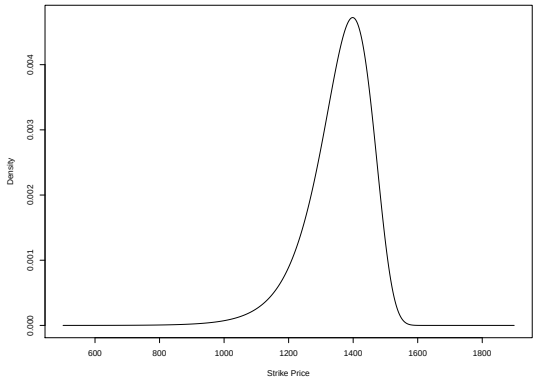


Figure 2.4: ASL

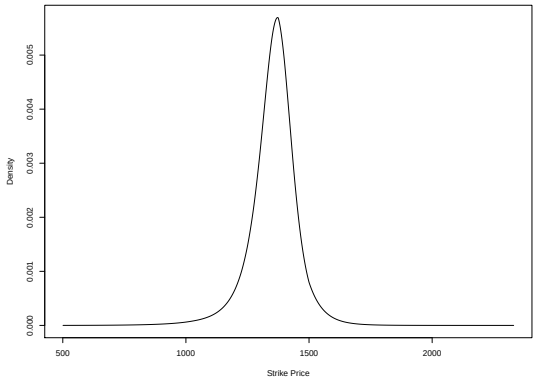


Figure 2.5: B-Spline

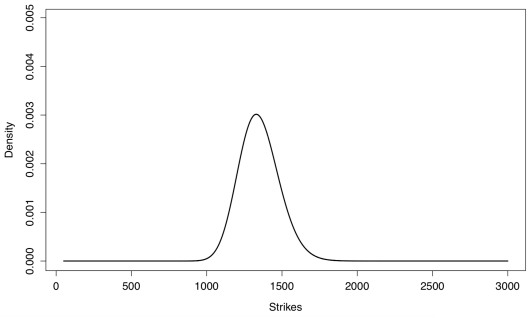


Figure 2.6: Single LNorm

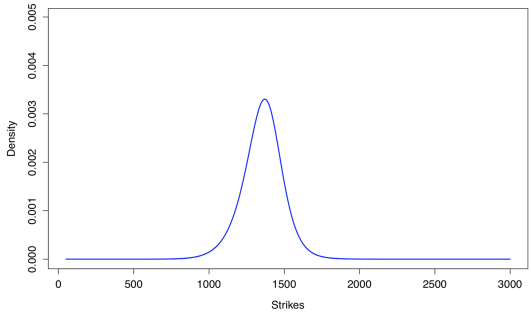


Figure 2.7: GenBeta

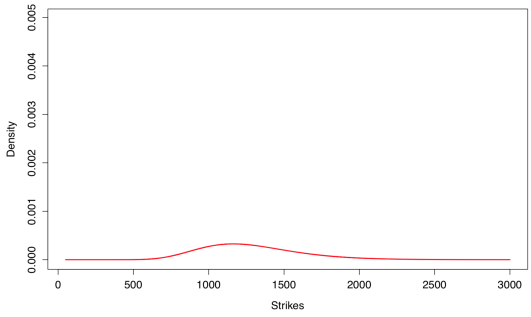


Figure 2.8: MixLNorm

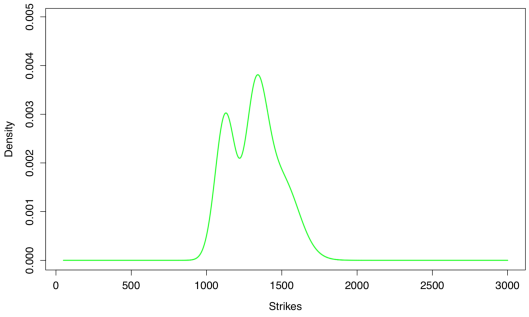


Figure 2.9: EW

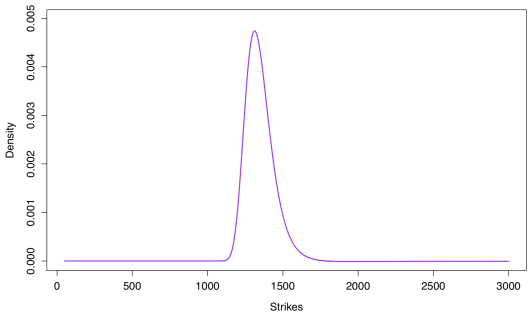


Figure 2.10: Shimko

Figures 2.4 through 2.10 illustrate the distributions interpolated using each method. The

mixed lognormal distribution does not produce a well-behaved distribution. This distribution almost looks like a uniform distribution with a slight hump around the current level of the S&P 500. The Edgeworth (EW) distribution looks much more like a proper distribution, but is bi-modal (although not perfectly). Bi-modality is interesting because it signals that the market believes in two scenarios likely to happen in the next three months or so: assuming a risk-neutral world, the market will either stay at its current level or move to a level that is at around $1,100. The rest of the graphs are all quite similar. They all look approximately normally distributed. Some of them seem to have more positive skews while some have more negative skews. This is why an application of these methodologies is necessary. Simply looking at these distributions, it is difficult to determine which one contains the ideal information set.

### *2.4.3 Forecast results – Recovery Theorem application*

One of the objectives of this paper is to illustrate the extent to which a scholar's extrapolation methodology matters. Too often, scholars use interpolation techniques, but fail to specify what method they used (see, as an example, Ross (2015) who provides no indication of how his risk neutral density was derived). In this section, I compare two interpolation methodologies in an application: 1) the interpolation methodology proposed by Aït-Sahalia and Lo, and 2) the interpolation method proposed in this paper. Note that here, I only compare the Aït-Sahalia and Lo to the methodology proposed in this paper (namely, the B-Spline and functional-based "mixed" interpolation methodology) because these are the only two methods that are intended to interpolate based on both the strike price and the TTM dimensions. The MOE interpolations are meant to be used on the strike price dimension but not on the TTM dimension. There is a limited number of interpolation techniques for the TTM dimension. Hence, I focus on my proposed method and the Aït-Sahalia and Lo method in this section of the paper. These two state price density estimation techniques are applied

to the Recovery Theorem (derived in section 2.3). As previously indicated, the interpolation methodology has a significant impact on the application since it, in effect, provides the basis for the information extracted from the market.

*Regression tables*

In this section, I compare my results to those using the interpolation method of Aït-Sahalia and Lo (hereafter referred to as "ASL"). The reader will notice that each table includes two columns. The first column represents the ASL method while the second set of results (column 2) corresponds to my methodology. The forecast regression is as follows:

$$R_t = \alpha + \beta_t E_{t-1}[R_t] + \epsilon_t \tag{2.36}$$

where $\alpha$ is the intercept, $\beta_t$ is the forecast coefficient, and $E_{t-1}[R_1]$ is the previous period's RT forecast. The forecast horizon is held to a quarter, so $t$ corresponds to 0.25 years. One of the criteria to evaluate the efficiency of the forecast is the forecast error. This error is defined as the residual, $\epsilon_t$, found in equation 3.37. Table 3.5 compares the results for the ASL and my proposed methodology. The sample for this first set of results is April 2009 to April 2013. To be clear, the column names "mixed method" in the upcoming tables is based on the interpolation methodology proposed in this paper. More specifically, the column named "proposed method" is based on the mixed interpolation methodology proposed in this paper (the function-based TTM dimension interpolation and the B-spline based strike price dimension interpolation).

| | ASL method (Apr 09–Apr 13) | Mixed method (Apr 09–Apr 13) |
|---|---|---|
| | (1) | (2) |
| Intercept | 0.00215 | 0.027675** |
| | (0.00561) | (0.009153) |
| Forecast | 0.23223*** | 0.338864*** |
| | (0.06020) | (0.070478) |
| Observations | 49 | 49 |
| $R^2$ | 0.240472 | 0.32970 |
| Adjusted $R^2$ | 0.22431 | 0.31544 |
| F statistic | 0.000348 | $1.6e^{-05}$ |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 2.4: Apr. 09–Apr. 13 subsample – Results

The two columns are quite different. The adjusted $R^2$ from the ASL method (column 1) is about 0.224 while it is 0.315 for my proposed method (column 2). The forecast coefficient for the ASL method is 0.232 whereas the forecast coefficient for the proposed method is about 0.339. The results in this first table illustrate precisely my point: the interpolation method has a significant impact on the results of an empirical exercise like this one (the RT). The interpolation method alone accounts for almost a 50% increase in the adjusted $R^2$ of the forecast regression. This is a clear indication that 1) my interpolation methodology outperforms the ASL method, and that 2) interpolation methodology has a significant impact on the resulting empirical results.

| | ASL method (Apr 96–Aug 15) | Mixed method (Apr 96–Aug 15) |
|---|---|---|
| | (1) | (2) |
| Intercept | −0.00287 | 0.004841 |
| | (0.00632) | (0.004626) |
| Forecast | 1.29653*** | 0.424605*** |
| | (0.14003) | (0.042434) |
| Observations | 235 | 235 |
| $R^2$ | 0.27068 | 0.30187 |
| Adjusted $R^2$ | 0.26752 | 0.29884 |
| F statistic | $1.46e^{-17}$ | $8.19e^{-20}$ |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 2.5: Full sample – Results

Table 3.6 compares the results for the ASL and my proposed methodology for April 1996 to August 2015. Here, however, notice the smaller difference between columns one and two. The interpolation increases the adjusted $R^2$ from 0.268 in the ASL case to 0.299 for my proposed method. The interesting part, however, is the difference in ASL coefficients across samples. In the smaller subsample, the ASL coefficient was relatively close to that of the proposed method. In this case, however, the coefficient is greater than one. In a world where rational expectations hold, we would expect the intercept to be equal to zero and the coefficient to be equal to one. Here, the ASL method has a result that is quite close to one.

| | ASL method (Apr 96–Apr 13) | Mixed method (Apr 96–Apr 13) |
|---|---|---|
| | (1) | (2) |
| Intercept | −0.00275 | 0.003038 |
| | (0.00706) | (0.005131) |
| Forecast | 1.34295*** | 0.450346*** |
| | (0.15186) | (0.047678) |
| Observations | 204 | 204 |
| $R^2$ | 0.278112 | 0.308482 |
| Adjusted $R^2$ | 0.274556 | 0.305025 |
| F statistic | $4.52e^{-16}$ | $9.59e^{-18}$ |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 2.6: Subsample I – Results

Table 3.7 shows the results for a random sample (selected by R). The sample is from the beginning of the available data (April 1996) to April 2013 (204 months). The results are consistent with the results discussed above. The forecast coefficients are still statistically significant at the $p¡0.001$ level. The adjusted $R^2$ increases from around 27% in the ASL model to around 31% for my method. The magnitude of the coefficients is also quite similar to previous results. The impact of the interpolation methodology on the adjusted $R^2$ ranges from around 3% to 9%. This difference may seem trivial, but it is important to note that, in asset pricing, a change in the adjusted $R^2$ of 3% to 4% is often considered to be a considerable one. Hence, I would argue that these results alone are significant enough to warrant a more thorough discussion about the importance of interpolation in forecasting models.

### *2.5 Conclusion*

In this paper, I developed a new methodology for interpolating option prices. Modern financial theory routinely assumes market completeness, but the methodology by which completeness is obtained is often ignored. The contributions of this paper are twofold: 1) I present a methodology that considers the difference between a strike price interpolation and a TTM interpolation, and 2) I show the impact of different interpolation methodologies on empirical applications. More specifically, I show that my methodology has two significant advantages when compared to other commonly used interpolation methodologies. First, I show that the volatility smirk/smile/skew that we need in the estimation of an implied volatility surface are satisfied with my methodology. Second, using an empirical application of the Recovery Theorem, I illustrate that my new interpolation methodology produces a better out-of-sample forecast than benchmark method. The empirical example also illustrates the importance of transparency about methodology in research papers.

Additional work should focus on implied volatility surfaces. As of now, most of the literature uses graph comparisons to judge the efficiency of interpolation methodologies. I believe that properly defined characteristics need to be derived for us to be able to judge the efficacy of various interpolation methodologies more accurately (see Bahra, 1997; Carr & Wu, 2003; Birru & Figlewski, 2012). Furthermore, more work needs to be done to determine how the option-based extraction of information is dependent on the inter/extrapolation used in the derivation of risk neutral densities (see Chernov & Ghysels, 2000; Bliss & Panigirtzoglou, 2004; Carr et al., 2003).

# Chapter 3

# RECOVERY THEOREM WITH A MULTIVARIATE MARKOV CHAIN

Ross's (2015) Recovery Theorem (RT) is a breakthrough in asset price forecasting. Using the RT, we can obtain the market's best estimate of future expected returns and risk aversions by separating the components of state prices (the discount rate, pricing kernel, and natural probability distribution). Not only does it allow us to use option prices to obtain an out-of-sample non-parameterized expected future distribution of an option's underlying asset, but it is one of the best asset forecasting models available today. However, it has certain shortcomings that this paper aims to address.

This paper's theoretical contribution is that it changes the original univariate model to a multivariate one. The original RT derived contingent state prices using a simple constrained linear regression, which assumed that the probability of transitioning to a new state was dependent on the previous state. But for option prices to truly reflect the conditional variance of the underlying asset (Engle & Mustafa, 1992), the transition path should control for volatility (Page et al., 2006). Controlling for the volatility in the transition path becomes even more important because of the nature of contingent state prices. These prices are not observed in the market. They are a function of observed state prices, which are used to infer prices for states that have not occurred. If contingent state prices were actually observed, they would already contain all available market information, including volatility. However, since we only observe state prices for the current state, it is crucial to derive the contingent state prices contingent on the observed underlying volatilities. Thus, including volatility in the derivation of the contingent state prices is critical to the proper specification of the

Recovery Theorem.

One of the key assumptions of the RT is that markets are complete. In reality, markets are not complete. To construct state prices that are complete and behave normally, it is necessary for the data to be as detailed as possible. The original RT was tested empirically using over-the-counter (OTC) data, which is richer[1] than publicly traded options data. However, it is unlikely that Ross's OTC dataset includes, for example, options with strike prices at every $1 interval. Moreover, contingent state prices require that we assume time homogeneity. To make this assumption, we must extrapolate option data based on time-to-expiration. I developed a methodology (see companion paper (Sanford, 2016b)) where I extrapolate readily available exchange traded option data on both the strike price and time-to-maturity dimensions by expanding on methods proposed by Figlewski (2008) and Chen (2011). This methodology makes the RT usable in any circumstance where we have sufficient data to estimate smooth splines.

I test the RT both in univariate and multivariate Markov chain settings. The forecast results indicate that the multivariate Markov chain produces results far superior to the univariate RT. Using quarterly forecasts (updated monthly) for the 1996-2015 period, the out-of-sample R-square of the RT increases from around 12% to 30%. Empirically, this paper constitutes one of the first exhaustive analyses of Ross's Recovery Theorem. This paper also provides an intuitive framework by which to understand both the univariate and the multivariate RT.

The chapter is divided into four main sections. Section 3.1 explains the univariate and multivariate RTs, and discusses the steps required to implement the theorem. It also walks through a simple numerical example for both the original univariate and the proposed multivariate RT. In section 3.2, I provide the setup by which we will be able to test the RT in a simulated environment. This section also argues that the inclusion of the implied volatility

[1]The notional amount for outstanding OTC equity-linked options is estimated to be $4.244 trillion while it is estimated to be $1.972 trillion for exchange traded options BIS (2012).

acts as a proxy for economic uncertainty. Section 3.3 introduces the data and presents the results. Finally, section 3.4 explores possible extensions and concludes.

## *3.1 Model*

The RT's ultimate goal is to obtain the natural probability distribution for asset returns (in this case, equity returns). It accomplishes this goal by deriving state prices using equity options. Using these state prices, we can then disentangle the discount rate, the risk-aversion parameter, and, ultimately, the natural probability distribution without making any parametric or utility function assumptions. I break down the RT into four steps:

1. construct the state prices,

2. construct the contingent state price matrix,

3. use the Perron-Frobenius (Meyer, 2000) theorem to extract the natural probability matrix, and

4. produce the natural marginal distributions, which can be used to obtain the recovered statistics (of which the recovered expected return and expected volatility are of particular interest).

To facilitate comparison, I adopt the same terminology and notation as Ross wherever possible. I do not present all of the proofs from the original RT since those can be found in Ross's paper. I limit the proofs in this paper to those that are new or crucial to the understanding of the model.

### *3.1.1 The Recovery Theorem*

Financial markets price assets as the present value of all future cash flows (Cochrane, 2009). However, if we are referring to risky assets, as is the case in this paper, prices are subject

to adjustments since future payoffs are not guaranteed and, by extension, are considered risky. We call this adjustment for the riskiness of the asset price the risk premium. The risk premium is defined as a function of the risk aversion and the overall level of risk of the asset being priced. We can refer to the price of an asset using the following equation (Cochrane, 2009):

$$p_t = E_t(m_{t+1}x_{t+1}) \tag{3.1}$$

where $p_t$ is the price of an asset at some time $t$, $E_t$ is the expectation operator, $m_{t+1}$ is the stochastic discount factor, and $x_{t+1}$ is the future cash flow of the asset. The variable $m_{t+1}$ in equation 3.1 is what gives us the risk premium because it is the adjustment to the price of an asset that makes it worthwhile for investors to purchase that asset given its level of risk. Part of the problem in pricing equities, however, is in defining the stochastic discount factor. In markets like the bond market, we can derive the forward rates. We obtain forward rates by comparing the yields of bonds with different expirations, which allows us to obtain the market's estimate of the stochastic discount factor. The same cannot be done with the equity market. So how can we estimate the risk premium? As Ross (2015) notes, we currently estimate the risk premium for equity markets by relying on historical returns or by using opinion polls. Historical returns assume that the past estimate of the risk premium is a good indicator of the future risk premium while opinion polls assume that the opinions of the analysts being polled reflect the entire market's overall sentiment. Both of these methodologies are flawed (Elton, 1999; Welch & Goyal, 2007; Campbell & Thompson, 2007).

In an effort to address these issues, Ross (2015) proposes to use options. Options, like forward rates, are forward-looking instruments with varying maturities. Hence, there is hope that we may use these securities to estimate the risk premium. That being said, option prices themselves do not explicitly depend on, or allow us to solve for, the risk premium. This is the question that motivates the original Recovery Theorem: how can we use option prices to obtain the risk premium? The RT provides a framework through which we can use options

to estimate state prices, which then allow us to estimate the underlying asset's risk premium.

### *3.1.2 Estimating state prices (S)*

Ross proposes that the starting point in deriving the equity risk premium is to obtain state prices from option prices. Why do we need state prices? We want a security that can be defined as a function of a pricing kernel and the true (or, as Ross calls them, "natural") probabilities. This is in essence a forward rate: a function of a pricing kernel and a probability. However, forward rates are not naturally found in equity markets, so we use option prices instead. Recall the definition for forward rates: today's rate for an asset that has a guaranteed payoff at some future point. Can these types of securities be obtained using equity options? An option can be defined as a function of the discount rate, the risk aversion parameter, and the probability of downside risk. However, we are not looking for an asset that is only a function of the left side of the returns distribution. Instead, we can construct a portfolio of options. We are going to call these portfolios "state prices." Formally, state prices correspond to the price of a security at some initial time, $t_0$, such that, at some future time $T$, the security pays a pre-specified amount (normalized to \$1) if the market is at a pre-specified state of the world and pays nothing otherwise. For example, assuming that the level of the S&P 500 today is 1,000, a state price would be the price of an asset that pays you 1\$ in, say, three months if the level of the S&P 500 is 1,500 at that time. The problem is that this type of security is not readily traded. Breeden & Litzenberger (1978) produce a method to derive state prices, beginning with the continuous time Black-Scholes-Merton equation (Black & Scholes, 1973; Merton, 1973) as follows:

$$Call(K,T) = \int_0^\infty [S_{t,p} - K]^+ p(S_{t,p}, T) dS_{t,p} = \int_K^\infty p(S_{t,p}, T) dS_{t,p}, \tag{3.2}$$

where $Call(K,T)$ is today's price for a call option with a strike price $K$ and time-to-maturity $T$. Taking the second derivative with respect to strike price $K$ gives the following result in

continuous time:

$$s(K,T) = Call''(K,T) \tag{3.3}$$

which is Breeden & Litzenberger's (1978) result. In discrete time, we can estimate equation 3.3 using a butterfly spread. A butterfly spread is a portfolio of three call options: buy a call option at strike price $K_1$, sell two call options at strike price $K_2$, and buy a call option at strike price $K_3$. Mathematically, this corresponds to the following equation:

$$s(K,T) \approx -Call_{K_1} + 2Call_{K_2} - Call_{K_3} \tag{3.4}$$

which, once standardized, gives a guaranteed payoff of \$1 at expiration $T$ if the market ends at $K_2$. Hence, we have defined and derived state prices. These state prices are the foundation of the Recovery Theorem.

Knowing the state price of a single state is not enough to solve for the natural probability distribution. We need $m$ equations but only have one set of equations, which implies that we cannot solve the system. However, if we knew the state prices for a complete set of states ($m$ states in this example), we would have $m$ equations and could solve the system of equations (see chapter two for more details). These $m$ equations will be obtained from the estimation of the contingent state prices.

**Numerical example** Before moving on to the derivation of the contingent state prices, let me introduce a simple numerical example that will be used throughout this paper. The goal of this example is twofold. First, it will provide the intuition behind the RT and its mechanics. Second, the example will show the importance of incorporating volatility in the derivation of contingent state prices (see section 3.1.3). The example will illustrate that a distribution that has a larger standard deviation will have a probability distribution function (pdf) that is wider (i.e., more probabilities in the tails) than one with a smaller standard

deviation. As a result, the probability of a given path is estimated more accurately when we consider volatility as a state variable in the model. This is especially true when we consider the probability of transitioning between states that are far away (e.g., the S&P 500 transitioning between a level of, say, 1,000 to a level of, say, 2,000 in a three-month period). These large movements are more likely to occur (higher probabilities) if the volatility is higher than if it is lower.

To begin, let us assume that we have a set of observable state prices in the economy. In particular, let us assume that we observe the following state prices:

| | 0.25 | 0.5 | 0.75 | 1 | 1.25 | 1.5 | 1.75 |
|---|---|---|---|---|---|---|---|
| **1,000** | 0.74 | 0.53 | 0.21 | 0.32 | 0.21 | 0.11 | 0.053 |
| **1,500** | 0.32 | 0.53 | 0.84 | 0.74 | 0.84 | 0.95 | 0.99 |

Table 3.1: State prices

where the rows represent the possible future state of the S&P 500 (in this case, I am assuming two possible states – 1,000 or 1,500) and the columns represent a specific timeframe (for example 0.25 corresponds to a state price in three months). The number highlighted in table 3.1 would be interpreted as the price of an asset that pays you 1$ if the S&P 500 ends at 1,000 in six months.

The example is purposefully kept very simple: there are only two possible future states (1,000 or 1,500). The entire system, up to this point, can be characterized for each time step using figure 3.1:

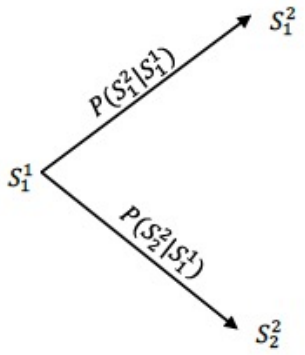


Figure 3.1: Generalized Setup

where $S_1^1$ represents the initial price or level of an underlying asset, $S_1^2$ and $S_2^2$ represent the two possible future states in our simplified world, and $P()$ represents the contingent state prices. For this example, I assume that the time-step, which is the difference between each time period in table 3.1, is set to three months. This simple world is one where we have a current level for the S&P 500, say $S_1^1 = \$1,000$, and where the possible future outcomes could be either $S_2^2 = \$1,000$ or $S_1^2 = \$1,500$. The next step involves the estimation of the contingent state prices $P()$.

### *3.1.3 Estimating contingent state prices (P)*

Contingent state prices are nothing more than state prices for initial states that are not currently observed in the market. This paper distinguishes between two derivations for these prices: the univariate (or "naïve") and multivariate contingent state prices.

*The univariate model*

In equation 3.1, I defined state prices as a function of a pricing kernel, $m$, and some future payout, $x$. Formally, contingent state prices are defined in the exact same way as state prices with the exception that these are now for states that are not observed in the market. We

can think of these as state prices for some future state, $i$, to some other future state, $j$. More intuitively, we can define the contingent state price matrix as an intermediate-step forward rate. In other words, it is the price of an asset in the future that guarantees a payoff of \$1 if the state of the world transitions from state $i$ to state $j$ at an intermediate time-step $t + \tau$, where $\tau > 0$. This is analogous to obtaining the forward rate at some future time-step. An intermediate time-step forward rate is the expected rate at time $t_0$ for rolling over a bond at some future time $t + \tau$ for a desired investment horizon that is at time $T$. This bond price is not known at the initial time, $t_0$. For example, if we assume an investment horizon of one year, we can decompose the forward rate into two six-month periods. We have the choice between investing in a one-year bond or investing in a six-month bond today and investing in another six-month bond in six months (rolling over the investment). The forward rate is thus the price at time zero (or the rate in this case) of the six-month bond that we will purchase six months from now for our total investment horizon of one year. The intuition for the contingent state price is the same. If we think about contingent state prices using the same horizons as the example for the forward rates, we have the price of a security that pays \$1 if the market starts at state $i$ in six months and expires at state $j$ in 12 months. Compared to the state prices estimated in the previous section, here we are estimating state prices for state levels that are hypothetical, rather than the current state level. For example, if we assume that the current level of the S&P 500 is, say, $1,000$ then we would calculate contingent state prices assuming a current level of the S&P 500 of, say, $1,500$ as well (in reality, we assume a large number of hypothetical current states). This understanding might seem trivial but it will be important later when I derive the multivariate Markov chain.

Before deriving the contingent state price matrix, I need to introduce an assumption that is crucial to its derivation.

**Assumption 1 (Time-Homogeneity)** *Time homogeneity implies that the contingent state price matrix, $P$, is not dependent on time.*

Using assumption 1, Ross (2015) estimates the contingent state price matrix using the following equation:

$$s_{t+1} = s_t P,\ t = 1, ..., m-1 \tag{3.5}$$

$$1 \geq P \geq 0$$

where $m$ is the number of states and $P$ is the contingent state price matrix. Assumption 1 allows me to obtain the contingent state prices using equation 3.5. Time homogeneity assumes that the contingent state prices are the same regardless of which time-step we are trying to estimate.

Now that I have derived the contingent state prices, I can rewrite equation 3.1 as follows:

$$p_{i,j} = \phi(\theta_i, \theta_j) f_{i,j} \tag{3.6}$$

where $p_{i,j}$ is a contingent state price, $\phi(\theta_i, \theta_j)$ is the kernel factor, and $f_{i,j}$ is the natural probability that we are ultimately trying to derive.

Once the contingent state price matrix has been obtained, the rest of the RT is derived using the Perron-Frobenius theorem along with some matrix algebra. At this point, we have all of the necessary components to solve for the natural probability matrix. However, a question still remains: can we improve on the estimation of the contingent state prices proposed by Ross? The next section extends the derivation of the contingent state prices to a multivariate Markov chain. This Markov chain controls for the volatility as well as the current level of the underlying asset.

**Numerical example (continued)** To maintain simplicity, I assume that there are only two possible hypothetical states of the world, $1,000$ and $1,500$. Let us assume the following univariate system:

| | 1000 | 1500 |
|---|---|---|
| 1000 | 0.32 | 0.74 |
| 1500 | 0.20 | 0.85 |

Table 3.2: Univariate contingent state price matrix

In table 3.2, the contingent state price of staying in state one $(1,000)$ is equal to 0.32 and the price of moving to state two $(1,500)$ is equal to 0.74. In other words, in this system, the price associated with transitioning from $S_1^1$ to $S_2^2$ is 0.74. Similarly, the price associated with transitioning from $S_1^1$ to $S_1^2$ is 0.85. In the first hypothetical state, $S_1^1$, investors believe that the market is more likely to switch to the a new state (state 2) over the next three months.

In table 3.2, the contingent state prices are not dependent on anything other than the initial state for that hypothetical world ($S_1^1$ or $S_2^1$). This is the major distinction between the setup of Ross and the setup proposed in this paper, and it will lead to a significant difference in the resulting expected natural distribution of returns. Before deriving the multivariate setup, I will derive the results for table 3.2 using the data in table 3.1. I will derive the value for $p_{1,1}$ using equation 3.5 as follows:

$$s_{t+1} = s_t \cdot p_{1,1} + \epsilon_t$$

where the vectors for $s_{t+1}$ and $s_t$ are defined respectively as:

$$s_{t+1} = \begin{bmatrix} 0.53 & 0.21 & 0.32 & 0.21 & 0.11 & 0.053 \end{bmatrix}'$$

$$s_t = \begin{bmatrix} 0.74 & 0.53 & 0.21 & 0.32 & 0.21 & 0.11 \end{bmatrix}'$$

such that we are solving the following equation:

$$\begin{bmatrix} 0.53 \\ 0.21 \\ 0.32 \\ 0.21 \\ 0.11 \\ 0.053 \end{bmatrix} = \begin{bmatrix} 0.74 \\ 0.53 \\ 0.21 \\ 0.32 \\ 0.21 \\ 0.11 \end{bmatrix} \cdot p_{1000,1000} + \epsilon_t$$

which gives us the solution we need, $p_{1,1} = 0.32$.

*The multivariate model*

Including the volatility in the derivation of the contingent state prices removes the assumption that volatility between periods is constant. This is the major contribution of this paper. Volatilities are different depending on the state path probability that we are trying to estimate. Hence, it becomes critical to control for these different changes in volatility in the contingent state price estimation. The intuition can be seen in the following figure:

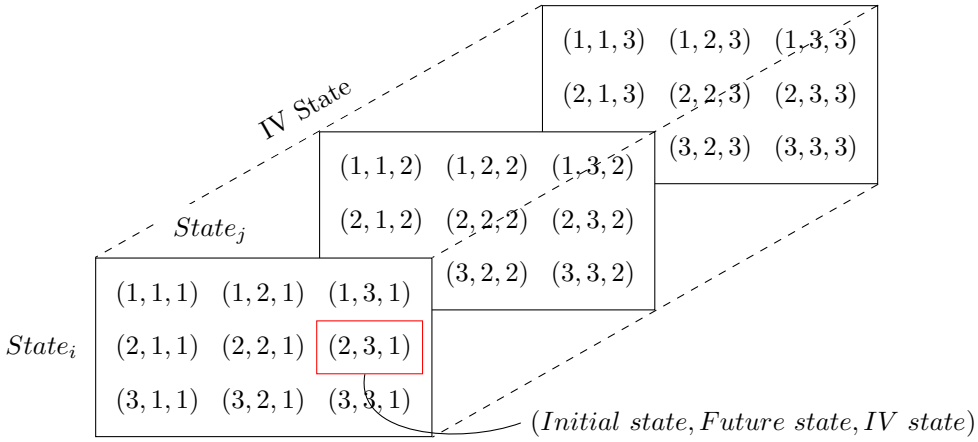


Figure 3.2: MVRT Intuition

where figure 3.2 shows the additional state variable, the implied volatility, added into the contingent state price specification.

Let us derive the multivariate Markov chain. The general specification for the multivariate Markov chain used in this paper was first introduced by Raftery (1985) and is as follows:

$$\min_{\lambda_{i,j}} \min_{P} [[\sum \lambda_{i,j} s_t P - s_{t+1}]_P] \tag{3.7}$$

where it must, by definition, be the case that:

$$1 \geq P \geq 0 \text{ and } \beta \geq 0$$

$$\sum \lambda_{i,j} = 1$$

More specifically, for the purposes of this paper, I can rewrite the general specification in equation 3.7 to a two-variable Markov chain as follows:

$$\min_{\lambda_{i,j}} \min_{P,\beta} [[\lambda_{i,j} s_t P + (1 - \lambda_{i,j}) \Phi_t \beta - s_{t+1}]_{P,\beta}] \tag{3.8}$$

$$1 \geq P \geq 0 \text{ and } \beta \geq 0$$

where $\Phi$ is an additional variable necessary for a more accurate derivation of the contingent state price matrix. A simple specification of the multivariate model is to assume that the contingent state price is solely defined by the state levels, but that we need to condition on the the volatility in the regression. This implies that we estimate the contingent state prices using a multivariate Markov chain as follows:

$$s_{t+1} = s_t P + Ivol_t \beta, \quad t = 1, ..., m-1 \tag{3.9}$$

where $Ivol_t$ is the implied volatility state at time $t$. In other words, equation 3.9 assumes that $\lambda = 1$ in equation 3.8. Implied volatility is used because it is the market's best estimate of the future volatility state. Equation 3.9 gives us a third dimension in the Markov chain and therefore results in a matrix of size $(m-1)^3$. Theoretically, we could add more variables to the regression equation. Since I estimate the Markov chain based on 11 states, however, it is best not to add too many variables to the regression equation because there will be too few degrees of freedom to consider the resulting contingent state price matrix reliable. Moreover, and this will be discussed in greater detail in section 3.2, volatility acts as a proxy for the uncertainty in the macroeconomy. Hence, controlling for volatility in contingent state prices gives us a better sense of the uncertainty of future state paths.

Including the volatility into the model, I solve the following equation:

$$\min_{P,\beta} \| s_{t+1} - s_t P - Ivol_t \beta \|^2 \tag{3.10}$$

where it must, by definition, be the case that:

$$1 \geq P \geq 0 \text{ and } \beta \geq 0 \tag{3.11}$$

Equation 3.11 includes a non-negativity condition in our regression such that $P \geq 0$. This is a necessary assumption for us to apply the Perron-Frobenius theorem in the next section. The assumption also makes intuitive sense since prices, by definition, are nonnegative. The upper bound on the contingent state price ensures that there are no prices that lead to arbitrage.

**Numerical example (continued)** The key insight from this paper is that the univariate state space model of the RT is not accurately specified. This idea is akin to one of an omitted variable bias. There may be a multitude of variables that affect the probabilities of transitioning from one state to another, but one of the most important variables is the volatility of the underlying asset. Volatility plays a critical role in specifying the probabilities of transitioning between states accurately. Extending the univariate example will show the impact of omitting volatility in deriving contingent state prices. Note that the time-step here is still $m$ (three months). In this example, there is only one possible volatility states, high. The resulting contingent state prices are now a function of both the initial state, $S_{1,1}$, and the volatility state, $\sigma_H$. I now assume that we have the following multivariate system:

| | 1000 | 1500 |
|---|---|---|
| 1000 | 0.21 | 0.84 |
| 1500 | 0.31 | 0.95 |

Table 3.3: Price | IVol

The contingent state price of $S_1^2$ given $S_1^1$ and $\sigma_H$, $P(S_1^2|S_1^1, \sigma_H)$, is 0.21. It is best to focus on what the contingent state prices represent and their intuition. For example, $P(S_2^2|S_1^1, \sigma_H)$ is equal to 0.31 because it is more likely that the market will transition to a far away state given a high volatility state. By contrast, the contingent state price of moving from an initial state one to the future state two is much more likely given a high volatility state. As such,

the contingent state price, $P(S_1^2|S_1^1, \sigma_H)$, is 0.84.

I will derive the value for $p_{1,1}$ using equation 3.5 as follows:

$$s_{t+1} = s_t \cdot p_{1000,1000} + IVol_t\eta_t + \epsilon_t$$

where the vectors for $s_{t+1}$ and $s_t$ are defined respectively as:

$$s_{t+1} = \begin{bmatrix} 0.53 & 0.21 & 0.32 & 0.21 & 0.11 & 0.053 \end{bmatrix}'$$

$$s_t = \begin{bmatrix} 0.74 & 0.53 & 0.21 & 0.32 & 0.21 & 0.11 \end{bmatrix}'$$

$$s_t = \begin{bmatrix} high & high & high & high & high & high \end{bmatrix}'$$

such that we are solving the following equation:

$$\begin{bmatrix} 0.53 \\ 0.21 \\ 0.32 \\ 0.21 \\ 0.11 \\ 0.053 \end{bmatrix} = \begin{bmatrix} 0.74 \\ 0.53 \\ 0.21 \\ 0.32 \\ 0.21 \\ 0.11 \end{bmatrix} \cdot p_{1000,1000} + \begin{bmatrix} high \\ high \\ high \\ high \\ high \\ high \end{bmatrix} \eta_t + \epsilon_t$$

which gives us the solution we need, $p_{1,1} = 0.21$. Once we have the contingent state prices, we can apply the RT to recover a natural probability distribution and our estimate of the expected return of an asset, as shown in the next section.

### *3.1.4 Estimating the natural probability distribution (F)*

At this point in the derivation, we are combining all of the elements from the previous sections to obtain the natural probability matrix. The natural probability matrix represents the market's best estimate of the future distribution of returns for the original option's underlying asset. This section describes the required theorem, assumptions, intuition, and methodologies to obtain the natural probability matrix. The first assumption is time-separable utility, which can be defined as follows:

**Assumption 2 (Time-Separable Utility)** *Time-separable utility implies that we can define the pricing kernel $\phi()$ as:*

$$\phi(\theta_i, \theta_j) = \delta \frac{U'(c(\theta_j))}{U'(c(\theta_i))} \tag{3.12}$$

*where $\delta$ is a discount rate such that $\delta \in (0, 1]$, and $U' > 0$ is the marginal utility for state $j$ or $i$.*

Intertemporal additive utility is assumed because it generates a transition independent kernel. It follows from the setup of an intertemporal model with a representative agent that has additive time-separable preferences. Once we have obtained the contingent state price matrix from section 3.1.3, we can apply Ross's RT (for proof, see Ross (2015)). Using a discrete time setup and assumption 2, I can rearrange equation 3.6 as:

$$U'_i p_{i,j} = \delta U'_j f_{i,j}, \tag{3.13}$$

where $U'_i$ is the marginal utility such that:

$$U'_i \equiv U'(c(\theta_i)) \tag{3.14}$$

which can then be written in terms of the normalized kernel:

$$\phi_j \equiv \phi(\theta_1, \theta_j) = \delta(\frac{U'_j}{U'_1}) \tag{3.15}$$

where $\theta_1$ is the current state. In continuous time, Ross defines the kernel as:

$$\phi(\theta_i, \theta_j) = \delta\frac{h(\theta_j)}{h(\theta_i)} \tag{3.16}$$

Using equation 3.16 and assuming transition independence, we have:

$$p(\theta_i, \theta_j) = \phi(\theta_i, \theta_j) f(\theta_i, \theta_j) = \delta\frac{h(\theta_j)}{h(\theta_i)} f(\theta_i, \theta_j) \tag{3.17}$$

where $h(\theta) = U'(c(\theta))$, and $p(\theta_i, \theta_j)$ is the state price transition function that was derived in section 3.1.3. From there, the objective is to solve the unknowns: the natural probability transition function $f(\theta_i, \theta_j)$, the kernel $\phi(\theta_i, \theta_j) = \delta\frac{h(\theta_j)}{h(\theta_i)}$, and the discount rate $\delta$. Back to the discrete time specification, we can rewrite equation 3.17 in matrix form as:

$$DP = \delta FD \tag{3.18}$$

where $P$ is the $m$ x $m$ state price matrix defined in section 3.1.2, $F$ is the $m$ x $m$ matrix that we are calling the natural probabilities and is the matrix of interest for this section, and $D$ is the diagonal matrix of undiscounted kernels or a diagonal of marginal rates of substitution as follows:

$$D = \frac{1}{U'_1}\begin{bmatrix} U'_1 & 0 & 0 \\ 0 & U'_i & 0 \\ 0 & 0 & U'_m \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & \phi_i & 0 \\ 0 & 0 & \phi_m \end{bmatrix}\frac{1}{\delta} \tag{3.19}$$

Rearranging equation 3.18, we get:

$$F = \frac{1}{\delta} DPD^{-1} \tag{3.20}$$

We obtained $P$ in section 3.1.3, so now $D$ must be estimated. Up to this point, the RT has not provided us with additional insight into disentangling the discount rate, pricing kernel, and natural probability distribution because there were not enough variables and equations to solve our system of equations. The key, however, is to notice that $F$ is a stochastic matrix which, by definition, implies that the rows of $F$ are transition probabilities and so they must sum to 1. Hence, we have the following equation:

$$Fe = e \tag{3.21}$$

where $e$ is simply a vector of ones. Substituting equation 3.21 into equation 3.20, we obtain:

$$Fe = \frac{1}{\delta} DPD^{-1}e = e \tag{3.22}$$

and if we define $z \equiv D^{-1}e$, we can rewrite equation 3.22 as:

$$Pz = \delta z \tag{3.23}$$

This still does not allow us to solve for $D$. However, we can make some assumptions about $P$ that will allow us to use the Perron-Frobenius Theorem (Meyer, 2000). Namely, we can assume that the option prices have no arbitrage opportunities (which, by definition, must be the case). No arbitrage implies that the contingent state price matrix will be nonnegative and less than one. Prices are, by definition, nonnegative, which was specified in the derivation of the contingent state price matrix in section 3.1.3. The second necessary assumption is that the matrix $P$ be irreducible. A matrix is said to be irreducible if we can reach any

state in $k$-steps. As Ross (2015) argues, even if some of the prices in $P$ correspond to a transition probability equal to zero, it should still be possible to reach the desired state via an intermediary state (or states). As such, since $P$ is nonnegative and irreducible, we can apply the Perron-Frobenius Theorem (Meyer, 2000), which states that all nonnegative and irreducible matrices have a unique positive characteristic root (eigenvector) $z$, and a Perron root $\delta$. This allows us to solve for $D$, which we can introduce in the natural probability distribution equation:

$$F = \frac{1}{\delta} DPD^{-1} \tag{3.24}$$

The previous paragraph explains the mechanics of obtaining the true distribution. But what has the application of the Perron-Frobenius theorem allowed us to accomplish? The Perron-Frobenius theorem provides us with two pieces of information critical to the derivation of the natural probability distribution: the discount factor ($\delta$) and the risk aversion ($D$). We obtain the discount factor and risk aversion using the marginal rate of substitution defined in equation 3.19. The components of the marginal rate of substitution are the marginal utilities of consuming today versus consuming tomorrow. The Perron-Frobenius theorem allows us to determine the single unique discount factor and marginal utilities that dictate the transition paths between states. In other words, under the assumptions of the Perron-Frobenius theorem, only one set of marginal utilities and one discount factor will hold. Basically, they are relating the discounted willingness for the representative agent to consume today versus consuming at some other period in the future given certain transition probabilities.

Once we have the true probability matrix, obtaining the market forecast becomes trivial. We divide state prices by the kernel to obtain the natural marginal probabilities. We multiply the natural marginal probabilities by the state levels to obtain an expected return for each time interval.

## *3.2 What does Implied Volatility Capture?*

In section 3.1, I alluded to the fact that the inclusion of the implied volatility in the derivation of the contingent state prices acted as a proxy for uncertainty in the macroeconomy. In this section, I show that including the implied volatility allows us to capture uncertainty in the business cycle. For example, intuitively, we should expect that, when the probability of a recession is high, the expected return would be low. The -0.5 correlation between the Federal Reserve's estimated U.S. Recession Probabilities (Chauvet & Piger, 2008) and the realized risk-premium illustrates that fact. Now, if we correlate the Federal Reserve's estimated U.S. Recession Probabilities with the univariate RT and the MVRT, we obtain correlations of 0.11 and -0.21 respectively. Hence, the MVRT seems to capture more of the uncertainty than the univariate RT (as it is much closer to -0.5).

To test the idea put forth in this section, I simulate data using Monte Carlo simulations similar to the ones proposed by Heston (1993). In this setup, we obtain the simulated stock price from a Geometric Brownian Motion (GBM) and the stochastic volatility from a stochastic process as in Cox et al. (1985). The parameters used in these simulations can be found in section 3.3.3. Figures 3.3 and 3.4 illustrate ten series of simulated stock prices and volatilities:

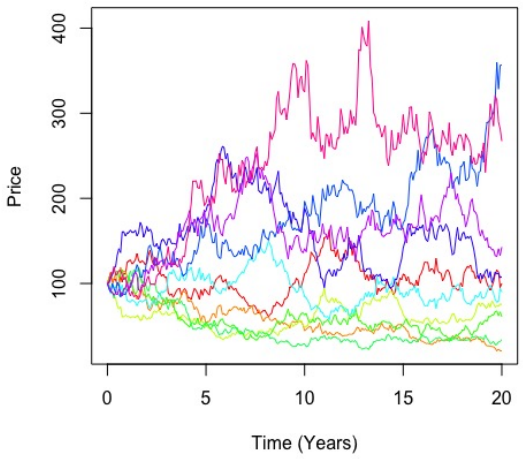


Figure 3.3: Stock Prices

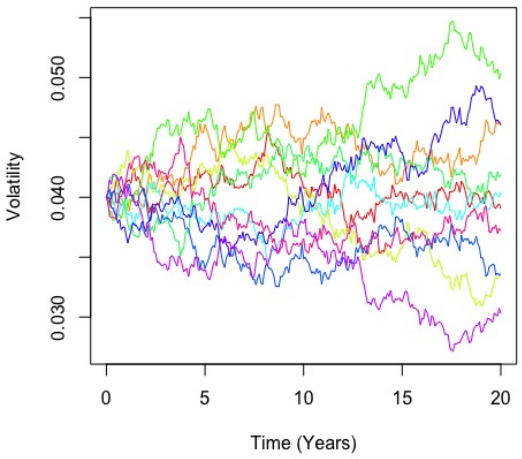


Figure 3.4: Stochastic Volatility

Once the data has been generated, I derive a binomial model with a representative agent that has heterogeneous habit formation (Campbell & Cochrane, 1999). The habit formation from Campbell & Cochrane (1999) is what generates the time-varying risk-premium. I start by defining the binomial model based off of Cox et al. (1979). We first define the initial stock prices as the state-dependent value of a stock as follows:

$$S = p_u \cdot S_u + p_d \cdot S_d \tag{3.25}$$

where $p$ is the risk-neutral probability of an up ($u$) or down ($d$) movement in the market and $S$ is defined as:

$$\begin{aligned} S_u &= u \cdot S_0 \\ S_d &= d \cdot S_0 \end{aligned} \tag{3.26}$$

where $S_0$ is the current stock price (or initial stock price), $u$ and $d$ represent up or down

movements in the market over a specific horizon and $S_u$ ($S_d$) represents the stock price after an a hypothetical up (down) movement. The up or down movements depend on whether we are trying to model the univariate or the multivariate RT. For the univariate RT, the movements are defined as:

$$\begin{aligned} u &= 1 + \sigma\sqrt{T} \\ d &= 1 - \sigma\sqrt{T} \end{aligned} \tag{3.27}$$

where $\sigma$ is the actual volatility observed in the market. For the multivariate RT, the movements are defined as:

$$\begin{aligned} u &= 1 + \sigma_{IVOL}\sqrt{T} \\ d &= 1 - \sigma_{IVOL}\sqrt{T} \end{aligned} \tag{3.28}$$

where the implied volatility is defined as the next period's volatility, $\sigma_{t+1}$, plus or minus an error term. The error term is a value taken from a standard normal distribution: $\epsilon_t \sim N(0,1)$. The implied volatility is defined as the market's best estimate of the future volatility. By taking the next period's volatility and adjusting it by some error term, I am suggesting that the market has some sense of future volatility, but that its estimation is imperfect.

Recall from equation 3.6 that we defined the price of an asset as:

$$p_{t+1} = \phi_{t+1} f_{t,t+1} \tag{3.29}$$

where $\phi_{t+1}$ is the intertemporal marginal rate of substitution and $f_{t,t+1}$ is the natural probability measure. In order to obtain a forecast, we must first derive the intertemporal marginal rate of substitution. This is done using defining preferences as a function of external habit formations. These habit formations are a function of aggregate consumption, $C_t^a$, and an

individual's habit, $X_t$, as follows:

$$S_t^a = \frac{C_t^a - X_t}{C_t^a} \tag{3.30}$$

which can be specified as the log surplus consumption ratio $s_t^a = ln\ S_t^a$ which evolves as a heteroskedastic AR(1) process:

$$s_t^a = (1-\omega)\bar{s} + \omega s_{t-1}^a + \lambda(s_{t-1}^a)(c_t^a - c_{t-1}^a - g) \tag{3.31}$$

where $\omega$ and $g$ are parameters from Campbell & Cochrane (1999) (summarized in section 3.3.3). Parameter $\bar{s}$ represents the log of the steady state surplus consumption ratio and is defined as:

$$\bar{S} = \sigma\sqrt{\frac{\gamma}{1-\omega}} \tag{3.32}$$

where $\gamma$ is the risk-aversion parameter. The sensitivity function, $\lambda(s_t^a)$, is defined as:

$$\lambda(s_t^a) = \begin{cases} \frac{1}{\bar{S}}\sqrt{1-2(s_t-\bar{s})} - 1, & s_t \leq s_{max} \\ 0, & s_t \geq s_{max} \end{cases} \tag{3.33}$$

where $s_{max}$ is defined as:

$$s_{max} = \bar{s} + \frac{1}{2}(1-\bar{S}^2) \tag{3.34}$$

Consumption growth is modeled as an i.i.d. lognormal process:

$$\Delta c_{t+1} = g + v_{t+1} \tag{3.35}$$

where $v_{t+1} \sim i.i.d.\ N(0, \sigma^2)$. The intertemporal marginal rate of substitution, in this case, is as follows:

$$\phi_{t+1} = \delta\Big(\frac{S_{t+1}}{S_t}\frac{C_{t+1}}{C_t}\Big)^{-\gamma} \tag{3.36}$$

which can then be used in equation 3.29. Once we have obtained the intertemporal marginal

rate of substitution, we can apply the RT derived in earlier sections to obtain the natural probability distribution $f_{t,t+1}$.

## *3.3 Data and results*

### *3.3.1 Overview of data*

I collected the data for this paper from the Wharton Research Data Services (WRDS) database. I use daily option prices on the S&P 500, the S&P 500's closing price, and the risk-free rate. The risk-free rate is the one-month Treasury Bill rate, which can be found in the Fama & French factors data. S&P 500[2] prices are from the CRSP dataset. The S&P 500 is generally thought to be the best proxy for the market portfolio. All of the option data are from OptionMetrics. The data are used to obtain forecasts at intervals that range from one day to one quarter. This paper covers the time period from January 1996 to July 2015, the entire timeframe included in the OptionMetrics database. I use this sample for two major reasons. First, one of the forecast horizons in this paper is quarterly. A quarterly forecast requires a large enough sample size to test the efficacy of the RT and this twenty-year sample provides me with approximately 80 data points. Second, it allows me to divide the sample into subsamples and test my model in periods that experience various shocks (such as the tech bubble and the recent financial crisis).

Strike prices on the options obtained from OptionMetrics are quoted for lots of 1,000 securities. The Black-Scholes-Merton equation requires strike prices that are on a per-stock basis, so I divided the strike price by 1,000. Time-to-maturity is converted from a date to a fraction of years to expiration, also a required input for the Black-Scholes-Merton equation. Option price is replaced with the midpoint of the bid-ask spread. This is consistent with Figlewski (2008), who argues that bid and ask prices are continuously quoted for almost all strikes regardless of whether a trade takes place. The alternative, transaction prices,

[2]SECID 108105

occurs irregularly (Figlewski, 2008) and would make it more difficult to extract a proper implied volatility curve. I compare my estimated implied volatilities to those provided by OptionMetrics. Since the difference between the two is negligible, I use my more complete set of estimates instead of the OptionMetrics data.

One of the difficulties of applying/replicating the RT is in constructing state prices. Ross (2015) uses over-the-counter data rather than the more limited publicly available data because it offers a significantly larger number of traded strikes and maturities. This paper uses readily available data from WRDS instead. Despite this difference in data source, I produce results that are very close to Ross's (see section 3.3). Another difficulty is that Ross (2015) does not explain how he derives state prices. Theoretically, state prices are easy to understand, but in practice, there is a lot of debate on how to construct them. Sanford (2016b) proposes a way to derive the extrapolated data required to construct state prices for this paper.

### *3.3.2 Empirical results*

This section presents the empirical results for the univariate and the multivariate recovery theorems. I divide the samples into three subsamples to show the impact of different volatility states on the results. The first set of results is for the entire sample (April 1996 to August 2015). The high volatility subsample is from April 1996 to April 2002. The low volatility subsample is from January 2004 to January 2007. I selected the subsamples by examining time series plots to determine which periods had high volatility and which had low volatility.

#### *Full sample results*

Table 3.4 compares the results of Ross (Ross UVRT – first column) with the results of the multivariate RT (MVRT – second column) proposed in this paper to illustrate the superiority of the MVRT. Please note that the univariate results are the closest possible proxy for the

results of Ross (since I did not have access to the data to replicate Ross's results exactly). All results presented in this section are out-of-sample. The very nature of the RT is such that in-sample results are not possible. Comparing the out-of-sample adjusted $R^2$, the MVRT method produces results superior to Ross's methodology.

| | Ross UVRT (Apr 09–Apr 13) | MVRT (Apr 09–Apr 13) |
|---|---|---|
| | (1) | (2) |
| Intercept | $-0.06054^{*}$ | $0.027675^{**}$ |
| | (0.035068) | (0.009153) |
| Coefficient | $5.710293^{**}$ | $0.338864^{***}$ |
| | (1.95258) | (0.070478) |
| Observations | 46 | 49 |
| $R^2$ | 0.2162744 | 0.329701 |
| Adjusted $R^2$ | 0.143715 | 0.315439 |
| F statistic | 0.005436 | $1.6e^{-05}$ |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.4: Ross Subsample - Summary Results

The tables below have four columns, each representing the result for a specific forecasting methodology. The first column is the univariate RT (UVRT), the proxy for Ross's original RT. The second column is the multivariate RT (MVRT), the new method proposed in this paper. The third column is the dividend-price ratio ($D/P$). The fourth column is the consumption-wealth ratio (CAY). The forecast regression equation is as follows:

$$R_t = \alpha + \beta E_{t-1}[R_t] + \epsilon_t \tag{3.37}$$

where $\alpha$ is the intercept, $\beta$ is the forecast coefficient, and $E_{t-1}[R_1]$ is the previous period's

RT forecast. The forecast horizon is held to a quarter (three months) so $t$ corresponds to 0.25 years. One of the criteria for forecast efficiency is the forecast error. This error is defined as the residual, $\epsilon_t$, found in equation 3.37 and graphed in section 3.3.2. The errors are used as a way to ensure that the model is accurately specified. In general, the smaller the errors, the better the forecast.

Table 3.5 presents the results for the entire sample (April 1996 to August 2015).

| | UVRT (Apr 96–Aug 15) | MVRT (Apr 96–Aug 15) | $D/P$ (Apr 96–Aug 15) | CAY (Apr 96–Aug 15) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Intercept | 0.01040 | 0.00482 | −0.00378 | 0.01936 |
| | (0.00930) | (0.00465) | (0.01557) | (0.00836) |
| Coefficient | 1.66110*** | 0.42471*** | 13.96761 | 0.65015 |
| | (0.29290) | (0.04259) | (9.45251) | (0.48717) |
| Observations | 235 | 235 | 235 | 78 |
| $R^2$ | 0.12267 | 0.30187 | 0.00928 | 0.02290 |
| Adjusted $R^2$ | 0.11885 | 0.29884 | 0.00503 | 0.01004 |
| F statistic | $4.244e^{-08}$ | $1.069e^{-19}$ | 0.14085 | 0.18601 |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.5: Results for the four methods, full sample

The MVRT clearly outperforms all other benchmark results presented in table 3.5. The out-of-sample adjusted $R^2$ is 0.29884 compared to the UVRT's adjusted $R^2$ of 0.11885. This significant increase is consistent across samples, indicating that the MVRT provides significantly better results than previous methods. The MVRT results are also significantly better than the results for other benchmark forecasting methodologies such as the dividend-price ratio and the CAY ratio.

The ideal coefficients in a forecast are for the intercept to be zero and the slope coefficient

to be one. In table 3.5, the slope coefficient in the MVRT is closer to one while maintaining the same level of significance as the UVRT. Both the UVRT and the MVRT seem to indicate that the intercept coefficient is equal to zero. Overall, the results look promising.

To test for robustness, the next set of results break down the original sample into smaller periods with either high or low volatilities. High-volatility subsamples represents periods where the volatility was constant at around 10% while low-volatility samples were periods where the volatility was around 5%. I also add periods (i.e., several months of data) of large changes in volatility to examine the effect on the forecast regression results. Based on the theory, the model should perform best when volatility remains relatively unchanged over time.

*High-volatility subsample results*

The first subsample is from April 1996 to April 2002. This subsample is the first period of time in the data where the volatility remains relatively high (and unchanged) throughout the sample ($\approx 8\%$).

| | UVRT (Apr 96–Apr 02) | MVRT (Apr 96–Apr 02) | $D/P$ (Apr 96–Apr 02) | CAY (Apr 96–Apr 02) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Intercept | 0.05871 | 0.00352 | −0.04873 | 0.02190 |
| | (0.01405) | (0.00675) | (0.02976) | (0.01884) |
| Coefficient | 3.15148*** | 0.58939*** | 58.51909** | 0.61102 |
| | (0.86772) | (0.06343) | (21.94959) | (1.26921) |
| Observations | 73 | 73 | 73 | 24 |
| $R^2$ | 0.15668 | 0.54871 | 0.09100 | 0.01042 |
| Adjusted $R^2$ | 0.14480 | 0.54236 | 0.07820 | -0.03456 |
| F statistic | 0.00053 | $6.814e^{-14}$ | 0.00950 | 0.63497 |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.6: Results for the four methods, April 1996 to April 2002

In table 3.6, the results for the MVRT are quite impressive. The out-of-sample adjusted $R^2$ is almost 55% compared to about 16% for the UVRT. This is quite large for a forecast, likely because there are very little changes both in the mean and the volatility of returns during this time period. We can see this by looking at the $D/P$ ratio, which also shows a significant forecasting ability. Normally, we would expect the dividend-price ratio to forecast long-term changes in asset prices. However, it seems to perform quite well during this period. Much like in the entire sample, the slope coefficient for the MVRT is getting closer to the desired coefficient of one. Moreover, the intercept does seem to be zero as we would hope.

*Low-volatility subsample results*

This next subsample, shown in table 3.6, is from April 2004 to January 2007. This period has a relatively low and constant volatility of around 4.7%.

| | UVRT (Jan 04–Jan 07) | MVRT (Jan 04–Jan 07) | $D/P$ (Jan 04–Jan 07) | CAY (Jan 04–Jan 07) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Intercept | 0.09510 | 0.01896* | 0.02177 | 0.02896 |
| | (0.14100) | (0.00770) | (0.02461) | (0.02706) |
| Coefficient | 5.64410 | 0.23165* | 2.46158 | 0.30591 |
| | (4.74010) | (0.08897) | (15.82837) | (1.97238) |
| Observations | 37 | 37 | 37 | 13 |
| $R^2$ | 0.03895 | 0.16225 | 0.00069 | 0.00218 |
| Adjusted $R^2$ | 0.01149 | 0.13832 | -0.02786 | -0.08853 |
| F statistic | 0.24170 | 0.01344 | 0.87731 | 0.87955 |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.7: Results for the four methods, January 2004 to January 2007

During this time period, all forecasting methodologies perform miserably with the exception of the MVRT. The best performance was from the UVRT which had an out-of-sample adjusted $R^2$ of about 1.1% while the MVRT's adjusted $R^2$ is about 14%. The statistical significance of the slope coefficient has decreased when compared with other sample periods. That being said, it is the only result during this period to achieve any level of statistical significance.

The following table examines what happens when I add months in the sample that have large changes in volatility. Using the sample from table 3.7 above as a starting point, I added eight months of data before and two years of data after. In total, the sample size went from 37 months to 73 months. Again, the purpose here is to study the impact of adding periods where the volatility changes on the results. These months changed the volatility for the period from about 4.7% to about 9%.

| | UVRT (Apr 03–Apr 09) | MVRT (Apr 03–Apr 09) | $D/P$ (Apr 03–Apr 09) | CAY (Apr 03–Apr 09) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Intercept | −0.04171 | −0.00404 | 0.08148* | −0.00540 |
| | (0.01940) | (0.00993) | (0.03112) | (0.01768) |
| Coefficient | 1.80831** | 0.33532** | −50.92840** | −1.36018 |
| | (0.63681) | (0.10376) | (18.57465) | (1.20377) |
| Observations | 73 | 73 | 73 | 25 |
| $R^2$ | 0.10200 | 0.12823 | 0.09574 | 0.05259 |
| Adjusted $R^2$ | 0.08935 | 0.11595 | 0.08301 | 0.01140 |
| F statistic | 0.00588 | 0.00187 | 0.00773 | 0.27016 |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.8: Results for the four methods, April 2003 to April 2009

From table 3.8, it is clear that the change in the volatilities has led to a decrease in the MVRT's forecasting ability. That being said, the difference is not substantial. The adjusted $R^2$ has decreased from around 14% to around 11.5%. The most dramatic change in this table appears in the other forecasting models. Specifically, the UVRT and the $D/P$ results have substantially improved. Intuitively, these results should not be surprising. The UVRT is not as affected by changes in the volatility levels as the MVRT. It takes time for the MVRT to improve after a substantial change in the volatilities. This is not necessarily the case for the UVRT. That being said, the MVRT still outperforms all of the benchmark forecasts presented in this table. So it is still performing quite well, just not as well as we might have hoped.

The next subsample is from April 2010 to the end of the sample period: August 2015. Much like the previous period, this subsample shows a relatively small volatility of about 5%.

| | UVRT (Apr 10–Aug 15) | MVRT (Apr 10–Aug 15) | $D/P$ (Apr 10–Aug 15) | CAY (Apr 10–Aug 15) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Intercept | 0.00130 | 0.01967** | 0.01480 | 0.03070 |
| | (0.00981) | (0.00604) | (0.02297) | (0.01746) |
| Coefficient | 2.00011*** | 0.24430*** | 9.80293 | 0.20893 |
| | (0.49340) | (0.04754) | (12.62916) | (0.75418) |
| Observations | 65 | 65 | 65 | 23 |
| $R^2$ | 0.20697 | 0.29538 | 0.00947 | 0.00364 |
| Adjusted $R^2$ | 0.19439 | 0.28420 | -0.00625 | -0.04380 |
| F statistic | 0.00014 | $2.892e^{-06}$ | 0.44053 | 0.78446 |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.9: Results for the four methods, April 2010 to August 2015

In table 3.9, both the UVRT and the MVRT perform quite well (although the MVRT does outperform the UVRT again). The out-of-sample adjusted $R^2$s were about 20% and 28% for the UVRT and MVRT respectively.

For this next subsample, I added 24 months to the subsample. The additional 24 months displayed higher volatility (from the financial crisis), which added a shift in the volatility to the sample. The volatility increased from about 5% to almost 9%.

| | UVRT (Apr 08–Aug 15) | MVRT (Apr 08–Aug 15) | $D/P$ (Apr 08–Aug 15) | CAY (Apr 08–Aug 15) |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Intercept | −0.01510 | 0.00342 | 0.00232 | −0.00596 |
| | (0.01521) | (0.00896) | (0.02938) | (0.01744) |
| Coefficient | 1.59471** | 0.34140* | 8.65916 | −1.47744 |
| | (0.52310) | (0.06878) | (15.38079) | (0.84335) |
| Observations | 89 | 89 | 89 | 31 |
| $R^2$ | 0.09651 | 0.22067 | 0.00363 | 0.09570 |
| Adjusted $R^2$ | 0.08613 | 0.21172 | -0.00782 | 0.06452 |
| F statistic | 0.00305 | $3.412e^{-06}$ | 0.57489 | 0.09037 |

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Table 3.10: Results for the four methods, April 2008 to August 2015

This last sample includes part of the financial crisis. As such, there was a major change in the volatility levels. This is reflected in the relatively worse results of the MVRT when comparing the results from table 3.10 to those from table 3.9. Moreover, the statistical significance of the slope coefficient substantially decreases despite the larger sample size.

*Varying the forecast horizon*

In the previous subsection, I showed the results for various time periods while keeping the forecast horizon the same. Here I show the results for a monthly, quarterly, and yearly forecast. In this section, however, the quarterly forecast is updated every quarter instead of every month as in the previous section. The overlap causes a slight upward bias on the adjusted $R^2$ results. This serves the purpose of showing that although there is bias, it is quite small. The results for the various forecast horizons are summarized in figures 3.5 and 3.6. Figure 3.5 shows the coefficients for the UVRT and the MVRT only. Both models perform

quite well (small errors) in the medium-term forecasts (monthly to quarterly) but the results start to deteriorate at the yearly forecast level. This is to be expected since options are not liquid at the annual time-to-maturity. This results in a forecast that is unreliable. Although the daily forecast result is not shown here, the forecast performs as poorly as the yearly forecast for the same reason.

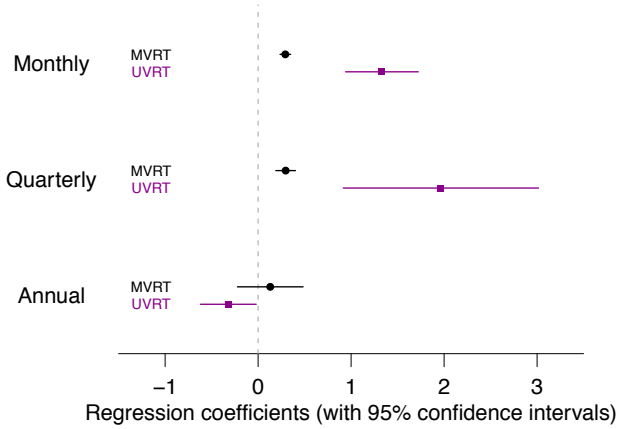


Figure 3.5: Regression Coefficients

Figure 3.6 shows the adjusted $R^2$ results at the various forecast horizons and compares those results to those of the dividend-price ratio. As was the case for the coefficients, the UVRT and the MVRT both perform well in the monthly and the quarterly forecast but are outperformed by the dividend-price ratio at the yearly forecast.

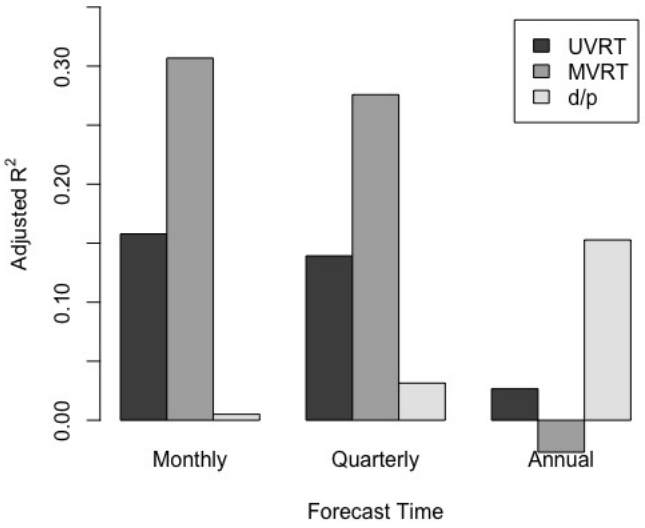


Figure 3.6: Adjusted $R^2$

### *3.3.3 Simulated results*

This section presents the results using simulated data (see section 3.2). The goal is twofold: to show 1) that the results are not merely a construct of the empirical data, and 2) that the MVRT captures some of the uncertainty in the business cycle. The uncertainty in the business cycle comes from the time-varying risk-premium. A model that successfully captures the uncertainty in the business cycle would be the model that has the highest predictive power. Table 3.11 below shows the values used for the parameters required in the simulations.

| Parameter | Variable | Value |
|---|---|---|
| Assumed: | | |
| Mean consumption growth (%)* | $g$ | 1.89 |
| Standard deviation of consumption growth (%)* | $\sigma$ | 1.50 |
| Log risk-free rate (%)* | $r^f$ | 0.94 |
| Persistence coefficient* | $\omega$ | 0.87 |
| Initial stock price | $S_0$ | 100 |
| Number of simulations | $n$ | 10000 |
| Volatility mean-reversion speed | $\kappa$ | 0.003 |
| Volatility of volatility | $\sigma(\sigma)$ | 0.009 |
| Correlation between stochastic volatility and spot prices | $\rho$ | -0.5 |
| Initial variance | $\sigma_0^2$ | 0.04 |
| Long-term variance | $\theta$ | 0.04 |
| Reproducibility seed | NA | 123 |

* Annualized values

Table 3.11: Parameters for simulations

Figures 3.7 and 3.8 below show the simulation results for the UVRT. Figure 3.7 shows the regression coefficient and figure 3.8 shows the adjusted $R^2$ for various risk-aversion parameters. The horizontal line represents the coefficient from the regression using empirical data. The goal is to determine which risk-aversion coefficient matches the empirical results. For the coefficient, the risk-aversion parameter that gives us the same results for the simulated data as the empirical data is between 7.5 and 15. The adjusted $R^2$ is presented for completeness. For some reason, it takes a very large risk-aversion parameter in order to be able to replicate the empirical forecastability results. Nevertheless, the model does seem to have forecasting power whenever a "realistic" risk-aversion parameter is considered.

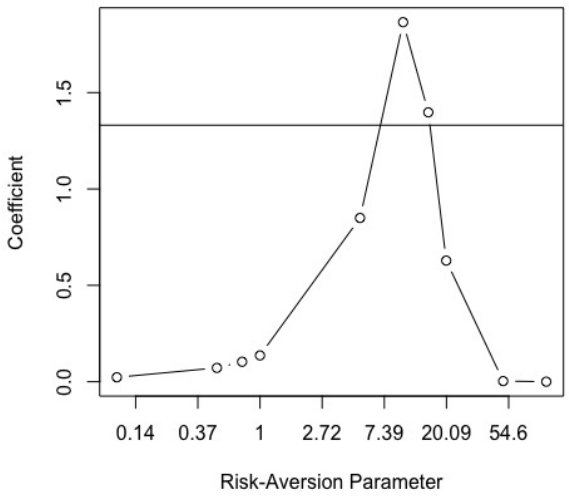


Figure 3.7: UVRT Simulations - Coefficient

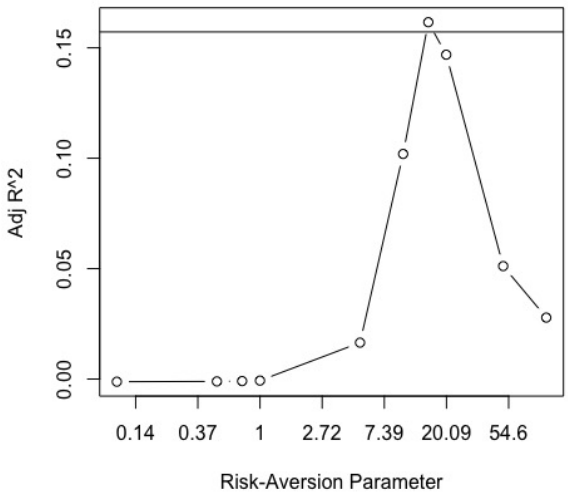


Figure 3.8: UVRT Simulations - Adj $R^2$

Figures 3.9 and 3.10 show the simulation results for the MVRT. The risk-aversion parameter where the simulated data and the empirical data converge is between 4.5 and 7.5. These values are much closer to what we would expect in reality than the UVRT values.

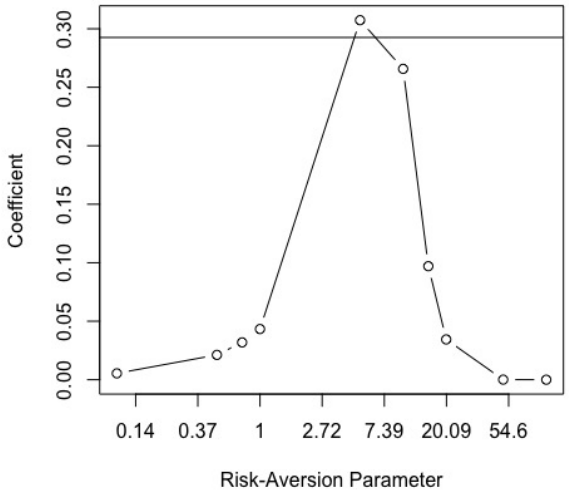


Figure 3.9: MVRT Simulations - Coefficient

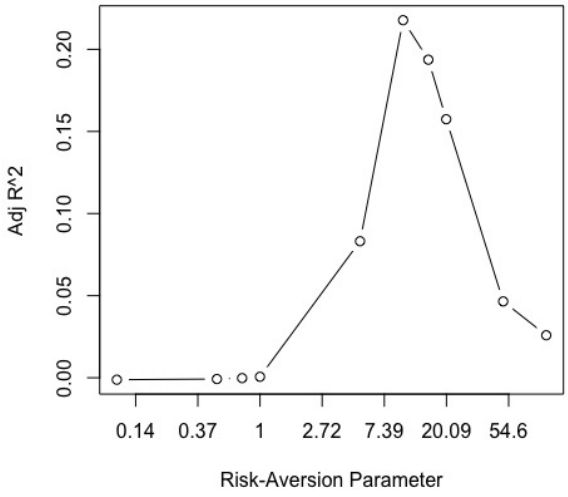


Figure 3.10: MVRT Simulations - Adj $R^2$

### *3.3.4 Market timing*

The true test of whether a forecasting model is valuable boils down to its applicability. In other words, can investors use the model to make money? This section illustrates how the multivariate RT performs when a simple trading strategy is implemented. I outline how the trading strategy was implemented and I present the results in the form of a cumulative returns plot as well as a time-series plot showing the profits generated by each trade for the strategy. This strategy is compared to the cumulative returns plot for a buy and hold strategy on the S&P 500.

The MVRT strategy has an initial investment of \$1. Each month, the MVRT gives the investor a signal to either buy (positive signal) or sell (negative signal) the S&P 500. If the signal is negative and the investor currently holds the asset, the asset is sold and shorted. Similarly, if the signal from the MVRT is positive and the investor is short, then the investor closes the current position and buys the asset. This exercise is repeated each time a new

signal is obtained (every month in this example). The MVRT occasionally outputs an error. If the signal is an error, then the signal on the following day will be used. In the interest of simplicity, trading costs are not considered. However, since the signals are only obtained once a month, there are a limited number of rebalances, which implies that there are also a limited number of trades. Hence, trading costs for this type of strategy would be negligible. The results can be seen in figures 3.11 and 3.12.

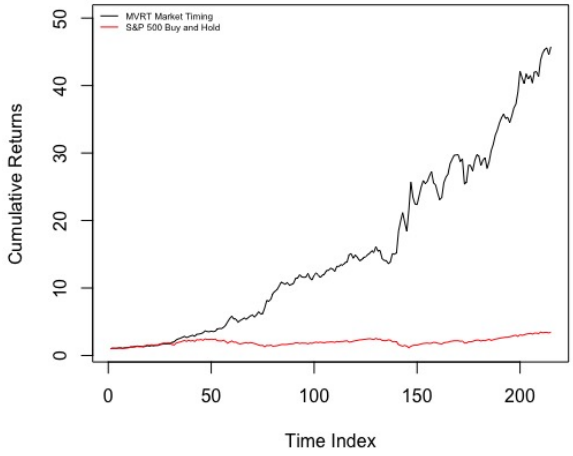


Figure 3.11: Cumulative Returns Plot

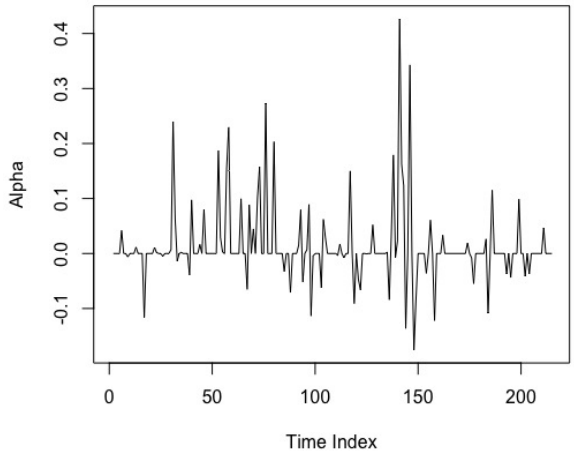


Figure 3.12: Profit and Loss Plot

Notice that, in figure 3.11, the cumulative returns from the MVRT (black line) outperform the S&P 500 buy and hold strategy (red line). This is accentuated by the fact that the cumulative returns consider compounding from reinvestment. A better depiction of the superiority of the MVRT can be seen in figure 3.12. Here we can see that, on average, the positive profits outnumber the negative profits. In fact, almost 57% of the trades are positive. Furthermore, the magnitude of the profits is substantially larger than that of the losses. The average profit is about 5% per trade compared to the average loss which is about 2.6% per trade.

### *3.4 Conclusion*

This paper aimed to improve the estimation of the natural probabilities derived from the Recovery Theorem (RT). Its major contribution is that it extends the RT by changing the univariate derivation of the contingent state price matrix to a multivariate one. By changing the derivation of the contingent state price matrix to a multivariate Markov chain, the inherent transition probabilities are more accurately defined. In the multivariate chain, I added the implied volatility, which results in significant improvements in the RT results. The out-of-sample forecast regression's adjusted $R^2$ increases from about 0.12 using Ross's specification to about 0.30 using the MVRT method. I show, using a simple numeric and intuitive example, that although the multivariate model performs better than the univariate model, it does much better whenever the changes in volatility are minimal. When changes in the underlying volatility occurs, it takes time for this new information to be fed into the model. As such, the multivariate model's performance does seem to suffer in instances when there are significant changes in volatility.

The Recovery Theorem was a giant leap forward in the estimation of financial market expectations. This paper improves on the original specification and will make it possible to use this methodology for other asset pricing endeavors. A number of extensions are possible. For example, since the multivariate RT extracts the market's true distribution of returns, we can extend this research to the question of hedging. A future research direction would be to explore whether firms change their hedging behavior in response to certain future expectations, where the expectations are derived from the RT's natural distribution (Fillebeen & Sanford (2016)).

The multivariate RT can also be used in portfolio construction applications. For instance, we could use the true distribution obtained from the multivariate RT as an actual returns distribution for a portfolio optimization problem. The portfolio weights can then be selected such that a measure that uses the distribution of returns (e.g. expected tail loss) is minimized

(see for example Sanford (2016a)). We may also want to use the exponential GARCH model (Bollerslev, 1986) to model the behavior of volatility. We can expect to obtain a better forecast if we incorporate a forward-looking volatility model rather than looking only at current implied volatility, as I do in this paper.

Finally, research should focus on whether the Recovery Theorem might apply in a setting where markets are incomplete. The RT assumes that the market is complete and, by extension, that it is possible to construct state prices. A natural question therefore arises: what assumptions would be necessary to apply the Recovery Theorem to an incomplete market? This would be a valuable extension to the current literature.

# Chapter 4

# MANAGING KNOWN UNKNOWNS: THE RESPONSE OF FIRM INVESTMENTS TO PURE UNCERTAINTY SHOCKS

## *4.1 Introduction*

This chapter answers the following question: what effect does uncertainty about the aggregate economy have on firm investment, holding news shocks constant? Traditionally, this question has been hard to answer because the effect of economic uncertainty on investment is generally intertwined with the effect of bad news more generally. Times of high economic uncertainty are typically also times of bad news (Bloom et al., 2007; Bloom, 2009; Baker et al., 2016). This chapter proposes a new methodology to untangle uncertainty and news shocks in stock return data. By using option prices to adjust abnormal returns for the time-varying risk premia, it is possible to estimate the impact of uncertainty shocks on firm investment while controlling for news shocks. We use option prices to estimate our news and uncertainty shocks in two different ways: parametric and nonparametric estimations.

The nonparametric methodology is based on ideas recently developed by Ross (2015). Ross' Recovery Theorem (RT) allows us to use state prices, which are portfolios of four options, to disentangle the three components of these prices (the pricing kernel, the risk-aversion parameter, and the natural probability distribution) nonparametrically. We then use the natural probability distribution as the expected return on the market. Assuming efficient markets, this measure of expected return includes all available information at the time investors are pricing the assets. In other words, we assume that the expected return obtained from the RT is the market's best estimate of what will happen in the future. We measure news by subtracting this expected return from the realized return. By doing so, we

capture the additional information added to the market between the time we estimated our expected return and the future time of interest. For example, if we are interested in finding out how much additional news was fed into the market between today and three months from now, we would subtract the realized return in three months from the expected return estimated today. This equation reflects the "surprise" news that investors did not perceive at time zero. We calculate the volatility of these news shocks to use as our measure of the uncertainty shocks. The parametric methodology, also based on option prices, follows Kadan & Manela (2017).

Using quarterly data on public firms from 1996 to 2015, we find that uncertainty shocks depress investment systematically, even after controlling for bad news. Lumpy investments reinforce the negative effect of uncertainty on investment, while better management systematically attenuates this negative effect.

The chapter is divided into five main sections. Section 4.2 presents the theory and derives the news and uncertainty shocks. Section 4.3 describes the data used in the analysis. Section 4.4 validates the derivation of our news and uncertainty shocks. Section 4.5 presents the results of panel regressions. Section 4.6 explores possible extensions and concludes.

## *4.2 Methodology*

### *4.2.1 Abnormal returns*

The rationale behind our measures of news and uncertainty shocks is similar to the logic of event studies (see, for example, Fama et al., 1969; Brown & Warner, 1985; Campbell et al., 1997; Kothari & Warner, 1997, 2007), which builds on the efficient market hypothesis. If stock prices reflect expectations of financial market participants on future fundamentals correctly, any changes in these expectations should induce changes in stock prices. These changes in stock prices are measured by stock returns $r_{t+1} = p_{t+1} - p_t$, where $p_t$ is the log of stock prices. However, this baseline logic needs to take into account that, even without any news,

we should expect stock prices to change predictably based on time and risk compensation, as captured by expected returns $E_\tau[r_{\tau+1}]$.

As a result, our actual object of interest are stock returns that are corrected to predictable returns, or abnormal returns, defined as:

$$r_{\tau+1} - E_\tau[r_{\tau+1}] \tag{4.1}$$

Abnormal returns capture information that surprised stock market participants, and are therefore the building blocks for our measures of news and uncertainty shocks. Positive abnormal returns signal better-than-expected news about future economic conditions, while negative abnormal returns signal worse-than-expected news.

To extract and separate news and uncertainty shocks, we use time aggregation. Abnormal returns are calculated weekly, while the investment data we are interested in is reported quarterly. We define the average abnormal return within a quarter as our measure of average news. The intuition behind this definition is the following: if positive surprises in one week are cancelled out by negative surprises in the next week, on average, we are not recording positive or negative news during the quarter. Formally, we define news shocks as:

$$\text{NEWS}_{t,t+1} = \frac{1}{N}\sum_{\tau=t}^{t+1}(r_{\tau+1} - E_\tau[r_{\tau+1}]) \tag{4.2}$$

In contrast, the dispersion of surprises defines uncertainty shocks within the quarter. To go back to the example presented above, where the positive surprises in one week are cancelled out by negative surprises in the following week, we would measure an uncertainty shock in that quarter. Formally, we define these shocks as:

$$\text{UNC}_{t,t+1} = SD_{t,t+1}[r_{\tau+1} - E_\tau[r_{\tau+1}]] \tag{4.3}$$

### *4.2.2 Normal returns*

Our methodology starts from the definition of expectational surprises based on abnormal returns. While realized returns are easy to measure, it is the construction of expected returns $E_t[r_{t+1}]$ that is the most important challenge in calculating abnormal returns. This is especially true in the presence of time variation in risk premia, which has been widely documented to be an important aspect of the data (Campbell & Cochrane, 1999; Cochrane, 2008).

One particularly attractive source of information for expected returns are option prices on stock returns, in our case options on the S&P 500 index. Information on the strike levels of option portfolios, combined with option prices, allows one to potentially recover data in state probabilities that can be used to construct expected returns on the underlying asset, in our case $E_t[r_{t+1}]$(Breeden & Litzenberger, 1978). Formally, let $f_{i,j}$ denote the state probabilities of moving from state $i$ to state $j$, such as a transition from an S&P 500 index value of $1,000$ to a value of $1,200$. The Breeden & Litzenberger (1978) methodology enables us to measure state prices $p_{i,j}$ that pay \$1 in state $j$ given that today's state is $i$, using a butterfly spread portfolio. State prices $p_{i,j}$ are not enough to calculate returns $E_t[r_{t+1}]$ since they combine the effects of differences in state probabilities with the effects of risk adjustment. To extract state probabilities $f_{i,j}$ from these state prices, we must use some functional form assumptions on the stochastic discount factor (SDF) to separate state probabilities from risk adjustment. We use two such methods here.

*Non-parametric recovery*

The first approach assumes time separability of risk preferences of a representative investor, and closely follows Ross (2015). Let $p_{i,j}$ denote state prices of future state $j$ to current state $i$.

$$p_{i,j} = \delta \frac{U'_j}{U'_i} f_{i,j} \tag{4.4}$$

Define $z_i = \frac{1}{U_i'}$ and impose $\sum_i f_{i,j} = 1$ as $f_{i,j}$ are probabilities. Then, we can rewrite this in matrix notation as:

$$Pz = \delta z \tag{4.5}$$

where $z = (\frac{1}{U_1'}, \frac{1}{U_2'}, ..., \frac{1}{U_S'})$. The recovery of state probabilities reduces to an eigenvector problem. For the matrix of state prices, $P$ is a sufficient statistic for the valuation effects of discounting, incorporating both the time discounting $\delta$ as well as the risk adjustment $\frac{U_j'}{U_i'}$ under separable preferences. Once the state price matrix is measured, and under the additional assumptions of time homogeneity and time separable preferences, recovering marginal utilities across states reduces to the stated eigenvector problem. After calculating the eigenvectors, we obtain the state probabilities from state prices through:

$$f_{i,j} = \frac{1}{\delta}\frac{z_j}{z_i}p_{i,j} \tag{4.6}$$

*Parametric recovery*

A number of researchers, such as Borovička et al. (2016), have pointed out that time homogeneity and separable preferences are very strong assumptions and that many popular models of stochastic discount factors or output-generating stochastic processes might violate these assumptions. A case of particular interest outside of the assumption of Ross (2015) are Epstein-Zin preferences.

We therefore follow Kadan & Manela (2017) to model the SDF of a representative agent with Epstein-Zin preferences and recover state probabilities using this specification as a robustness check.

$$p_{i,j} = \delta^{\frac{1-\gamma}{1-\rho}} exp(\frac{\rho-\gamma}{1-\rho}r_j)E[exp(-\rho(\frac{1-\gamma}{1-\rho})g_j)]f_{i,j} \tag{4.7}$$

with $\gamma$ as the intertemporal elasticity of substation, $\rho$ the coefficient of relative risk aversion, $\delta$ the one period discount factor, $r_j$ the return to wealth and $g_j$ the growth rate of consumption.

Under additional assumptions that (i) the return to wealth is proportional to stock returns, (ii) consumption growth is separable in fundamental and expectation-related components, (iii) expected consumption growth rates are independent, and (iv) return of wealth and consumption growth rate are proportional to each other, they show that state probabilities can be recovered by:

$$f_{i,j} = \frac{exp(\gamma r_j)p_{i,j}}{\sum_k exp(\gamma r_k)p_{i,k}} \tag{4.8}$$

In other words, under Epstein-Zin preferences, with the simplifying assumptions (i)-(iv), state probabilities can be recovered from state prices by applying an exponential tilt, depending on the parameter $\gamma$.

### *4.2.3 Investment regression specifications*

We build on the empirical literature on the determinants of firm-level investments, especially Eberly et al. (2012) for our empirical model. Specifically, our regression specification is of the form:

$$\frac{I}{K}_{i,t+1} = \beta_0 + \beta_1 \frac{I}{K}_{i,t} + \beta_2 lnQ_{i,t} + \beta_3 ln(CFA)_{i,t} + D_i + \beta_N NEWS_{t,t+1} + \beta_U UNC_{t,t+1} + \epsilon_{i,t} \tag{4.9}$$

where $\frac{I}{K}_{i,t}$ is the investment rate as a percentage of capital for firm $i$ at time $t$ and is the dependent variable of interest. Our main independent variables of interest are measure of news and uncertainty shocks, $NEWS_{t,t+1}$ and $UNC_{t,t+1}$. We measure these shocks in the quarter up to the reporting time of the dependent variable to approximate the contemporaneous impact of news and uncertainty shocks on investment.

We include a number of different control variables to account for important determinants that past research has focused on. $lnQ_{i,t}$ measures the log of Tobin's $Q$ to account for standard Neoclassical investment theory. $ln(CFA)_{i,t}$ is a measure of cash flows as a percentage of current assets. $D_i$ are firm fixed effects. We also estimate specifications with

lagged investment, as captured by $(I/K)_{i,t}$, to account for accelerator effects documented by Eberly et al. (2012). Since specifications with lagged dependent variables also feature firm fixed effects, we estimate Arellano/Bond-type (1991) dynamic panel models to identify these models correctly.

It is important to note that the firm-level data consists of large, public firms. Such firms have been shown to exhibit much smoother investment patterns than smaller firms whose investments tend to be lumpier (Cooper & Haltiwanger, 2006). Since the degree of lumpiness in firm investment patterns varies a lot, we address the lumpy nature of investment explicitly in extensions below.

## *4.3 Data*

We obtained all data from the Wharton Research Data Services (WRDS) database (see subsections below for additional details). The option and stock data generally originate from the OptionMetrics and CRSP datasets respectively, while firm-level data generally come from the Compustat dataset.

### *4.3.1 Option price data*

As mentioned in section 4.2.2, we use daily option prices to estimate expected returns on S&P 500 prices. This chapter covers the time period from January 1996 to July 2015, the entire timeframe included in the OptionMetrics database at the time of data collection. We must make adjustments to the OptionMetrics data for them to be usable in the Black-Scholes equation: (i) we adjust OptionMetrics prices to be on a per-stock basis instead of lots of $1,000$ securities, (ii) time-to-maturity is converted from a date to a fraction of years to expiration, and (iii) option price is replaced with the midpoint of the bid-ask spread. This is consistent with Figlewski (2008), who argues that bid and ask prices are continuously quoted for almost all strikes regardless of whether a trade takes place. The alternative,

transaction prices, occurs irregularly (Figlewski, 2008) and would make it more difficult to extract a proper implied volatility curve. We compare our estimated implied volatilities to those provided by OptionMetrics. Since the difference between the two is negligible, we use the more complete set of estimates instead of the OptionMetrics data.

### *4.3.2 Firm-level data*

Firm-level data was constructed in accordance with Eberly et al. (2012) and Gulen & Ion (2015). The time period used for the firm-level data is the same as above: January 1996 to July 2015 or the maximum available data, whichever is largest. The investment variable is defined as expenditures on property, plant, and equipment (data item 30 in the CRSP/Compustat merged database). The cash flow variable is defined as income before extraordinary items (data item 123) plus depreciation and amortization (data item 125) plus minor adjustments.[1] Capital stock is calculated using the Salinger & Summers (1983) method, whereby the initial value of the capital stock is equal to the book value of capital and the subsequent values are calculated as:

$$K_{i,t} = (K_{i,t-1}\frac{P_{K,t}}{P_{K,t-1}} + I_{i,t})(1 - \delta_j) \tag{4.10}$$

where $K_{i,t-1}$ is the previous period's capital stock, $\frac{P_{K,t}}{P_{K,t-1}}$ is the ratio of the current price of capital to the previous period's price of capital, $I_{i,t}$ is the expenditures on property, plant, and equipment, and $\delta_j$ is the industry level's implied depreciation rate (indexed by $j$) .[2] Tobin's $Q$ is calculated as the previous period's market value of equity plus the previous

[1]Minor adjustments are calculated as: extraordinary items and discontinued operations (data item 124) plus deferred taxes (data item 126) plus equity in net loss (data item 106) plus the sale of property, plant, and equipment and sale of investments (data item 213) plus funds from operations (data item 217) plus accounts receivables (data item 302) plus inventory change (data item 303) plus accounts payable and accrued liabilities (data item 304) plus income taxes accrued (data item 305) plus assets and liabilities (data item 308). All of the above should be equal to the operating activities (data item 308).

[2]The implied depreciation rate is calculated as half of the industry's useful life of capital goods.

period's debt minus the previous period's inventories, the total of which is divided by the current capital stock.

In addition to estimating the effect of news and uncertainty shocks on investment, we also examine the impact of lumpy investments and management practices. Gulen & Ion (2015) use a function of rent expense, depreciation expense, and sale of PPE as a proxy for lumpy investments. Higher values of the lumpy investment index are associated with higher levels of sunk costs, which is associated with higher levels of investment irreversibility. This methodology for estimating lumpy investment is based on the industrial organization literature (see Kessides, 1990; Farinas & Ruano, 2005).

Finally, we obtained management practices data by three-digit SIC sector from the World Management Survey. We multiply management score by the sector-based revenue variable to create a time-varying measure of the efficacy of the management practices for individual firms.

#### *4.3.3 Other data*

To estimate realized return and the risk-neutral density of option prices, we use S&P 500[3] prices from the CRSP database. The risk-free rate is obtained from the Treasury Bill rate, which can be found in the Fama & French factors dataset also available on WRDS.

### *4.4 Validation*

In this section, we report validation exercises that showcase the advantages of our empirical approach. First, we test whether our measures of news and uncertainty shocks Granger-cause investments or vice versa. Second, we analyze how important our definition of abnormal returns is in the construction of news and uncertainty shocks by constructing measures that assume zero expected returns.

[3]SECID 108105

### *4.4.1 Granger causality*

We begin with an analysis of the predictive value of our measures of news and uncertainty shocks. As discussed in section 4.2.1, our measures of news and uncertainty are based on surprise movements in stock prices. If the variation in returns identifies surprises to market participants' information sets correctly, two implications follow.

First, because these movements are surprises, they should be unforecastable using other lagged variables, such as investments. In other words, investments should not Granger-cause our measures of news and uncertainty shocks. Second, because the surprises (and therefore our measures of news and uncertainty shocks) capture movements in expectations of future fundamentals, our shocks should have predictive value for investment. In other words, our news and uncertainty shocks should Granger-cause investments.

Tables 4.1 through 4.4 summarize the results of Granger causality tests for our measures of news and uncertainty shocks, using Ross's (2015) non-parametric recovery method to calculate normal returns.[4] The tests show that we cannot reject the hypothesis that news and uncertainty shocks Granger-cause investments. On the flipside, lagged investment has no predictive value for our measures of news and uncertainty shocks, so the hypothesis that investments Granger-cause news or uncertainty shocks is rejected.

| L = lags | F-Stat | P-Value |
|---|---|---|
| 1 | 577.7 | $< 2.2e-16^{***}$ |
| 2 | 229.59 | $< 2.2e-16^{***}$ |
| 3 | 123.23 | $< 2.2e-16^{***}$ |
| 5 | 51.632 | $< 2.2e-16^{***}$ |
| 10 | 13.338 | $< 2.2e-16^{***}$ |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.1: News shocks do not Granger-cause I/K

| L = lags | F-Stat | P-Value |
|---|---|---|
| 1 | 0.0883 | 0.7663 |
| 2 | 0.0467 | 0.9544 |
| 3 | 0.1188 | 0.9491 |
| 5 | 0.0924 | 0.9934 |
| 10 | 0.122 | 0.9996 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.2: I/K does not Granger-cause news shocks

[4]We also conducted Granger causality tests for our parametric recovery approach to calculate normal returns, with similar results. Results available from the authors.

| L = lags | F-Stat | P-Value |
|---|---|---|
| 1 | 1802.1 | $< 2.2e-16^{***}$ |
| 2 | 719.57 | $< 2.2e-16^{***}$ |
| 3 | 386.94 | $< 2.2e-16^{***}$ |
| 5 | 163.12 | $< 2.2e-16^{***}$ |
| 10 | 41.455 | $< 2.2e-16^{***}$ |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.3: Unc. shocks do not Granger-cause I/K

| L = lags | F-Stat | P-Value |
|---|---|---|
| 1 | 0.7648 | 0.3818 |
| 2 | 0.6962 | 0.4985 |
| 3 | 0.465 | 0.7067 |
| 5 | 0.6269 | 0.6793 |
| 10 | 0.5629 | 0.8454 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.4: I/K does not Granger-cause unc. shocks

### *4.4.2 Recovery-based shocks vs. "realized volatility"*

While the statistical tests of Granger causality are consistent with the view that our shock measures really represent changes in the information set of financial market participants, another natural question is whether we need to use abnormal returns to measure these expectation changes. In other words, how different would the results be if one were willing to assume that the predictable part of realized returns is zero (i.e. $E_\tau[r_{\tau+1}] = 0$)? This assumption, which has been used in related work (Berger et al., 2017), would imply that shock prices follow a random walk. While a random walk assumption is a good approximation of stock returns in the very short run, longer-run stock returns are much more predictable (Cochrane, 2008). Whether normal returns are needed to adjust realized stock returns is therefore an empirical question, depending on the frequency of the data under consideration.

To address this question, we construct test measures of news and uncertainty shocks as specified in equations 4.2 and 4.3, but under the condition that $E_\tau[r_{\tau+1}] = 0$. Table 4.5 compares summary statistics of these test measures with our baseline measures of news and uncertainty shocks. Since these test measures only use realized return data, their summary statistics are displayed in the "return" section. The first two columns present summary statistics for the parametric recovery. The third and fourth columns present summary statistics for the nonparametric recovery. Finally, the fifth and sixth columns present the summary

statistics for the shocks obtained using actual recovery and volatility. The return news shock is almost equal to zero. This indicates that the return is almost equal to the expectation and, based on this measure of expectation, no additional news was incorporated into the market. The average news shock from the parametric recovery is quite a bit larger than averages from the other methods. This, along with a relatively low standard deviation, indicates that the news shocks from the parametric recovery are tightly distributed and mostly positive. As for the uncertainty shocks, the nonparametric recovery has the highest average shock with the largest standard deviation by a significant margin. Clearly, the Recovery Theorem seems to capture a more significant amount of variation in the risk premium.

We use these test measures of news and uncertainty shocks in investment regressions to compare our results with our main results later.

| | Parametric Recovery | | Nonparametric | | Return | |
|---|---|---|---|---|---|---|
| | News Shock | Unc. Shock | News Shock | Unc. Shock | News Shock | Unc. Shock |
| Mean | 0.03192 | 0.06675 | 0.01054 | 0.14096 | 0.00621 | 0.03250 |
| Std Dev. | 0.06956 | 0.03464 | 0.18097 | 0.22762 | 0.02684 | 0.01882 |

Table 4.5: Summary Statistics

As documented in tables 4.6 and 4.7, the regression coefficients on news and uncertainty shock measures that are only based on realized returns have either the wrong sign (see table 4.6) or they are unstable (see table 4.7).

We are concerned that the significantly negative impact of average returns on investment may be reflecting the fact that a large part of the realized average return is in fact the risk premium. This interpretation would be entirely consistent with the predictions of economic theory regarding how the risk premium influences investment activity.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | 0.06912*** | 0.07504*** | 0.15276*** | 0.11999*** | 0.17189*** |
| | (0.00033) | (0.00098) | (0.00071) | (0.00133) | (0.00075) |
| $(I/K)_{t-1}$ | 0.31625*** | | | | 0.17857*** |
| | (0.00279) | | | | (0.00264) |
| $Ln(Q_t)$ | | 0.02999*** | | 0.02129*** | 0.01601*** |
| | | (0.0008174) | | (0.00085) | (0.00039) |
| $Ln(CashFlow/K)_t$ | | | 0.02125*** | 0.01753*** | 0.05178*** |
| | | | (0.00036) | (0.00035) | (0.00022) |
| S&P 500 Return | −0.23958*** | −0.30935*** | −0.20479*** | −0.24520*** | −0.12653*** |
| | (0.00783) | (0.00857) | (0.00859) | (0.00848) | (0.00676) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.6: Return Shock

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | 0.06569*** | 0.06403*** | 0.14277*** | 0.11095*** | 0.16636*** |
| | (0.00047) | (0.00107) | (0.00081) | (0.00084) | (0.00083) |
| $(I/K)_{t-1}$ | 0.31812*** | | | | 0.17380*** |
| | (0.00284) | | | | (0.00268) |
| $Ln(Q_t)$ | | 0.02931*** | | 0.02053*** | 0.01600*** |
| | | (0.00081) | | (0.00084) | (0.00039) |
| $Ln(CashFlow/K)_t$ | | | 0.02150*** | 0.01797*** | 0.05217*** |
| | | | (0.00036) | (0.00035) | (0.00022) |
| Std Dev S&P 500 Return | 0.04246*** | 0.30386*** | 0.28147** | 0.28498*** | 0.17798*** |
| | (0.01137) | (0.01204) | (0.01216) | (0.01199) | (0.00974) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.7: Standard deviation of Return Shock

Table 4.8 shows results when the news and uncertainty shocks measured using returns

are both included in the regression. Again, we are concerned about the effect of news shocks on investments. Intuitively, we should expect a positive effect. If the news are better than expected, managers should want to increase investments. Instead, we observe a negative relationship between news shocks measured by returns and investment. Furthermore, the uncertainty shock (the volatility of the news shock) is unstable, as it was in tables 4.6 and 4.8. This is problematic because we would expect these uncertainty shocks to have a negative effect on investment decisions. Intuitively, increased uncertainty dampens the ability of managers to "predict" the future and encourages them to wait to make investment decisions. In other words, the patterns that we observe in tables 4.6 and 4.8 should be consistent when both shocks are in the model. The relationship between news shocks and investments should be positive and the relationship between uncertainty shocks and investment should be negative. However, it is not the case when we use returns to estimate our shock measures. In the next section, we test whether our parametric and nonparametric shock estimation methods produce coefficients that are more in line with expectations.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | 0.07027*** | 0.06869*** | 0.14542*** | 0.11336*** | 0.16658*** |
| | (0.00049) | (0.00107) | (0.00083) | (0.00139) | (0.00083) |
| $(I/K)_{t-1}$ | 0.31741*** | | | | 0.17184*** |
| | (0.00282) | | | | (0.00267) |
| $Ln(Q_t)$ | | 0.02989*** | | 0.02117*** | 0.01675*** |
| | | (0.00082) | | (0.00084) | (0.00039) |
| $Ln(CashFlow/K)_t$ | | | 0.02126*** | 0.01756*** | 0.05170*** |
| | | | (0.00036) | (0.00035) | (0.00022) |
| S&P 500 Return | −0.24527*** | −0.26519*** | −0.15462*** | −0.19841*** | −0.10534*** |
| | (0.00804) | (0.00912) | (0.00911) | (0.00898) | (0.00690) |
| Std Dev S&P 500 Return | −0.03712** | 0.18819*** | 0.21428*** | 0.19887*** | 0.14623* |
| | (0.01158) | (0.01281) | (0.01293) | (0.01271) | (0.00993) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.8: Both Shocks

## *4.5 Main results*

This section documents our main results, covering first the impact of news shocks on investment, then the impact of uncertainty shocks, and finally the simultaneous impact of news and uncertainty shocks. Throughout this section, we present results from the non-parametric (RT) and parametric recovery methods in turn.

### *4.5.1 News shocks*

If we measured news shocks (and therefore changes in average expectations of future economic conditions) accurately, we would expect firms to respond positively to positive news about the economy.

This is exactly what we find in table 4.9. The estimated coefficient of news shocks is always positive and relatively stable, as expected. The coefficient is similar across models 1 through 4, but not in model 5. The most likely explanation for this difference is that

Tobin's $Q$ and cash flows capture some of the variations in the expected news about future profitability. Hence, it would seem likely that, when these coefficients increase, the news coefficient would decrease, all else equal.

All of the results are highly statistically significant. The results are similar when including alternative news shock measures that use parametric assumptions on the stochastic discount factor (SDF) to construct abnormal returns (see table 4.10).

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | 0.06300*** | 0.07342*** | 0.14646*** | 0.11915*** | 0.17367*** |
| | (0.00036) | (0.00108) | (0.00081) | (0.00145) | (0.00083) |
| $(I/K)_{t-1}$ | 0.30702*** | | | | 0.17569*** |
| | (0.00314) | | | | (0.00296) |
| $Ln(Q_t)$ | | 0.02672*** | | 0.01828*** | 0.01035*** |
| | | (0.00092) | | (0.00094) | (0.00041) |
| $Ln(CashFlow/K)_t$ | | | 0.02100*** | 0.01802*** | 0.05152*** |
| | | | (0.00041) | (0.00040) | (0.00025) |
| $News_t$ | 0.08382*** | 0.08414*** | 0.07377*** | 0.07066*** | 0.03902*** |
| | (0.00166) | (0.00203) | (0.02371) | (0.00206) | (0.00142) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.9: Nonparametric News Shock

Much like the results for the nonparametric recovery presented in table 4.9, results in table 4.10 are highly statistically significant and positive, as we would expect. The major difference between these results and the nonparametric results lies in the size of the coefficient. The coefficients here are quite a bit smaller than those in table 4.9. Smaller coefficients, however, are to be expected given that, in table 4.5, we showed that the parametric news shocks were, on average, quite a bit larger than the nonparametric news shocks. Finally, the fact that our measure of news shocks produces a positive coefficient in all specifications confirms that we

were successful in controlling for the effects of time-varying risk premia.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | 0.066880*** | 0.07221*** | 0.15018*** | 0.11876*** | 0.17197*** |
| | (0.00035) | (0.00099) | (0.00072) | (0.00134) | (0.00076) |
| $(I/K)_{t-1}$ | 0.31961*** | | | | 0.18487*** |
| | (0.00282) | | | | (0.00264) |
| $Ln(Q_t)$ | | 0.02923*** | | 0.02050*** | 0.01493*** |
| | | (0.00081) | | (0.00084) | (0.00039) |
| $Ln(CashFlow/K)_t$ | | | 0.02157*** | 0.01807*** | 0.05242*** |
| | | | (0.00036) | (0.00035) | (0.00022) |
| $News_t$ | 0.0190 | 0.05871*** | 0.06107*** | 0.05860*** | 0.02952*** |
| | (0.00321) | (0.00364) | (0.00378) | (0.00368) | (0.00272) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.10: Parametric News Shock

### *4.5.2 Uncertainty shocks*

The baseline results for uncertainty shocks are documented in tables 4.11 (nonparametric method) and 4.12 (parametric method).

As expected, uncertainty shock coefficients in table 4.11 are negative and highly statistically significant.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | $0.07695^{***}$ | $0.08390^{***}$ | $0.15408^{***}$ | $0.12646^{***}$ | $0.17668^{***}$ |
| | (0.00038) | (0.00113) | (0.00081) | (0.00147) | (0.00082) |
| $(I/K)_{t-1}$ | $0.30835^{***}$ | | | | $0.17950^{***}$ |
| | (0.00308) | | | | (0.00295) |
| $Ln(Q_t)$ | | $0.02640^{***}$ | | $0.01828^{***}$ | $0.01015^{***}$ |
| | | (0.00093) | | (0.00095) | (0.00041) |
| $Ln(CashFlow/K)_t$ | | | $0.02055^{***}$ | $0.01758^{***}$ | $0.05051^{***}$ |
| | | | (0.00041) | (0.00040) | (0.00041) |
| $Uncertainty_t$ | $-0.11576^{***}$ | $-0.07822^{***}$ | $-0.06470^{***}$ | $-0.06224^{***}$ | $-0.04163^{***}$ |
| | (0.00137) | (0.00165) | (0.00162) | (0.00161) | (0.00122) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.11: Nonparametric Uncertainty Shock

Table 4.12 shows results using the parametric uncertainty shocks. As was the case with the nonparametric uncertainty shocks, the coefficients are negative and, in general, highly statistically significant. Please note that the coefficients in table 4.11 and table 4.12 are not that different. Given that the summary statistics showed that the uncertainty shocks has a larger mean and standard deviation for the nonparametric estimation, one might have expected the size of the coefficients to be different. Instead, they are quite similar. In addition, the coefficients for the control variables are very similar. It is unclear, at this time, why these coefficients are so similar. It would be interesting to investigate the distributions for these shocks further rather than examining only the first two moments.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | $0.07413^{***}$ | $0.07931^{***}$ | $0.15338^{***}$ | $0.12305^{***}$ | $0.017269^{***}$ |
| | (0.00049) | (0.00104) | (0.00080) | (0.00136) | (0.00079) |
| $(I/K)_{t-1}$ | $0.32211^{***}$ | | | | $0.18274^{***}$ |
| | (0.00281) | | | | (0.002639) |
| $Ln(Q_t)$ | | $0.02955^{***}$ | | $0.02066^{***}$ | $0.01486^{***}$ |
| | | (0.00082) | | (0.00084) | (0.00039) |
| $Ln(CashFlow/K)_t$ | | | $0.02155^{***}$ | $0.01796^{***}$ | $0.05241^{***}$ |
| | | | (0.00036) | (0.00035) | (0.00022) |
| $Uncertainty_t$ | $-0.10951^{***}$ | $-0.08208^{***}$ | $-0.017799^{**}$ | $-0.04281^{***}$ | 0.00626 |
| | (0.00554) | (0.00606) | (0.00625) | (0.00602) | (0.00472) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.12: Parametric Uncertainty Shock

The next section investigates the effect of news and uncertainty shocks together on investments. Intuitively, we should expect that the results will not be systematically different when both shocks are included in the model. Namely, we would expect that the effect of the news shock on investment be positive and that the effect of the uncertainty shock on investment be negative.

### *4.5.3 Combination of uncertainty and news shocks*

Our baseline results regarding the impact of uncertainty shocks on investment while controlling for news shocks are documented in tables 4.13 (nonparametric method) and 4.14 (parametric).

Comparing the results from table 4.13 to those from tables 4.9 and 4.11, where only one shock was included in the model at a time, we notice that the coefficients and their level of significance do not change much when both shocks are present. The coefficients in table 4.9 ranged from about 0.08 to 0.04. In table 4.13, the coefficients range from about 0.088 to

0.018. The coefficients for the uncertainty shocks in table 4.11 ranged from about $-0.12$ to $-0.042$ while those from table 4.13 range from about $-0.11$ to $-0.033$. All control variables in models one through five are very similar. Hence, the results seem to be quite stable and consistent regardless of the specification.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | $0.07634^{***}$ | $0.08064^{***}$ | $0.15035^{***}$ | $0.12317^{***}$ | $0.17531^{***}$ |
| | (0.00040) | (0.00114) | (0.00082) | (0.00148) | (0.00083) |
| $(I/K)_{t-1}$ | $0.30786^{***}$ | | | | $0.17780^{***}$ |
| | (0.00308) | | | | (0.00295) |
| $Ln(Q_t)$ | | $0.02613^{***}$ | | $0.01811^{***}$ | $0.01031^{***}$ |
| | | (0.00093) | | (0.00095) | (0.00041) |
| $Ln(CashFlow/K)_t$ | | | $0.02038^{***}$ | $0.01744^{***}$ | $0.05050^{***}$ |
| | | | (0.00041) | (0.00040) | (0.00025) |
| $News_t$ | $0.08850^{***}$ | $0.04965^{***}$ | $0.04707^{***}$ | $0.04491^{***}$ | $0.01763^{***}$ |
| | (0.00197) | (0.00235) | (0.00244) | (0.00239) | (0.00171) |
| $Uncertainty_t$ | $-0.11151^{***}$ | $-0.05776^{***}$ | $-0.04541^{***}$ | $-0.04386^{***}$ | $-0.03310^{***}$ |
| | (0.00167) | (0.00191) | (0.00191) | (0.00187) | (0.00148) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 4.13: Nonparametric Both Shocks

Most of the observations about the nonparametric results in table 4.13 hold true for the parametric results as well. Coefficient size and significance in this specification are similar to the one-type-of-shock-at-a-time results in tables 4.10 and 4.12. The main difference between parametric and nonparametric results is that some of the coefficients are not as statistically significant as those using the nonparametric shocks. For example, in model 1, the coefficient of the news shock is not significant, just like before.

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 |
|---|---|---|---|---|---|
| Constant | 0.07427*** | 0.07673*** | 0.15033*** | 0.12043*** | 0.17135*** |
| | (0.00051) | (0.00106) | (0.00082) | (0.00138) | (0.00080) |
| $(I/K)_{t-1}$ | 0.32189*** | | | | 0.18468*** |
| | (0.00282) | | | | (0.00265) |
| $Ln(Q_t)$ | | 0.02943*** | | 0.02052*** | 0.01491*** |
| | | (0.00081) | | (0.00084) | (0.00039) |
| $Ln(CashFlow/K)_t$ | | | 0.02157*** | 0.01800*** | 0.05246*** |
| | | | (0.00036) | (0.00040) | (0.00022) |
| $News_t$ | −0.00305 | 0.05359*** | 0.06090*** | 0.05649*** | 0.03018*** |
| | (0.00321) | (0.00365) | (0.00380) | (0.00368) | (0.00273) |
| $Uncertainty_t$ | −0.10994*** | −0.06835*** | −0.00230 | −0.02826*** | −0.0108* |
| | (0.00555) | (0.00607) | (0.00625) | (0.00600) | (0.00474) |
| Num. obs. | 142322 | 142322 | 142322 | 142322 | 142322 |

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, ˙$p < 0.1$

Table 4.14: Parametric Both Shocks

## *4.6 Conclusion*

In this chapter, we test the effect of uncertainty shocks on firm-level investments. We answer the question: what effect does uncertainty about the aggregate economy have on investment, holding news shocks constant? We propose a nonparametric estimation method (using the methodology proposed by Ross (2015)) and a parametric estimation method of news and uncertainty shocks (using the methodology proposed by Kadan & Manela (2017)). Using quarterly data on public firms from 1996 to 2015, we find that uncertainty shocks depress investment systematically, even after controlling for bad news.

# Chapter 5

# CONCLUSION

The first two chapters in this dissertation were written as companion papers. The first chapter, which derives the complete risk-neutral density, illustrates the importance of documenting our interpolation methodology in empirical applications. When we consider that research is considered to be a step forward when we are able to increase the $R^2$ by 3-4% in an out-of sample forecast, we should also consider the black boxes that are routinely used blindly in finance. The second chapter is an extension of the RT. The paper argues that, when deriving the natural probability distribution, we must also consider other possible state variables. By adding a proxy for the expected uncertainty of the market as a state variable, we can almost double the out-of-sample $R^2$ of the model (from about 15% to about 30%). The third and final paper illustrates one of the many ways in which we can use the results from Ross (2015). In this paper, we used the RT to obtain news and uncertainty shocks (the difference between the realized return and the expected return obtained from the RT), which then explained variations in certain firm-level variables. More specifically, we can estimate the impact of uncertainty shocks on firm investment while controlling for news shocks. Using quarterly data on public firms from 1996 to 2015, we find that uncertainty shocks systematically depress investment, even after controlling for bad news. Additionally, the chapter shows that lumpy investments reinforce the negative effect of uncertainty on investment, while better management attenuates this negative effect systematically.

## *5.1 Limitations*

I made efforts to discuss the limitations of my work in the relevant sections above. That being said, certain limitations are worth mentioning again. The most important issue in the first and second chapters is probably generalizability. More specifically, chapter one depends on the fact that there is enough liquid data to be able to use the spline and functional methods. If options are not heavily traded (meaning that there is little to no volume on any given day), the interpolation methodology presented in chapter one will not produce reliable results. If the density estimated from chapter one is unreliable, it follows that the results obtained in chapter two, meaning in the RT, will also be unreliable. Hence, it is important to know one's data before using the methods presented in this dissertation.

The most important theoretical limitation of the RT is its reliance on two strong assumptions: time homogeneity and time-invariant kernel. I do not believe that we have the mathematical tools necessary to solve the RT without the time-invariant kernel currently. However, it may be possible to remove the time homogeneity assumption. As mentioned earlier in this dissertation, time homogeneity assumes that the contingent state prices are the same regardless of which time-step we are trying to estimate. In other words, we are assuming that the market characteristics are the same whether we are comparing options that expire in 3 months to options that expire in 6 months, or options that expire in 27 months to options that expire in 30 months. Clearly, the market information about options that expire in 27 months is different from the information for options that expire in 3 months. Hence, the time homogeneity assumption needs to be problematized to create a model that is more realistic. A recent working paper attempts to address the concern of time homogeneity (Jensen et al., 2017). It will be interesting to see where the model goes without this assumption.

## *5.2 Broader Implications and Future Research*

Why do we care about the work presented in this dissertation? Will the world be better off because of this work? For one, the pursuit of a model that provides a reliable estimate of expected returns on stock market assets is valuable practically because it allows us to make money. It is true that this "broader implication" may not seem like a noble one, but ultimately the purpose of investments is to increase our wealth.

Beyond this practical implication, my dissertation advances the scholarly discussion surrounding asset pricing: I provide a new methodology to extract a full risk-neutral density, extend the RT so that it incorporates information about additional state variables in the derivation of the natural probability distribution, and provide a new nonparametric measure of uncertainty shocks which allows us to explain the impact of uncertainty shocks on firm investment while controlling for news shocks.

Where do we go from here? It may be necessary to remove some of the assumptions in the RT. For example, what happens to the model (and the results more generally) if we remove some of its unrealistic assumptions (Borovička et al., 2016)? Researchers have started to examine this question (Jensen et al., 2017), but more work is certainly needed. The multivariate RT proposed in chapter two can also be used in a multitude of applications. Since the model provides us with a forward-looking estimate of market expectations, any application that uses market-based expectations could benefit from the results of the multivariate RT. For example, portfolio theory often uses historical returns as the basis for asset selections. It may be worthwhile to solve portfolio questions using a forward-looking measure like the one proposed by the multivariate RT. The model could also be used beyond the stock market, for instance for the commodities, currencies, or bond markets.

As far as estimating risk-neutral densities more accurately, it may be worthwhile to turn to machine learning to determine if algorithms would produce more accurate representations of the risk-neutral density. As an input in many empirical models, we want the risk-neutral

density to be as accurate as possible. With more powerful computers, we may be able to borrow techniques from computer scientists to obtain a more accurate picture of the information contained in options.

# BIBLIOGRAPHY

Aït-Sahalia, Y., & Lo, A. W. (1998). Nonparametric estimation of state-price densities implicit in financial asset prices. *The Journal of Finance*, *53*(2), 499–547.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The Review of Economic Studies*, *58*(2), 277–297.

Bahra, B. (1997). Implied risk-neutral probability density functions from option prices: theory and application. *Working Paper*.

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636.

Benaim, S., & Friz, P. (2009). Regular variation and smile asymptotics. *Mathematical Finance*, *19*(1), 1–12.

Berger, D., Dew-Becker, I., & Giglio, S. (2017). Uncertainty shocks as second-moment news shocks. Tech. rep., National Bureau of Economic Research.

Birru, J., & Figlewski, S. (2012). Anatomy of a meltdown: The risk neutral density for the S&P 500 in the fall of 2008. *Journal of Financial Markets*, *15*(2), 151–180.

BIS (2012). BIS quarterly review, June 2012. *Bank for International Settlements*.

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

Bliss, R. R., & Panigirtzoglou, N. (2004). Option-implied risk aversion estimates. *The Journal of Finance*, *59*(1), 407–446.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, *77*(3), 623–685.

Bloom, N., Bond, S., & Van Reenen, J. (2007). Uncertainty and investment dynamics. *The Review of Economic Studies*, *74*(2), 391–415.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Borovička, J., Hansen, L. P., & Scheinkman, J. A. (2016). Misspecified recovery. *Journal of Finance*.

Breeden, D. T., & Litzenberger, R. H. (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business*, *51*(4), 621–651.

Brigo, D., & Mercurio, F. (2002). Displaced and mixture diffusions for analytically-tractable smile models. *Mathematical Finance*, *Bachelier Congress 2000*, 151–174.

Brown, S. J., & Warner, J. B. (1985). Using daily stock returns: The case of event studies. *Journal of Financial Economics*, *14*(1), 3–31.

Campbell, J. Y., & Cochrane, J. H. (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of Political Economy*, *107*(2), 205–251.

Campbell, J. Y., Lo, A. W.-C., MacKinlay, A. C., et al. (1997). *The econometrics of financial markets*, vol. 2. Princeton University Press.

Campbell, J. Y., & Thompson, S. B. (2007). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

Carr, P., Geman, H., Madan, D. B., & Yor, M. (2003). Stochastic volatility for lévy processes. *Mathematical Finance*, *13*(3), 345–382.

Carr, P., & Wu, L. (2003). What type of process underlies options? A simple robust test. *The Journal of Finance*, *58*(6), 2581–2610.

Chauvet, M., & Piger, J. (2008). A comparison of the real-time performance of business cycle dating methods. *Journal of Business & Economic Statistics*, *26*(1), 42–49.

Chen, T. (2011). Improve OVDV long-term volatilities. *Bloomberg Research*.

Chernov, M., & Ghysels, E. (2000). A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *Journal of Financial Economics*, *56*(3), 407–458.

Ching, W.-K., Ng, M. K., & Fung, E. S. (2008). Higher-order multivariate Markov chains and their applications. *Linear Algebra and its Applications*, *428*(2), 492–507.

Cochrane, J. H. (2008). The dog that did not bark: A defense of return predictability. *Review of Financial Studies*, *21*(4), 1533–1575.

Cochrane, J. H. (2009). *Asset Pricing (Revised Edition)*. Princeton University Press.

Cont, R., Da Fonseca, J., et al. (2002). Dynamics of implied volatility surfaces. *Quantitative finance*, *2*(1), 45–60.

Cooper, R. W., & Haltiwanger, J. C. (2006). On the nature of capital adjustment costs. *The Review of Economic Studies*, *73*(3), 611–633.

Cox, J. C., Ingersoll Jr, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, *53*(2), 385–407.

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

Dumas, B., Fleming, J., & Whaley, R. E. (1998). Implied volatility functions: Empirical tests. *The Journal of Finance*, *53*(6), 2059–2106.

Eberly, J., Rebelo, S., & Vincent, N. (2012). What explains the lagged-investment effect? *Journal of Monetary Economics*, *59*(4), 370–380.

Elton, E. J. (1999). Presidential address: Expected return, realized return, and asset pricing tests. *The Journal of Finance*, *54*(4), 1199–1220.

Engle, R. F., & Mustafa, C. (1992). Implied ARCH models from options prices. *Journal of Econometrics*, *52*(1), 289–311.

Fama, E. F., Fisher, L., Jensen, M. C., & Roll, R. (1969). The adjustment of stock prices to new information. *International Economic Review*, *10*(1), 1–21.

Farinas, J. C., & Ruano, S. (2005). Firm productivity, heterogeneity, sunk costs and market selection. *International Journal of Industrial Organization*, *23*(7-8), 505–534.

Fengler, M. R. (2009). Arbitrage-free smoothing of the implied volatility surface. *Quantitative Finance*, *9*(4), 417–428.

Figlewski, S. (2008). Estimating the implied risk neutral density. In T. Bollerslev, J. R. Russell, & M. Watson (Eds.) *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*. Oxford: Oxford University Press.

Fillebeen, T., & Sanford, A. (2016). Do small firms hedge: Forward looking beliefs using the recovery theorem. *Work in Process*.

Gulen, H., & Ion, M. (2015). Policy uncertainty and corporate investment. *The Review of Financial Studies*, *29*(3), 523–564.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, *6*(2), 327–343.

Hull, J. C., & Basu, S. (2016). *Options, futures, and other derivatives*. Pearson Education India.

Jackwerth, J. C., & Rubinstein, M. (1996). Recovering probability distributions from option prices. *The Journal of Finance*, *51*(5), 1611–1631.

Jarrow, R., & Rudd, A. (1982). Approximate option valuation for arbitrary stochastic processes. *Journal of Financial Economics*, *10*(3), 347–369.

Jensen, C. S., Lando, D., & Pedersen, L. H. (2017). Generalized recovery. *Available at SSRN 2674541*.

Jiang, G. J., & Tian, Y. S. (2007). Extracting model-free volatility from option prices: An examination of the vix index. *Working Paper*.

Jondeau, E., Poon, S.-H., & Rockinger, M. (2007). *Financial modeling under non-Gaussian distributions*. Springer Science & Business Media.

Kadan, O., & Manela, A. (2017). Estimating the value of information. *Working Paper*.

Kessides, I. N. (1990). Market concentration, contestability, and sunk costs. *The Review of Economics and Statistics*, *72*(4), 614–622.

Kothari, S., & Warner, J. (2007). Econometrics of event studies. *Handbook of Empirical Corporate Finance*, *1*, 3–36.

Kothari, S., & Warner, J. B. (1997). Measuring long-horizon security price performance. *Journal of Financial Economics*, *43*(3), 301–339.

McDonald, R. L. (2006). *Derivatives markets*, vol. 2. Addison-Wesley Boston.

Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, *4*(1), 141–183.

Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*, vol. 2. Philadelphia, PA: SIAM: Society for Industrial and Applied Mathematics.

Page, S. E., et al. (2006). Path dependence. *Quarterly Journal of Political Science*, *1*(1), 87–115.

Raftery, A. E. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological)*, *47*(3), 528–539.

Ross, S. (2015). The recovery theorem. *The Journal of Finance*, *70*(2), 615–648.

Salinger, M., & Summers, L. H. (1983). Tax reform and corporate investment: A microeconometric simulation study. In *Behavioral simulation methods in tax policy analysis*, (pp. 247–288). University of Chicago Press.

Sanford, A. (2016a). Forward-looking expected tail loss: An application of the recovery theorem. *Working Paper*.

Sanford, A. (2016b). State price density estimation with an application to the recovery theorem. *Working Paper*.

Sanford, A. (2017). Recovery theorem with a multivariate Markov chain. *Working Paper*.

Shimko, D. (1993). Bounds on probability. *Risk*, *6*, 33–47.

Stoll, H. R. (1969). The relationship between put and call option prices. *The Journal of Finance*, *24*(5), 801–824.

Welch, I., & Goyal, A. (2007). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.