Recovery Theorem with a Multivariate Markov Chain[*]

Anthony Sanford[†]

December 19, 2016

For the latest version, **click here**

Draft. Please do not circulate or distribute without permission from the author.

Abstract

The objective of this paper is to estimate three functions—the natural probabilities, pricing kernel, and discount rate—from observed state prices more accurately. Since these three functions appear as a single piece of information in the marketplace, we are unable to discern each function's specific contribution to a change in price. To solve this problem, Ross (2015) proposed the univariate Recovery Theorem (RT), where the transition probability matrix is based on the current level of the S&P 500. In contrast, I derive a transition probability matrix using a multivariate Markov chain. 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. This multivariate method improves significantly upon the univariate RT because the latter does not include volatility in the state transition, which makes its transition probabilities less precise. The forecast results indicate that the multivariate Markov chain produces superior results over 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%.

[*] This work was completed in part with computing resources provided by the University of Washington's Center for Studies in Demography and Ecology. All errors are my own.

[†] Department of Economics, University of Washington. Email: sanfoan@uw.edu

Contents

1 Introduction

Ross's (2015) Recovery Theorem (RT) is a breakthrough in asset 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). As such, 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 constitutes one of the best asset forecasting models available today. However, it has certain shortcomings that this paper aims to address.

Borovička et al. (2016) have shown that the RT's transition matrix may not be unique. This paper's theoretical contribution is that it changes the original univariate model to a multivariate one. The original RT derived the transition matrix using a simple constrained linear regression which assumed that the probability of transitioning to a new state was dependent on the previous state. Yet, a growing literature argues for volatility persistence in financial markets (Patton and Sheppard, 2015). If volatility is indeed persistent and option prices reflect the conditional variance of the underlying asset (Engle and Mustafa, 1992), then the transition matrix should control for volatility (Page et al., 2006).

Including volatility into the transition probability estimation also makes intuitive sense. It is analogous to a basic continuous time process thought to model asset prices well: a geometric Brownian motion. The geometric Brownian motion has two major parameters: a drift term and a diffusion term. The drift term is a deterministic component and can be thought of as the overall trend in the return process. The diffusion term is a random component and can be thought of as the volatility of the return process. In a world characterized by a geometric Brownian motion, the path that an asset takes is highly dependent on the random components as well as the overall trend. For example, if we assume that the drift and diffusion parameters are both equal to 10%, McDonald (2006) shows that the diffusion term contributes approximately four times more change in the asset price than the drift term for one-month price changes. The path which the asset follows is driven more by the randomness—the diffusion (or the volatility) process—than by the overall trend. The key insight for us here is that, although trend (measured by the level) is important for long-term forecasting, the majority of market movements are dictated by the volatility for shorter-term forecasts. This is why I argue that including volatility in the derivation of the transition probabilities 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, the transition matrix requires that we assume time homogeneity. To make this assumption, we must extrapolate option data based on time-to-expiration. This paper uses a methodology that I developed (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. Because the extrapolation method proposed is not directly comparable to the results of Ross (no access to his OTC data), I compare my results to those derived using Aït-Sahalia and Lo's (1998) method. I test both in multivariate and univariate Markov chain settings. The forecast results indicate that the multivariate Markov chain (using my proposed extrapolation methodology) produces superior results over 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

[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).

Recovery Theorem. In addition, it compares the efficacy of the multivariate Markov chain RT at various forecasting horizons. Results indicate that, surprisingly, the RT at shorter-term horizons (e.g. daily and weekly forecasts) is less reliable, and is more effective in the one-to-three-month forecast range.

The paper is divided into five main sections. Section 2 explains the original and multivariate RTs, and discusses the steps required to implement the theorem. Section 3 introduces the data. Section 4 presents the results. Section 8 simulates data using Monte Carlo simulations and tests both the univariate and multivariate RTs on this artificial data. Finally, section 5 explores possible extensions and concludes.

2 Model

The RT's ultimate goal is to obtain the natural probability distribution for equity returns. It accomplishes this by first 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. To understand this paper and its contributions, it is necessary to briefly introduce and provide the intuition behind the original RT. I break down the original RT into four major steps:

1. construct the state prices,

2. construct the transition probability matrix,

3. use the Perron-Frobenius (Meyer, 2000) theorem to extract what Ross calls the "natural probability transition matrix," and

4. produce what Ross calls the "natural marginal distributions," which can then 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. Once I have provided the background for the original RT, I move on to the intuition and derivation for the Recovery Theorem with a multivariate Markov chain proposed in this paper.

2.1 The Univariate 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, these prices are subject to adjustments since the 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 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):

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

where p_t is the price of an asset at some time t, E_t is the expectation operator, m_{t+1} is a stochastic discount factor, and x_{t+1} is the future cash flow of the asset. The variable m_{t+1} in equation 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 this 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 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.

In an effort to address these issues, Ross (2015) uses 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 use options to estimate state prices, which then allow us to estimate the underlying asset's risk premium.

2.1.1 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? A put 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 and Litzenberger (1978) produce a method to derive state prices, beginning with the continuous time Black-Scholes-Merton equation (Black and Scholes, 1973; Merton, 1973) as follows:

$$C(K,T) = \int_0^\infty [S_{t,p} - K]^+ p(S_{t,p},T) dS_{t,p} = \int_K^\infty p(S_{t,p},T) dS_{t,p}, \tag{2}$$

where $C(K,T)$ 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:

$$s(K,T) = C''(K,T) \tag{3}$$

which is Breeden and Litzenberger's (1978) result. In discrete time, we can estimate equation 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:

$$s(K,T) \approx -C_{K_1} + 2C_{K_2} - C_{K_3} \tag{4}$$

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.

Recall that I defined state prices as a function of the discount rate, risk aversion, and natural probability distribution. This can be expressed as follows:

$$s^{i,j} = p^{i,j} \phi^{i,j} \tag{5}$$

where $p^{i,j}$ is the state price transition matrix, $\phi^{i,j}$ represents the pricing kernel (or the stochastic discount factor), and $s^{i,j}$ represents the state prices. All of the components in equation 5 have already been defined with the exception of the state price transition matrix (see section 2.1.2).

According to Ross (2015), knowing the state price of a single state is not enough to be able to solve equation 5. 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 start solving the system of equations (see appendix A for more details).

Below, I provide an example of a state price matrix with 11 state levels. Examples for the state price transition matrix and the natural probability distribution are presented in section 2.2, where I also detail my proposed improvements to the RT.

State Price Results - An Example I obtain state prices by summing the butterfly spreads (with \$1 strike price increments) between $\frac{s_{i-1}+s_i}{2}$ and $\frac{s_i+s_{i+1}}{2}$. As a numerical example, in table 1, for a state 676, $\frac{s_{i-1}+s_i}{2}$ is 649 and $\frac{s_i+s_{i+1}}{2}$ is 708.[2] Hence, the values in table 1 can be interpreted as follows: the greyed out price, \$0.42, represents the price of an asset that guarantees a \$1 payoff if the market ends between 649 and 708 in 6 months (expiration).

State/ TTM (mths)	3	6	9	12	15	18	21	24	27	30	33	36
476	0.002	0.000	0.000	0.000	0.001	0.017	0.040	0.000	0.000	0.000	0.001	0.001
504	0.002	0.001	0.001	0.001	0.002	0.006	0.011	0.000	0.000	0.000	0.000	0.000
538	0.007	0.007	0.008	0.011	0.016	0.022	0.030	0.000	0.000	0.000	0.000	0.000
577	0.023	0.023	0.024	0.025	0.028	0.031	0.036	0.000	0.000	0.000	0.000	0.000
622	0.150	0.150	0.145	0.139	0.131	0.117	0.099	0.000	0.000	0.000	0.000	0.000
676	0.524	0.420	0.363	0.321	0.282	0.238	0.190	0.000	0.000	0.000	0.000	0.000
740	0.272	0.324	0.330	0.322	0.305	0.277	0.240	0.000	0.000	0.000	0.000	0.000
816	0.018	0.060	0.096	0.129	0.161	0.193	0.225	0.000	0.001	0.001	0.001	0.001
910	0.001	0.009	0.023	0.041	0.064	0.092	0.124	0.000	0.000	0.001	0.001	0.001
1023	0.000	0.001	0.003	0.005	0.007	0.007	0.004	0.000	0.997	0.996	0.995	0.994
1162	0.000	0.004	0.006	0.005	0.002	0.000	0.000	0.998	0.000	0.000	0.000	0.001

Note: 0.000 corresponds to a non-zero value; 0 corresponds to a zero value.

Table 1: State Prices on the S&P 500 for 1 April 1996

[2] $\frac{s_{i-1}+s_i}{2} = \frac{622+676}{2} = 649$ and $\frac{s_i+s_{i+1}}{2} = \frac{676+740}{2} = 708$.

2.1.2 Transition Probability Matrix (P)

In equation 5, I defined state prices as being a function of a pricing kernel, m, and a state price transition matrix, p. Formally, this is the probability of transitioning from a previous state, i, to a new state, j. More intuitively, we can define the state price transition 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 it 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 state price transition matrix is the same. If we think about the state transition price 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 actual state level. This understanding might seem trivial but it will be important later when I derive the multivariate Markov chain. The analogy to a forward rate will be used to demonstrate that the proper derivation of the transition matrix is multivariate rather than univariate.

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

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

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

$$s_{t+1} = s_t P,\ t = 1, ..., m-1 \tag{6}$$

$$\sum_{i=1}^{m} P_i = 1 \text{ and } P \geq 0$$

where m is the number of states and P is the state price transition probability matrix. Assumption 1 allows me to obtain the state price transition probabilities using equation 6. Time homogeneity assumes that the transition probabilities are the same regardless of which time-step we are trying to estimate. Jensen et al. (2015) propose a methodology to remove this assumption. The impact of removing this assumption in the multivariate case is not examined in this paper.

Ross (2015) notes that he adds a unimodality condition to his regression ($p_{i,i} \geq p_{i,j}$). He argues that this is necessary to minimize the error (however, it is unclear what error is being minimized). The unimodality condition ensures that the probability on the diagonal of the transition probability matrix is the largest. This condition implies that the market's best estimate of the future state is the current state. Although this might seem intuitive, it may not always be the case. Suppose, for example, that there is a strongly held belief that an event at time $t+1$ will adversely affect the S&P 500. It should be possible for market participants to believe that the most likely scenario would be for the index to decline. However, imposing unimodality removes this possibility. Since our goal is to capture as much market information as possible, it seems sensible to remove this assumption. Results show that the market often assumes that the highest transition probability is for the market to remain at its current level. However, some situations exist where it thinks that the market might move to the next state. If we imposed unimodality, we would remove this possibility.

Now that I have derived the state price transition matrix, I can rewrite equation 5 as follows:

$$p_{i,j} = \phi(\theta_i, \theta_j) f_{i,j} \tag{7}$$

where $p_{i,j}$ is a state price transition probability, $\phi(\theta_i, \theta_j)$ is the kernel factor, and $f_{i,j}$ is the natural probability that we are ultimately trying to derive. This equation stems directly from the fact that state price transition probabilities are simply state prices at some point in the future. They can therefore be defined similarly to equation 5 with the exception that, here, we are referring to terminal probabilities instead of transition probabilities. This is why we change the transition probability, $p_{i,j}$, to a natural probability, $f_{i,j}$.

Once the transition 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.

2.1.3 Natural Probability Transition Matrix (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:

$$\phi(\theta_i, \theta_j) = \delta \frac{U'(c(\theta_j))}{U'(c(\theta_i))} \tag{8}$$

where δ 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. More details can be found in Ross (2015). Once we have obtained the transition matrix from section 2.1.2, we can apply Ross's RT. The proof is available in Ross (2015).

Using a discrete time setup and assumption 2, I can rearrange equation 7 as:

$$U'_i p_{i,j} = \delta U'_j f_{i,j}, \tag{9}$$

where U'_i is the marginal utility such that:

$$U'_i \equiv U'(c(\theta_i)) \tag{10}$$

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

$$\phi_j \equiv \phi(\theta_1, \theta_j) = \delta(\frac{U'_j}{U'_1}) \tag{11}$$

where θ_1 is the current state. In continuous time, Ross defines the kernel as:

$$\phi(\theta_i, \theta_j) = \delta \frac{h(\theta_j)}{h(\theta_i)} \tag{12}$$

Using equation 12 and assuming transition independence, we have:

$$p(\theta_i, \theta_j) = \phi(\theta_i, \theta_j) f(\theta_i, \theta_j) = \delta \frac{h(\theta_j)}{h(\theta_i)} f(\theta_i, \theta_j) \tag{13}$$

where $h(\theta) = U'(c(\theta))$, and $p(\theta_i, \theta_j)$ is the state price transition function that was derived in section 2.1.2. 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 δ. Back to the discrete time specification, we can rewrite equation 13 in matrix form as:

$$DP = \delta F D \tag{14}$$

where P is the m x m state price matrix defined in section 2.1.1, 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:

$$D = \frac{1}{U_1'} \begin{bmatrix} U_1' & 0 & 0 \\ 0 & U_i' & 0 \\ 0 & 0 & U_m' \end{bmatrix} = \begin{bmatrix} \phi_1 & 0 & 0 \\ 0 & \phi_i & 0 \\ 0 & 0 & \phi_m \end{bmatrix} \frac{1}{\delta} \tag{15}$$

Rearranging equation 14, we get:

$$F = \frac{1}{\delta} D P D^{-1} \tag{16}$$

We obtained P in section 2.1.2, 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 (risk aversion), 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, be definition, implies that the rows of F are transition probabilities and so they must sum to 1. Hence, we have the following equation:

$$Fe = e \tag{17}$$

where e is simply a vector of ones. Substituting equation 17 into equation 16, we obtain:

$$Fe = \frac{1}{\delta} D P D^{-1} e = e \tag{18}$$

and if we define $z \equiv D^{-1}e$, we can rewrite equation 18 as:

$$Pz = \delta z \tag{19}$$

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 transition matrix will be nonnegative. Probabilities are, by definition, nonnegative and this was specified in the derivation of the state price transition matrix in section 2.1.2. 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 δ. This then allows us to solve for D, which we can introduce in the true distribution equation:

$$F = \frac{1}{\delta} D P D^{-1} \tag{20}$$

The description provided in the previous paragraph provides the mechanics of obtaining the true distribution. However, the question still remains as to what the application of the Perron-Frobenius theorem has allowed us to accomplish? As previously mentioned, the Perron-Frobenius theorem provides us with two critical pieces to the derivation of the true distribution: the discount factor and the risk-aversion. The discount factor is characterized by δ whereas the risk-aversion is characterized by D through the marginal rate of substitution which was defined in equation 15. The components to the marginal rate of substitution are simply the marginal utilities between consuming today versus consuming tomorrow. What the Perron-Frobenius theorem allows us to do is to determine the single unique discount factor and marginal utilities that dictates the transition paths between states. In others words, under the assumptions necessary for the Perron-Frobenius theorem to hold, there is only one set of marginal utilities and a discount factor that 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.

Please note that the derivation of the RT described above also applies to the multivariate Markov chain (MVMC) derivation described in section 2.2.1. The key difference between the univariate and the multivariate Markov chain is that, in the multivariate case, we are introducing an additional variable for the derivation of the state price transition matrix. For the purposes of this paper, the added variable is implied volatility (see section 2.2.1 for more details). By introducing volatility into the derivation, we can obtain a much more robust and precise transition matrix. As Page et al. (2006) describe, path dependence is critical in producing an accurate transition matrix. If the transition matrix is inaccurate in the first place, the RT will be inaccurate as well.

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. Similarly, we can use the probability and state levels to derive expected standard deviations. I present an example below.

True Probability Matrix Results - An Example The true probabilities are derived by applying the RT to the transition probabilities. Table 2 shows the expected distribution of returns for the Δ-forward period (three months in this case). I highlighted the 1.000 value in table 2 because it might appear to be an absorption state. However, this is only the case because of rounding in the table. It is true that the probability of any other state occurring if the market is at -0.29 is highly unlikely but it is not impossible.

State/State	-0.35	-0.29	-0.23	-0.16	-0.08	0	0.09	0.19	0.3	0.41	0.54
-0.35	0.271	0.326	0.366	0.080	0.000	0.000	0.000	0.000	0.000	0.000	0.000
-0.29	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
-0.23	0.000	0.435	0.468	0.038	0.053	0.000	0.000	0.000	0.000	0.000	0.000
-0.16	0.000	0.279	0.302	0.313	0.068	0.036	0.000	0.000	0.000	0.000	0.000
-0.08	0.000	0.110	0.119	0.124	0.254	0.395	0.000	0.000	0.000	0.000	0.000
0	0.001	0.001	0.003	0.010	0.091	0.564	0.281	0.022	0.002	0.000	0.003
0.09	0.000	0.000	0.000	0.000	0.171	0.405	0.388	0.000	0.000	0.000	0.000
0.19	0.033	0.037	0.040	0.041	0.056	0.097	0.160	0.387	0.007	0.007	0.016
0.3	0.004	0.005	0.006	0.006	0.008	0.014	0.014	0.018	0.891	0.006	0.014
0.41	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.548	0.530
0.54	0.020	0.022	0.023	0.022	0.030	0.050	0.045	0.052	0.086	0.077	0.575

Note: 0.000 corresponds to a non-zero value; 0 corresponds to a zero value.

Table 2: True Probabilities on the S&P 500 for 1 April 1996

Forecast Summary - An Example From the true distribution above, we divide the state prices by the kernel to obtain a marginal distribution. Multiplying the marginal distribution by the state levels, we obtain the forecast results seen below for April 1, 1996:

Statistic/ Horizon	3	6	9	12	15	18	21	24	27	30	33	36
Mean	0.021	0.052	0.080	0.099	0.112	0.127	0.146	0.298	0.927	0.926	0.928	0.929
Sigma	0.019	0.044	0.070	0.087	0.098	0.107	0.119	0.096	0.049	0.055	0.065	0.084

Table 3: Expected Return and Standard Deviation on the S&P 500 for 1 April 1996

Note that these results are not annualized, and therefore correspond to the expected returns and standard deviation for the number of months indicated in the table. As an example of the interpretation of this table, the grey value means that the RT forecasts the 3-month return to be 2.1%. The sigma row represents the expected standard deviation for the forecast horizon.

2.2 The Multivariate Recovery Theorem

2.2.1 Multivariate Transition Probability Matrix (P)

Why should we use a multivariate Markov chain in the specification of the transition matrix for the RT? The reasons are twofold. First, if markets are inefficient, it makes sense to include variables for which we may not have accounted completely through the state prices observed in the market. Second, if I argue that the transition matrix can be thought of as a forward rate, then our measure of the transition is actually for a time in the future, $t + \tau$. Although markets are likely to incorporate a lot of information, we cannot observe all possible future paths of the so-called forward rates. As such, it becomes necessary to account for these possible path dependencies in the regression equation. I discuss these two arguments in greater detail below.

The first argument for including volatility in the derivation of the transition matrix concerns market inefficiencies. If options were priced rationally, prices should reflect all observable data or information. However, whether options are actually priced accurately remains a question. Aıt-Sahalia et al. (2001) compare the cross-section of the time-series of the S&P 500 index and the state price densities (SPD) and find that options are not priced accurately. Specifically, in continuous time, they find noticeable differences between the diffusion process of the SPDs and that of the S&P 500. As such, we should expect that the $\delta\frac{U'(c_j)}{U'(c_i)}$ (equivalent of the stochastic discount factor (SDF) in this paper) is not accurately specified if we do not control for other variables, such as volatility. In other words, since we derive the kernel, which is the marginal rate of substitution (MRS), from the state price transition matrix, the resulting kernel may be misspecified if the state price transition matrix did not include volatility.

In this paper, I propose to add volatility to the estimation of the transition probability matrix to account for market inefficiencies and path dependence of option prices. By deriving the transition matrix using the state level and volatility, I am controlling for changes in the distribution of the implied volatility. The objective is to capture some of the characteristics in the distribution of implied volatilities that lead to mispricings.[3]

The second, and most important, argument for including volatility in the derivation of the transition matrix is that we are pricing securities at some point in the future. Recall, that a forward rate is defined as 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. Given the assumption of time homogeneity, we assume that the transition matrix for an asset priced today is the same as the transition matrix for an asset priced tomorrow. However, given the persistence of volatility, it is likely that the implied volatilities will be different in the future. Hence, implied volatilities should be included in the derivation of the transition probabilities. Furthermore, since these "forward rates" are not actually observed in the market, we cannot assume that the state prices today would accurately reflect state prices tomorrow.

For example, assume that we are deriving the "forward rate" today for a security that we would purchase in six months and that expires in one year. In pricing these "forward rate," we assume that the information set of investors today accurately reflects all information in the market (market efficiency). Thanks to our time-homogeneity assumption, we can use the state prices that we observe today for our rates in six months. However, information/uncertainty between today and our investment date in six months is not included in today's state prices. Some of this information can be captured through implied volatilities. By including Markov states for implied volatilities, we are not only controlling for implied volatilities, but also for the various paths that these volatilities can take. We are adding information that the market could price into state prices if it was able to observe them between today and our investment date six months from now. This becomes particularly important when the forecast horizon is shorter since the proper specification of the transition probability matrix is more dependent on volatilities (or randomness) than on the overall trend.

It is easier to see the importance of the transition path by assuming a model of

[3] Aıt-Sahalia et al. (2001) note that it may be useful to estimate the implied volatility distribution by defining the volatility distribution using higher moments and/or jump diffusion processes.

asset prices. Here, I propose to use a geometric Brownian motion because it is fairly simple to understand and is generally thought to model asset prices quite well. The transition probabilities being estimated in the RT are for states of the world which have not actually occurred (theoretical states of nature). In other words, we are estimating probabilities for possible paths (like the forward rate described above) which an asset can take at a future point in time.

The geometric Brownian motion has two major parameters: a drift term and a diffusion term. The drift term is a deterministic component and can be thought of as the overall trend in the return process. The diffusion term is a random component and can be thought of as the volatility of the return process. The key for a geometric Brownian motion is that the path that an asset takes, especially in the shorter-term, is highly dependent on the random components rather than simply the overall trend. Hence, the path that describes the asset's evolution is more heavily influenced by the short-term noise rather than the long-term valuation (the drift).

Assume that we characterize the changes in asset prices using a geometric Brownian motion (GBM) written as follows:

$$dS(t) = \mu S(t)dt + \sigma S(t)dZ(t) \tag{21}$$

where the percentage change in asset prices is normally distributed with an instantaneous mean, μ, and an instantaneous variance, σ^2. We can think of the instantaneous mean as a the trend of the process over time. The variance term can be thought of as the randomness of the process over time. Asset prices follow a general trend, but deviations from that trend can be attributed to randomness. The drift term is $\mu S(t)dt$ and the diffusion term is $\sigma S(t)dZ(t)$.

I use the GBM here for two major reasons: 1) it models stock prices in the Black-Scholes model quite well (Hull, 2006), and 2) it will be used later to simulate artificial option prices. There are, however, a few shortcomings to the model. The most notable of these issues is that the GBM assumes that the underlying process is normally distributed. The second issue is that it assumes that volatility is constant and that there are no jumps in stock prices. That being said, it still provides insight into the importance of volatility in modeling asset returns.

Now that we have chosen a model for our asset prices, I must demonstrate that the transition probability matrix is dependent on volatilities. In the short term, the primary driver of changes in asset prices is volatilities rather than the overall trend, which is important since we are applying the RT to relatively short-term forecasts (forecast intervals of less than one year).

Hypothesis 1. For a short-term forecasting period, the primary driver of changes in asset prices is the diffusion (volatility) rather than the drift (trend). The opposite is true for longer-term forecasts.

Proof. First, assume a discrete counterpart to the geometric Brownian motion as follows:

$$S(t+\tau) - S(t) = \mu S(t)\tau + \sigma S(t)\sqrt{\tau}$$

Now, if we define the ratio of the diffusion and the drift as follows:

$$\frac{\sigma S(t)\sqrt{\tau}}{\mu S(t)\tau} = \frac{\sigma}{\mu\sqrt{\tau}}$$

For simplicity, I assume that $\sigma = \mu$. If $\tau < 1$, the process that drives the change in asset prices is dominated by the diffusion (volatility). If $\tau > 1$, the drift process dominates the changes in asset prices. This result can be summarized as follows:

If $\tau < 1$	diffusion process $\sigma S(t)\sqrt{\tau}$ dominates drift process $\mu S(t)\tau$
If $\tau = 1$	diffusion process $\sigma S(t)\sqrt{\tau}$ and drift process $\mu S(t)\tau$ contribute equally
If $\tau > 1$	drift process $\mu S(t)\tau$ dominates diffusion process $\sigma S(t)\sqrt{\tau}$

□

In other words, for forecast horizons that are shorter than one year, the path taken by asset prices is more heavily dependent on the volatility rather than on the level. The ratio, $\frac{\sigma}{\mu\sqrt{\tau}}$, means that, as long as $\tau < 1$, the numerator (diffusion process) will be the more important factor in the path of asset prices. A simple numerical example should provide additional insight into the theorem.

Assume that a drift term and a diffusion term are both equal to 10% ($\mu = \sigma = 10\%$). If we are interested in a five-year forecast, we would have $\frac{0.1 \cdot \sqrt{5}}{0.1 \cdot 0.5} = 0.447$. Here, the result is less than one, which indicates that the drift process (denominator) is dominating the changes in the asset prices. If, on the other hand, I assume that we are interested in a three-month forecast, $\tau = 0.25$, which implies that $\frac{0.1 \cdot \sqrt{0.25}}{0.1 \cdot 0.25} = 2$. For the quarterly forecast, the diffusion process (and by extension the volatility) dominates the change in asset prices.

Since I focus primarily on a three-month forecast in this paper, the result from this simple numerical example closely aligns with the overall improvements generated by the multivariate model I propose. For instance, where the drift and diffusion of the process are similar, we see an improvement in the model by a scale of approximately two. This simplified example reinforces the link between the inclusion of volatility in the estimation of the transition probability matrix and the end result.

Now let us derive the multivariate Markov chain. Mathematically, this paper argues that:

$$p_{i,j} = P(s_{t+1} = i_0 | s_t = i_1, \Phi_t = i_1) \tag{22}$$

where $i_t, ..., i_0 \in \{1, ..., m\}$ is a state, s is the state price, and Φ is an additional variable necessary for a more accurate derivation of the state price transition matrix. The transition probability for s_{t+1} is not dependent only on the previous period's state price (s_t), but also on other variables. In this specific case, volatility is the other variable in question.

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

$$\min_{\lambda_{i,j}} \min_{P} [[\sum \lambda_{i,j} s_t P - s_{t+1}]_P] \tag{23}$$

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

$$\sum_{i=1}^{m} P_i = 1$$
$$P \geq 0 \text{ and } \beta \geq 0$$
$$\sum \lambda_{i,j} = 1$$

More specifically, for the purposes of this paper, I can rewrite the general specification in equation 23 to a two-variable Markov chain as follows:

$$\min_{\lambda_{i,j}} \min_{P,\beta} [[\lambda_{i,j} s_t P + (1 - \lambda_{i,j}) \Phi_t \beta - s_{t+1}]_{P,\beta}] \tag{24}$$

$$\sum_{i=1}^{m} P_i = 1 \text{ and} \sum_{i=1}^{m} \beta_i = 1$$
$$P \geq 0 \text{ and } \beta \geq 0$$

A simple specification of the multivariate model is to assume that the transition is solely dependent on state prices, but that we need to control for the the volatility in the regression. This implies that we estimate the transition matrix using a multivariate Markov chain as follows:

$$s_{t+1} = s_t P + vol_t \beta, \quad t = 1, ..., m-1 \tag{25}$$

where vol_t is the volatility state at time t. In other words, equation 25 assumes that $\lambda = 1$ in equation 24. This gives us a third dimension in the Markov chain and therefore results in a matrix of size $(m-1)^3$. Note that, in equation 25, I add the volatility variable, but we could just as easily add some other variable that affects state prices.

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 transition probability matrix result reliable.

Ideally, we would include a forecast of future volatility (see section 5) in the model because we want to estimate future state prices. However, present-day volatility forecasting models are not ideal. As a result, I include current volatility as a measure of future volatility in the transition probability regression based on the assumption that volatilities are persistent over short periods. However, please note that whenever there is a fundamental shift in volatility, the current volatility (equation 25) may not incorporate this “new” information.

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

$$\min_{P,\beta} \| s_{t+1} - s_t P - vol_t \beta \|^2 \tag{26}$$

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

$$\begin{gathered} \sum P = 1 \text{ and} \sum \beta = 1 \\ P \geq 0 \text{ and } \beta \geq 0 \end{gathered} \tag{27}$$

Equation 27 holds that the rows in our transition matrix should sum to one. Since P corresponds to probabilities, this is merely ensuring that the probability for each state transition sums to one. We also include 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 probabilities, by definition, are nonnegative.

Transition Matrix Results - An Example Here, I present a portion of the transition probability matrix obtained from the multivariate Markov chain using equation 26. It is based on the volatility (level) for April 1st, 1996. In reality, we have a three-dimensional table that contains a number of volatility state levels (similar to the state levels presented in table 1, but for volatility instead of underlying asset levels). Presenting a full three-dimensional table is not practical, so I present the section of the table based on the current volatility level. In section 2.1.2, I noted that I removed the unimodality condition ($p_{i,i} \geq p_{i,j}$). If table 4 included the unimodality condition, the diagonal of the matrix would be greater than or equal to any of the components in the row. As the table shows, without the unimodality condition, the diagonal is not always the largest value. The value next to it is the largest in the row in some cases (in bold). This result indicates a market belief that there will likely be a slight movement in the S&P 500 over the next three months.

State/State	-0.35	-0.29	-0.23	-0.16	-0.08	0	0.09	0.19	0.3	0.41	0.54
-0.35	0.096	**0.447**	0.340	0.220	0.029	0.005	0.016	0.054	0.114	0.001	0.000
-0.29	0.069	0.344	0.212	0.145	0.023	0.006	0.011	0.027	0.045	0.000	0.000
-0.23	0.007	0.206	0.310	**0.345**	0.077	0.023	0.037	0.063	0.083	0.001	0.000
-0.16	0.000	0.000	0.266	0.483	0.148	0.052	0.070	0.073	0.052	0.002	0.000
-0.08	0.000	0.000	0.093	0.591	0.347	0.170	0.213	0.132	0.000	0.005	0.001
0	0.002	0.002	0.007	0.023	0.150	0.524	0.272	0.018	0.001	0.000	0.000
0.09	0.000	0.000	0.115	0.599	0.404	0.229	0.317	0.246	0.000	0.009	0.003
0.19	0.000	0.064	0.323	0.545	0.236	0.087	0.143	0.227	**0.290**	0.005	0.002
0.3	0.022	0.239	0.368	0.472	0.136	0.039	0.074	0.153	0.237	0.002	0.001
0.41	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.501	**0.666**
0.54	0.182	0.692	0.530	0.473	0.041	0.000	0.029	0.207	0.655	0.002	0.000

Table 4: Transition Probabilities on the S&P 500 for 1 April 1996

The interpretation of table 4 is as follows: the market believes that there is a 0.447 probability of transitioning from a market level of -0.35 of the current S&P 500 to -0.29 in a three-month period. This particular case is an example of a situation where the diagonal is not the largest value. For an initial state of -0.35 of the current S&P 500 level, the most probable transition is -0.29 rather than the same future level of -0.35. Note also that the sum of the rows is equal to one (since rows correspond to probabilities and the sum of the probabilities should equal one).

2.2.2 MVRT - Natural Probability Matrix (F)

The derivation of the natural probability matrix for the MVRT is the same as for the univariate RT. The only difference is that the estimation of the natural probabilities will be more accurate if we start from a transition matrix that is more accurate. In section 4, I compare the results of using the univariate chain with those of a multivariate chain.

As was the case in section 2.1.3, the natural probability matrix is derived using the following equation:

$$F = \frac{1}{\delta} DPD^{-1} \tag{28}$$

where P is obtained using the multivariate Markov chain derived in section 2.2, δ is the discount rate estimated using the Perron root, and D is estimated using the characteristic root of P.

In the next two sections, I present the data and results for this paper.

3 Data

Data for this paper are available from the Wharton Research Data Services (WRDS) database. I use daily options 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[4] prices are from the CRSP dataset. The S&P 500 is generally thought to be the best proxy for the market portfolio. I obtained all of the option data from OptionMetrics through the Wharton Research Data Services (WRDS) database. The data is 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

[4] SECID 108105

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, occurs irregularly (Figlewski, 2008) and would make it more difficult to extract a proper implied volatility curve (see section 2.1.1). 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. Summary statistics appear in appendix A.

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, one of the benchmarks tested here obtains results that are very close to the results produced by Ross (see section 4). 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. Appendix A proposes a way to derive the extrapolated data required to construct state prices for this paper.

4 Results - Real-World Data

4.1 Forecast Results

Table 5 compares the results of Ross with my results, providing an overall summary of what the reader can expect in the upcoming section. The first column of table 5 shows Ross's results and the second column shows the results for the RT methodology proposed in this paper (Sanford multivariate method[5]). At first glance, the Sanford method seems to provide superior results compared to the methodology proposed by Ross.

	Ross method (Apr 09–Apr 13)	Sanford method (Apr 09–Apr 13)
	(1)	(2)
Intercept	-0.06054^{*}	0.027675^{**}
	(0.035068)	(0.009153)
Forecast	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 5: Ross Subsample - Summary Results

Please note that all of the results presented in this section are out-of-sample. Careful readers will notice that the very nature of the RT is such that in-sample results are not possible since there is no need for calibration. This fact about the RT makes the results even more powerful.

[5] The method proposed in this paper uses the new extrapolation methodology proposed in appendix A and the multivariate RT derived earlier.

4.1.1 Regression Tables

In this section, I compare results from Ross (2015) to results using the extrapolation method of Aït-Sahalia and Lo and to the results for my proposed method. Each results table is divided into four columns. The first column shows Ross's results for his specific subsample. In the second column, the model is identical to the one proposed by Ross in his paper with the exception that the option prices are extrapolated using the Aït-Sahalia and Lo (ASL) method. The transition matrix is derived using a univariate Markov chain. The third column represents the results of extrapolating with the Aït-Sahalia and Lo method and including a multivariate Markov chain in the derivation of the transition matrix (ASLMV). The final set of results (column 4) corresponds to the multivariate Recovery Theorem proposed in this paper (Sanford method). The forecast regression is as follows:

$$R_t = \alpha + \beta_t E_{t-1}[R_t] + \epsilon_t \tag{29}$$

where α is the intercept, β_t is the forecast coefficient, and $E_{t-1}[R_1]$ is the previous period's RT forecast. In the first iteration of these results, the forecast horizon is held to a quarter so t corresponds to 0.25 year. One of the criteria for the efficiency of the forecast is the forecast error. This error is defined as the residual, ϵ_t, found in equation 29 and graphed in section 4.1.2. The errors are used as a way to ensure that the model is accurately specified. Moreover, the errors are also used to compare the overall fit with other models. In general, the smaller the errors, the better the forecast.

	Ross method	ASL method	ASLMV method	Sanford method
	(Apr 09–Apr 13)	(Apr 09–Apr 13)	(Apr 09–Apr 13)	(Apr 09–Apr 13)
	(1)	(2)	(3)	(4)
Intercept	-0.06054^{*}	0.00103	0.00215	0.027675^{**}
	(0.035068)	(0.00144)	(0.00561)	(0.009153)
Forecast	5.710293^{**}	0.12816^{**}	0.23223^{***}	0.338864^{***}
	(1.95258)	(0.03905)	(0.06020)	(0.070478)
Observations	46	49	49	49
R^2	0.2162744	0.1832232	0.2404717	0.329701
Adjusted R^2	0.143715	0.1662070	0.224311	0.315439
F statistic	0.005436	0.001929	0.000348	$1.6e^{-05}$

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

Table 6: Results for the four methods, using the Ross subsample

In table 6, the sample sizes are slightly different from those of Ross. I suspect that some of the months in his sample were left out, since there are 49 months between April 1996 and April 2013. That being said, it seems unlikely that three months of additional data would significantly change his or my results. Since the dates were reported by Ross, I decided to keep the same dates and report a difference in the number of observations because I do not know which months were not included in his sample.

Since I do not have access to the same data as Ross, it was important to find a benchmark that could produce results similar to those of the original Ross RT. The results from columns 1 and 2 of table 6 are very similar, which indicates that the ASL method could be an appropriate benchmark to proxy Ross's method.

The adjusted R^2 is about 0.144 for Ross and 0.166 for the ASL. The only major difference between the two sets of results is the forecast coefficient: 5.710 for Ross and 0.128 for ASL. I suspect that the large difference in the scale of the coefficient is a result of a scaling that Ross may have done. Since the sign of the coefficient and the overall fit seems to be quite similar, I argue that this method is a suitable benchmark for the RT with the original methodology proposed by Ross (2015).

When I add the multivariate Markov chain to the ASL method (column 3), the coefficient increases both in scale and in significance. Moreover, the adjusted R^2 increases by almost 6%. This significant increase is consistent across samples and indicates that the multivariate Markov chain provides significantly better results than previous methods. This is also supported by the results for the Sanford method (column 4). Here, the adjusted R^2 increases to almost 32% all while increasing the overall statistical significance of the model. In this set of results, there is a large increase in the adjusted R^2 between the ASL method and the Sanford method. Based on this sample, it would appear that the extrapolation methodology I propose has a significant impact on the overall performance of the model. In comparison, the full sample, as will be discussed shortly, shows that the most significant improvement is from the multivariate component. This is further indication that the Sanford model as a whole significantly increases the forecasting ability of the RT.

In table 7, I reduce the sample to include time-series data from October 2002 to August 2015. I chose to use this subsample because it excludes some of the original data where a pattern could be observed in the residual regression plot (see figure 2).

	Ross method	ASL method	Sanford method
	(Apr 09–Apr 13)	(Oct 02–Aug 15)	(Oct 02–Aug 15)
	(1)	(2)	(3)
Intercept	-0.06054^{*}	-0.00104	0.008863
	(0.035068)	(0.00106)	(0.005871)
Forecast	5.710293^{**}	0.146706^{***}	0.336237^{***}
	(1.95258)	(0.03509)	(0.052532)
Observations	46	154	154
R^2	0.2162744	0.103107	0.212301
Adjusted R^2	0.143715	0.097206	0.207118
F statistic	0.005436	$4.9e^{-05}$	$1.82e^{-05}$

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

Table 7: Results for the four methods, using the Ross subsample

Table 7 still indicates a significant improvement over the original Ross results. The adjusted R^2 for the Sanford method (column 3) is still almost 1.5 times larger than Ross's original results. The striking difference here is between column 2 and column 3 where the adjusted R^2 more than doubles.

The following table (table 8) extends the analysis to the full sample of S&P 500 options data available in OptionMetrics. The general patterns apparent in table 6 seem to persist throughout the entirety of the sample.[6] Here, however, we can see an even greater difference between the univariate and the multivariate models in columns 2 and 3. The adjusted R^2 increases from 0.11850 in the univariate case to 0.267522 in the multivariate case. The results for my proposed method (column 4) are even stronger: they more than double Ross's R^2 despite the increase in sample size. In all columns, the forecast coefficient maintains its statistical significance.

	Ross method (Apr 09–Apr 13)	ASL method (Apr 96–Aug 15)	ASLMV method (Apr 96–Aug 15)	Sanford method (Apr 96–Aug 15)
	(1)	(2)	(3)	(4)
Intercept	-0.06054^{*}	-0.00103	-0.00287	0.004841
	(0.035068)	(0.00092)	(0.00632)	(0.004626)
Forecast	5.710293^{**}	0.16609^{***}	1.29653^{***}	0.424605^{***}
	(1.95258)	(0.02928)	(0.14003)	(0.042434)
Observations	46	232	232	232
R^2	0.2162744	0.122664	0.270679	0.302375
Adjusted R^2	0.143715	0.11850	0.267522	0.299355
F statistic	0.005436	$4.24e^{-08}$	$1.46e^{-17}$	$8.19e^{-20}$

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

Table 8: Results for the four methods, using the full OptionMetrics sample (except column 1)

Table 9 shows the results for a random sample selected by **R**.[7] The sample shown is from the beginning of the available data (April 1996) to April 2013, or 204 months. The forecast coefficients are still highly statistically significant ($p < 0.001$). The adjusted

[6] Please note that the results for the Ross column (column 1) are still the same as in the previous table because it is impossible for me to extend Ross's analysis without his data.

[7] I will include more subsamples in the final paper.

R^2 increases from around 12% in the original benchmark (ASL) to around 31% for my proposed method. The magnitude of the coefficients are also quite similar to those from previous samples. To provide a visual representation of the results, I describe time series plots in the following section.

	Ross method (Apr 09–Apr 13)	ASL method (Apr 96–Apr 13)	ASLMV method (Apr 96–Apr 13)	Sanford method (Apr 96–Apr 13)
	(1)	(2)	(3)	(4)
Intercept	-0.06054^{*}	-0.00113	-0.00275	0.003038
	(0.035068)	(0.00092)	(0.00706)	(0.005131)
Forecast	5.710293^{**}	0.16484^{***}	1.34295^{***}	0.450346^{***}
	(1.95258)	(0.03136)	(0.15186)	(0.047678)
Observations	46	204	204	204
R^2	0.2162744	0.120316	0.278112	0.308482
Adjusted R^2	0.143715	0.115961	0.274556	0.305025
F statistic	0.005436	$3.72e^{-07}$	$4.52e^{-16}$	$9.59e^{-18}$

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

Table 9: Results for the four methods, using a random subsample (except column 1)

4.1.2 Plots for Sanford Method (Proposed Method)

The results presented in this section are limited to my proposed model (Sanford method) and the benchmarks I included in section 2. I am unable to present visual representations for the results of Ross's exact specification because I do not have access to the data he used.

The time series plots compare the actual time series for the quarterly return of the S&P 500 (calculated monthly) to the corresponding forecasted recovered quarterly returns (Sanford method). In other words, the results of the RT are fed through the regression equation from the previous section to obtain predicted values that are plotted against the actual returns. For all of the time series graphs discussed in this paper, the red line corresponds to the specific method of the section (in this case, the Sanford method) while the blue line represents the S&P 500's actual return. Figure 1 is the time series of the predicted values from my proposed method plotted against the actual S&P 500 returns for the entire sample. The peaks and troughs of the forecasted and actual S&P 500 returns line up quite nicely, apart from a few exceptions. In particular, around June 2012 and May 2014, there are large spikes in the forecast that are not accompanied by spikes in the actual return. It would be interesting to examine what happened during those times to see if perhaps there were some macroeconomic announcements that may have caused overreaction by market participants.

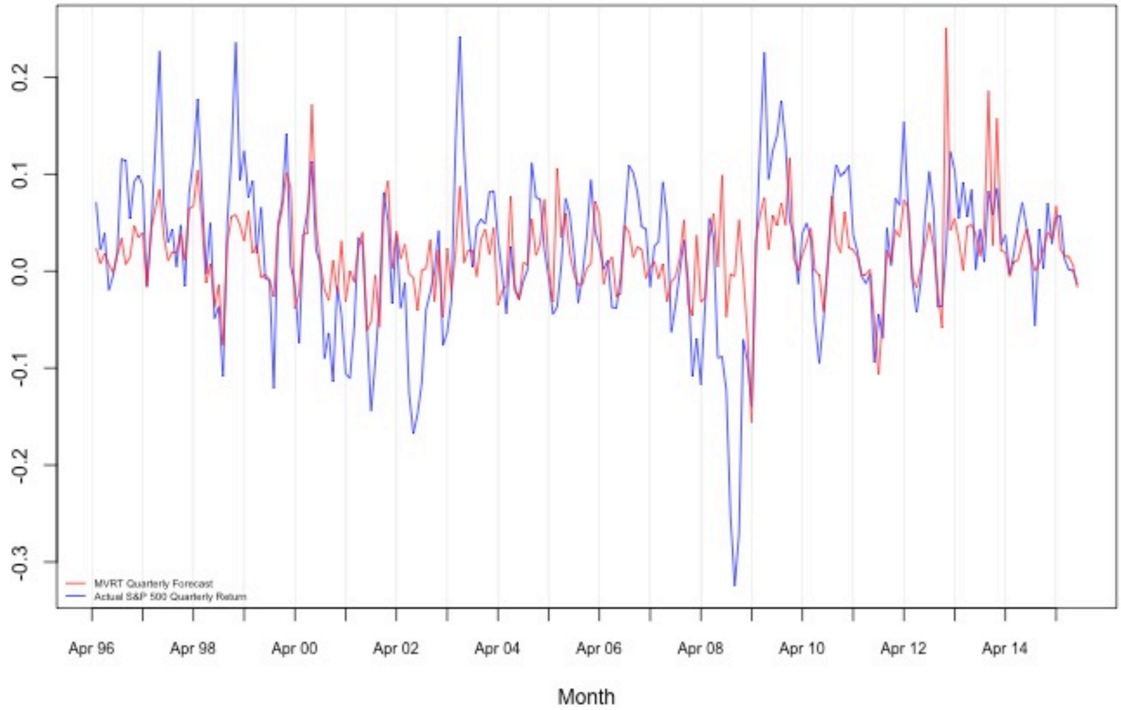

Figure 1: Sanford method vs S&P 500 return

Figure 2 shows the residuals from the regression equation. More specifically, it is the graphical representation of the following:

$$\epsilon_t = \alpha + \beta_t E_{t-1}[R_t] - R_t \tag{30}$$

which implies that a negative value occurs whenever the actual return is larger than the forecasted return. In figure 2, there does appear to be a pattern from the beginning of the sample, in April 1996, to about November of 2002. The pattern is a slight downward slope in the residuals. Along with the ACF/PACF plots discussed next, this may be an indication that there is something odd about the model specification for the first few years of the analysis. In order to verify this, I will analyze the same graphs excluding the time-series from mid-1996 to late-2002. Beyond 2002, most of the points seem to cluster near or around the zero line and the points are quite symmetric around zero. However, the negative residuals are larger than the positive residuals, which means that the forecast does not forecast extreme values very well.

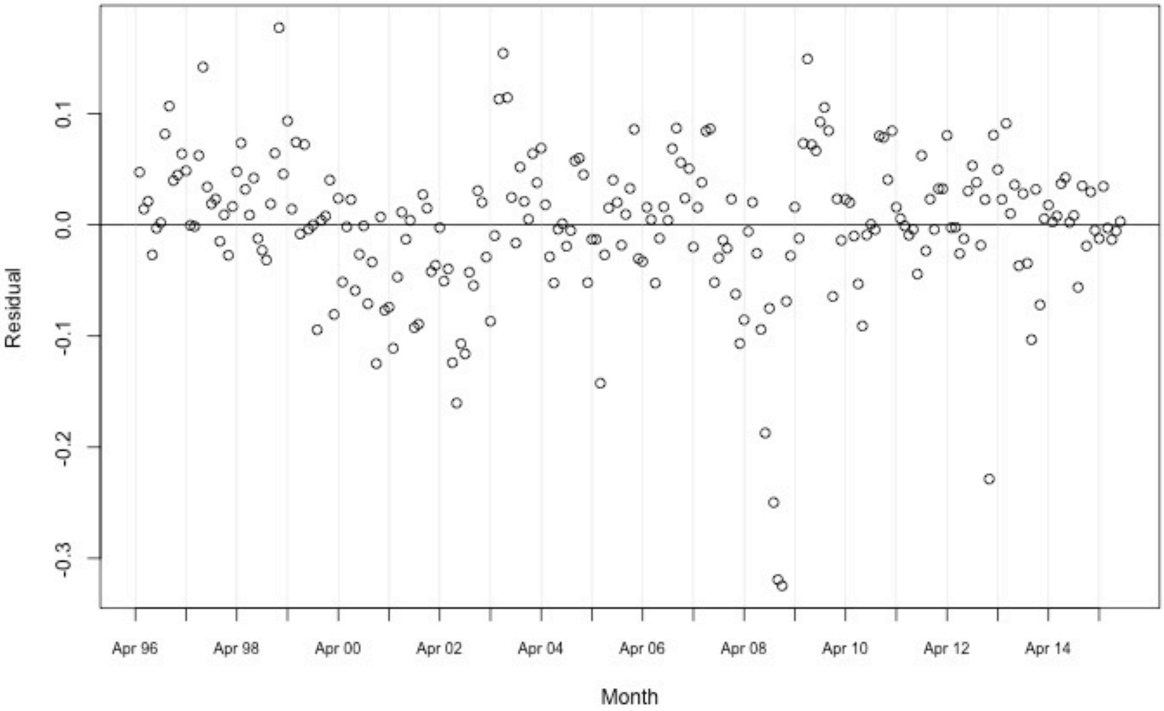

Figure 2: Residual plot: Sanford method vs actual S&P 500 return

In order to assess the overall quality of the model, I also present the ACF and PACF plots for the regression residuals for all models. The ACF and PACF plots can help us determine if there is any autocorrelation in the residuals. Figure 3 shows both the autocorrelation function (ACF) and the partial autocorrelation function (PACF) for the regression residuals. Results from this figure indicate that some autocorrelation persists through the fifth lag. In order to confirm this finding, I conducted a Durbin-Watson (Durbin and Watson, 1951) test and a Breush-Godfrey (Breusch, 1978; Godfrey, 1978) test (LM test for serial correlation). For both tests, I reject the null hypothesis that the autocorrelation of the residuals is equal to zero. This could indicate that there is still some information that is unexplained and that could be exploited in order to obtain a better forecast. Please note that the autocorrelation persists even with the inclusion of lags into the regression equation.

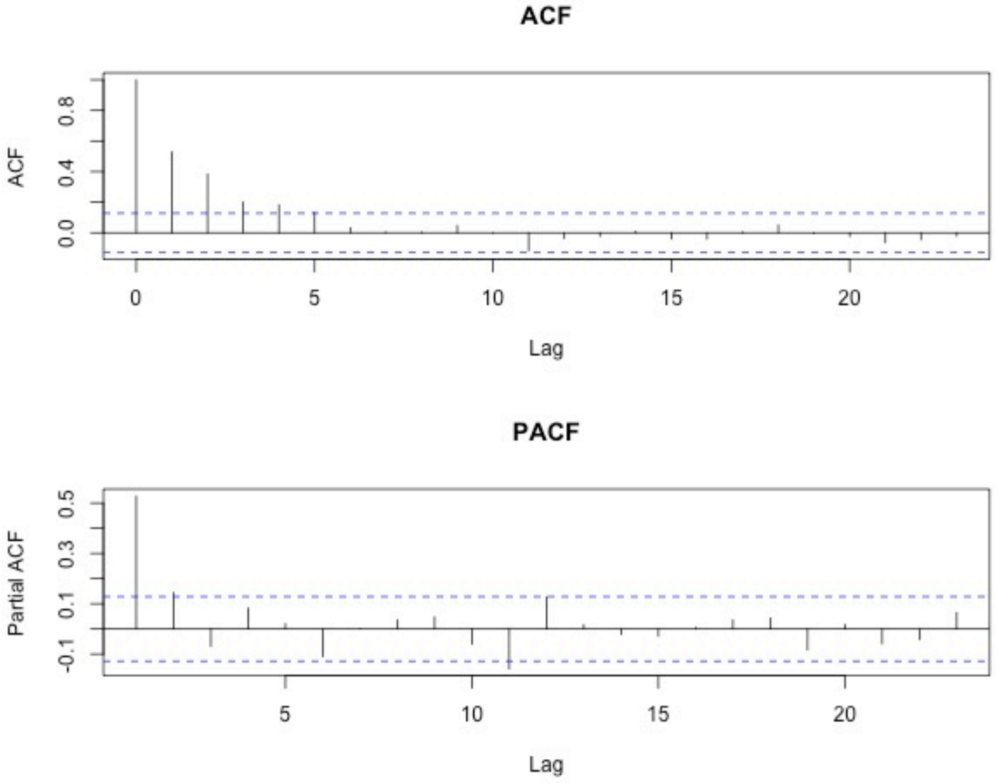

Figure 3: Regression residual ACF and PACF: Sanford method vs S&P 500 return

In the following two figures, I present the results for a subsample (from October 2002 to August 2015) that excludes the dates that exhibited a downward pattern in the residual plot. I do not, however, include the time series plot since it is identical to the one in figure 2.

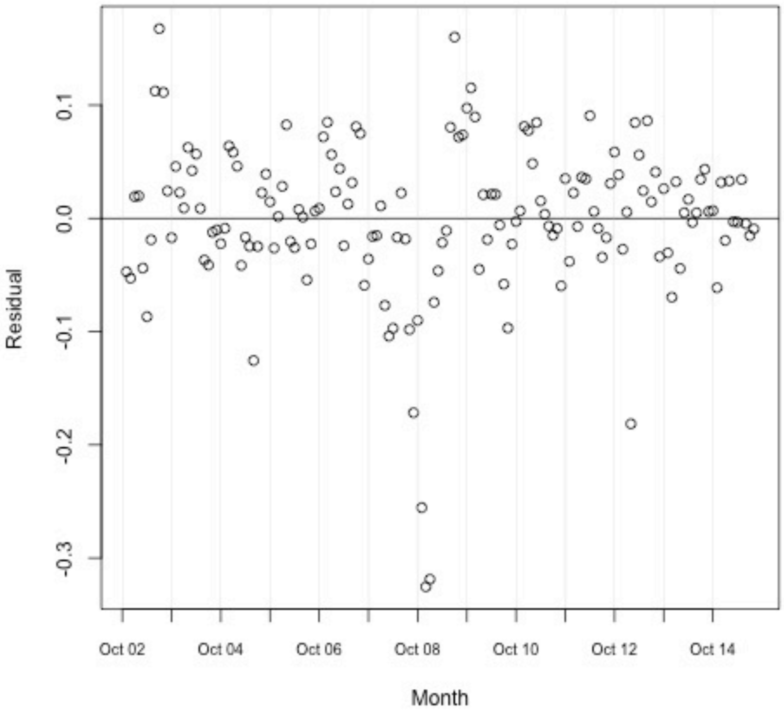

Figure 4: Regression residual (Oct 02–Aug 15): Sanford method vs S&P 500 return

As expected, figure 4 no longer exhibits any patterns throughout the subsample. Apart from a few outliers, everything seems to be behaving normally.

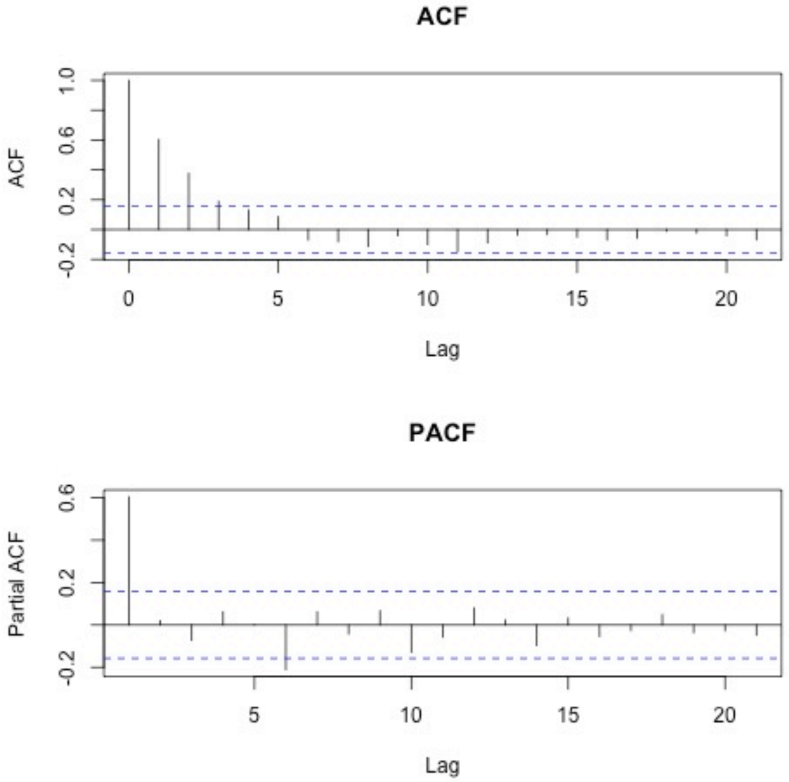

Figure 5: Regression residual ACF and PACF (Oct 02–Aug 15): Sanford method vs S&P 500 return

Comparing figure 5 to figure 3, there is clearly less autocorrelation. However, there is still autocorrelation up to the third lag. Again, this indicates that there is room for improvement in this model.

Below, I present the same figures as above, but for an even smaller subsample. I chose the sample dates of April 2009–April 2013 because they are the dates that Ross chose to use in his presentation of the RT.

Figure 6 gives a better picture of how well the Sanford method performs. Although the Sanford method does not forecast the exact levels of the S&P 500 returns, it does track the overall trends quite well. As was the case in the full sample (figure 1), the Sanford method has an occasional overreaction, which, in this subsample, occurs around March 2013. The residual plot illustrates this finding more clearly. One possible explanation of these overreactions might be that they are caused by structural breaks in the underlying asset's prices.

The graphs for the ASL and ASLMV models can be found in appendix A. Most of the figures are quite similar to the ones found in this section. There is, however, one exception: the ACF graph. In the ACF graph for the ASLMV (figure 15), I find no indication of autocorrelation of the residuals. This is one of the biggest differences between figures 15 and 3. It seems that the extrapolation method I use impacts the final results since we can see autocorrelation in the residuals. As of now, I do not have an explanation for this finding.

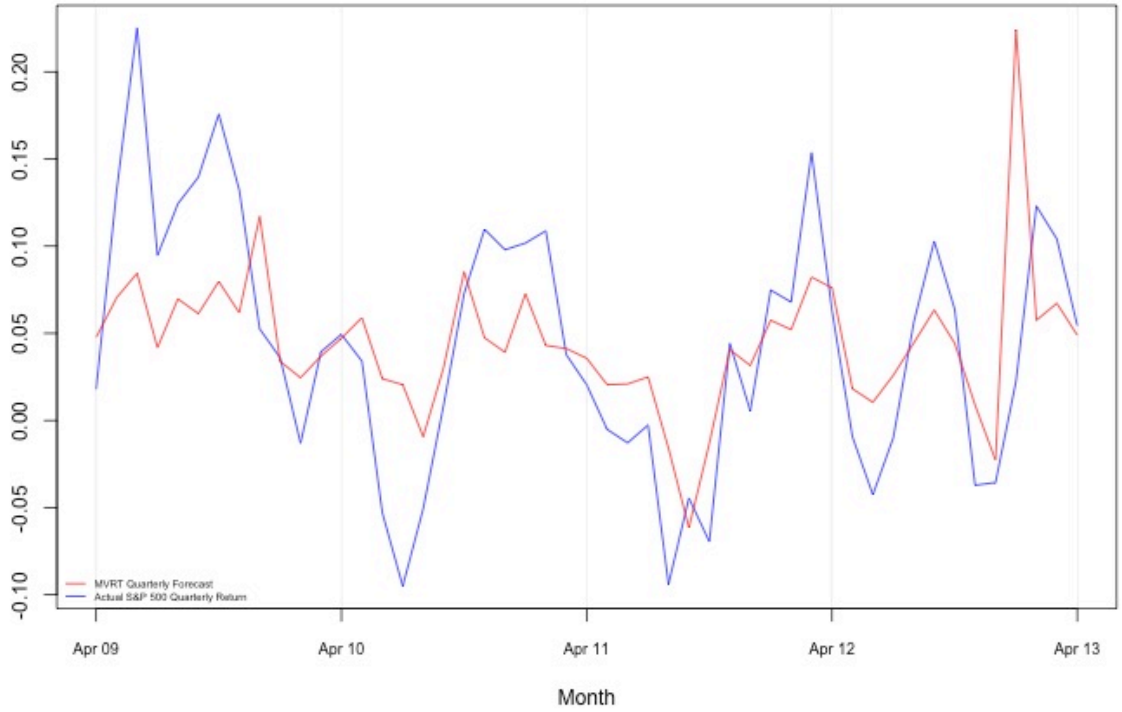

Figure 6: Sanford method vs S&P 500 return (Ross subsample)

Figure 7 is the residual plot for the forecast regression equation for the Ross subsample. The observations for this subsample are, once again, similar to observations for the entire sample (figure 2). There are no obvious patterns in the residuals and the distribution of points is fairly symmetric around the zero line. One distinction is that, if we connect the outer dots in this subsample, there appears to be a very slight decreasing trend. However, the trend is very slight, so I do not believe that this is a major issue. The occasional outliers lead me to believe that the forecasting model does not accurately forecast large movements in the underlying asset (in this case the S&P 500).

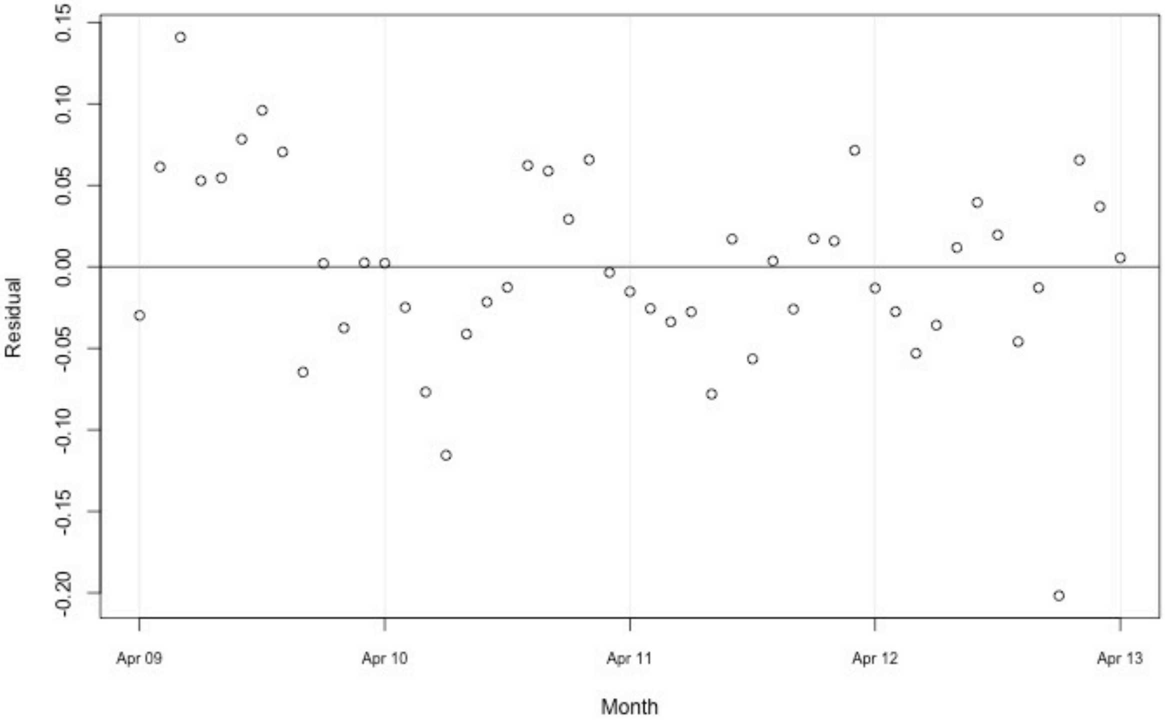

Figure 7: Sanford method regression residual (Ross subsample)

Figure 8 shows the ACF and PACF for the Ross subsample. The ACF does not exhibit as much autocorrelation with the lags as it did in the full sample. In figure 3, we could discern autocorrelation until a lag of five, whereas the autocorrelation tapers off starting after the first lag here. This is a sign that, at least for this subsample, the model is correctly specified.

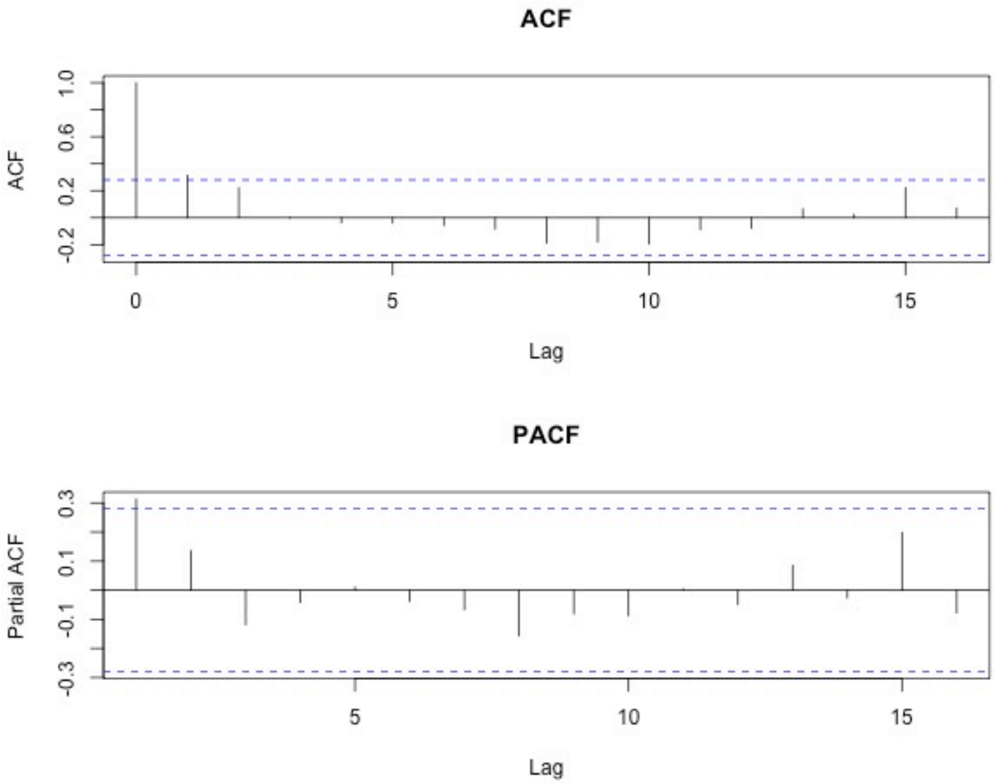

Figure 8: Sanford method regression residual ACF and PACF (Ross subsample)

4.1.3 Comparing the Sanford method to the Dividend-Price Ratio

In table 10, I compare the Sanford method to the dividend-price ratio (Cochrane, 2008).[8] Although it is not directly related to the work in this paper, the dividend-price ratio is often considered to be a benchmark for equity market forecasting. For this reason, I have included a comparison in this paper. In order to obtain table 10, I used the Matlab code that Cochrane provides on his website (Cochrane, 2008). The first column shows the results for the dividend price ratio, column 2 presents the results for the log of the dividend price ratio, and column 3 presents the results for the Sanford method proposed in this paper. The Sanford method is a significant improvement over the dividend/price ratio model in terms of R^2. This is not so surprising since

[8] Readers may want to consult Cochrane (2008) and Campbell and Thompson (2008) for more information and comparisons of the various forecasting models.

the forecast horizon here is three months. The Sanford model seems to explain a much larger fraction of return variation. It would be interesting to see if, like the dividend/price ratio, the Sanford method performs better at longer forecast horizons when I add lags into the forecast regression.

Relating back to the importance of adding volatility into the transition matrix estimation, in long-run forecasting models like the dividend price ratio, we are not so interested in the volatility estimation because we are mostly concerned with the long-run drift term, which gives us an idea of the overall trend for the underlying process. However, when we are attempting to estimate shorter-term asset prices, the volatility term becomes very important because it contributes to the likelihood of certain paths. This explains why the results for the dividend price ratio presented here are so dismal. These models are meant to capture long-run trends rather than short-term movements.

One important consideration, however, is the fact that the dividend price ratio is substantially faster to compute when compared to any of the RT methods presented in this paper. In part, this is because extrapolation of option prices is necessary (unless I use the OTC data obtained by Ross) and takes a significant amount of time. Furthermore, estimating the individual components required by the RT is considerably longer (in terms of computation time) compared to a simple regression between two variables.

	D_t/P_t	$d_t - p_t$	Sanford method
	monthly sample	monthly sample	monthly sample
	(Jan 96–Jan 15)	(Jan 96–Jan 15)	(Jan 96–Jan 15)
	(1)	(2)	(3)
Intercept	13.9252	0.4055	0.004841
	(10.6287)	(0.9395)	(0.004626)
Forecast	1.3102**	0.4316	0.424605***
	(0.0076)	(0.2405)	(0.042434)
Observations	232	232	232
R^2	0.0299	0.0368	0.302375
Adjusted R^2	N/A	N/A	0.299355
F statistic	N/A	N/A	$8.19e^{-20}$

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

Table 10: Comparison of dividend-price ratio and Sanford method RT

4.1.4 Comparing the Sanford Method to the CAY Variable

In table 11, I show the comparison between the results for the Sanford method and the consumption-wealth ratio (CAY) of Lettau and Ludvigson (2001). Their paper shows that the ratio of aggregate consumption to wealth has forecasting power of stock returns. Moreover, and in contrast to the dividend-price ratio presented above, the CAY ratio is meant to forecast shorter-term fluctuations in stock prices. Hence, it constitute a good benchmark to compare the forecast obtained by the Recovery Theorem. The data for the CAY ratio was obtained from the companion webpage from the authors of the paper (Lettau and Ludvigson, 2001) and is for the period that matches the available data from OptionMetrics (from April 1996 to August 2015). The forecast is a quarterly forecast updated quarterly (which is different from the quarterly forecast updated monthly in the previous tables). The forecast is updated quarterly since that is the frequency at which the CAY ratio is obtained. Clearly, the Sanford method introduced in this paper is significantly better than the CAY method. However, for this period, the CAY method does not produce great results. That being said, even if I compare the Sanford results to those presented in the original paper by Lettau and Ludvigson (2001), the results for the RT are still significantly better. More specifically, the adjusted R^2 for the entire sample presented by Lettau and Ludvigson (2001) is 0.09 which is still about only about a third of the adjusted R^2 from the Sanford method.

	CAY quarterly sample (Apr 96–Aug 15)	Sanford method quarterly sample (Apr 96–Aug 15)
	(1)	(2)
Intercept	0.01936	0.014709
	(0.008357)	(0.010762)
Forecast	0.650148	0.617909***
	(0.487169)	(0.147358)
Observations	78	78
R^2	0.022898	0.235753
Adjusted R^2	0.010041	0.222346
F statistic	0.186011	$9.70e^{-05}$

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

Table 11: Comparison of CAY and Sanford method RT

5 Conclusion

The purpose of this paper was to improve the estimation of the natural probabilities derived from the Recovery Theorem (RT). The major contribution of this paper is that it extends the RT by changing the univariate transition probability matrix to a multivariate one. By changing the derivation of the transition matrix to a multivariate Markov chain, I argue that the transition probabilities are more accurately defined. I also add the variable of volatility, which results in significant improvements in the RT results. I show, using a geometric Brownian motion, that the inclusion of volatility in the multivariate chain improves the model. For shorter-term forecasts, the path of asset prices is mostly dominated by the diffusion process rather than the drift. Since the diffusion process is, in essence, the volatility, the inclusion of volatility in the multivariate RT is critical. The forecast regression's R^2 increases from 0.144 using Ross's specification to 0.315 using the Sanford method (outlined in this paper).

The Recovery Theorem was a giant leap forward in the forecasting of asset returns. This paper improves on those results and will make it possible to use this methodology for other asset pricing endeavors. A number of extension 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 may be to explore whether firms change their hedging behavior in response to certain shocks, where the shocks are derived from the true distribution (Fillebeen and Sanford (2016)).

The multivariate RT could 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 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 literature.

References

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

Aıt-Sahalia, Y., Wang, Y., and Yared, F. (2001). Do option markets correctly price the probabilities of movement of the underlying asset? *Journal of Econometrics*, 102(1):67–110.

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

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

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

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

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

Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, 17(31):334–355.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531.

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

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.

Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression. ii. *Biometrika*, 38(1/2):159–177.

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

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

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

Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica: Journal of the Econometric Society*, pages 1293–1301.

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. (2006). *Options, futures, and other derivatives*. Pearson Education.

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

Jensen, C. S., Lando, D., and Pedersen, L. H. (2015). Generalized recovery. *Available at SSRN 2674541*.

Lettau, M. and Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *the Journal of Finance*, 56(3):815–849.

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

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

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

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

Patton, A. J. and Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3):683–697.

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

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

Rubinstein, M. (1994). Implied binomial trees. *The Journal of Finance*, 49(3):771–818.

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*.

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

A Appendix A - Implied Volatility Extrapolation

In this section, I introduce my proposed implied volatility extrapolation method and show how extrapolated prices lead to a dense set of option prices. I then briefly define and derive the benchmark extrapolation method used in this paper: the Aït-Sahalia and Lo model. For more information on the extapolation methodology defined in this section, see Sanford (2016b).

A.1 Strike Price Extrapolation

The first step for the MVRT involves extrapolating the volatility surface with respect to two dimensions: strike prices and time-to-maturity. We extrapolate in terms of strike prices because there are only a certain number of strikes that are traded on any given day. For example, table 12 shows the (unique) strike prices for call options on the S&P 500 for 1 April 1996. However, for this specific day, we would need a set of strike prices ranging from about 350 to 1,200 in order to produce a complete volatility surface. Thus, extrapolation is necessary.[9]

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 12: Strike Prices on S&P 500 call options for 1 April 1996

The strike price extrapolation is based on a slightly modified risk-neutral density estimation methodology proposed by Figlewski (2008). Figlewski (2008) 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. That being said, I have found that using smoothed B-splines rather than quartic splines provides a better overall fit. This is what I used in this paper.

[9] Extrapolation based on strike price is common practice in the volatility surface literature (Jackwerth and Rubinstein, 1996; Rubinstein, 1994; Figlewski, 2008).

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

$$\min_{\beta} ||C - G\beta||^2 + \lambda\beta'\Omega\beta \tag{31}$$

where

$$G_{i,j} = g_j(\sigma_{IV,i}), \quad i, j = 1, ..., n \tag{32}$$

$$\Omega_{i,j} = \int g_i''(t)g_j''(t)dt, \quad i, j = 1, ..., n \tag{33}$$

where n is the number of knots, x is the actual knot, $g()$ are the B-spline basis functions, Ω is the penalty matrix, and λ is the smoothing parameter. Next, we need to define what we mean by a B-Spline basis function.[10] We can define the B-Spline function as follows:

$$G_{i,j} = \sum_{i=1}^{n+1} B_j(\sigma_{IV,i})G_i, \qquad \sigma_{IV,min} \leq \sigma_{IV,i} < \sigma_{IV,max} \tag{34}$$

where G_i corresponds to the control points, $B()$ is the basis function of order j, and x corresponds to the knots. Then, we can define the basis function from the B-spline as follows:

$$B_{i,1}(\sigma_{IV}) = \begin{cases} 1, & \text{if } \sigma_{IV,i} \leq \sigma_{IV} < \sigma_{IV,(i+1)} \\ 0, & \text{otherwise} \end{cases} \tag{35}$$

$$B_{i,j}(\sigma_{IV}) = \frac{\sigma_{IV} - \sigma_{IV,i}}{\sigma_{IV,(i+j-1)} - \sigma_{IV,i}} B_{i,j-1}(\sigma_{IV}) + \frac{\sigma_{IV,(i+j)} - \sigma_{IV}}{\sigma_{IV,(i+j)} - \sigma_{IV,(i+1)}} B_{i+1,j-1}(\sigma_{IV}) \tag{36}$$

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

$$\hat{r}(C) = \sum_{j=1}^{n} \hat{\beta}_j g_j(\sigma_{IV}) \tag{37}$$

[10] Note that the notation here is slightly different from traditional notation in order to be consistent with the notation in the rest of the paper.

A.2 Time-to-Maturity Extrapolation

Table 13 shows the TTM on S&P 500 call options for 1 April 1996 in number of years. The time interval between each of the TTMs is not constant. Therefore, I need to extrapolate the data such that TTM follows a constant interval (for now, this interval is set to a constant three-months).[11]

0.05	0.13	0.23	0.47	0.72	0.97	1.22	1.72

Table 13: Time-to-maturity on S&P 500 call options for 1 April 1996

For the TTM extrapolation, I use a method devised by Bloomberg (Chen, 2011) as an extension of Heston (1993). First, let us define the extrapolated call price as follows[12]:

$$C(T, K) = \sum_{l=1}^{N} p_l(T) \cdot BSP(\xi_l(T) S_{0,p}, K, r_f, \Sigma_l(T)/\sqrt{T}) \tag{38}$$

where BSP corresponds to the traditional Black-Scholes equation (Black and Scholes, 1973) where each variable is a regular Black-Scholes input with certain parameters adjusted for extrapolation. The extrapolation details and the parameters in equation 38 are discussed in greater detail later in this section.

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

$$\varphi(t) = \frac{T_{i+1} - t}{T_{i+1} - T_i} \tag{39}$$

$$\eta_l(t) = log(\frac{\xi_{l+1}(t)}{\xi_l(t)}) \tag{40}$$

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 time-to-maturities.

[11] Later in the paper, I test various interval lengths.

[12] Note that it is trivial to show that extrapolating the option price is the same as extrapolating the option price as long as the inputs for the equation are the same but where the volatility is, in fact, the implied volatility.

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:

$$D(t) = 1 - Q(t) = \sum_{l} p_l(t) = e^{-\Lambda(t)t} \tag{41}$$

where the initial $\Lambda(t)$ is obtained from the Bloomberg survival probability data. Once we have the benchmark hazard rate and survival probability, we need to estimate four equations (the new $\Lambda()$, $p_l()$, $\eta_l()$, and $\Sigma_l()$) and use the values as inputs for equation 38. The specific equations are dependent on whether we are extrapolating between TTMs, we are doing a shorter-term TTM extrapolation (less than three months), or a longer-term TTM extrapolation (greater than six months).[13] Each of these is derived and discussed in its own section below.

Shorter-Term Extrapolation A shorter-term extrapolation is an extrapolation that occurs either within three months of an available datapoint, or an extrapolation 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 $\lambda(t)$ in order to obtain $p_l(t)$. This is obtained as follows:

$$\Lambda_{new} = \Lambda e^{\frac{x_m^2 - x^2}{2T_t}} \tag{42}$$

$$\hat{\Lambda}_{new} = \Lambda_{new} e^{\frac{x^2}{2}(\frac{1}{T_0} - \frac{1}{t})} \tag{43}$$

where $x_m = K_{min}/F(T_i)$, $x = K/F(T_i)$, 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. Once we have adjusted this hazard rate, we can easily obtain $p_l(t)$ by ensuring that

[13] The longer-term extrapolation is used only occasionally since we usually have data within six months of extrapolations of interest.

its weights have the same ratio as what we would have at the lowest TTM.[14] Then, we can obtain the time-dependent standard deviation of the l-th lognormal, $\Sigma_l(t)$, and the means of each lognormal as:

$$\Sigma_l(t) = \frac{\Sigma_l(T_1)t}{T_1} \tag{44}$$

$$\eta_l(t) = \eta_l(T_1)\sqrt{\frac{t}{T_1}} \tag{45}$$

Now, we have all of the necessary components to solve equation 38 (Black and Scholes, 1973).

Extrapolation between Time-to-Maturities Here, we need to extrapolate between available TTMs. First, we derive the dampened hazard rate using equation 42. The only difference is that we adjust K_{min} by defining it as follows:

$$K_{min} = \varphi(t)K^i_{min} + (1 - \varphi(t))K^{i+1}_{min} \tag{46}$$

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:

$$p_l(t) = (\frac{p_l(T_i+1)}{D(T_{i+1})}\frac{\sqrt{t}-\sqrt{T_i}}{\sqrt{T_{i+1}}-\sqrt{T_i}} + \frac{p_l(T_i)}{D(T_i)}\frac{\sqrt{T_{i+1}}-\sqrt{t}}{\sqrt{T_{i+1}}-\sqrt{T_i}})D(t) \tag{47}$$

$$\Sigma^2_l(t) = (1 - \varphi(t))\Sigma^2_l(T_{i+1}) + \varphi(t)\Sigma^2_l(T_i) \tag{48}$$

$$\eta^2_l(t) = (1 - \varphi(t))\eta^2_l(T_{i+1}) + \varphi(t)\eta^2_l(T_i) \tag{49}$$

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

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

time-dependent weights as:

$$p_l(t) = p_l(T_n)\frac{D(t)}{D(T_n)} \tag{50}$$

where T_n is the largest available datapoint with respect to TTM and recalling that we define the survival probability, $D(t)$, using equation 41. We then obtain the time-dependent volatility as:

$$\Sigma_l^2(t) = \Sigma_l^2(T_n)\frac{t}{T_n} \tag{51}$$

Finally, we need to derive the means as follows:

$$\eta_l(t) = \eta_l(T_n)\sqrt{\frac{t}{T_n}} \tag{52}$$

A.3 Implied Volatility Surface and Option Prices

Implied Volatility Surface Figure A.3 illustrates the skew of the extrapolated implied volatilities on 1 April 1996. 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 (although in this case it is almost a volatility smirk). The figure confirms that the extrapolated produced the desired characteristics.

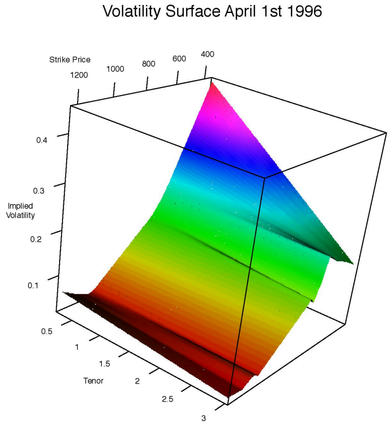

Figure 9: Implied Volatility Surface, 1 April 1996

Option Prices Once we have obtained a matrix with implied volatilities at the required strike prices (outlined in section A.1)[15] and TTMs (outlined in section A.2), we can proceed to obtain option prices by inputting the data in the Black-Scholes-Merton equation (Black and Scholes, 1973):

$$C(S_{0,p}, t) = N(d_1)S_{0,p} - N(d_2)Ke^{-r_f(T-t)} \tag{53}$$

where

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

where $N()$ is a value from the normal distribution. The above produces a matrix of call prices at our required strike prices and TTMs.

A.4 Aït-Sahalia and Lo Model Extrapolation

The method proposed by Aït-Sahalia and Lo (1998) is a non-parametric option pricing/volatility extrapolation methodology. For reference, the implied volatility extrapolation equation is the following:

$$\hat{\sigma}(F_t, K, \tau) = \frac{\sum_{i=1}^{n} k_F\big(\frac{F_t - F_{t_i}}{h_F}\big)k_K\big(\frac{K-K_i}{h_K}\big)k_\tau\big(\frac{\tau-\tau_i}{h_\tau}\big)\sigma_{IV,i}}{\sum_{i=1}^{n} k_F\big(\frac{F_t - F_{t_i}}{h_F}\big)k_K\big(\frac{K-K_i}{h_K}\big)k_\tau\big(\frac{\tau-\tau_i}{h_\tau}\big)} \tag{54}$$

where K is the strike price, τ 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 recommend that the interested reader refer to the original article (Aït-Sahalia and Lo, 1998).

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

B Appendix B

B.1 Time Series and Residual Plots for ASL Method (Aït-Sahalia and Lo, Univariate Markov Chain)

The following figures present the results for the extrapolation method of Aït-Sahalia and Lo. The same figures as in section 4.1.2 are presented: a time series graph, a regression residual graph, and an ACF/PACF graph. For the sake of brevity, I have opted to display and discuss the graphs for the entire sample. The Ross sample graphs are omitted because they show the same general patterns as for the entire sample.

Figure 10 shows a comparison between the forecasted return using the extrapolation method of Aït-Sahalia and Lo (red line) and the actual monthly return of the S&P 500 (blue line) for the entire sample (April 1996 to April 2015). These results do not include the multivariate Markov chain that this paper proposes. As discussed above (see table 8), the ASL methodology produces results that are closest to the original results of Ross. From figure 10, it is clear that, although the RT does not perfectly forecast the level of the underlying asset's return, it does manage to forecast the peaks and troughs quite accurately. Intuitively, this makes sense since we are using market prices at time t to forecast returns at time $t + 1$. Since there is a significant amount of noise between time t and time $t + 1$, we would expect some kind of over/underreaction to expected market movements. While general sentiment or perceived value seems fairly accurate using the RT, over/underreaction is to be expected.

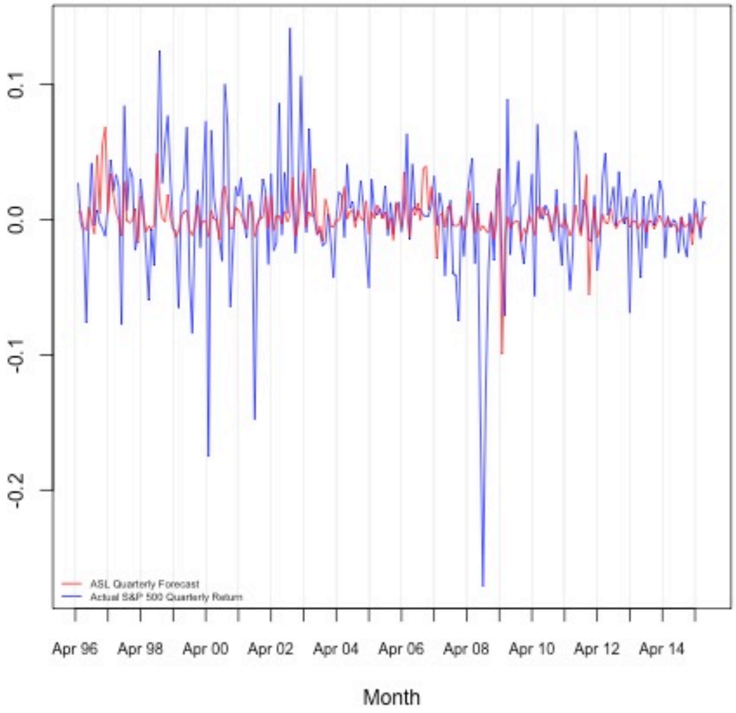

Figure 10: ASL RT Forecast vs Actual S&P 500 Return

In the residual plot (figure 11), a negative value means that the RT forecast was smaller than the actual return. We can see that the forecast residuals are clustered around the zero line as we would hope. Moreover, there does not appear to be a clear pattern, which is another desirable property of residual plots. One difference between this graph and the graphs of the results for the Sanford method is that the value of the residuals is smaller for the outlier residuals. Also, there are fewer outliers here compared to the Sanford results.

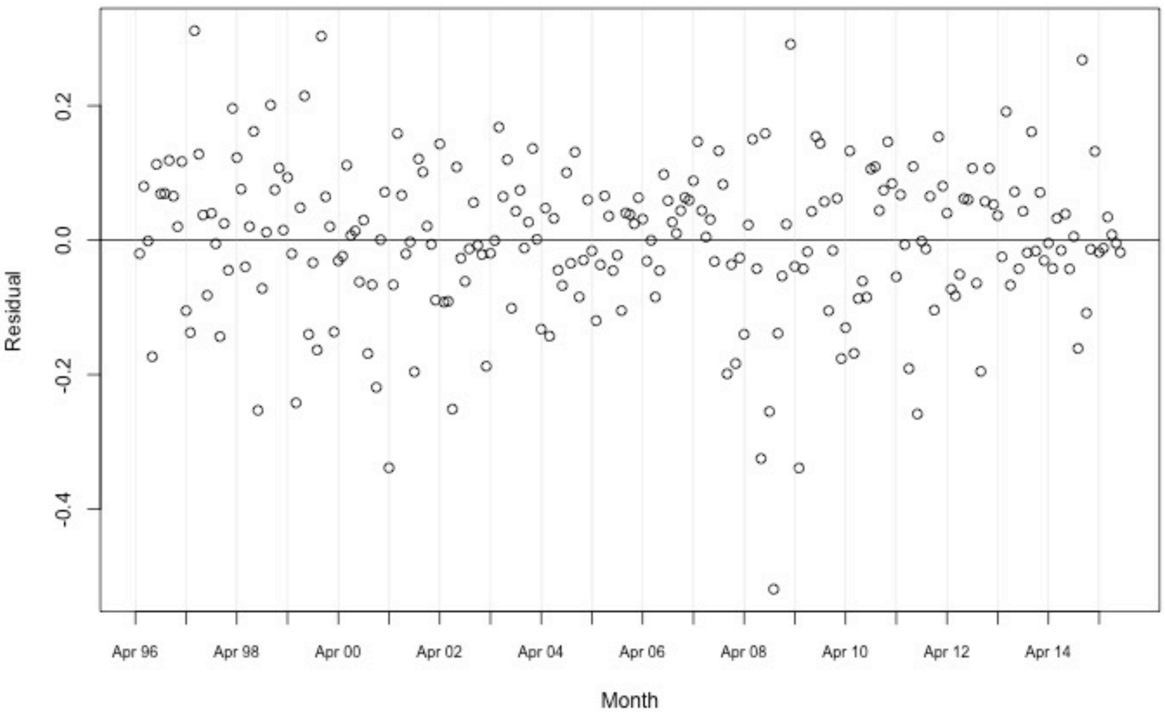

Figure 11: ASL RT Forecast vs Actual S&P 500 Return - Residual Plot

In figure 12, both the ACF and the PACF behave the way we would expect them to behave. They do not appear to exhibit any autocorrelation in the residuals which means that I have appropriately defined the forecasting model.

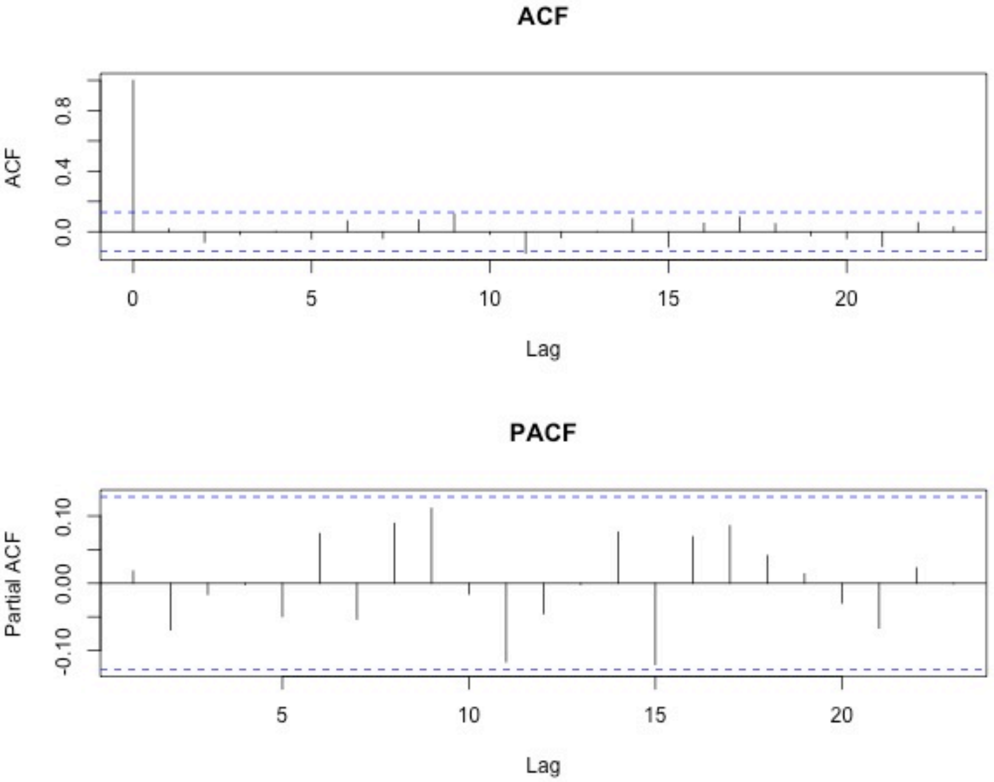

Figure 12: ASL vs S&P 500 Return Regression Residual ACF and PACF

B.2 Time Series and Residual Plots for ASLMV Method (Aït-Sahalia and Lo, Multivariate Markov Chain)

Figure 13 shows the results for the Aït-Sahalia and Lo extrapolation technique with the multivariate Markov chain. The figure illustrates that the multivariate component significantly improves the forecast. Yet, the results from the multivariate Markov chain seem more susceptible to exaggerations. This method appears more sensitive to the market's perceptions of large future changes, especially for negative events. This observation will become more apparent when I compare figure 14 and figure 11.

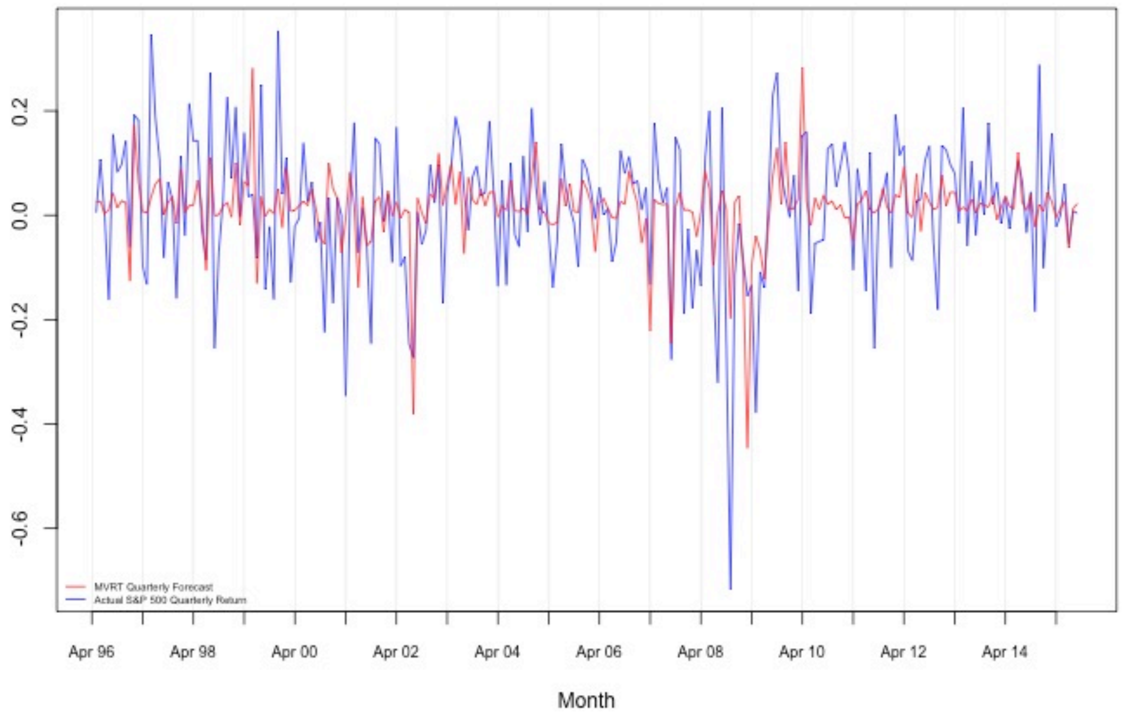

Figure 13: ASLMV Forecast vs Actual S&P 500 Return

The first, and probably most obvious, observation when we compare residual plots across the two Aït-Sahalia and Lo models (figures 14 and 11) is the fact that the the regression residuals are slightly more widely spread when the transition matrix is obtained using a multivariate Markov chain. This appears to be caused by the fact that the RT with a univariate Markov chain includes fewer fluctuations in the time series when compared to the multivariate Markov chain. In other words, the time series with the univariate chain is more of a straight line through zero rather than actually forecasting the levels of the S&P 500. As such, even if the residuals are smaller for the ASL method, its forecasting is actually inferior to the ASLMV method. Other than that, the residuals do appear to be behaving randomly as expected.

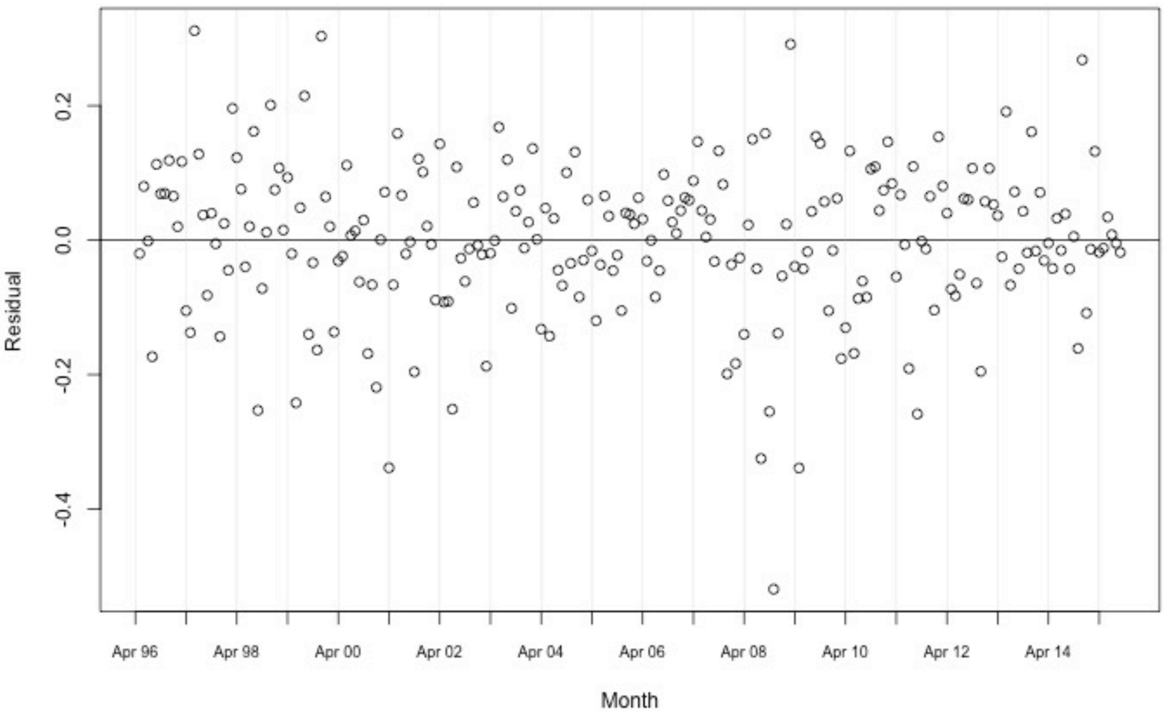

Figure 14: ASLMV Forecast vs Actual S&P 500 Return - Residual Plot

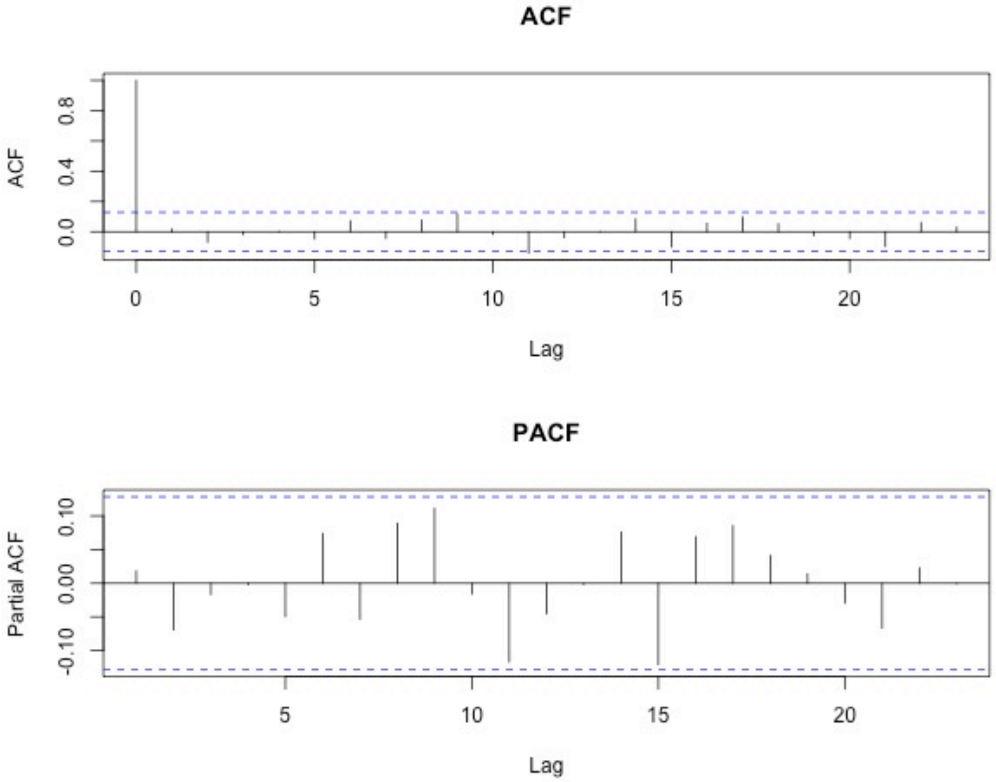

Figure 15: ASLMV vs S&P 500 Return Regression Residual ACF and PACF