High-Frequency Financial Volatility and the Pricing of Volatility Risk

by

Natalia Sizova

Department of Economics
Duke University

Date: _____

Approved:

_____
Tim Bollerslev, Chair

_____
George Tauchen

_____
A. Ronald Gallant

_____
Shakeeb Khan

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Economics
in the Graduate School of Duke University
2009

Abstract

High-Frequency Financial Volatility and the Pricing of Volatility Risk

by

Natalia Sizova

Department of Economics
Duke University

Date: _____

Approved:

_____
Tim Bollerslev, Chair

_____
George Tauchen

_____
A. Ronald Gallant

_____
Shakeeb Khan

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Economics in the Graduate School of Duke University
2009

Abstract

The idea that integrates parts of this dissertation is that high-frequency data allow for more precise and robust methods for forecasting financial volatility and elucidating the role of volatility in forming asset prices. Thus, the first two chapters compare the performance of model-free forecasts specifically designed to employ high-frequency data with the performance of "classical" forecasts developed for daily data. The final chapter of the dissertation incorporates high-frequency data to verify the predictions of asset pricing models about the risk-return relationships at the very shortest horizons. The results are arranged in the following order.

Chapter 1 presents the analytical comparison of feasible reduced-form forecasts designed to employ high-frequency data and model-based forecasts updated to use high-frequency data. The prediction errors of both forecast groups are calculated using the ESV-representation of Meddahi (2003), which allows one to generalize the statements from this analysis to a wider class of volatility processes. The results show that reduced-form forecasts outperform model-based forecasts at longer horizons and perform just as well for day-ahead forecasts.

Chapter 2 expands the conclusions from Chapter 1 to economic measures of forecast performance. These performance measures are constructed within a microeconomic framework that mimics the decision-making process of a variance trader who uses volatility forecasts to predict the future profitability of a trade. The results support the theoretical predictions of Chapter 1.

Chapter 3 is co-authored with Professor Tim Bollerslev and Professor George Tauchen. It extends the "long-run risk" model of Bansal and Yaron(2004) to consistently price volatility risks and to be applicable to high-frequency data. The hypothesis at the outset is that while financial volatility is a long-memory process (it exhibits long-range dependence), its own variance (volatility-of-volatility) is a short-memory one. Then the presented model implies that the volatility premium (the measure of the difference between option-implied and expected variances) should be short-memory as well. This insight is confirmed by studying cross-correlations of returns and volatility measures. Horizons at which cross-correlations are considered are unique for the literature; they start at intra-day values, as short as five minutes.

Contents

List of Tables

List of Figures

Acknowledgements

I am grateful to the members of my current and past dissertation committees: my advisor, Professor Tim Bollerslev, Professor George Tauchen, Professor Bjørn Eraker, Professor Shakeeb Khan, Professor A. Ronald Gallant, and Professor Craig Burnside for patient and thoughtful supervising. I am also thankful to Professor Tim Bollerslev for engaging me as a research assistant on the project that became a part of this dissertation.

I thank the participants of the Duke Econometrics and Finance Lunch Group and the UNC Kenan-Flagler Brown Bag Seminars for help in preparing my ideas for presentations at professional meetings.

In preparing my thesis I got invaluable advice regarding my writing style from Sarah Zubairy, Michael R. Dalton, and the managing editor of the journal "History of Political Economy", Paul Dudenhefer.

During my PhD studies, my research was funded by fellowships from the Duke Graduate School and a research assistantship from Professor Tim Bollerslev. Thanks!

1

Integrated Variance Forecasting: Model-Based versus Reduced-Form

1.1 Introduction

Traditionally, researchers who wanted to extract and forecast financial volatility had to rely on data recorded at only moderate intervals: daily, for instance, or even monthly. But recently, data at much more frequent intervals – high-frequency data – has become increasingly available.

The increasing availability of high-frequency data allows researchers to improve on the techniques used to forecast volatility. Two types of these techniques are model-based, which construct efficient volatility forecasts relying on the model for returns, and reduced-form, which construct simple projections of volatility on past volatility measures. Both types had been developed before the availability of high-frequency data. However, only model-based techniques were initially considered somewhat reliable, as they performed more accurately for daily data. For instance, two classes of volatility models – ARCH models [1] by [26] and stochastic volatility models [2] –

[1] ARCH and GARCH models are reviewed by [7]

[2] Reviewed by [96].

were conventionally preferred for extracting the volatility series. These model-based approaches gave more accurate estimates for latent volatility and provided better ways to form forecasts of the future volatility. Reduced-form techniques, on the other hand, relied on excessively noisy proxies of the volatility, e.g. daily squared returns.

Now that high-frequency data are available, is it still the case that model-based forecasts are better than reduced-form ones? Is it possible that when using high-frequency data, reduced-form forecasts are just as good as, or perhaps better than model-based forecasts? Employing high-frequency data increases the efficiency of extracting latent spot variances by model-based techniques; hence model-based techniques can perform better. On the other hand, for reduced-form techniques, high-frequency data allows researchers to define better proxies for the daily volatility. For example, one of non-parametric proxies – the realized variance – is a sum of squared returns. It is observable, and hence it admits a direct modeling, as in e.g. [40], and [11].

Apparently, the availability of high-frequency data starts a new chapter in the contest between reduced-form and model-based approaches as the most efficient for forecasting volatility. In this paper, we investigate this comparison for observed data and simulated data. Furthermore, we investigate this comparison analytically.

The object of interest is a forecast of the integrated variance. Integrated variance is a natural descriptor of the volatility of daily returns. It is an analog of the variance of a daily return in a discrete-time model, e.g. extracted variance in the GARCH model. Integrated variance can be used in VaR models of risk management, as an input to option pricing models and for variance hedging in trading; see [8].

The primary goal of the current study is to compare analytically feasible model-based and reduced-form forecasts of integrated variance. The defining word here is "feasible", i.e. the comparison is carried out under "realistic" conditions, assuming

that the true data generating process is unknown, and forecast inputs are estimated with errors.

In order to guarantee results that have the greatest generality, we formulate this comparison analytically within the framework of Meddahi's ESV-models. The path-breaking result of Meddahi is that any square-integrable variance process can be decomposed into the sum of simple processes. This decomposition allows us to write down many of our results in analytical form. Therefore, rather than resorting to time consuming simulations to compare forecasts, we can plug parameters into a formula and immediately evaluate the comparative performance.

The forecasts to be compared are briefly defined as follows. The first one is model-based. To implement this forecast, we model returns by a stochastic volatility (SV) model. E.g. daily data can be used to estimate the parameters of the model. Then, we use this model to predict the future integrated variance. For example, SV-models that can be used to form the forecast include affine models by [67] and [45], CEV-models, and exponential models as in [38].

The second forecast is reduced-form. In this case, the predictor of the future integrated variance is a linear function of its historic values. This form of the forecast is based on theory developed by [82].

The current study is logically connected to the paper by [14]. They compared the performance of "infeasible" model-based and reduced-form forecasts. Their interesting finding is that the reduced-form forecast performs remarkably well even though the "infeasible" model-based forecast minimizes forecast error by definition. In their study, the authors assumed that a true model is known, and all the inputs to the forecasts are observed. The major difference between their work and the current study is that we take things a step further and consider the case in which the model and parameters are unknown.

Another paper that relates to the current study is the paper by [21]. Among

other things, the authors looked at the problem of extracting the integrated variance using prior knowledge about a model. They compared this model-based estimate of the integrated variance to the reduced-form estimate – realized variance. They found that the results of the comparison are affected by the following parameters: sampling frequency and volatility-of-volatility.

Similar to the aforementioned work we show that the sampling frequency and the volatility-of-volatility parameters are major factors in comparison between model-based and reduced-form forecasts of integrated variance. Furthermore, we present the theoretical explanation why these parameters affect the comparison.

The paper is organized as follows. In Section 1.2, the model framework is set up and the forecasts to be compared are defined. In the same section, we will introduce definitions for error components, that will be used throughout the paper. Section 1.3 studies analytical comparison of feasible forecasts. In Section 1.4 the theoretical findings from the previous part will be evaluated for observed data. Section 1.5 concludes the paper.

1.2 One Model and Two Forecasts

1.2.1 Model

Throughout the paper we assume the following dynamics for the logarithm of prices s_t. Denote x_t as an n-dimensional vector of independent states. The $W_{i,t}$, $i = \overline{1,n}$, are independent standard Brownian motions. W_t^s is a standard Brownian motion which may correlate with $W_{i,t}, i = \overline{1,n}$, i.e. $\text{corr}(dW_t^s, dW_{i,t}) = \rho_i(x_t)$

The dynamics of states x_t and log-prices s_t are described by the following system of equations:

$$\begin{bmatrix} ds_t \\ dx_{1,t} \\ ... \\ dx_{n,t} \end{bmatrix} = \begin{bmatrix} \mu(x_t) \\ \kappa_1(x_{1,t}) \\ ... \\ \kappa_n(x_{n,t}) \end{bmatrix} dt + \begin{bmatrix} \sigma(x_t)dW_t^s \\ \Lambda_1(x_{1,t})dW_{1,t} \\ ... \\ \Lambda_n(x_{n,t})dW_{n,t} \end{bmatrix} \tag{1.1}$$

where $\sigma^2(x_t)$ is referred to as a spot variance. The functions $\Lambda_i(x_{i,t})$ stand for diffusion terms in the dynamics of the states x_t, and the functions $\kappa_i(x_{i,t})$ are drifts. For notational simplicity, denote $\sigma^2(x_t) \equiv \sigma_t^2$.

We are interested in forecasting the ex-post variability of returns as measured by the integrated variance:

$$IV_t^{t+T} \equiv \int_t^{t+T} \sigma_s^2 ds, \tag{1.2}$$

which is assumed to be well-defined. Intuitively, IV defines the variance of the T-period return, i.e. $\text{Var}_t(s_{t+T} - s_t)^2 = E_t(IV_t^{t+T})$, if the drift in prices $\mu_(x_t)$ is a predictable process of finite variation. In general, the same relation holds approximately, since the variation in the drift $\mu_(x_t)$ is negligible in comparison to the variation in the diffusion part $\sigma_t dW_t^s$. This implies that by forecasting the integrated variance, we aim to forecast the variability of the asset price over the next T periods. For example, variances that are inputs to Sharpe ratios of the assets can be taken from the values of IV_t^{t+T}.

Note that the system (2.27) has a very general form. Any known Markov continuous dynamics of stochastic volatility can be represented in this form, e.g. affine, GARCH-SV, and log-volatility models, to be defined more formally below.

For the purpose of later derivations, we will use another representation of the same system (2.27). This representation is referred to as the ESV-representation and

was introduced by [80]. He showed that any square integrable spot variance $\sigma^2(x_t)$ appearing in the SDE system (2.27) admits an Eigenfunction Stochastic Volatility (ESV) representation:

$$\sigma^2(x_t) = a_0 + \sum_{i=1}^{\infty} a_i P_i(x_t), \tag{1.3}$$

where processes $P_i(x_t)$ are called factors. These are square-integrable processes with the following properties:

(i) zero-mean: $EP_i(x_t) = 0$;

(ii) uncorrelated, with a unit variance: $\text{Cov}(P_i(x_t), P_j(x_t)) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$;

(iii) if discretely observed, each factor follows an AR(1) process:

$$E(P_i(x_{t+T})|x_\tau, \tau \leq t) = e^{-k_i T} P_i(x_t). \tag{1.4}$$

In the following discussion, a process will be referred to as a one-factor process if $a_i = 0, \forall i > 1$, and two-factor if $a_i = 0, \forall i > 2$. In general, for a p-factor process the variance can be represented by the sum, $\sigma^2(x_t) = a_0 + \sum_{i=0}^{p} a_i P_i(x_t)$. It should be emphasized that the above representation is equivalent to the representation (2.27), and was introduced only because it facilitates further analytical derivations.

At this point, we proceed to the definition of two types of forecasts under study. The first one will be called "model-based" or simply the "best" forecast. The second will be referred to as "model-free" or "reduced-form".

1.2.2 The Model-Based Forecast

In this section, we define the model-based forecast. Additionally, we show how the prediction error from this model-based forecast can be decomposed into three

components: "genuine" forecast error, error from estimates of the parameters and states, and error from model misspecification.

The model-based forecast is defined as the "best" forecast in terms of the mean squared error (MSE). That is, given the information set F_t at time t, the model-based forecast of the integrated variance IV_t^{t+1} minimizes the square-loss function:

$$IV_{t+1,t}^{model} = \arg \min_{IV_{t+1|t} \in F_t} E(IV_t^{t+1} - IV_{t+1|t}|F_t)^2. \tag{1.5}$$

Mean-squared error is easy to analyze and is the most popular loss-function in the literature. Despite a controversy surrounding the choice of the most appropriate performance measure (see [88, 89]), mean squared error remains a common reference point for a forecast performance.

The function that satisfies condition (1.5) is the expectation of IV_t^{t+1} conditional on information at time t, i.e. $E(IV_t^{t+1}|F_t)$. For the model described by system (2.27), let F_t be a σ-algebra generated by prices and states up to time t, i.e. $p_\tau, x_\tau, \forall \tau \leq t$. Therefore, the model implies that $IV_{t+1,t}^{model} = E(IV_t^{t+1}|p_\tau, x_\tau, \forall \tau \leq t)$. Moreover, due to the Markovian dynamics of log-prices and variances given by the system (2.27), the model-based forecast depends only on the latest realization of the state:

$$IV_{t+1|t}^{model} = E(IV_t^{t+1}|x_t). \tag{1.6}$$

The above forecast depends on the current state x_t and the model specification (2.27). The latter classifies it as "model-based".

The ESV-framework (1.3) yields a closed-form representation for the model-based forecast:

$$IV_{t+1|t}^{model} = \int_t^{t+1} [a_0 + \sum_{i=1}^{\infty} a_i e^{-k_i(s-t)} P_i(x_t)] ds = a_0 + \sum_{i=1}^{\infty} a_i \frac{1 - e^{-k_i}}{k_i} P_i(x_t). \tag{1.7}$$

As was anticipated in (1.6), the model-based forecast depends only on the latest state realization x_t.

However, the model-based forecast described by (1.7) is infeasible, since it is based on an unknown model, parameters and states. To define the feasible version of the same forecast, it is common to proceed in the following steps. In the first step, we choose a model and derive a corresponding formula for the model-based forecast. In the next step, we can estimate the parameters and unobserved states using a general method such as MLE, exact or an approximation. And finally, we can plug recovered states and parameters into the formula for the model-based forecast. This resulting structure is referred to as a feasible model-based forecast.

The feasible version of the model-based forecast based on the ESV-representation will take the form:

$$IV_{t+1|t}^{model} = \hat{a}_0 + \sum_{i=1}^{\infty} \hat{a}_i \frac{1 - e^{-\hat{k}_i}}{\hat{k}_i} P_i(\hat{x}_t). \tag{1.8}$$

The feasible model-based forecast given by (1.8) and the corresponding forecast error are objects of interest for the rest of this subsection.

The starting point is the ideal case, under which the exact model, parameters and latent states are known. In this case, the only error of the model-based forecast is the "genuine" forecast error:

$$\text{GFE}^{model} = E[IV_t^{t+1} - IV_{t+1|t}(\Theta, x_t)]^2, \tag{1.9}$$

where Θ are the parameters in the model, and x_t are the factors in the model. The "best" forecast minimizes the above value (1.9) by definition.

However, the total forecast error of the feasible model-based forecast involves two extra parts. The first comes from not knowing the states and coefficients. This part will be referred to as the error in forecast $IV_{t+1|t}$ due to errors in the parameters and

states:

$$F(\hat{x}_t - x_t, \hat{\Theta} - \Theta) = E[IV_{t+1|t}(\Theta, x_t) - IV_{t+1|t}(\hat{\Theta}, \hat{x}_t)]^2, \tag{1.10}$$

where $\hat{\Theta}$ are estimated parameters in the model, and $\hat{x}_t$ are estimated states in the model. Hence, the total error of predicting IV_t^{t+1} – the total mean-squared prediction error (MSPE) – has two components:

$$\text{Total MSPE} = E[IV_t^{t+1} - IV_{t+1|t}(\hat{\Theta}, \hat{x}_t)]^2 = \text{GFE}^{model} + F(\hat{x}_t - x_t, \hat{\Theta} - \Theta). \tag{1.11}$$

The two components of the error are uncorrelated, since the "genuine" error is unpredictable based on information at time t, i.e. $E[IV_t^{t+1} - IV_{t+1|t}|F_t] = 0$, and the error in parameters and states is a function of the information available at time t, i.e. $IV_{t+1|t}(\Theta, x_t) - IV_{t+1|t}(\hat{\Theta}, \hat{x}_t) \in F_t$. The total MSPE is thus simply a sum of two variances.

The second part of the error comes from model misspecification. Model misspecification is practically unavoidable, since the true model is generally not known for any observed data set. Hence, any chosen model is at best an approximation of the true one.

This second additional component of the error is important, since each step to form a feasible forecast incorporates the knowledge of the model. First, we use the model to define the functional form of the model-based forecast. Then we use the model to estimate the parameters and extract the spot values of states, e.g. by MLE and particle filters, EMM and the reprojection method, or by Bayesian methods. (See [71, 75, 95, 59, 90]) Hence, the model misspecification is a critical factor that affects all the steps above and can weaken the performance of the model-based forecast.

To summarize, although the model-based forecast minimizes the mean-squared error in (1.5), it does so only under the assumption that the true model with the parameters is known, and all the states are observable. However, it may not do so under a more realistic assumption that many inputs are to be estimated. In this

latter case, the total error consists of several parts and the "feasible" model-based forecast as defined by (1.8) may not perform the best under the mean-squared error loss.

We are going to investigate when the above factors may outweigh the advantages of model-based forecasting. As a natural competing forecast, we use a model-free reduced-form forecast, as originally advocated by [11] and formally analyzed by [82].

1.2.3 The Reduced-Form Forecast

In this section, we define a benchmark reduced-form forecast. This forecast can be derived starting from formula (1.7). The formula for the model-based forecast (1.7) reports the conditional expectation for the integrated variance IV_t^{t+1} based on the ESV-representation. The right-hand side of this formula is a sum of p autoregressive processes. This property has an important implication; it implies that for a p-factor ESV model, the integrated variance IV_t^{t+1} is a sum of p AR(1) processes and a white noise term:

$$IV_{t+1,t} = a_0 + \sum_{i=1}^{p} a_i \frac{1 - e^{-k_i}}{k_i} P_i(x_t) + \varepsilon_{t+1},$$

where $E(\varepsilon_{t+1}|F_t) = 0$. This decomposition suggests that the integrated variance is an ARMA(p,p)- process.

[82] derives the coefficients of ARMA models for IV_t^{t+1} in the case of one-factor and two-factor models. In general, ARMA(p,p) for the integrated variance will take the form:

$$\prod_{i=1}^{p}(1 - e^{-k_i}L)(IV_t^{t+1} - \theta) = \eta_{t+1} - \sum_{i=1}^{p} \beta_i \eta_{t+1-i}, \tag{1.12}$$

where η_t is heteroscedastic white noise, and $k_1, ...k_p$ are mean-reversions in the ESV-representation (1.4).

We can find the parameters of the ARMA-model (1.12) if we know the parameters of the base model (2.27). However, we may also estimate the same parameters from a reduced-form model. That is, since IV_t^{t+1} can be described as a reduced-form ARMA process, we may simply fit a linear time-series model to IV_t^{t+1}. The resulting coefficient estimates will be "model-free". Therefore, the forecast based on this reduced-form time-series model yields a model-free IV_t^{t+1} predictor based on the past realizations of the integrated variance.

Definition: The reduced-form forecast of the integrated variance IV_t^{t+1} is a linear projection of IV_t^{t+1} onto the space generated by its past realizations $IV_\tau^{\tau+1}$, $\tau \leq t$:

$$IV_{t+1|t}^{rf} = P(IV_t^{t+1}|IV_\tau^{\tau+1}, \tau \leq t).$$

The above definition describes an infeasible version of the reduced-form forecast. This forecast is not feasible, since the integrated variance is not observed. However, we may construct a feasible version of the same forecast, using a close proxy of the integrated variance – realized variance:

Definition Suppose log-prices s_t are observed at discrete times 0, h, 2h, etc. Then the realized variance over the period $[t, t+1]$ is defined as

$$RV_t^{t+1} \equiv \sum_{i=1}^{1/h} (s_{t+ih} - s_{t+(i-1)h})^2. \tag{1.13}$$

The realized variance is a directly observable measure of the intra-day variance of the price. The asymptotic behavior of RV and its consistency as a proxy for IV is discussed by [**?**, 22].

To construct a feasible version of the reduced-form forecast, we project RV_t^{t+1} on its past values $\hat{P}(RV_t^{t+1}|RV_\tau^{\tau+1}, \tau \leq t)$. This projection can be constructed absolutely model-free using only the data that are observable. For zero drifts in returns, this forecast will be equivalent to the projection of IV_t^{t+1} on the past values of RV, i.e. $\hat{IV}_{t+1|t}^{rf} = \hat{P}(IV_t^{t+1}|RV_\tau^{\tau+1}, \tau \leq t)$. This follows from the fact that if there is no drift in returns, then the difference $RV_t^{t+1} - IV_t^{t+1}$ is unpredictable. (See [21].) The same equivalence holds approximately for intra-day data, since the drift in asset prices is negligibly small between finely sampled observations.

The ARMA-representation for RV follows from the ARMA-representation of integrated variance(see [82]):

$$\prod_{i=1}^{p}(1 - e^{-k_i}L)(RV_t^{t+1} - \theta) = \eta_{t+1}(h) - \sum_{i=1}^{p}\beta_i(h)\eta_{t+1-i}(h), \tag{1.14}$$

where $\eta_t(h)$ is heteroscedastic white noise. In contrast to the ARMA representation of the integrated variance, the ARMA representation for RV depends on the distance between observations h. Coefficients for the above ARMA-representation are derived by [82] for one-factor and two-factor models with no leverage, i.e. when the contemporaneous shocks to the factors x_t and to the return are uncorrelated: $\rho_i(x_t) = 0, \forall i = \overline{1,n}$ in (2.27) .

If the parameters of the ARMA-representation (1.14) are known, then the "genuine" error of the reduced-form forecast is the difference between the variance of the

shock η_{t+1} in (1.14) and the variance of the noise term $RV_t^{t+1} - IV_t^{t+1}$:

$$\begin{aligned} \text{GFE}^{rf} &= E\left[IV_t^{t+1} - P(IV_t^{t+1}|RV_\tau^{\tau+1}, \tau \leq t)\right]^2 = \\ &= E\left[RV_t^{t+1} - P(RV_t^{t+1}|RV_\tau^{\tau+1}, \tau \leq t) - (RV_t^{t+1} - IV_t^{t+1})\right]^2 = \\ &= E\left[RV_t^{t+1} - P(RV_t^{t+1}|RV_\tau^{\tau+1}, \tau \leq t)\right]^2 - E\left[RV_t^{t+1} - IV_t^{t+1}\right]^2. \end{aligned} \tag{1.15}$$

The decomposition above follows from the fact that the difference between the realized and integrated variance is unpredictable.

If the parameters of the ARMA-representation are unknown, then the total Mean Squared Prediction Error of the reduced-form forecast includes an additional part:

$$\text{Total MSPE}^{rf} = \text{GFE}^{rf} + F^{rf}(\hat{\Theta}^{rf} - \Theta), \tag{1.16}$$

where the second part $F^{rf}(\hat{\Theta}^{rf} - \Theta)$ comes from the errors in the coefficients and Θ^{rf} are the parameters in the regression of RV_t^{t+1} on its past values. Note that the covariance term is absent from the above decomposition. Since the infeasible error is orthogonal to the linear space spanned by $RV_\tau^{\tau+1}$, $\tau \leq t$, therefore it is orthogonal to the error from parameter estimation. When expressing the reduced-form forecast as

$$\hat{IV}_t^{t+1rf} = \hat{\theta} + \sum_{i=1}^{\infty} \hat{\alpha}_i \left(RV_{t-i}^{t-i+1} - \theta\right), \tag{1.17}$$

the error term due to parameter estimation error uncertainty takes the form

$$F^{rf}(\hat{\Theta}^{rf} - \Theta) = E\left[\hat{\theta} + \sum_{i=1}^{\infty} \hat{\alpha}_i \left(RV_{t-i}^{t-i+1} - \hat{\theta}\right) - \theta - \sum_{i=1}^{\infty} \alpha_i \left(RV_{t-i}^{t-i+1} - \theta\right)\right]^2. \tag{1.18}$$

It is worth noting that the total Mean Squared Prediction Error of the feasible reduced-form forecast does not include errors from estimating the instantaneous unobservable states x_t.

1.3 Analytical Comparison of Feasible Forecasts

In this section we present a simple framework that allows us to compare feasible model-based and reduced-form forecasts analytically. Special attention will be given to the effects of errors in the state estimates $\hat{x}_t$ and model misspecification.

Throughout this section, we will assume that the data is generated from a multi-factor model. Formally, this implies that for the ESV representation (1.3, 1.4) $a_2 \neq 0$. This assumption was tested for financial time-series and the hypothesis of a one-factor model ($H_0 : a_2 = 0$) was consistently rejected. For example, [38] reject the one-factor hypothesis for stock-index data using χ^2-statistics within the EMM estimation. Also, [30] also reject the one-factor hypothesis within the GMM estimation that matches conditional first and second moments of the realized variance for foreign exchange rates.

The success of multi-factor models is explained by their dual implications; they can simultaneously generate a large variability in the variance and a high persistence over long horizons. This combination of properties is an attribute of asset price series. Since any multi-factor process is a mixture of at least two components, one of these components adds to the variability of the variance, and the other accounts for the high persistence. Using the ESV notation (1.3), this implies that for a two-factor model with $a_1 \neq 0$, $a_2 \neq 0$ and $a_i = 0, i > 2$, the mean-reversion of one factor k_1 is significantly higher than $k_2 \approx 0$. One example of such a specification is the model by [69], which mimics the S&P 500 index behavior. In their model, one volatility component has mean-reversion with a half-life of 1/2 days and the other component has mean-reversion with a half-life of two years.

Nonetheless, in the literature, models with only one component are often used instead of multi-component models . For instance, the models that are listed in Table A.1 are often chosen due to the simplicity of their handling and estimation.

Therefore, although the true generating process may be governed by several components, econometricians often assume simple one-component processes. In general, there are few research papers that focus on models with two components and even fewer that consider models with three components. (See [21].) We may expect that the true number of components is not limited to two or even three for real data, implying that the number of components is underestimated in many applications. Hence, the effect of the underestimation of the number of factors(components) can be a common source of forecasting error.

In this section, we consider a simple case in which an econometrician assumes a one-factor model of the general form:

$$d\sigma_t^2 = k(\theta - \sigma_t^2)dt + \Lambda(\sigma_t)dW_t. \tag{1.19}$$

In the above model, the only factor is the spot variance. Therefore, we can make a link between the model (1.19) and the ESV-representation (1.3) by defining the ESV-parameters $a_0 = \theta, \quad a_1 = \sqrt{\text{Var}\sigma_t^2}, \quad P_1(x_t) = \frac{\sigma_t^2 - a_0}{a_1}$, and $k_1 = k$.

Hence, the model-based forecast can be derived as a special case of the general formula (1.7):

$$E_t IV_t^{t+1} = \theta + \frac{1 - e^{-k}}{k}(\sigma_t^2 - \theta). \tag{1.20}$$

The model-based forecast is linear in the last realized spot variance and the slope is a function of the mean-reversion coefficient k.

Formula (1.20) defines an infeasible form of the model-based forecast. To implement this forecast in practice the following problems are to be solved: parameter estimation and estimation of the latent spot variances σ_t^2.

The first problem – estimation of the parameters – will not be fully addressed in this paper. We assume that a large span of data is available, thus bringing errors in the parameters to zero. In the following analysis, we are concerned only with

the errors that arise from the finite number of observations per unit of time (infill asymptotic).

Nevertheless, not knowing the parameters still poses a problem for the model-based approach, since if a true model were known, then regardless of the estimation procedure, the estimates would converge to their true unique values. However, in our case the model is misspecified. Consequently the limits of the estimated parameters may depend heavily on the estimation procedure. For example, for a dummy-model:

$$y_t = \sigma\varepsilon_t,\ \varepsilon_t \propto N(1,1) \tag{1.21}$$

MLE-estimate of the parameter σ converges to the root of the following equation:

$$1 = E\left[\left(\frac{y_t}{\sigma} - 1\right)\frac{y_t}{\sigma}\right]. \tag{1.22}$$

On the other hand, a GMM-estimate based on the second moment converges to the variance Var y_t. Unless the above model is correct, those two different moment conditions will generally yield different parameters.

Therefore, the estimation procedure for the model parameters also plays an important role in the feasible model-based forecast. Our first choice is a procedure that leaves forecast (1.20) unbiased irrespective of the model. This is achieved by a GMM procedure that matches the expectation

$$E(IV_t^{t+1} - \theta) = 0, \tag{1.23}$$

and the regression slope:

$$\frac{\mathrm{cov}(IV_t^{t+1}, \sigma_t^2)}{\mathrm{Var}\sigma_t^2} = \frac{1 - e^{-k}}{k}. \tag{1.24}$$

The case when parameters are defined by the conditions above will be further referred to hereafter as the "no-bias" case. However, instead of the above moments, other

moment conditions can be used to define k and θ. In general, estimation procedures based on those other moments may introduce a forecast bias. Here, we derive the contribution of this bias to the total MSPE.

To keep the algebra simple, we will assume that for all estimation procedures θ matches the unconditional mean of the integrated variance IV_t^{t+1} as in (1.23). In this case, irrespective of the choice for the other GMM moments, the "genuine" forecast error will be a sum of two parts

$$E\left[IV_t^{t+1} - \theta - \frac{1-e^{-k}}{k}(\sigma_t^2 - \theta)\right]^2 =$$

$$= E\left[IV_t^{t+1} - P_t(IV)\right]^2 + \mathrm{Var}\sigma_t^2\left[\frac{\mathrm{cov}(IV_t^{t+1}, \sigma_t^2)}{\mathrm{Var}\sigma_t^2} - \frac{1-e^{-k}}{k}\right]^2,$$

where the first term includes a linear projection of the integrated variance on the most recent spot variance:

$$P_t(IV_t^{t+1}) = E\sigma_t^2 + \frac{\mathrm{cov}(IV_t^{t+1}, \sigma_t^2)}{\mathrm{Var}\sigma_t^2}\left[\sigma_t^2 - E\sigma_t^2\right] \tag{1.25}$$

Therefore, once we allow for model misspecification, the "genuine" forecast error includes two terms. The first term is a forecast error from the linear forecast based on the last observed spot. The second term is the "bias". This term is absent if the forecast is based on the parameters defined by (1.24), i.e. in the "no-bias" case.

The second problem of implementing the model-based forecast is filtering the state σ_t^2. As a result of this next step, another error will be added to the total prediction error. As we plug the estimate $\hat{\sigma}_t^2$ into formula (1.20), we get the following decomposition of the error for the "no-bias" case:

$$IV_t^{t+1} - \theta - \frac{1-e^{-k}}{k}(\hat{\sigma}_t^2 - \theta) =$$

$$= \left[IV_t^{t+1} - P_t IV_t^{t+1}\right] + \left[\frac{\mathrm{cov}(IV_t^{t+1}, \sigma_t^2)}{\mathrm{Var}\sigma_t^2}(\sigma_t^2 - \hat{\sigma}_t^2)\right] \tag{1.26}$$

The first part of the above expression is the infeasible forecast error of the "last-spot" forecast, i.e. the linear forecast based on σ_t^2. The second part is due to the error in the spot variance.

For consistent estimates of the spot variance $\hat{\sigma}_t^2$, the error from the second part in (1.26) converges to zero as the sampling frequency increases to infinity, e.g. [2]. However, at very high frequencies, microstructure effects may blur the results. To avoid such microstructure effects the minimum distance between observations is often not lower than 5 minutes for financial time-series studies. Therefore, despite the asymptotic negligibility of the error in spot variance, it remains a nontrivial part of the total error in (1.26) under typical conditions. The error in the spot variance will also depend on the particular filtering technique used to extract the spot variance.

There are several ways to extract the spot variance σ_t^2 from past prices s_{t-i}, $i \geq 1$. In this paper, we will focus on the efficient ARCH-filters of [86]. ARCH-filters give consistent estimates of the variance under very general conditions, most importantly even for misspecified models. (See [85].) For example, for GARCH-SV models with $\Lambda(\sigma_t^2) = \eta\sigma_t^2$ in (1.19), the efficient ARCH-filter takes the form of a discrete-time GARCH(1,1):

$$\hat{\sigma}_{t+h}^2 = \phi_h + a_h\hat{\sigma}_t^2 + b_h\xi_{t+h} \tag{1.27}$$

where $\xi_{t+h} = \frac{s_{t+h}-s_t-\mu h}{\sqrt{h}}$ is a normalized innovation in prices. The parameters of the filter ϕ_h, a_h, and b_h are chosen optimally based on an estimated model. In particular, they will depend on the estimates of the persistence k and the volatility-of-volatility $\lambda = \frac{Var\sigma^2}{E\sigma^4}$.

In the Appendix, we prove the following statement:

Proposition 1. *Let σ_t^2 be square integrable with the correlation function: $corr(\sigma_{t+h}^2, \sigma_t^2) = \frac{\sum_{i=1}^p a_i^2 e^{-k_i h}}{\sum_{i=1}^p a_i^2}$. Suppose we apply the efficient ARCH-filter of the form (1.27) to extract*

the spot variance. The log price process is described by (2.27) with no leverage effect and zero drift. Then the comparison of the reduced-form forecast and the forecast based on a one-factor model depends only on the following set of parameters Ξ*:*

- *Volatility-of-volatility –* $\lambda = \frac{Var\sigma_t^2}{E\sigma_t^4}$*;*
- *Relative weights of factors:* $\frac{a_i}{a_1}, i = \overline{2,p}$*;*
- *Persistence of factors:* $k_i, i = \overline{1,p}$*;*
- *Sampling frequency –* h*.*

Notably, the assumption about the correlation structure of the variance process is very general, since ESV-representation for any square-integrable process satisfies this assumption. Hence, the comparison of the forecasts is simplified within the ESV-framework.

All the steps of the analytical comparison are given and proved in the Appendix. In the main body of the paper we present the following corollaries of the proof. First, we will derive asymptotic errors for $h \approx 0$, i.e. arbitrarily high sampling frequencies. In particular, we will characterize the effect of decreasing the sampling frequency on the model-based and the reduced-form forecasts. Second, we will quantify the effect of the misspecification on the model-based forecast. Finally, we will continue the GARCH-SV example from [14]. In particular, we will replicate the comparison of the infeasible forecasts and update it by comparison of the feasible forecasts.

1.3.1 The Effect of Sampling Frequency

To form the feasible version of the model-based forecast, the state σ_t^2 is to be estimated. In this subsection, we show how the errors in the estimates of σ_t^2 depend on the sampling frequency. Furthermore, we summarize the effects of sampling fre-

quency on model-based and reduced-form forecasts for $h \approx 0$, i.e infinitely active sampling.

The asymptotic behavior of an MLE-estimated spot variance is discussed by [20], and [62, 63]. In this paper we derive the exact formula for the error in the estimate $\hat{\sigma}_t^2$ extracted by the ARCH-filter (1.27). The formula is given by (A.22) in the Appendix and can be further simplified through approximation around $h \approx 0$. For simplicity, let's assume a "no-bias" case, i.e. the parameter $\hat{k}$ converges to $\tilde{k}$, which satisfies the moment condition(1.24). Then, from (A.22) it follows that the error in the estimate $\hat{\sigma}_t^2$ equals approximately:

$$E(\hat{\sigma}_t^2 - \sigma_t^2|k)^2 \approx hVar\sigma^2 \frac{\bar{k}}{b_h} + b_h E\sigma^4 \tag{1.28}$$

$$\bar{k} = \frac{\sum_{i=1}^{p} a_i^2 k_i}{\sum_{i=1}^{p} a_i^2} \tag{1.29}$$

The formula above clarifies the trade-off in estimating $\hat{\sigma}_t^2$. The choice is between using as much data as possible to make the estimator more efficient, or using an estimation window as narrow as possible to reduce the bias. Both characteristics are functions of the same parameter b_h and the error is minimized by the following value [3] of the GARCH-parameter $b_h = \sqrt{\lambda \bar{k} h}$. However, if one assumes a one-factor model and estimates the persistence parameter k from (1.24), then the chosen GARCH-parameter is:

$$b_h = \sqrt{\lambda \tilde{k} h}, \tag{1.30}$$

where $\tilde{k}$ satisfies the moment condition in(1.24). The resulting error in the variance estimate equals

$$E(\hat{\sigma}_t^2 - \sigma_t^2)^2 \approx \mathrm{Var}\sigma_t^2 \sqrt{\frac{h}{\lambda}} \left(\frac{\bar{k}}{\sqrt{\tilde{k}}} + \sqrt{\tilde{k}}\right). \tag{1.31}$$

[3] The parameter of the efficient GARCH-filter is derived by [86] for a one-factor GARCH-SV process. Here we extend the result by deriving the parameters of the efficient GARCH-filter for an arbitrary ESV-model with no leverage.

The total MSPE of the model-based forecast will include the part proportional to the above term and also the term that is proportional to the covariance between the error $\sigma_t^2 - \hat{\sigma}_t^2$ and the "genuine" forecast error. (See (A.20) in the Appendix.) However, the latter term is of the order $O(h)$. Therefore, approximation of the total MSPE around $h \approx 0$ will contain the "genuine" forecast error, which includes the error from model misspecification, plus the part coming from the term in(1.31):

$$\frac{\text{Total MSPE}}{\text{Var}\sigma_t^2} =$$

$$= 2\overline{\left[\frac{1}{k} - \frac{1-e^{-k}}{k^2}\right]} - \overline{\left[\frac{1-e^{-k}}{k}\right]}^2 + \sqrt{\frac{h}{\lambda}}\left(\frac{\bar{k}}{\sqrt{\tilde{k}}} + \sqrt{\tilde{k}}\right)\left[\frac{1-e^{-\tilde{k}}}{\tilde{k}}\right]^2 + O(h).$$

The analogous Taylor decomposition of the reduced-form forecast error given by equation (A.27) in the Appendix yields:

$$\frac{\text{Total MSPE}}{\text{Var}\sigma_t^2} = \frac{\text{Var}[IV_t^{t+1} - P(IV_t^{t+1}|IV_{t-\tau}^{t+1-\tau}, \tau = 1, 2, ...)]}{\text{Var}\sigma_t^2} + O(h). \qquad (1.32)$$

The $O(h)$ term appears in the above equation, since to predict IV_t^{t+1}, we use past realizations of RV instead of the latent IV-series.

It is worth noting that the two errors above have different convergence rates with respect to the distance between observations h; the reduced-form forecast is of a stochastic order $O(h)$, while the model-based forecast of a stochastic order $O(\sqrt{h})$. Hence, we can conclude that

Corollary 1. *The decrease in sampling frequency has a higher negative effect on the performance of the model-based forecasts, than on the performance of the reduced-form forecasts.*

Another interesting observation is that the increase in the volatility-of-volatility λ affects the forecast comparison in favor of the model-based forecast. This can

be explained by the following consideration. Volatility series are filtered from noisy squared returns. The level of noise in the squared return is proportional to $E\sigma^4 = \text{Var}\sigma + E^2\sigma^2$. Hence, a reduction in the mean $E\sigma$ would reduce the noise in the returns while keeping the same variability of the spot variance. Therefore, an increase in the volatility-of-volatility makes square returns better proxies for the spot variance.

Corollary 2. *As the sampling frequency decreases, the comparative performance of the model-based forecasts deteriorates less for higher levels of volatility-of-volatility* $\lambda = \frac{Var_\sigma}{E\sigma^4}$.

1.3.2 The Effect of Model Misspecification

In the above discussion, model misspecification resulted from application of one-factor models to the modeling of multi-factor processes. Consider the case of a two-factor process. Within the ESV-framework the true variance process includes two components with different mean-reversions:

$$\sigma_t^2 = a_0 + a_1 P_1(x_t) + a_2 P_2(x_t),$$

where $a_1 > 0$ and $a_2 > 0$. However, the econometrician wrongly assumes that either $a_1 = 0$ or $a_2 = 0$. Therefore, the measure of the model misspecification is the ratio $\ln(\frac{a_1}{a_2})$. This ratio is equal to $\pm\infty$ when the model is truly one-factor, and thus the contribution of one of the factors is zero. At the other extreme, this ratio is zero when both factors are of equal importance, i.e. $a_1 = a_2$.

In this subsection, we investigate the effect of the model misspecification as measured by $\ln(\frac{a_1}{a_2})$ on the comparison between the model-based and reduced-form forecasts. The comparison is carried out for different values of the other parameters that influence the outcome: mean-reversions k_1 and k_2, volatility-of-volatility $\lambda = \frac{\text{Var}_{\sigma_t^2}}{E\sigma_t^4}$ and the sampling frequency.

First, we will fix the estimates for k_1 and k_2 from studies that fitted two-factor models to financial series. These studies are summarized in Table A.2 and include four examples: three papers estimated coefficients of multi-factor models for foreign exchange rates and one dealt with stock indices. We derived the parameters of interest – mean-reversions k_i, model misspecification $\ln a_1/a_2$, and volatility-of-volatility λ – from the parameters reported in those papers.

For each pair of k_1 and k_2 from Table A.2 , Figure A.1 shows the regions in the space $\left(\ln \frac{a_1}{a_2}, \lambda\right)$ where the reduced-form forecast outperforms the model-based forecast, i.e. total MSPE is lower for the reduced-form forecast. The regions corresponding to different sampling frequencies are depicted in different intensities of grey. Small square marks inside the graph indicate representative parameter values taken from the studies in Table A.2. For example, for the model of [3] the coefficients are $k_1 \approx 0.9$, $k_2 \approx 0.02$, $\lambda \approx 0.6$, and $\ln(a_1/a_2) \approx -1$. From the bottom left panel in Figure A.1, we see that the intersection of 0.6 for ordinates and -1.0 for abscissas is dark-grey, which corresponds to 15-minute sampling frequencies. This implies that the model-based forecast renders smaller errors for 5-minute sampling, but cedes efficiency to the reduced-form forecast for 15-minute and 30-minute frequencies.

Figure A.1 illustrates the statements from the previous sections. First, the graph shows that the decrease in sampling frequency adversely affects the model-based forecast, making it less appealing in comparison to the simpler alternative. As the distance between observations increases from 5 to 30 minutes, a larger area of parameter sets yields MSPE(reduced-form) $<$ MSPE(model-based).

Second, the graph confirms that the effect of finite sampling on the performance of the model-based forecast is lower for high volatility-of-volatility λ. For instance, consider the bottom-left panel with $k_1 = 0.9$ and $k_2 = 0.02$ and choose the level of misspecification to be $\ln(a_1/a_2) = -1$. For $\lambda = 0.1$, the model-based forecast results

in higher errors for all the chosen frequencies. For $\lambda = 0.4$, the model-based forecast is the most efficient for 5-minute but not for 15-minute and 30-minute samplings. Finally, for $\lambda = 0.8$, the model-based forecast dominates for the highest frequencies (5-minute and 15-minute) and retreats only for the 30-minute sampling.

Third, a higher persistence favors the reduced-form forecast. For each of the four panels, it holds that k_1 is higher than k_2. The "average" mean-reversion of the variance is equal to $a_1^2(a_1^2 + a_2^2)^{-1}k_1 + a_2^2(a_1^2 + a_2^2)^{-1}k_2$, implying that the mean-reversion of the variance is increasing in $\ln \frac{a_1}{a_2}$. Figure A.1 shows that the areas with a better performance of the reduced-form forecast are located to the left from the symmetric case $\ln \frac{a_1}{a_2} = 0$. Thus, other factors being equal, a higher persistence of the volatility process gives an advantage to the reduced-form forecast.

Finally, Figure A.1 demonstrates how the model-misspecification affects the forecast comparison. Model-misspecification is measured by the ratio $|\ln \frac{a_1}{a_2}| \in [0, +\infty)$, with no-misspecification cases located at $\ln \frac{a_1}{a_2} = \pm\infty$, Notably, the areas where the reduced-form forecast is better (shown in grey) are centered at $|\ln \frac{a_1}{a_2}| \approx 0$ for three of the four graphs, implying that the model misspecification acts in favor of the reduced-form forecast.

It can be formally shown that the "genuine" error in the model-based forecast (A.13) is quadratic [4] in the ratio $\frac{a_1^2}{a_2^2}$ and achieves the global maximum at a finite value of $|\ln \frac{a_1}{a_2}|$. That is, the predictive power of the model-based forecast is at its minimum when the spot variance is comprised of two components: slow-moving and fast-moving, and the contributions of each of these components are non-trivial, i.e. in the case of misspecification.

From the results of the last two subsections for day-ahead forecasts, it follows that the reduced-form forecast performs very close to the model-based forecast due

[4] For comparison, the "genuine" error is linear in $\frac{a_1^2}{a_2^2}$ for a correctly specified two-factor model(see (A.14)), and achieves maximum for $\ln \frac{a_1^2}{a_2^2} = -\infty$.

to two effects. First, the efficiency of the "best forecast" (model-based) fades away for finite sampling frequencies. The second effect comes from model misspecification, which explains why the reduced-form forecast eventually outperforms the model-based forecast for certain parameter values.(See Figure A.1.) This is a general result that holds for all data-generating processes described by (2.27).

1.3.3 Numerical Example

In this section we continue a numerical example from [14]. The example compares the model-based and reduced-form forecasts for the one-factor GARCH-SV process calibrated to daily series of the spot DM/USD exchange rate from 1987 to 1992. This exercise assumes that states are observable and the GARCH-SV is the true model for the data. We extend this example, first by relaxing the assumption that the states are observable, and second by relaxing the assumption that the true model is one-factor. Instead, we assume that the true model is two-factor, with parameters calibrated to the same DM/USD series.

Therefore, we proceed with the same assumption from the previous subsection: an econometrician employs a one-factor model of the type (1.19). Specifically, following [14], he estimates the parameters to be $k = 0.035$, $\theta = 0.636$ and $\Lambda(\sigma_t) = \sigma\sigma_t^2$, where σ is chosen so that the volatility-of-volatility is $\lambda = 0.296$. However, the true data-generating process is two-factor and described by either of the following two candidates.

The first candidate model is estimated by [30] for high-frequency 5-minute spot returns on DM/USD from 1986 until 1996. They suggest the following two-factor

affine model:

$$\begin{aligned} ds_t &= \sqrt{x_{1t} + x_{2t}} dW_t^s, \\ dx_{1t} &= k_1(\theta_1 - x_{1t})dt + \sigma_1\sqrt{x_{1t}} dW_{1t}, \\ dx_{2t} &= k_2(\theta_2 - x_{2t})dt + \sigma_2\sqrt{x_{2t}} dW_{2t}, \\ \text{corr}(dW_t^s, dW_{1t}) = 0,\ &\text{corr}(dW_t^s, dW_{2t}) = 0, \end{aligned} \tag{1.33}$$

where $\theta_1 = 0.3257$ and $\theta_2 = 0.1786$ are factors' means, $k_1 = 0.5708$ and $k_2 = 0.0757$ are mean-reversions, and $\sigma_1 = 0.2286$ and $\sigma_2 = 0.1096$ define the volatility-of-volatility.

The second candidate is estimated by [21]. They fitted their CEV-SV model to 5-minute return data on the DM/USD series during the period 1986-1996. The resulting volatility dynamics is represented by the two-factor model:

$$\sigma_t^2 = x_{1t} + x_{2t}, \tag{1.34}$$

$$Ex_{it} = 0.509\ w_i, \tag{1.35}$$

$$\text{Var}x_{it} = 0.461\ w_i, \tag{1.36}$$

where $w_2 = 0.212$, $w_1 = 0.788$, and the corresponding mean-reversions for the factors are $k_1 = 3.74$ and $k_2 = 0.0429$. The resulting volatility-of-volatility λ and the ratio $\ln \frac{a_1}{a_2}$ for both of the models are reported in Table A.2.

Summarizing, we have three different scenarios. Under the first scenario, the returns follow the one-factor GARCH-SV model. This case corresponds to the first row in Table A.3. The second row in Table A.3 is for the version where the true model is two-factor, as suggested by [30]. Finally, the third row of the table corresponds to the case when the true model is from [21]. The data reported in Table A.3 are the performances of the forecast based on the one-factor GARCH-SV model (the first two columns), and the reduced-form forecast (the last two columns).

The table presents the forecast comparison from two perspectives. On the one hand, it illustrates the effect of misspecification. In particular, for the first row the model is specified correctly, but in the second and third rows the model is misspecified.

On the other hand, Table A.3 uncovers the error-in-latent-states effect. In particular, the first and the third columns correspond to the case when the spot variance is observable, while the second and the fourth columns are for the case when the spot variance is latent. The results in the table assume a 5-minute distance between observations for "feasible" forecasts and are calculated using formulas derived in the Appendix (A.28 - A.30).

Table A.3 demonstrates the following results. First, the model-based forecast is the most efficient if the model is correct and the variance is observed. This is in accordance with the definition of the model-based forecast. In this case, the MSPE is a mere 2.3%, while the reduced-form forecast yields an error of 4.3%. Second, when the model is correct but the variance is unobserved, the quality of the model-based forecast deteriorates but this approach remains the most efficient with an MSPE of 6.2%.

Third, when the model is wrong, the performance of the model-based forecast is affected much more strongly than in the case of unobserved variance. For example for the 2F-SR-SV(II) model, instead of an MSPE of 2.3%(if the model were correct), the model-based forecast delivers an MSPE of 67.2%. However, the model-based forecast keeps the leading position, as the reduced-form forecast gives 68.6%. Finally, the combination of two effects – misspecification in the model and unobserved variance – drives the performance of the model-based forecast below the reduced-form forecast. Despite its simplicity, the reduced-form forecast gives 68.8% MSPE versus 115.6% MSPE of the model-based forecast.

In the case of the 2F-SR-SV(I) model, the results are qualitatively similar, and

the reduced-form forecast renders smaller errors, although the difference in errors is less dramatic: 35.4% error in the reduced-form forecast versus 39.3% error in the model-based forecast.

The most remarkable conclusion from Table A.3 is that irrespective of the case we considered, Mincer-Zarnowitz $R^2 s$ [5] of the feasible model-based forecast and the feasible reduced-form forecast are generally close. This is an evidence that the day-ahead reduced-form forecast is successful in capturing the same information that is available to the model-based forecast. Moreover, in terms of the MSPE, the reduced-form forecast can be significantly more accurate, since it is unbiased by construction.

To summarize, when comparing the model-based forecast and reduced-form forecast, both the possible model misspecification and the finite frequency of price observations should be taken into account. The combination of these factors causes a different ranking of the forecasts. This example serves as an illustration of how the comparative ranking of the model-based and the reduced-form forecasts can switch under realistic assumptions.

1.3.4 Multi-Period Forecasts

It can be of separate interest to investigate how model-based and model-free approaches perform for longer-horizons. Multi-factor models may exhibit long-memory-like properties, but one-factor models cannot. Therefore, a different behavior can be expected from one-factor and multi-factor models for longer-horizon predictions.

[5] To construct R^2, data on realizations of IV_t^{t+1} are regressed on the forecasts $IV_{t+1|t}$. The R^2-statistic of this regression is a measure of the forecast efficiency. There is a link between R^2 and MSPE, which is a main performance measure in this paper:

$$\frac{\text{MSPE}}{\text{Var}IV_t^{t+1}} = 1 - R^2 + \frac{[P(IV_t^{t+1}|IV_{t+1|t}) - IV_{t+1|t}]^2}{\text{Var}IV_t^{t+1}},$$

where $P(IV_t^{t+1}|IV_{t+1|t})$ is the projection of IV_t^{t+1} on its forecast $IV_{t+1|t}$. It follows that R^2 is a valid measure of the forecast performance that does not, however, take into account the bias part: $P(IV_t^{t+1}|IV_{t+1|t}) - IV_{t+1|t}$.

In this section, we consider the forecasting of the integrated variance IV_t^{t+T} if $T = 5$, i.e. weekly forecasts. A multi-period model-based forecast for the ESV-models is a generalization of the formula (1.7):

$$E_t IV_t^{t+T} = a_0 + \sum_{i=1}^{p} a_i \frac{1 - e^{-k_i T}}{k_i} P_i(x_t). \tag{1.37}$$

For example, for a one-factor model, the multi-period forecast takes the form:

$$E_t IV_t^{t+T} = \theta + \frac{1 - e^{-kT}}{k}(\sigma_t^2 - \theta). \tag{1.38}$$

The formulas above imply that as the horizon T increases, the only error (whose effect may be amplified) is the error from misspecification. Indeed, the error from the estimates of latent states remains constant, since the input $\hat{x}_t$ is the same for all T. Moreover, as $T \to \infty$, its contribution to the MSPE fades away. On the other hand, as T increases the true forecast converges to

$$\lim_{T\to\infty} E_t IV_t^{t+T} = a_0 + \sum_{i=1}^{p} \frac{a_i}{k_i} P_i(x_t). \tag{1.39}$$

Therefore, the factors with the smallest mean-reversions, k_i, will dominate the forecast. If the number of factors is underestimated, the components that play a dominant role for weekly, monthly, and longer-horizon forecasts will not be properly extracted. For example, for one-factor models, they are hidden within the estimate of the spot variance $\hat{\sigma}_t^2$.

We extend the analysis from Section 1.3.2 to find the parameters for which the multi-period reduced-form forecast outperforms the model-based forecast. Algebraic computations in the general case are given in the Appendix, and the results for $T = 5$ are presented in Figure A.2. As in the one-period case, the parameter sets for which the reduced-form forecast performs the best are colored in grey(light grey, dark grey and black), starting from black for 5-minute frequencies.

Quite remarkably, all of the actual model estimates from Table A.2, indicated by squares, now fall into this dark-shaded area, suggesting that for multi-period forecasts the reduced-form forecast is more efficient even for the finest frequencies. It follows that for all these models, the underestimation of the number of factors weakens the power of the model to predict for longer horizons, while the performance of the reduced-form forecast is affected much less.

1.4 Empirical Example

We have demonstrated that the efficiency of the model-based forecast dissipates, if we consider its feasible version. This was established theoretically for the models that are calibrated to observed data. In this section we demonstrate the same effect with actual DM/USD exchange rate 5-minute returns from December 2, 1986, through June 30, 1999. The chosen market is open twenty-four hours a day, yielding a total of 288 observations per day. This data has been extensively studied before, and for a thorough description, one can refer to [6]. One of the features of the DM/USD data set is the presence of strong intra-day patterns in volatility. (See [5].) The same paper shows that the intra-day seasonal component in DM/USD volatilities may be described as a function of only the time of day and does not depend on the level of volatility. In this study, we use raw data to form RV series. For filtration of states in the model-based forecasts, we adjust the intra-day returns by the seasonal component [6] :

$$s^*_{t+ih} - s^*_{t+(i-1)h} = \frac{s_{t+ih} - s_{t+(i-1)h} - \mu h}{\hat{f}(i,h)} + \mu h,$$

$$\hat{f}(i,h)^2 = \frac{\sum_{t=1}^{N}(s_{t+ih} - s_{t+(i-1)h} - \mu h)^2}{\sum_{t=1}^{N}\sum_{i=1}^{1/h}(s_{t+ih} - s_{t+(i-1)h} - \mu h)^2}.$$

We assess the performance of two model-based forecasts. The first one is con-

[6] We also verified that the results of this section do not change when using raw data.

structed using a one-factor model. The other is based on a two-factor model. The goal of this section is to compare these forecasts to the reduced-form forecast. The object of the forecasting in all these cases will be the realized variance, RV.

1.4.1 One-Factor Model-Based Forecasts

For the simplest case – a one-factor model – the log-price is assumed to follow the dynamics

$$\begin{aligned} ds_t &= \mu dt + \sigma_t dW_t^s, \\ d\sigma_t^2 &= k(\theta - \sigma_t^2)dt + \Lambda(\sigma_t)dW_t, \\ \text{cov}(W_t^s, W_t) &= 0. \end{aligned} \tag{1.40}$$

In the above system, the volatility-of-volatility $\Lambda(.)$ is left unspecified. This generality serves two related purposes. First, if the data are indeed generated from a one-factor model, then regardless of $\Lambda(.)$, there is no model misspecification. Second, the only type of misspecification possible in this setup is the underestimation of the number of factors.

In order to estimate the parameters in (1.40) we rely on the method-of-moments that is independent of the specification for $\Lambda(.)$. Specifically, we match the mean of the T-period realized variance $\theta = \frac{1}{T} E RV_t^{t+T}$, the correlation structure of the realized variance $e^{-kT} = \text{corr}\left(RV_{t-T}^t, RV_t^{t+T}\right) * *(\text{corr}\left(RV_{t-T}^t, RV_{t+T}^{t+2T}\right))^{-1}$, and the variance of the realized variance to get the volatility-of-volatility $\lambda = \text{Var}\sigma_t^2 / E\sigma_t^4$ from $\text{Var}(RV_t^{t+T})$ in (A.3). To avoid any biases in favor of the reduced-form forecast, we report parameters and the forecast performance for different horizons T. Specifically, we will work with parameters calibrated to daily ($T = 1$), weekly ($T = 5$), monthly($T = 20$), and quarterly data ($T = 60$).

Table A.4 illustrates that different choices of T do indeed yield quite different sets of parameters. For instance, for daily observations $k \approx 0.25$ in daily units, which

implies a half-life of less than 3 days. On the other hand, for quarterly correlations with $k \approx 0.018$ the half-life of a shock is almost 2 months. This dependence of the parameters on T is a special case of a more general result, that parameter estimates for a misspecified model depend on the estimation technique. If the underlying model were truly one-factor, then the estimated profiles $k(T)$ and $\lambda(T)$ would be flat. However, for the observed data the degree of mean-reversion $k(T)$ is decreasing in T. This pattern is easy to replicate by estimating a one-factor model on the data simulated from a multi-factor model.

For the financial volatility, the mean-reversion parameter k is typically within the $0.01 - 0.03$ range in daily units. (See e.g. [4, 50].) As follows from Table A.4, the mean-reversion corresponding to the quarterly data is within this range.

After estimation, we can proceed to the next step involving filtering of the latent state, i.e. the spot variance. Just as in the analytical section, we rely on the efficient ARCH-filter of Nelson given by (1.27) with the optimal choice of filter parameters given by (A.23),(A.24) in the Appendix. The resulting forecasts are then given by:

$$\widehat{IV}_t^{t+1} = \hat{\theta} + \frac{1 - e^{-\hat{k}}}{\hat{k}}(\hat{\sigma}_t^2 - \hat{\theta}). \tag{1.41}$$

The performance of this forecasting procedure is reported in the first half of Table A.5, which contains Mincer-Zarnowitz R^2s and normalized MSPEs for different parameter sets in Table A.4. The forecast performance was evaluated using the whole data sample excluding the first year, i.e. for the period of 1987 - 2007. The last line of the table corresponds to the performance of the reduced-form forecast that is a simple AR-model for RV, with the number of lags chosen by the BIC-criterion. The reduced-form forecast is purely out-of-sample, as the number of lags and values of parameters are calculated using only the data available at the time of the forecast. Just as for the model-based forecast, the performance was evaluated using the whole

data set excluding the first year.

The Mincer-Zarnowitz R^2 of the model-based forecast is 35.6% for parameters calibrated to quarterly data. This result is close to the corresponding statistic for the reduced-form forecast (35.7%). Moreover, the difference is slightly in favor of the latter. For the other parameter values, the model-based forecast performs noticeably worse. In terms of the MSPE, the reduced-form forecast has a clear advantage with the error of 64.6% versus 76.5%.

1.4.2 Two-Factor Model-Based Forecast

The sensitivity of the estimates for k to T also indirectly suggests that the one-factor model is misspecified. A more advanced way to model the variance dynamic is to consider two-factor models. This class of models can capture short-run and long-run dynamics in volatility. The two-factor model considered here is the affine model, which has been already described in Section 1.3.3. The parameters of this model were estimated by [30] with the same data set on the DM/USD exchange rates, using first and second conditional moments of the realized variance within the GMM framework.

In contrast to the one-factor case, there are two latent states (x_{1t}, x_{2t}) in this model, the first being the fast mean-reverting volatility component, and the second being the slow-moving volatility component. To extract these components, we choose among the methods available for estimation of these two latent series. The most popular and efficient methods are particle filtering and Bayesian methods. In contrast to them, the Kalman filter provides a simple linear updating scheme that does not employ any information about the volatility-of-volatility term. If the functional form for the volatility-of-volatility is known (e.g. in the model (1.33) it is $\sigma_1\sqrt{x_{it}}$ for a factor $x_{it}, i = \overline{1,2}$), we are better off using the particle filtering. Nevertheless, if the model is misspecified, then any advantage of using the particle filter can be

annihilated and the Kalman-filter may produce better proxies for latent states. To avoid a bias in favor of the reduced-form forecast, we apply both methods – the Kalman-filter and the particle-filter – to extract latent states following [21], and [46], respectively.[7]

The second half of Table A.5 reports the forecasting performance of the two-factor model. The table shows that in terms of R^2 the model-based forecast with particle filtering appears to be the best choice, explaining 37.0% of the variance of RV. However, the reduced-form forecast is very close, with an R^2 of 35.7%. Moreover, in terms of the mean-squared error, the reduced-form forecast is better than the model (64.6% vs. 69.8%). This difference in rankings is a matter of how R^2 and MSPE account for forecast biases. The Mincer-Zarnowitz R^2 depends only on the variance of the forecast error and not on its mean, i.e. the conditional bias is irrelevant. On the other hand, MSPE includes both the variance of the error and its expected mean.

The model-based forecast with Kalman filtering performs slightly worse in terms of the R^2 (35.6%) than both the reduced-form forecast and the forecast with the particle filtering. But in terms of the MSPE, it is the best option out of the model-based forecasts (66.2 %). In order to explore the reason why the particle filter and the Kalman filter give rise to different MSPEs, we examined the dynamics of the forecasts for both of these types. We found that particle filtering delivers visibly over-smoothed predictors.

The conclusion from Table A.5 is that, despite a better fitting of the data, the two-factor model produces forecasts that do not significantly outperform the one-factor model in terms of R^2. This could be due to the fact that the large part of variation in the next-day realized variance is explained by the short-run components, and therefore one-factor models are sufficient to capture the dynamics of the variance

[7] In this study, the states x_{1t}, x_{2t} are extracted using 5-minute, 15-minute and hourly returns. Here we report results only for 15-minute frequencies, since they led to the smallest forecast errors.

over short horizons. As was indicated in the previous section and as also follows from Table A.5, the reduced-form forecast succeeds in capturing the same information, and therefore renders the close R^2 result. Moreover, owing to its unbiasedness, it gives the lowest MSPE (64.6%).

1.4.3 Multi-Period Forecasts

In Section 1.3, the analytical results were extended to the comparison of multi-period forecasts. In particular, we demonstrated that multi-period forecasts put more weight on the factors with higher persistence, thus increasing the role of the model misspecification in the forecast comparison. This change in importance between the factors explains the results highlighted by Table A.6. In this table we continue the empirical example from the previous subsection and study the performance of one-factor models, the two-factor model, and the reduced-form forecast for the realized variance of the DM/USD spot rate. We focus on longer forecast horizons: weekly ($T = 5$) and monthly ($T = 22$).

Table A.6 reports the Mincer-Zarnowitz R^2 and the total MSPE. The one-factor forecast in the table is "quarterly"-calibrated, i.e. k equals 0.018. This parametrization proved to be the most successful for day-ahead forecasting. However, for longer horizons, its performance gradually decays, falling to an R^2 of 27.7%.

In contrast to the one-factor models, the drop in the quality of the reduced-form forecast and of the two-factor model is less dramatic. The R^2 for the two-factor model remains at the levels of 31.6% (particle filtering) and 31.1%, (Kalman filtering). The reduced-form forecast achieves an R^2 of 32.3% for the month-ahead forecast.

For the two-factor model (1.33), the multi-period forecast (1.37) can be represented in the form:

$$E_t IV_t^{t+T} = \sum_{i=1}^{2} \left[\theta_i + \frac{1 - e^{-k_i T}}{k_i} (x_{it} - \theta_i) \right]. \tag{1.42}$$

Like the day-ahead forecasts, its implementation requires time-consuming estimation of the parameters and states $x_{1,t}, x_{2,t}$. In return, it performs only on a par with the reduced-form forecast, which is much simpler to implement. Moreover, its MSPE of 71.1%/82.2% is objectively higher than that of the reduced-form forecast, which is 68.1%. Thus, based on the results in this section, we find that even more sophisticated two-factor models, which are better at explaining the data, yield higher forecasting errors than the reduced-form forecasts.

1.5 Conclusion

In this paper, we compare the performances of model-based and reduced-form forecasts of integrated variance assuming availability of intra-day data. We show that when it comes to the feasible versions of the forecasts, reduced-form forecasts can outperform model-based forecasts. Since the model-based forecast requires the knowledge of both the true model and the latent instantaneous volatility, model misspecification and the errors in instantaneous volatility estimates can in effect combine to make the model-based perform worse than the reduced-form forecast. We also confirmed these results with actual high-frequency foreign exchange rates and several popular stochastic volatility models.

This paper challenges the conventional wisdom that models always render the most efficient forecasts, and simpler approaches are certainly destined to fail in comparison with this benchmark. That is we challenge the belief that, though estimation and forecasting within SV-models are notoriously time-consuming, the resulting gain in efficiency justifies it. On the contrary, based on the analytical and empirical evidence in this study, we conclude that reduced-form forecasts perform similar to supposedly efficient forecasts and are often better. And though our analysis has been limited to particular assumptions, the latter are still quite general to foresee that our results may carry over to more complex settings.

2

Variance Forecast Performance Measures: An Economic Approach

2.1 Introduction

Statistical forecast performance measures such as mean-squared error (MSE), mean-absolute error (MAE), median-absolute error, and similar statistics are sometimes found to be unsatisfactory from an economic point of view. For instance, [77] found that MSE and MAE are uncorrelated with profit-losses derived from forecasting of interest rates. [53]found inconsistencies between statistical measures and profits for evaluation of money supply forecasts. [100], and also [52] give further evidence that the MSE-criterion is not the best measure of forecast performance.

Statistical criteria also exhibit certain features, such as symmetry, that do not pertain to the forecasts reported by forecasting agencies. For example, [89]found that the economy growth forecasts reported by the Federal Reserve Board are downward biased, i.e. the FRB employs an asymmetric criterion.

As alternative to the statistical criteria, forecast performance can be measured by utility-based loss-functions, which naturally follow from applications of forecasts and

are consistent with internal preferences of forecasters. For example, the loss-function based on utility of the portfolio manager who uses variance forecasts for mean-variance optimization is suitable to measure the accuracy of variance-covariance matrices; see [55, 56], [100]. Another utility-based approach is designed to measure the accuracy of forecasts for asset prices. It is based on profits that are derived from trading these assets; see [77], [65], [76], and [74]. Under this type of "profit" loss-function, the success of a forecast depends not only on the forecast and the actual value, but also on a third variable – market price. From no-arbitrage considerations a market price represents a forecast itself. Therefore, "profit" loss-functions favor forecasts that can outperform this "aggregate" forecast, i.e. that can correctly identify mispricing in the market. However, statistical measures of forecast performance do not account for this third variable. They consider the forecast and actual value in isolation from other external factors.

We extend the latter type of utility-based performance measure to the case of univariate volatility forecasting. We develop a performance measure that tracks the mispricing of the variance in the market. To define the market price for the variance, we need to find securities that enable betting on volatility, e.g. call and put options. Future volatility plays so important a role in option trading defining the price and volumes of option markets, that the phrase "to buy and sell volatility" has become a standard way to refer to buying and selling options. The information about variance that is entering option prices can be synthesized to create a pure tool to bet on volatility – variance swap. It can be expected that sensible forecast methods should benefit variance swap trading strategies. Thus, a loss-function based on the utility of a variance swap trader who employs a particular forecast can be a valid measure of forecast comparison.

In this study we compare two broad groups of forecasts of variance that we will refer to as model-based and reduced-form respectively. The first group includes

generally sophisticated methods that form " efficient" variance forecasts based on fully specified models for returns, e.g. ARCH and SV models as reviewed by [7] and [96]. Forecasts in the second group are based on linear regressions for observable variance proxies, such as HAR models by [40] and ARFIMA models by [11]. Despite their simplicity, the forecasts in the latter group perform very often on par and sometimes better than the best forecasts in the first group; see [14], [94]. Our goal is to determine if the same finding holds for utility-based performance measures.

We show that the forecasts from both groups provide economic value for risk-averse traders. In terms of compensations, the advantage from using a forecast is worth a fee up to 5% - 10% of the contract price relative to the maximum utility from the strategies that do not rely on forecasts. We demonstrate that to be able to compete with reduced-form forecasts, a model-based forecast has to employ high-frequency data and exhibit a sufficiently rich structure. Such a forecast and a simple linear reduced-form forecast are close in performance, the first being marginally better for low-uncertainty periods and the latter being slightly better for high-uncertainty periods. In terms of compensations, less risk-averse traders are willing to pay no more than 1 % of the contract price to switch from a model-based forecast to a reduced-form forecast, and more risk-averse traders are willing to pay no more than 0.5% of the contract price to make the opposite switch.

The paper is organized as follows. Section 2.2 defines the variance swap. In Section 2.3, we propose a model of a variance swap trader. In Section 2.4, we define loss functions based on variance swap trading. In Section 2.5, we introduce the forecasts to be compared and describe estimation methodologies. In Section 2.6, we compare those forecasts for the observed data on S&P 500 futures and VIX data. Section 2.7 replicates the comparison using simulated data, and section 2.8 summarizes the main findings.

2.2 Variance Swap as a Benchmark for Forecasts

The loss function to be introduced in this paper compares the ability of variance forecasts to predict that the variance will be larger or lower than the level implied by derivative prices. For example, this level can be set by prices of variance swaps, the derivatives that summarize the variance information contained in options. Variance swap is a forward contract that allows to hedge against future volatility. It has two "legs". One "leg" of the swap will pay the realized variance over a specified period $[t, t+H]$, say RV_t^{t+H}. The other "leg" of the swap will pay a fixed amount, the strike price P_t. All the cash flows are exchanged on the maturity of the contract at $t+H$. At this time the buyer of the swap receives the difference $C_{t+H} = RV_t^{t+H} - P_t$. Therefore, if initially the buyer was exposed to the volatility in the underlying security, then after buying such a contract he reduces his exposure. His losses are bounded by the price P_t, since RV_t^{t+H} is positive.

The value for the realized variance is defined in the following way. Suppose that the log-price of the underlying security s_t (e.g. stock index) is observed at discrete times $t, t+h, t+2h, ..., t+H$. Then, the realized variance over the period $[t, t+H]$ is defined as the sum of squared returns:

$$RV_t^{t+H} = \sum_{j=1}^{H/h} \left(s_{t+jh} - s_{t+jh-h}\right)^2 . \tag{2.1}$$

The realized variance, as defined above, is a proxy for the integrated variance that is a measure of the return variability; see [6], and [21]. Therefore, the realized variance (RV) itself is a valid measure of the financial volatility, and the variance swaps are pure tools to bet on the volatility. For instance, if one expects variance to soar, he may buy a variance swap to profit from this expectation. [1]

[1] Variance betting is not the only reason for trading variance swaps. Besides speculation, they

In the variance swap contract there are two participating counterparts: variance seller and variance buyer. In the following, we will assume that variance buyer is a large investor with a diversified portfolio who for a given price P_t has a fixed demand K for variance swaps. In contrast, variance seller is a trader who does not hold a diversified portfolio. In the next section we build a model for a variance swap seller who facing the order of size K evaluates the chances that the realized variance will be sufficiently lower than the price P_t, thus giving him a premium for undertaking the risk.

2.3 Model of Variance Seller

The variance seller caters to large clients who are willing to hedge their investment portfolios against rough movements of the stock market by buying inelastically K variance swap contracts with a maturity H. Therefore, the seller shorts variance and his profit-losses are equal to $K(P_t(H) - RV_t^{t+H})$, where $P_t(H)$ is the price of the contract. For providing a sell-side for this transaction, the trader certainly requires a premium, since he takes on the risk of unbounded losses. The seller can refrain from trading if he thinks that the offered premium is not high enough to compensate for the risk associated with the trade.

The minimum price that is required by the trader depends on his own forecast of the future variance. Overestimation of the future variance results in lost trading opportunities, and underestimation results in losses. Therefore, a more accurate forecast enables a more accurate decision rule for wether trade or not to trade. Consequently, for each forecast there is a corresponding utility of a trader who uses this forecast to form his trading strategy.

In this section, we build a model of a variance seller with a given utility over his

can be used to hedge against strong downward movements in the market. This hedge is possible due to the robust correlation between directional and volatility risks; see [29], and [42] for a general equilibrium explanation of this phenomenon.

cash flows. Based on this model, a loss-function will be defined. The model will be presented in the following steps. First, we will define the trader's cash flows. Second, we will find the optimal trading rule for an arbitrary utility. Third, we will specify the family of utility functions. Finally, we will specify the exogenous market price of the variance swap.

Let us start by defining the trader's cash flows. If the trader agrees to sell the variance at the price $P_t(H)$, then after H periods, he will receive the difference between this price and the realized variance $K(P_t(H) - RV_t^{t+H})$. If he refuses to participate in this contract, he will be idle for this period and his cash flows from variance swap selling on the date $t+H$ will be zero. Therefore, his action space is an indicator of an undertaken trade $I_t \in \{0, 1\}$. The corresponding cash flows at time $t + H$ are equal to

$$C_{t+H} = \left[P_t(H) - RV_t^{t+H}\right] K I_t. \tag{2.2}$$

Cash flows of the trader are positive for low realized variance and negative for high realized variance. This variability induces a risk of large losses. Therefore, in general a trader will require a risk premium to trade, i.e. the trader will sell the contract only if the price $P_t(H)$ is high enough to remunerate the trader for the risk. We can give analytical content to this statement if we specify the utility of the trader over his cash-flows.

Suppose, the trader is myopic and his objective function is given by the risk-adjustment operator $V_t(C_{t+H})$. For a given objective, the optimal decision of the trader to trade or not to trade is defined through the optimization:

$$\begin{aligned} I_t^* &= \arg \max_{I_t \in \{0,1\}} V_t(C_{t+H}) \\ s.t.\ C_{t+H} &= \left[P_t(H) - RV_t^{t+H}\right] K I_t. \end{aligned} \tag{2.3}$$

We assume the strict monotonicity of the operator $V_t(.)$. If $V_t(.)$ is strictly increasing, then there exists a reservation value $\bar{P}_t(H)$ such that the trader trades if and only if

$P_t(H) > \bar{P}_t(H)$, i.e.

$$I_t^* = I(P_t(H) > \bar{P}_t(H)), \tag{2.4}$$

and this value can be found from the rationality constraint:

$$V_t(0) = V_t(K[R_t(H) - RV_t^{t+H}]). \tag{2.5}$$

Hence, in this model, a trader will not swap the variance if and only if his internal valuation of the variance is higher than the one implied by the market's P_t. This can happen for the following three reasons. First, his risk-aversion can be higher than the current risk aversion of the market participants on average (the "aggregate investor"). Second, the inelastic transaction size K(notional) can be higher than typically traded on the market, making the trader more cautious about undertaking the risk. Third, he may employ a forecasting method that gives him an advantage in comparison to the forecasting methods used by the market participants, which allows him to avoid unprofitable trades.

2.3.1 Utility Specification

Prior to specifying the class of utilities, we will define the properties that are required from this class given the objectives of this paper. First, our focus is only on the prediction of the realized variance level. Because most of the literature deals with the prediction of RV, we believe considering only the RV forecasts is the most relevant first step. [2] Considering that our focus in on the prediction of the realized variance level, the current study is only concerned only with utilities that are linear in cash-flows.

Second, as will be shown in the results section, only a sufficiently risk-averse trader enjoys an advantage from employing statistical forecasts. For risk-neutral traders,

[2] A natural extension could be forecasts for logarithms of RV or for realized volatilities $\sqrt{RV_t^{t+H}}$, that are also evaluated in the econometric literature. However, generalization to arbitrary functions of RV is not undertaken in this paper.

uninterrupted trading ($I_t = 1, \forall t$) will appear to be optimal. Therefore, the chosen utility class should exhibit risk-aversion. Third, given the above constraints we will consider a general class of utilities to demonstrate that our results are preserved under various specifications of the trader's risk-aversion.

For example, the quadratic utility [3] that incorporates a negative attitude towards uncertainty may be suitable for our purposes. The corresponding objective function depends on two conditional moments:

$$V_t(C_{t+H}) = E_t\left(C_{t+H}\right) - \frac{\alpha}{2}\mathrm{Var}_t\left(C_{t+H}\right), \tag{2.6}$$

with the quadratic utility itself to be ex-post evaluated in the following way:

$$u^q(C_{t+H}) = C_{t+H} - \frac{\alpha}{2}\left(C_{t+H} - E_t\left(C_{t+H}\right)\right)^2. \tag{2.7}$$

Later on we will assume that conditional expectations are not observed. Therefore, we will modify the above ex-post utility by substituting in an arbitrary estimate of the future cash-flows $\tilde{E}_t(C_{t+H})$:

$$u^q(C_{t+H}) = C_{t+H} - \frac{\alpha}{2}(C_{t+H} - \tilde{E}_t(C_{t+H}))^2 \tag{2.8}$$

Furthermore, to focus only on prediction of the RV level, the above utility will be modified to substitute the quadratic term by an estimate $\widetilde{\mathrm{Var}}_t(C_{t+H}) = \tilde{E}_t\left(C_{t+H} - \tilde{E}_t(C_{t+H})\right)^2$, that is known in advance:

$$u(C_{t+H}) = C_{t+H} - \frac{\alpha}{2}\widetilde{\mathrm{Var}}_t(C_{t+H}). \tag{2.9}$$

Note that if $\tilde{E}_t\left(C_{t+H} - \tilde{E}_t(C_{t+H})\right)^2$ is equal to $E_t\left(C_{t+H} - \tilde{E}_t(C_{t+H})\right)^2$, then $Eu^q(C_{t+H}) = Eu(C_{t+H})$, i.e. comparisons based on (2.8) and (2.9) are equivalent

[3] [68]

on average. Therefore, to focus only on predictions for the levels of RV we may consider only the latter specification for the utility. To check that our results are not severely affected by this simplification, we will also present the results for the classic quadratic utility in (2.8).

The quadratic utility is, of course, not the only preferences that yields risk-aversion. We also experiment with a class of utilities of the following form:

$$u(C_{t+H}) = C_{t+H} - \frac{\alpha}{2}\left(\tilde{E}_t\left|C_{t+H} - \tilde{E}_t(C_{t+H})\right|^m\right)^n. \tag{2.10}$$

Under linearity of the operator $\tilde{E}_t(.)$, the latter part is equal to

$$\begin{aligned} &\frac{\alpha}{2}\left(\tilde{E}_t\left|C_{t+H} - \tilde{E}_t(C_{t+H})\right|^m\right)^n = \\ &= \frac{\alpha}{2} I_t K^{m+n}\left(\tilde{E}_t\left|RV_{t+H} - \tilde{E}_t(RV_{t+H})\right|^m\right)^n \equiv I_t K \pi_t, \end{aligned} \tag{2.11}$$

where π_t from now on will be referred to as the trader's premium. The utility in (2.10) is strictly increasing in cash flows and concave in K if $C_{t+H} = K(P_t(H) - RV_t^{t+H})$ for $\alpha > 0$ and $m + n > 1$. Under these conditions, the relative risk-aversion of the trader is increasing in the notional K and the preference parameter α. Therefore, a higher notional size requires a more discrete and precise decision rule on when to trade, that is there is a higher need for accurate forecasting for higher K.

We are now in a position to define precisely the reservation value of a trader from the rationality constraint (2.5). For the quadratic utility (2.6) the corresponding reservation value is equal to

$$\bar{P}_t(H) = E_t RV_t^{t+H} + \frac{\alpha K}{2}\mathrm{Var}_t RV_t^{t+H}. \tag{2.12}$$

In general, we may define the reservation value by

$$\bar{P}_t(H) = E_t RV_t^{t+H} + \pi_t. \tag{2.13}$$

Thus, the reservation value is composed of the variance forecast and the premium, demanded by the trader.

In what follows we will use the parametrization of utilities that expresses the risk-aversion of the trader in volatility premium units. A volatility premium is a measure of the risk premium on the variance evident in option and variance swap prices. It is measured in annualized percentage units, e.g. for RV over the next month calculated using returns in percentage units and $P_t(22)$ in analogous units, the average market volatility premium is equal to:

$$\text{Market Volatility Premium} = \frac{1}{T}\sum_{t=1}^{T}\left[\sqrt{12P_t(22)} - \sqrt{12RV_t^{t+22}}\right].$$

The estimated value for the volatility premium is reported to be around $3\% - 3.3\%$, see [49]. We will employ similar units for measuring the "trader's" volatility premium, which is a measure of the difference between the reservation value $\bar{P}_t$ and the expected realized variance:

$$\text{Trader's Volatility Premium} = \frac{1}{T}\sum_{t=1}^{T}\left[\sqrt{12\bar{P}_t(22)} - \sqrt{12E_tRV_t^{t+22}}\right],$$

Substituting for the reservation value from equation (2.13) and assuming that the premium π_t is much smaller than the forecast $E_tRV_t^{t+22}$, we approximate the trader's volatility premium by $\frac{\sqrt{3}}{T}\sum_{t=1}^{T}\pi_t$. Therefore volatility premium demanded by the trader is approximately proportional to the average premium π_t. Note, that from (2.10) π_t is strictly increasing in α. Therefore, for each specification of the premium (choice m and n) trader's volatility premium is strictly increasing in α, and, thus, there is a one-to-one correspondence between trader's risk-aversion and trader's volatility premium.

2.3.2 Variance Swap Price

The final element in the model is the price for the variance swap P_t. Suppose that the trader cannot bargain about the price of the contract, as the buyer has an outside option of hedging with options. Therefore, due to "no-arbitrage" considerations, the price of the variance swap is equal to $P_t(H) = E_t^Q RV_t^{t+H}$, where the expectation is taken under a risk-neutral measure.(For example, see [37].)

This value can be synthesized from the call and put option prices for different strikes of any traded security. If it were possible to observe the call and put option prices for a continuum of strikes, then this would be the exact fair-value of the variance swap. In practice, this fair value can be approximated from the prices of only those options that are available for trading.

In our empirical exercise we will make use of widely-used series that are available for $E_t^Q RV_t^{t+H}$ of the S&P 500 index for the maturity of one month, i.e. $H = 22$ days. These data are reported by CBOE as VIX-series. Specifically, $E_t^Q RV_t^{t+22} \approx 30/365(VIX_{CBOE,t}/100)^2$. For notational simplicity, denote $P_t(22) = VIX_t^2$, assuming that all the necessary conversions were taken into account.

The market consistently offers a premium for variance swap sellers with the ratio $\frac{E(VIX_t^2 - RV_t^{t+22})}{ERV_t^{t+22}}$ to be around $0.2-0.3$, or in volatility percentage units $E(\sqrt{12VIX_t^2}) - \sqrt{12RV_t^{t+22}}) \approx 3.0\%$, see e.g. [37], [98], [29]. As was mentioned, the trader will trade if his premium is lower than the one implied by the market price of the variance swap.

2.4 Variance Forecast Loss-Function

Before presenting the loss-function that is motivated by the model of the variance trader, we will summarize this model. The trader observes the market price of the swap $P_t(H)$ and sells a contract with a notional K and duration H if and only if $P_t(H) > \bar{P}_t(H)$, where the reservation value is given by $\bar{P}_t(H) = E_t RV_t^{t+H} + \pi_t$.

Thus, his expected utility at time t is equal to:

$$V_t = K\left(P_t(H) - E_t RV_t^{t+H} - \pi_t\right) I\left(P_t(H) - E_t RV_t^{t+H} > \pi_t\right). \tag{2.14}$$

Assume that the trader observes his premium π_t, but does not observe the expectation $E_t RV_t^{t+H}$. Consequently, he estimates it by a forecast $\widehat{RV}_t^{t+H}$ to form his reservation value $\bar{P}_t = \widehat{RV}_t^{t+H} + \pi_t$. Therefore, his expected utility equals:

$$V_t = K\left(P_t(H) - E_t RV_t^{t+H} - \pi_t\right) I\left(P_t(H) - \widehat{RV}_t^{t+H} > \pi_t\right). \tag{2.15}$$

Thus the quality of the forecasts affects the trader's utility through the trading decision $I_t^* = I(P_t(H) - \widehat{RV}_t^{t+H} > \pi_t)$, but not through the expected outcome of the trade $[P_t(H) - E_t RV_t^{t+H} - \pi_t]$.

The comparison of two forecasting systems that yield predictors $\widehat{RV}_{1,t}^{t+H}$ and $\widehat{RV}_{2,t}^{t+H}$ can be carried out by comparing the expected utilities (2.15) from strategies based on these predictors. For both strategies, we assume the same premiums $\pi_{1,t} = \pi_{2,t} = \pi_t$, e.g. in the case of quadratic utility (2.9) this implies the equality of the following predictors: $\widetilde{\mathrm{Var}}_{1,t} RV_t^{t+H} = \widetilde{\mathrm{Var}}_{2,t} RV_t^{t+H}$. This mirrors the literature on return predictability. In this literature, for two forecasts to be compared based on Sharpe ratios $ER^e/\sigma(R^e)$, the forecasts for the return $\widehat{E}R^e$ are taken from the models being compared, but the forecast of the variance $\hat{\sigma}(R^e)$ is the same for all models (See [65].) This distinction allows studying the forecasts of the levels separately from the forecasts of higher moments. In our case, this ensures that the difference in the performance is all due to the accuracy of the forecasts and not to the difference in risk aversion between traders caused by different evaluations of π_t.

Simple representation for the loss-function can be given, if we denote π_t^m to be the expected variance premium in the market:

$$\pi_t^m \equiv P_t(H) - E_t RV_t^{t+H}.$$

Then the trader's expected utility follows from (2.15):

$$V_t = K(\pi_t^m - \pi_t)I(\pi_t^m - \pi_t > \varepsilon_t), \tag{2.16}$$

where ε_t is the expected error in the forecast $\widehat{RV}_t^{t+H} = E_t RV_t^{t+H} + \varepsilon_t$. Therefore, the corresponding loss-function is equal to:

$$\mathcal{L} = -E[(\pi_t^m - \pi_t)I(\pi_t^m - \pi_t > \varepsilon_t)]. \tag{2.17}$$

Equivalently, the same loss function can be derived in terms of the ex-post market variance premium $\pi_t^{m*} = VIX_t^2 - RV_t^{t+H}$ and ex-post forecast error $\varepsilon_t^* = \widehat{RV}_t^{t+H} - RV_t^{t+H}$. The ex-post error includes ε_t plus the unpredictable part $E_t RV_t^{t+H} - RV_t^{t+H}$. We may rewrite the loss-function in (2.17) as

$$\mathcal{L} = -E[(\pi_t^{m*} - \pi_t)I(\pi_t^{m*} - \pi_t > \varepsilon_t^*)], \tag{2.18}$$

and define its in-sample analog by the following sample mean:

$$\widehat{\mathcal{L}} = -\frac{1}{T}\sum_{t=1}^{T}(\pi_t^{m*} - \pi_t)I(\pi_t^{m*} - \pi_t > \varepsilon_t^*). \tag{2.19}$$

The loss-function is equal to the average of the difference in the market and trader's variance premiums in the periods where this difference is larger than the forecast error.

2.4.1 Properties of the Loss-Function

For the derived loss-function (2.17), its one-period " realization" equals

$$\mathcal{L}_t = -(\pi_t^m - \pi_t)I(\pi_t^m - \pi_t > \varepsilon_t). \tag{2.20}$$

The loss-function defined above exhibits certain properties that are usually found in economic loss-functions. First, it is asymmetric and favors biased forecasts. Second,

similar to the median AE and in contrast to MSE and MAE, it is robust to outliers. Finally, contrary to MSE , MAE and Median AE, it depends not only on the unconditional distribution of the forecasting error, but also on its conditional distribution given the current level of uncertainty.

Focusing on the latter property, note that during the periods when the difference between market and trader's premiums is large, small errors ε_t do not force the forecaster into wrong action. That is, for $\pi_t^m - \pi_t \gg 0$ and $\pi_t^m - \pi_t \ll 0$ it holds that $I(\pi_t^m - \pi_t > \varepsilon_t) = I(\pi_t^m - \pi_t > 0)$ if ε_t is small. Therefore, the loss from an error ε_t depends on the difference in the premiums. Denote the difference in the premiums by $\Delta\pi_t$. Then,

$$\mathcal{L} = -E\left[\Delta\pi_t I(\Delta\pi_t > \varepsilon_t)\right], \tag{2.21}$$

or integrating over the forecasting error ε_t

$$\mathcal{L} = -E\left[\Delta\pi_t F_{\varepsilon_t|\Delta\pi_t}(\Delta\pi_t)\right] \tag{2.22}$$

where $\Delta\pi_t = \pi_t^m - \pi_t$, and $F_{\varepsilon_t|\Delta\pi_t}(.)$ is the CDF of the expected forecast error given the difference in premiums. For the case of constant π_t,the loss-function simplifies to:

$$\mathcal{L} = -E\left[(\pi_t^m - \pi)F_{\varepsilon_t|\pi_t^m}(\pi_t^m - \pi)\right]. \tag{2.23}$$

For the case of π_t that is proportional to the market variance premium, i.e. $\pi_t^m = \kappa_0\pi_t$, the loss-function takes the form:

$$\mathcal{L} = -(1-\kappa_0^{-1})E\left[\pi_t^m F_{\varepsilon_t|\pi_t^m}\left(\frac{\pi_t^m}{1-\kappa_0^{-1}}\right)\right]. \tag{2.24}$$

In the examples above, the performance of a forecast depends on the distribution of the error ε_t conditionally on π_t^m rather than its unconditional distribution. We may expect that π_t^m is a function of uncertainty in the market. Many studies have

found that different volatility measures co-move, e.g. jumps and continuous part of variance, long-run and short-run variances, volatility and default probabilities. Therefore, we may conclude that the performance of forecasts depends on how errors are distributed in the periods of high and low volatilities.

2.5 Data and Forecasts

To implement the forecast evaluation using (2.19) as a loss-function, we collect data of three types: a series of the realized variance forecasts $\widehat{RV}_t^{t+H}$, a series of the actual realized variance RV_t^{t+H}, and a series of market prices P_t for which we use VIX_t^2. Since the latter corresponds to monthly contracts on volatility of the S&P 500 index, we take the forecast horizon to be equal to $H = 22$ trading days and calculate the realized variances for S&P 500 returns.

Out of these series, VIX_t^2 is the the simplest to construct. Daily closing quotes for VIX are reported by CBOE/CFE exchanges. [4] The data run from 2 January, 1990, until today and are updated by CBOE daily.

The second by simplicity of construction is the realized variance RV_t^{t+22}. To construct this series we use the intra-day 5-minute futures return data provided by TickWrite. This frequency is chosen based on signature plots of the realized variances, showing that the data constructed using higher frequencies may be excessively contaminated by microstructure noise. The time stamps at which the index value was recorded start at 8:35 a.m. and finish at 15:15 p.m. Eastern Time, yielding 80 intra-day returns per day, and one overnight return calculated from the 15:15 p.m. quote of the previous day to the 8:35 a.m. quote of the current day. Therefore, the

[4] We convert the data from CBOE back to monthly units, i.e. we undo the normalization carried out by CBOE: $VIX_t^2 = 30\ VIX_{CBOE,t}^2/365$.

daily realized variance for the day $t+1$ is equal to

$$RV_t^{t+1} = (s_{t+1,open} - s_{t,close})^2 + \sum_{j=1}^{1/80} \left(s_{t+1,j/80} - s_{t+1,(j-1)/80}\right)^2. \tag{2.25}$$

and the corresponding multi-period realized variance is a sum of daily RV:

$$RV_t^{t+H} = \sum_{i=1}^{H} RV_{t+i-1}^{t+i} \tag{2.26}$$

We choose the span starting January, 1992, and ending October, 2007. This range includes the high-volatility periods in 1998-1999, in the beginning of the 2000s and during the last period of high volatility in August, 2007.

Finally, we construct the forecast series. In this paper, we evaluate a set of forecasts focusing primarily on the comparison of model-based and reduced-form methods. The interest in the comparison between model-based forecasts and reduced-form forecasts is motivated by the trade-off between the efficiency of the first and the simplicity of the latter [5].

Table 2.1 summarizes the list of the forecasts that are featured in the current study: one-factor and two-factor model-based forecasts, the model-based forecast derived from the SV-model with jumps, the reduced-form forecast based on the autoregressive model for RV_t^{t+1}, exponential smoothing, and finally, the simplest constant variance forecast. Each of these forecasts is characterized by a set of variables, that it employs to predict the realized variance, i.e. predictors. For example, for the one-factor model-based forecast the sole predictor is the estimated spot variance. For the reduced-form AR7 forecast, predictors are the past realizations of RV. Table 2.1 links the forecasts to the corresponding predictors.

In the next subsections, we will give a short description for each of the forecasts. More detailed descriptions are given in the Appendix.

[5] For more details on model-based and reduced-form forecasts see [14], and [94].

Table 2.1: Forecasts and Corresponding Predictors

Forecasts	Predictors
Model-Based SV-J	Spot variance $\hat{\sigma}_t^2$
Model-Based 2F	Factors $\hat{x}_{1,t}$,and $\hat{x}_{2,t}$: $\sigma_t^2 = x_{1,t} + x_{2,t}$
Model-Based 1F	Spot variance $\hat{\sigma}_t^2$
Reduced-Form AR7	RV_{t-i}^{t-i+1}, $i = \overline{1,7}$
Exponential Smoothing	RV_{t-i}^{t-i+1}, $i \geq 1$
Random Walk	RV_{t-H}^{t}

2.5.1 Model-Based Forecasts: SV-CJ, Two-Factor and One-Factor

The SV-CJ forecast is based on the model by [50]:

$$\begin{bmatrix} ds_t \\ d\sigma_t^2 \end{bmatrix} = \begin{bmatrix} \mu \\ \kappa(\theta - \sigma_t^2) \end{bmatrix} dt + \begin{bmatrix} \sigma_t dW_t^s \\ \sigma_v \sigma_t dW_t^v \end{bmatrix} + \begin{bmatrix} \xi_t^s \\ \xi_t^v \end{bmatrix} dJ_t, \tag{2.27}$$

where dW_t^v and dW_t^s are increments of standard Brownian motions with the correlation ρdt, J_t is a Poisson process with the intensity λ, ξ_t^s and ξ_t^v are jump sizes: ξ_t^v is exponential with the mean $1/\mu_v$, and ξ_t^s is conditionally normal with the mean $\mu_j + \rho_j \xi_t^v$ and the variance σ_j^2. For each day t, we estimate the model on daily close-to-close data up to t and construct the forecast of RV_t^{t+22}. We repeat this procedure for the period of August, 2000, to October, 2007. That gives us 1780 out-of-sample monthly forecasts.

This approach is based on the daily data and disregards the intra-day information. Although this may be seen as a drawback if the intra-day data contain extra information, this may be seen as an advantage of the method if the intra-day data is contaminated with microstructure noise. Moreover, taking this model to higher-frequency data may require a more elaborate dynamics, e.g. short-term or seasonal components, see [5].

Another feature of this approach is that it separates the fast-reverting part of the variance into the Poisson component. The rest of the variance is the slow-reverting Gaussian process σ_t^2. Alternatively, the same separation of the slow-moving and fast-moving components can be achieved by a two-factor model.

As a more elaborate alternative to SV-CJ, we will consider a two-factor model-based forecast that employs intra-day data. To form this forecast, we assume a general two-factor ESV-model (see [82]), estimate the parameters by matching the correlation structure of the variance, and use the Kalman-filter to extract states. This procedure is not as efficient as MLE. Nevertheless, if the model is indeed two-factor, then, regardless of the exact form of the model, this procedure delivers the consistent estimate of $E_t RV_t^{t+H}$. Therefore, it is robust to certain kinds of misspecification.

Another alternative is a one-factor model-based forecast. It is constructed in the same way as the two-factor model-based forecast. However, in this case, we do not disentangle fast and slow moving components in the variance. A more detailed description of the model-based methods is given in the Appendix.

2.5.2 Reduced-Form Forecasts: Autoregressive, Exponential Smoothing, and Random Walk

From the class of reduced-form forecasts, we selected the autoregressive, exponential smoothing, and the random walk forecasts. The autoregressive model is the projection of RV on its own past values:

$$RV_{t+1|t} = \beta_0^{(1)} + \sum_{i=1}^{I} \beta_i^{(1)} RV_{t-i}^{t+1-i}. \tag{2.28}$$

The above model is the simplest in the class of linear models for RV. A more elaborate example is the ARFIMA model by [11]. We chose the autoregressive model because of its computational simplicity. Iterating the above equation, we obtain the form of

the H-period forecast:

$$RV_{t+H|t} = \beta_0 + \sum_{i=1}^{I} \beta_i RV_{t-i}^{t+1-i}. \tag{2.29}$$

Several prior studies have shown that the part of the realized variance coming from large jumps in prices has a weak predictive power with respect to future variances. This finding motivated the HAR-RV-J and the HAR-RV-CJ models by [9] and [13], respectively. Here, we again choose the simplest way to improve the forecasting power by using bi-power variations (BV) on the right-hand side of (2.29) instead of RV:

$$RV_{t+H|t} = \beta_0 + \sum_{i=1}^{I} \beta_i BV_{t-i}^{t+1-i}, \tag{2.30}$$

$$BV_t^{t+1} = \sum_{j=2}^{1/h} \frac{\pi}{2} \left| s_{t+jh} - s_{t+jh-h} \right| \left| s_{t+jh-h} - s_{t+jh-2h} \right|$$

Intuitively, BV strips off jump-components from RV that have a weak correlation with the future dynamics of the variance. Therefore, using BV as a predictor may improve the forecasting power of this AR model.

Exponential smoothing is a reduced-form filter for RV, that is given in the iterative form:

$$\begin{aligned} RV_{t+1|t} &= \alpha RV_{t|t-1} + (1-\alpha) RV_{t-1}^{t}, \\ RV_{t+H|t} &= H\left(\alpha RV_{t|t-1} + (1-\alpha) RV_{t-1}^{t}\right). \end{aligned} \tag{2.31}$$

The simplest forecast out of the reduced-form forecasts is the random-walk that assumes that "variance will not change":

$$RV_{t+H|t} = RV_{t-H}^{t}. \tag{2.32}$$

We will compare two methods for estimation of (2.30) and (2.31), namely OLS and WLS. For instance, for an AR-7 forecast given by (2.30) with $I = 7$, OLS estimates

will minimize the sum of squared residuals:

$$\min_{\beta} \sum_{t} (RV_t^{t+H} - \beta_0 - \sum_{i=1}^{7} \beta_i BV_{t-i}^{t-i+1})^2,$$

and WLS will minimize the following sum:

$$\min_{\beta} \sum_{t} \frac{(RV_t^{t+H} - \beta_0 - \sum_{i=1}^{7} \beta_i BV_{t-i}^{t-i+1})^2}{w_t}, \tag{2.33}$$

where the weights take the form $w_t = \exp(\gamma_0 + \sum_{i=1}^{7} \gamma_i BV_{t-i}^{t-i+1})$. The exponent in the definition of w_t prevents it from turning negative. The coefficients γ_0, γ_i are obtained through a two-step procedure.

Note that for the SV-CJ model we use MLE. However, for the two-factor and one-factor model-based forecasts we perform a regression in the final step (see the Appendix). Therefore, these two forecasts will also require a choice between OLS and WLS.

2.6 Results

In this section we apply the economic loss-function defined in this paper towards comparison of the following forecasts: SV-CJ, two-factor model-based, one-factor model-based, reduced-form AR-7, exponential smoothing, and random walk. Formally, we may classify two-factor, one-factor and SV-CJ forecasts as model-based since they approximate the conditional expectation using the model for prices. All the other forecasts can be classified as reduced-form forecasts. We focus our attention on the comparison of these two groups. As we vary the premium specification, we may potentially see different rankings of these forecasts. As a conclusion, we want to determine if the difference in the performances between these two groups is economically and statistically significant under certain premium specifications. We

also want to investigate if the ranking based on the utility-based loss-function is drastically different from the ranking based on statistical loss-functions.

2.6.1 Statistical Loss-Functions

We start by assessing the statistical performance of the forecasts. To mimic the swap contract, we consider the forecast horizon of one month. The estimation period includes the data from January, 1992. The out-of-sample performance of the forecasts is evaluated on data from August, 2000, to October, 2007. Table 2.2 reports the mean-squared error, bias, mean absolute error and median absolute error. All four are routinely used in the forecasting literature.

The lowest MSE corresponds to the SV-CJ forecast. It yields an average squared error of 50.79%, down from 53.46% for the two-factor and down from 53.65% for the AR-7 forecasts. All the other forecasts perform considerably less successfully. Exponential smoothing yields the smallest bias.

Both MAE and Median AE, measures that put less weight on large errors compared to MSE, favor reduced-form forecasts: AR-7 and exponential smoothing respectively. Comparing this ranking to the one based on MSE, we conclude that the first rank of the SV-CJ forecast by MSE may be explained by several large errors admitted by reduced-form forecasts.

Overall, based on statistical loss functions, the performances of the reduced-form and model-based forecasts are quite close. The model-based forecasts lead in MSE, and the reduced-form forecasts lead in MAE and MedAE measures. In the next subsection, we ask if the results from the statistical comparison carry over to economic loss-functions.

Table 2.2: Statistical Comparison of Variance Forecasts for Actual Data

Forecasts	MSE normalized	BIAS	BIAS normalized	MAE	Median AE
OLS: Model-Based 2F	53.46 %	-0.46	0.027 %	11.72	6.21
OLS: Model-Based 1F	80.20 %	-2.10	0.549 %	16.16	12.08
OLS: Reduced-Form AR7	53.65 %	-1.30	0.211 %	11.59	6.72
OLS: Exp Smoothing	70.28 %	0.01	0.000 %	12.06	4.46
WLS: Model-Based 2F	58.19 %	0.83	0.087 %	11.73	4.61
WLS: Model-Based 1F	84.51 %	-2.25	0.633 %	12.89	6.68
WLS: Reduced-Form AR7	57.09 %	-0.08	0.001 %	10.96	4.56
WLS: Exp Smoothing	76.06 %	-0.01	0.000 %	11.85	4.22
Random Walk	78.47 %	0.15	0.003 %	13.49	4.82
SV-CJ	50.79 %	-1.14	0.162 %	11.55	6.84

Note: Statistical comparison of forecasts by the normalized MSE $= \frac{\sum_{t=1}^{T}(RV_t - \widehat{RV}_t)^2}{\sum_{t=1}^{T}(RV_t - \overline{RV}_t)^2}$, bias $= \sum_{t=1}^{T}(\widehat{RV}_t - \overline{RV}_t)/T$, and the normalized bias $= \frac{[\sum_{t=1}^{T}(\widehat{RV}_t - \overline{RV}_t)/T]^2}{[\sum_{t=1}^{T}(RV_t - \overline{RV}_t)^2/T]}$, where $\overline{RV}_t$ is the sample average of the realized variances, and $\widehat{RV}_t$ is the forecast. MAE is the sample average of the absolute errors $|\widehat{RV}_t - \overline{RV}_t|$.

2.6.2 Economic Loss-Functions for Constant Risk Premiums

The simplest benchmark case is the one with a constant premium $\pi_t = \pi(\alpha)$ that increases with the risk-aversion parameter α. For simplicity and without loss of generalization, we may parameterize the trader's premium to be $\pi_t = \alpha$. For each choice of α and a forecasting system, we form trading activity of the variance swap seller and calculate the corresponding utility. Figure B.2 shows sample average utilities corresponding to several forecasting systems. Axis X in Figure B.2 corresponds to the risk-aversion of the trader measured in volatility premium units, i.e. it is approximately the sample average of $\sqrt{3}\pi_t$. The risk-aversion of the trader in its classical form grows from the left to the right. Utilities are normalized by the average price of the contract, thus they can be interpreted in a compensation manner. For example,

consider a trader who demands a volatility premium of 1% on average, i.e. the mean of $\sqrt{12\bar{P}_t} - \sqrt{12E_t\widehat{RV}_t^{t+H}}$ is equal to 1. If he trades always (i.e. $I_t = 1, \forall t$), then his average utility is around 20.0%. If he never trades (i.e. $I_t = 0, \forall t$), then his average utility is zero. Therefore, he will be willing to pay up to 20 % of the contract price to be able to participate in the trade.

There are five lines in the figure that are utilities derived from employing a two-factor model-based forecast, one-factor model-based forecast, SV-CJ model-based forecast, AR-7 reduced-form forecast, and the strategy of "trading always", i.e. always agreeing to trade. All the forecasts are calculated using weighted least-squares, as their OLS versions perform uniformly weaker.

Figure B.2 can be divided into three regions. In the first region the seller would prefer to trade always, because the utility from this strategy exceeds the maximum utility from strategies based on forecasting. That is, in this region forecast errors prevent sellers from carrying out profitable trades, rather than alerting them against future high uncertainty in the market. This region lies below the risk-aversion with the volatility premium of 2%, implying that a nearly risk-neutral trader derives no advantage from using statistical forecasts.

The second region, above 10%, is the region where the market volatility premium is not high enough to encourage betting on volatility. In this region, traders can be better off by declining to be on the sell side of the variance swap trade. In this region, forecast errors push forecasters into risky trading when they would be better off by refraining from participation altogether.

Finally, the most interesting region, that lies between 2% and 10% is the region where the variance seller profits from forecasting. In this region, the reduced-form forecast (solid dark line) uniformly dominates all the other forecasts. The forerunner is the two-factor model-based forecast, which employs high-frequency data. Other

forecasts also result in positive gains. The point at which the trader is willing to pay the highest fees for forecasts is at around 3.5 % of the trader's volatility premium. At this point, the trader is indifferent between always participating in trading and never participating, and the compensation for the access to statistical forecasts reaches 7% (AR-7 forecast).

It follows from Figure B.2 that out of three depicted model-based forecasts the one that is based on high-frequency data and incorporates two components in the variance is the most successful. SV-CJ also includes a pair of volatility components (the Gaussian part and the Poisson part) but is based on daily data. One-factor forecast performs the worst despite using the high-frequency data, which is in accordance with prior research on the performance of one-factor models for long-term forecasting, see [94].

Table 2.3: Utility-Based Comparison of Variance Forecasts for Actual Data : Constant Premium

Forecasts	3%	5%	7%
OLS: Model-Based 2F	5.38 %	1.37 %	0.85 %
OLS: Model-Based 1F	5.54 %	-0.28 %	-3.70 %
OLS: Reduced-Form AR7	7.09 %	2.47 %	1.76 %
OLS: Exponential Smoothing	5.66 %	0.87 %	0.06 %
WLS: Model-Based 2F	6.14 %	2.42 %	1.72 %
WLS: Model-Based 1F	6.68 %	1.31 %	-0.98 %
WLS: Reduced-Form AR7	**7.71 %**	**3.55 %**	**2.88 %**
WLS: Exponential Smoothing	6.86 %	1.65 %	0.43 %
Random Walk	2.97 %	-1.75 %	-1.41 %
SV-CJ	5.63 %	1.52 %	1.44 %
Always trade	4.05 %	-14.18 %	-34.26 %

Note: The table reports utility of the variance trader for different levels of risk-aversion in volatility premium units (see Section 2.3.1). The utilities are expressed as percentages from the contract price. Rows correspond to different forecasts that are used to form optimal trading strategies.

A larger set of forecasting systems is compared in Table 2.3, where the cuts of

the graph at 3%, 5% and 7% of the risk-aversion are taken. We see again that the reduced-form AR-7 forecast performs the best among the whole set of forecasts. In terms of economic significance, the trader would be willing to switch from the AR-7 to the best of the model-based forecasts for not less than 1.1% of the contract price. Figure B.3 shows the difference in utility between two identical traders, one of whom employs the reduced-form forecast, while the other employs the model-based forecast. The grey area represents the 95 % confidence intervals around this difference, which are built using the asymptotic theory of [60]. It follows from the figure, that the difference is statistically significant at the 5% level for trader's volatility premiums of less than 5%.

Now we can compare the rankings of the forecasts in Table 2.3 and Table 2.2. First, the utility-based measure in Table 2.3 and MSE in Table 2.2 do not agree on the choice between OLS and WLS estimations for forecast parameters. Judging by the out-of-sample MSE, the forecasts with OLS-parameters perform better than the same forecasts with WLS parameters. For example, the two-factor model-based forecast yields the MSE of 58.19% and 53.46% for WLS and OLS estimators, respectively. This result is expected, as OLS parameters are found through the minimizing of squared residuals in-sample, what translated into the minimizing of squared residuals out-of-sample. However, for the utility-based loss function, WLS-versions of forecasts perform better for all levels of risk aversion. For example, at trader's volatility premiums of 7% the trader is willing to pay around 0.9% of the contract price to switch from using the OLS-type two-factor model-based forecast to the WLS-type forecast.

Utility-based and MSE rankings disagree on the performance of the SV-CJ model. This forecast is the most efficient based on MSE, but loses to the reduced-form forecast based on the utility-based performance measure. Overall, median absolute error seems to give the ranking that is the closest to the one based on the trader's

Table 2.4: Model for Higher Moments of Variance: Parameter Estimates

m	1	2	3	1	2	3
	$\hat{x}_{1t} < 0$			$\hat{x}_{1t} > 0$		
β_0	2.68 (0.33)	5.91 (0.68)	9.77 (0.65)	2.46 (0.12)	6.20 (0.23)	10.52 (0.35)
β_1	10.03 (5.52)	15.81 (8.46)	17.27 (17.18)	1.03 (0.24)	0.69 (0.37)	-0.16 (0.83)
β_2	40.34 (24.94)	63.67 (42.51)	51.83 (106.88)	-0.25 (0.11)	-0.02 (0.15)	0.36 (0.35)
β_3	57.15 (32.35)	92.00 (59.44)	65.74 (154.20)	0.03 (0.01)	$3e^{-3}$ (0.01)	-0.03 (0.04)

Note: The table reports the estimates of coefficients in the formula (2.34). The model was estimated in-sample for the years of 1992-2007 using non-linear least squares. Newey-West standard errors are shown in parenthesis.

utility. The property that possibly links these two measures is the robustness to outliers.

Note that the utility-based loss-function from this section is still quite close to the statistical measures, since it assumes that the risk of the transaction is not changing over time. The next step is to consider more natural specifications for the trader's premium, where the premium is related to the current uncertainty about the future variance.

2.6.3 Economic Loss-Function for Variable Risk Premiums

To define the premium in this section, we construct a proxy for the conditional moments of realized variance $\tilde{E}_t|RV_t^{t+H} - \tilde{E}_t RV_t^{t+H}|^m$ using the two-factor model-based forecast for $\tilde{E}_t RV_t^{t+H}$ and then fitting an exponential model of the form:

$$\tilde{E}_t|RV_t^{t+H} - \tilde{E}_t RV_t^{t+H}|^m = e^{\beta_0+\beta_1\hat{x}_{1,t}+\beta_2\hat{x}_{1,t}^2+\beta_3\hat{x}_{1,t}^3}, \tag{2.34}$$

where $\hat{x}_{1t}$ is the estimated slow-reverting component of volatility from the two-factor model. (See the Appendix.) Table 2.4 reports resulting the coefficients.

Although there are many possible ways to specify this proxy, e.g. casting it within

the reduced-form forecast framework, the two-factor model-based offers certain simplifications. Within this framework, all the conditional moments are the functions of only two states, x_{1t} and x_{2t}. The choice was motivated mostly by simplicity of including higher orders of the regressor $x_{1,t}$ into the formula. Note, that $x_{2,t}$ is not included, as the slopes on this component were found to be insignificant. The exponent in (2.34) ensures that the trader's premium is positive.

Subsequently, we varied the two parameters $m \in 1, 2, 3$ and $n \geq \frac{1}{m}$ to collect a set of premiums $\pi_t = \alpha K^{m+n-1} \left[\hat{E}_t \left| RV_t^{t+H} - RV_{t+H|t} \right|^m \right]^n$. We start with the quadratic-utility case for $m = 2$ and $n = 1$, and report the results for other premium specifications in the Appendix.

Figure B.4 shows average utilities that result from trading strategies based on different forecasting systems. As in the case of the constant premium, the X axis in Figure B.4 corresponds to the risk-aversion of the trader measured in volatility premium units, i.e. it is approximately the sample average of $\sqrt{3}\pi_t$. Risk-aversion increases from the left to the right. Unlike the constant premium case, the risk-aversion depends not only on the preferences of the trader, but also on the size of the contract K: $\pi_t = \pi_t(\alpha, K)$. This variance premium is strictly increasing in α and the notional K. Therefore the utility patterns across different values on the X axis have a dual interpretation: on the one hand they can be interpreted as utilities of traders with different absolute risk-aversions, on the other hand they can be interpreted as utilities of the same trader for different contract sizes, that are exogenously given by the hedging demands of a buyer.

Similar to the case of a constant trading premium, the graph can be divided into three areas. The first one is the area where traders should prefer to trade regardless of their forecasts. This area lies below the 0.9% level of the trader's volatility premium. The second area, above 12% of the volatility premium, is the area where they should

always hedge their position. The area where forecasts bring substantial profits to the trader lies between these two marks.

The utility from using statistical forecasts increased for all levels of risk-aversion and all the forecasts, with the exception of the one-factor model-based forecast. Now the maximum value from forecasting is reached at trader's volatility premium of 2.4 % and is above 10 % of the contract price (for AR-7 forecast). The weakest of the presented forecasts – the one-factor forecast – is still informative for the trader yielding positive profits for risk aversions of less than 4%. Interestingly, for the premium that changes with uncertainty in the market, the two-factor model-based forecast has a clear advantage in comparison to the SV-CJ, which employs only daily data, and to the simpler one-factor model.

From Figure B.4 it follows that for the quadratic utility case, the two-factor model-based and reduced-form forecasts perform equally well. In terms of their comparative performance, the reduced-form forecast is slightly better for less risk-averse traders or alternatively, for smaller sizes of contracts, and the model-based forecast is slightly better in the opposite case of high risk-aversion and larger contract sizes. The threshold lies at around 4% of the volatility premium. Table 2.5 reports the results for the whole set of forecasts at certain values of risk-aversion. Among the reduced-form forecasts, exponential smoothing performs the best for all three columns. It is the best among all the forecasts for lower risk-aversion, and yields less than 0.25% of the contract price to the two-factor model-based forecast for higher risk-aversions.

Therefore, we found that for the risk-premium π_t that is proportional to the current uncertainty in the market, the comparison between reduced-form forecasts and model-based forecasts depends on the risk-aversion of the trader. In particular, for smaller risk-aversions, AR-7 and exponential smoothing dominate the two-factor model-based forecast, but for larger risk-aversions they yield respectively up to 0.2%

Table 2.5: Utility-Based Comparison of Variance Forecasts for Actual Data: Variable Premium

Forecasts	3%	5%	7%
OLS: Model-Based 2F	4.09 %	2.52 %	0.73 %
OLS: Model-Based 1F	-4.92 %	-3.33 %	-1.71 %
OLS: Reduced-Form AR7	2.41 %	1.21 %	-0.04 %
OLS: Exponential Smoothing	3.56 %	1.48 %	0.47 %
WLS: Model-Based 2F	6.90 %	**4.15 %**	**2.24**%
WLS: Model-Based 1F	2.14 %	0.40 %	-0.16 %
WLS: Reduced-Form AR7	7.16 %	3.76 %	1.87 %
WLS: Exponential Smoothing	**7.18%**	3.97 %	1.96 %
Random Walk	3.16 %	2.03 %	2.19 %
SV-CJ	4.23 %	1.92 %	1.16 %
Always trade	-9.12 %	-39.66 %	-32.72 %

Note: The table reports utility of the variance trader for different levels of risk-aversion in volatility premium units (see Section 2.3.1). The utilities are expressed as percentages from the contract price. Rows correspond to different forecasts that are used to form optimal trading strategies.

and 0.5% of the contract price. This outcome is different from the case of the constant π_t, under which the reduced-form forecast was uniformly better than the two-factor model-based forecast.

Figure B.5 explains the difference between the cases of the constant π_t and the variable π_t. The figure reports the averages of the trade indicator I_t, cash flows C_t, and utilities u_t within seven equally-sized periods between August, 2000, and October, 2007. During this period, the first four years 2000-2003 are characterized by high volatility. (See Figure B.5.) The following years of 2004-2006 were relatively stable.

The first two panels of Figure B.5 show how often the trader participated in selling swaps during each year. Note that under quadratic utility, the probability to trade in low-uncertainty times is higher than the same probability in times of high uncertainty. Thus, the variable premium down weights periods with high volatility

and puts more weight on low volatility periods. Intuitively, traders will avoid trading in the turbulent years, and will participate in trade during the less volatile years even though the market may offer them moderate expected profits.

The second and the third pairs of panels in Figure B.5 report two alternative measures of the success for the forecasts: average cash-flows and utilities. Note that the reduced-form forecast is better for the periods with high uncertainty and slightly worse in the periods with low uncertainty. Thus, for the variable premium the advantage from using the reduced-form forecast disappears as the importance of the first part of the sample diminishes. The higher the risk aversion of the trader, the less chance that the trader will trade in the high-uncertainty times. Hence, eventually the model-based forecast will become slightly better, as the only period that will matter for the forecaster will be the years of 2004 - 2006.

Several robustness checks are presented in the Appendix. First, we show that the results of this section hold for the "classic" quadratic preferences as given by (2.8). Second, we report the forecast comparison for other choices of the premium parameters.

To summarize our results for the quadratic utility, the performances of the reduced-form forecasts and model-based forecasts for variable trader's premium are very close. The reduced-form forecast is still better for moderate risk aversions. The dependence of the winning forecast on the risk aversion suggests that none of the forecasts can successfully model the difference in the dynamics between low volatility and high volatility periods.

2.7 Simulation Study

In this section we reproduce the results of the forecast comparison on simulations with the SV-CJ dynamics by [50]. The SV-CJ model was used to form one of the model-based forecasts in the previous section. Under this model the process for

log-prices s_t is described by the next system of equations:

$$\begin{bmatrix} ds_t \\ d\sigma_t^2 \end{bmatrix} = \begin{bmatrix} \mu \\ \kappa(\theta - \sigma_t^2) \end{bmatrix} dt + \begin{bmatrix} \sigma_t dW_t^s \\ \sigma_v \sigma_t dW_t^v \end{bmatrix} + \begin{bmatrix} \xi_t^s \\ \xi_t^v \end{bmatrix} dJ_t. \tag{2.35}$$

The parameters for simulations were estimated on the daily S&P 500 futures returns over the sample of 1992 - 2007, and take the following values: mean return $\mu = 0.032$, volatility mean-reversion $k = 0.0216$, mean variance $\theta = 0.730$, variance parameter of volatility $\sigma_v^2 = 0.0178$, jump intensity $\lambda = 0.012$, and the leverage effect $\rho = -0.725$. Also, jumps in the variance ξ_t^v are distributed exponentially with the parameter $\mu_v = 2.09$, and jumps in returns ξ_t^s are conditionally normal $N(\mu_j + \rho_j \xi_t^v, \sigma_j^2)$, where $\mu_j = -0.995$, $\rho_j = -1.52$, and the variance of jumps σ_j^2 is equal to 2.66.

The data was simulated using the Euler discretization scheme at one second frequencies. From one-second log-prices we constructed 5-minute returns. As in the observed data, each trading day in these simulations lasts for 6.5 hours, plus one overnight return. For simplicity, the distribution of the overnight returns is the same as for 5-minute day-time returns. Simulations include 5 data sets each of a length of 4000 days, that is approximately equal to 5 times 16 years. Additionally, we simulated 60000-day series to estimate the reduced-form model for the realized variance: vector-autoregression for RV and BV. Table B.1 in the Appendix reports the estimates of the reduced-form model.

To form the model-based forecast from (2.35) we take the parameters of the model as given and use particle filtering to extract the monthly RV forecasts. To form close-to-close VIX^2-series, we estimated the following HAR-model on observed data that links VIX^2 to past realized variance over different horizons:

$$VIX_t^2 = \beta_0 + \sum_{i=1}^{5} \beta_i RV_{t-i}^{t-i+1} + \beta^{mon} RV_{t-20}^{t} + \sum_{i=1}^{3} \beta_i^{qrt} RV_{t-60i}^{t-60(i-1)} \tag{2.36}$$

The resulting R^2 of the above regression is 82.6%. In simulations, VIX^2 was con-

structed using formula (2.36) with the estimates reported in Table B.2 in the Appendix.

Similar to the observed data, we asses the performance of the model-based forecast and the reduced-form forecast, first using statistical measures, and then the new utility-based measure. The statistical performance of the forecasts is reported in Table 2.6. The utilities derived from application of the same forecasts are shown in Figure B.6. By statistical performance, the model-based forecast is slightly leading, giving the minimum of MSE (45.02%), MAE (7.26) and Median AE (5.42). The model-based forecast under no misspecification is the most efficient by construction, but the difference from the reduced-form forecast is minimal. The reduced-form forecast yields the MSE of 46.13 %, MAE of 7.44 and Median AE of 5.66.

Table 2.6: Statistical Performance of Variance Forecasts for Simulated Data: True Model.

	MSE	MAE	Median AE
Reduced-Form	46.13%	7.44	5.66
Model-Based	45.02%	7.26	5.42

Note: The table reports MSE $= \frac{\sum_{t=1}^{T}(RV_t-\widehat{RV}_t)^2}{\sum_{t=1}^{T}(RV_t-\overline{RV}_t)^2}$, Mean AE $= \frac{1}{T}\sum_{t=1}^{T}|RV_t-\widehat{RV}_t|$, and Median AE statistics for the reduced form forecast based on VAR, and the forecast based on the model (2.35). The data consists of five simulations from the same model (2.35) of the length 4000 days each.

Figure B.6 shows the utilities for different risk-aversions of the variance trader derived from using the model-based and the reduced-form forecasts. These two graphs visibly coincide both in Panel A, for the constant risk premium, and in Panel B, for the quadratic utility.

To summarize, if prices follow the SV-CJ dynamics and are observed at 5-minute intervals, then the reduced-form forecast performs very close to the most efficient model-based forecast in terms of MSE and other statistical measures. The same

holds for the utility-based loss-function; for each level of the trader's risk aversion, utilities derived from using the reduced-form forecast almost coincide with the utilities derived from the efficient forecast based on the true model. Therefore, the reduced-form forecast,which does not make any assumptions about the underlying process for returns and is ultimately simple to construct, practically attains the maximum efficiency among all the possible forecasts.

2.7.1 Misspecification Case

To see why the model-based approach may eventually fail in comparison to the reduced-form one, we expand the original model by adding a new component to the variance $v_{2,t}$ that has a higher mean-reversion than the original variance. Thus, the new model includes three components – two Gaussian ones and one that is a Poisson jump:

$$\begin{bmatrix} ds_t \\ dv_{1,t} \\ dv_{2,t} \end{bmatrix} = \begin{bmatrix} \mu \\ \kappa(\theta - v_{1,t}) \\ \kappa_2(\theta - v_{2,t}) \end{bmatrix} dt + \begin{bmatrix} \sqrt{v_{1,t} + v_{2,t}} dW_t^s \\ \sigma_v \sqrt{v_{1,t}} dW_t^v \\ \sigma_{v,2} \sqrt{v_{2,t}} dW_t^v \end{bmatrix} + \begin{bmatrix} \xi_t^s \\ \xi_t^v \\ 0 \end{bmatrix} dJ_t, \qquad (2.37)$$

where $\kappa_2 = 1.5$. We specify the other parameters as in the original model, and pick the volatility-of-volatility parameters in such a way that $\sigma_v^2/\kappa = \sigma_{v,2}^2/\kappa_2$. Finally, the return data are adjusted by a scalar to ensure that the average market volatility premium is equal to 3.23%, as in the data for the S&P 500 and VIX.

To construct the model-based and the reduced-form forecast, we use 60000 days of simulated data to estimate the parameters of the SV-CJ model and vector autoregression (VAR) for RV and BV. Second, we generate five data sets of the length 4000 days, and form 3878 forecasts based on the SV-CJ model that employs 5-minute returns, reduced-form forecasts, VIX series and actual monthly RV series.

The results of the statistical performance of the forecasts are presented in Table 2.7. This table demonstrates that inclusion of the additional fast-reverting compo-

nent in the variance, that is not a jump, "confuses" the model-based forecast, so it performs much worse in comparison to the case with no misspecification (70.27% MSE vs. 45.02% MSE) and also in comparison with the reduced-form forecast (54.75% MSE). The same is true for the mean and median absolute errors.

Table 2.7: Statistical Performance of Variance Forecasts For Simulated Data: Misspecified Model.

	MSE	MAE	Median AE
Reduced-Form	54.75%	3.91	3.06
Model-Based	70.27%	5.04	4.66

Note: The table reports MSE $= \frac{\sum_{t=1}^{T}(RV_t-\widehat{RV}_t)^2}{\sum_{t=1}^{T}(RV_t-\overline{RV}_t)^2}$, Mean AE $= \frac{1}{T}\sum_{t=1}^{T}|RV_t-\widehat{RV}_t|$, and Median AE statistics for the reduced form forecast based on VAR for RV and BV, and the forecast based on the model (2.35). The data consists of five data sets simulated from the model (2.37) of 4000 days each.

Figure B.7 shows the results of the utility-based comparison. In contrast to the case when we constructed the forecast based on the true model, in Figure B.7 we see that now there is a difference in performance between the reduced-form and the model-based forecasts; for lower risk-aversions the reduced-form forecast is about 2.5 % better than the model-based forecast. This difference vanishes as the risk-aversion increases.

Note also that the utility patterns in Panels A and B are very similar and qualitatively the same. The reason for this similarity is that the data simulated from (2.37) does not exhibit clear periods of large and low volatility. The next logical step would be to consider the model with regime-switching, that could give this property, e.g. as in the paper by [70].

Summarizing this section, we can conclude that for a correctly specified model,the reduced-form forecast performed very close to the model-based forecast that is the most efficient by construction. Furthermore, in the case of misspecification, it out-

performed the model-based forecast both for statistical and utility-based measures.

2.8 Conclusion

We proposed a new variance forecast loss-function based on variance trading. In contrast to conventional forecast performance measures, the suggested loss-function evaluates the performance of a forecast versus information that is already included in the market prices.

We showed that the performance measure that is relevant for designing trading strategies is state-dependant. In particular, it puts less weight on the forecast errors during periods of turmoil.

For this new loss function, we examined the out-of-sample performances of reduced-form and model-based forecasts for variances of the S&P 500. We demonstrated that a simple reduced-form forecast is not outperformed by more sophisticated techniques, such as model-based forecasts. This paper demonstrates this fact for utility-based loss-functions.

Regarding the differences between statistical and utility-based measures, we found that for utility-based measures, the performance of the reduced-form forecast vs. the model-based approach may be improved even further by using non-linear techniques that may be more successful in capturing the difference in variance dynamics during periods of high and low uncertainty.

3

Volatility in Equilibrium: Asymmetries and Dynamic Dependencies

3.1 Introduction

Modeling and forecasting of stock market return volatility has received unprecedented attention in the academic literature over the past two decades. The three most striking empirical regularities to emerge from this expanding literature arguably concern: (i) the highly persistent own dynamic dependencies in the volatility;[1] (ii) the existence of a typically positive volatility risk premium as manifested by the variance swap rate exceeding the corresponding expected future volatility;[2] (iii) the apparent asymmetry in the lead-lag relationship between returns and volatility.[3] Despite these now well-documented and generally accepted empirical facts, no formal

[1] The historically low volatility over several years proceeding the Fall 2008 financial crises and the subsequent sustained heightened volatility provide anecdotal evidence for this idea.

[2] The preponderance of options traders "selling" volatility to gain the premium indirectly supports the notion of volatility carrying a risk premium.

[3] Again, anecdotal evidence like the heightened volatility following Russia's default and the LTCM debacle in September 1998, the relatively low volatility accompanying the rapid run-up in prices during the tech bubble, as well as the recent sharp increase in volatility accompanying the Fall 2008 financial crises and sharp market declines are all consistent with this asymmetry.

model yet exists for explaining these features within a coherent economic framework. This paper fills that void by developing an entirely self-contained equilibrium based explanation for the asymmetry and volatility risk-premium that also accommodates long-run dependencies in the underlying volatility process.

Before discussing the model any further, it is instructive to illustrate anew the empirical regularities we seek to explain. To that end, the top most solid line in Figure 3.1 shows the sample autocorrelations for the aggregate market volatility out to a lag length of ninety days, based on daily data for the squared options-implied volatility index VIX over the past two decades; further details concerning the data and different volatility measures are given in Section 3.4. The autocorrelations in Figure 3.1 decay at a very slow rate and remain numerically large and statistically significant for all lags. Consistent with these highly persistent own dynamic dependencies in the volatility, it is now widely accepted that the typical rate of decay is so slow as to be best described by a fractionally integrated long-memory type process; for some of the earliest empirical evidence along these lines see, e.g., [91], [41] and [15].

The VIX index in effect represents the market's expectation of the volatility for S&P 500 index over the next month together with any premium for bearing the corresponding volatility risk.[4] Isolating the variance risk premium, the second line in Figure 3.1 shows the daily autocorrelations for the difference between the squared VIX index and the one-month-ahead forecasts *from a simple reduced form time series model* for the *actually observed* daily realized variation of the S&P 500 returns; further details concerning the high-frequency based realized volatility series and the construction of the model forecasts are again deferred to Section 3.4. Although the autocorrelations indicate non-trivial dependencies for up to several weeks, the pre-

[4] The variance risk premium is formally defined as the difference between the expected future volatility under the risk-neutral and the actual probability measures.

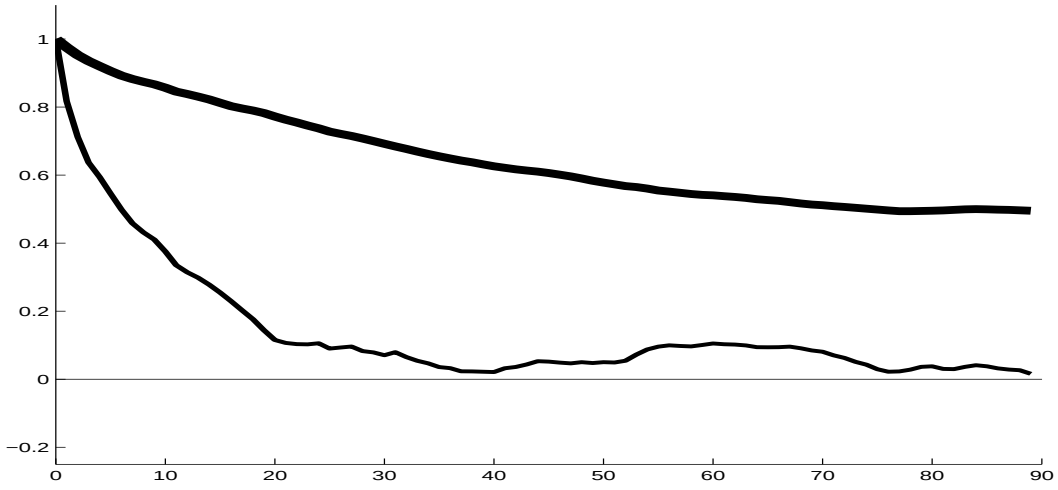

FIGURE 3.1: Sample Autocorrelations. The top most solid line shows the sample autocorrelations for the VIX^2 volatility index to a lag length of 90 days. The lower line shows the sample autocorrelations for the variance risk premium. The calculations are based on daily data and variable definitions as described in more detail in Section 3.4.1.

mium is clearly not as persistent as the volatility process itself. Again, this is not a new empirical result per se, but one that any satisfactory economic model must be able to accommodate. For instance, the empirical analysis in [27] also supports the idea of relatively fast mean reversion in the volatility risk premium, as does the evidence of fractional co-integration between implied and realized volatility presented by, e.g., [16] and [87].

Next, to highlight the aforemention return-volatility asymmetry, the first panel in Figure 3.2 plots the very high frequency cross-correlations between leads and lags of the S&P 500 returns and the corresponding squared options-implied VIX volatility index. [28] have previously demonstrated the advantage of using high-frequency intraday data for more effectively estimating and analyzing the lead-lag relationship between returns and volatility.[5] We therefore follow their lead in using five-minute observations for both the VIX and the S&P 500 returns in calculating the sample

[5] [31] have also argued that the asymmetry tend to be stronger for implied than realized volatility.

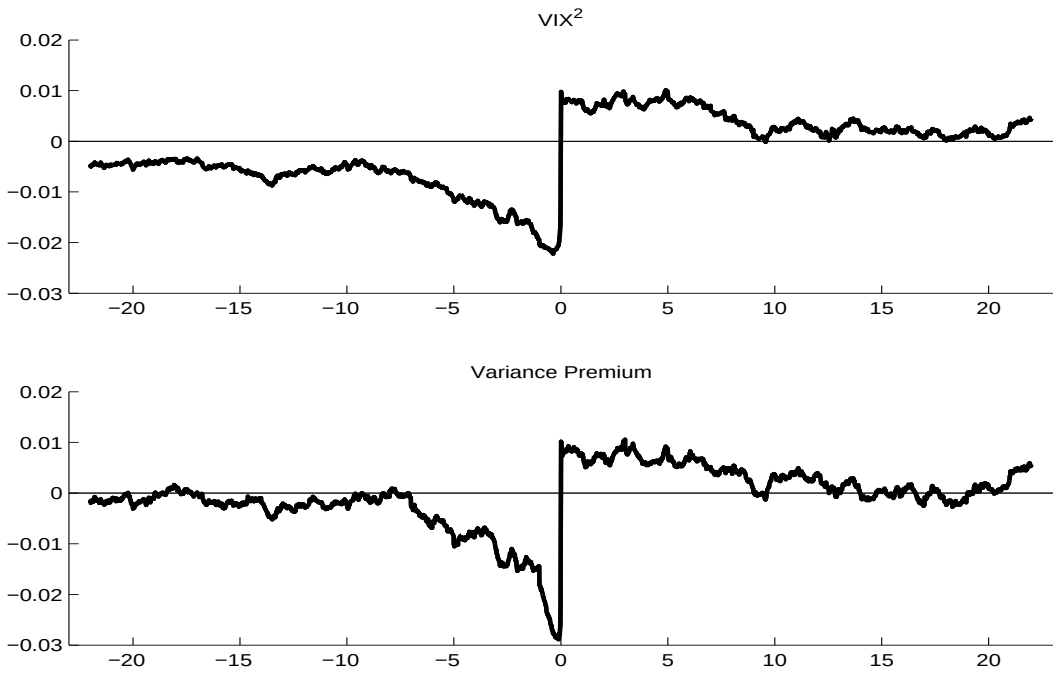

FIGURE 3.2: Sample Cross-Correlations. The top panel shows the sample cross-autocorrelations between the VIX^2 and lags and leads of the returns ranging from -22 to 22 days. The bottom panel shows the sample cross-autocorrelations between the variance risk premium and the returns. The calculations are based on high-frequency five-minute data and and variable definitions as further detailed in Section 3.4.1.

cross-correlations for leads and lags ranging up to 22 days, or 1,716 leads and lags at the five-minute sampling frequency. High-frequency data for the VIX have only been available since 2003, so that the cross-autocorrelations depicted in Figure 3.2 are necessarily based on a somewhat shorter five-year calendar-time span compared to the longer sample of daily observations used for illustrating the longer-run own volatility dependencies in the previous Figure 3.1. Nonetheless, this shorter high-frequency sample still reveals a striking negative pattern in the correlations between the volatility and the lagged returns, lasting for several days. On the other hand, the correlations between the volatility and the future returns are all positive, albeit close to zero. Following [25], who first observed the apparent asymmetry in the

return-volatility relationship, the left part of the figure and the negative correlation between lagged returns and current volatility is now commonly referred to as a "leverage effect,"[6] while the right part of the figure and the positive correlation between current volatility and future returns has been termed a "volatility feedback effect." This systematic pattern in the high-frequency based cross-correlations is directly in line with the empirical evidence from numerous empirical studies based on coarser lower frequency daily data and specific parametric models, including the early influential work by [57], [84], [61] and [35].

Taking the analysis one step further, the bottom panel in Figure 3.2 shows the cross-correlations between the five-minute S&P 500 returns and the variance risk premium, where as before the variance risk premium is defined as the difference between the squared VIX index and the corresponding forecast constructed from a simple reduced form time series model for the observed realized volatilities. Comparing this new picture to the return-volatility dependencies in the top panel, the signs of the cross-correlations generally coincide. However, there is a noticeable faster decay toward zero in the magnitude of the cross-correlations between the variance risk premium and the lagged returns compared to the decay in the corresponding cross-correlations between the squared VIX index and the lagged returns. Interestingly, this difference directly mirrors the difference in the shape and the rate of decay in the standard sample autocorrelations for the two daily volatility series depicted in Figure 3.1.

The key empirical return-volatility patterns and dynamic dependencies illustrated in the figures indicate that volatility carries a risk premium. Standard equilibrium based models build around a representative agent with time-separable utility rules out priced volatility risk; see, e.g., the discussion in [18]. Instead, following the liter-

[6] It is now also widely agreed that the negative correlations actually have little if anything to do with changes in financial leverage; see, e.g., [54].

ature on long-run risk models pioneered by [18], we assume that the representative agent is equipped with Epstein-Zin-Weil preferences. Importantly, in this situation the Stochastic Discount Factor (SDF) will depend not only on the consumption growth rate, but also on the future investment opportunities. Consequently, the aggregate market return will be a function of the expected growth in the economy, as in the traditional time-separable utility case, as well as the uncertainty about the future economic growth; see, e.g., [34]. Intuitively, this also explains why investors may be willing to pay an uncertainty premium, and in turn why the VIX may differ from the corresponding actual return volatility and why the corresponding variance risk premium may serve as a separately priced risk factor.

Our model is cast in continuous time. The Epstein-Zin-Weil preference structure was first employed in a continuous-time asset pricing setting by [44]. However, in contrast to the expression for the SDF involving the compensator function derived in that paper, we find it more convenient to work with the discount factor expressed in terms of the consumption growth rate and the market return. As formally show below, this expression in turn results in continuous-time Euler equations analogous to the discrete-time equations originally derived by [47], and utilizing a log-linearization allows for closed-form tractable solutions of the model for values of the Intertemporal Elasticity of Substitution (EIS) different from unity.

The model is most directly related to the equilibrium long-run risk model recently developed by [29], in which the volatility-of-volatility in the economy is determined by its own separate stochastic process, and the model of [42], in which the expected growth rate in consumption and the volatility of consumption growth are both allowed to "jump." In comparison to the specific discrete-time formulations employed in those papers, however, the continuous-time formulation adopted here has the advantage of allowing for the calculation of internally consistent model implications across different sampling frequencies and return horizons, and it is able to accommo-

date much richer longer-run volatility dependencies including fractional integration. Importantly, the continuous-time formulation also permits a stringent definition of the VIX volatility index and avoids the inherent problem in traditional discrete-time long-run risk models with GARCH type errors that the (conditional) variance is known one period in advance and therefore formally can't generate a variance premium. A related long-run risk type model in which the economic uncertainty, or the volatility of consumption growth, is allowed to "jump" in continuous-time has also recently been explored by [49], in an attempt to explain the existence of a (on average) positive volatility risk premium.

Other recent studies concerned with the equilibrium pricing of volatility risk include [58] and [99], both of whom analyze the implications of rare disasters, and [78] who emphasize the role of low frequency movements in macroeconomic uncertainty for explaining low frequency multi-year movements in stock market valuations. Several studies more squarely rooted in the options pricing literature have also explored the equilibrium implications of allowing for richer volatility dynamics and non-standard preference structures; see, e.g., the recent papers by [24], [51] and [93] and the references therein. However, the empirical focus of the present paper is distinctly different from all of these other studies, and to the best of our knowledge, no other coherent economic equilibrium-based explanation for the volatility asymmetries and dynamic dependencies depicted in Figures 3.1 and 3.2 is yet available in the literature.

The plan for the rest of the paper is as follows. The new theoretical model is formally defined and solved in Section 3.2. The equilibrium implications from the model in regards to the return-volatility asymmetries and the own dynamic volatility dependencies are presented in Section 3.3. The data used in the construction of the figures discussed above and the model's ability to reproduce these basic empirical features are the subject of Section 3.4. Section 3.5 concludes. Most of the

mathematical proofs are deferred to two Appendixes.

3.2 Volatility in Equilibrium

The classic continuous-time Intertemporal CAPM of [83] is often used to justify the existence of a traditional volatility risk premium in aggregate market returns. However, the model is unable to explain the leverage effect and asymmetric return-volatility dependencies actually observed in the data. The continuous-time endowment economy developed here instead builds on the discrete-time long-run risk model pioneered by [18], and the extensions of that same model in [29] and [42]. We begin in the next subsection by a description of the basic continuous-time model setup and solution under short-memory dynamics. We subsequently show how the model may be extended to incorporate empirically relevant long-memory dependencies.

3.2.1 Basic Model Setup and Assumptions

Let the local geometric growth rate of consumption C_t in the economy be denoted by $g_t \equiv \frac{dC_t}{C_t}$. To simplify the analysis and explicitly focus on the role of time-varying volatility, we rule out any predictability in g_t by assuming that it follows the continuous martingale

$$g_t \;=\; \mu_g dt \;+\; \sigma_{g,t} dW_t^c, \tag{3.1}$$

where μ_g denotes the constant mean growth rate, $\sigma_{g,t}$ refers to the conditional volatility of the growth rate, and W_t^c is a standard Wiener process.[7] Further, we assume that the volatility dynamics in the economy are governed by the following continuous-time affine processes,

$$d\sigma_{g,t}^2 \;=\; \kappa_\sigma(\mu_\sigma - \sigma_t^2)dt \;+\; \sqrt{q_t} dW_t^\sigma, \tag{3.2}$$

$$dq_t \;=\; \kappa_q(\mu_q - q_t)dt \;+\; \varphi_q \sqrt{q_t} dW_t^q, \tag{3.3}$$

[7] The growth rate of consumption is identically equal to the dividend growth rate in this Lucas-tree economy.

where the two Wiener processes W_t^σ and W_t^q are independent and jointly independent of W_t^c, and the parameters satisfy the non-negativity restrictions $\mu_\sigma > 0$, $\mu_q > 0$, $\kappa_\sigma > 0$, $\kappa_q > 0$, and $\varphi_q > 0$.[8] The stochastic volatility process $\sigma_{g,t}^2$ represents time-varying economic uncertainty in consumption growth, with the volatility-of-volatility process q_t in effect inducing an additional source of temporal variation in that same process. As discussed further below, both of these processes play an important role in generating empirically realistic time varying volatility risk premia. Note that the assumption of independent innovations across all three equations means the internal structure of the model itself must explain the return-volatility correlations, and it rules out any correlations that might otherwise arise via purely statistical channels.

We assume that the representative agent's consumption and investment decisions are based on the maximization of Epstein-Zin-Weil recursive preferences. As formally shown in Appendix A this implies that the equilibrium relationship between the inter-temporal marginal rate of substitution, M_t, consumption, C_t, and the cumulated return on the aggregate wealth portfolio, R_t, must satisfy the relation derived in equation (C.9),

$$d\log M_t + \frac{\theta}{\psi} d\log C_t + (1-\theta) d\log R_t = -\rho\theta dt, \tag{3.4}$$

where ρ denotes the instantaneous subjective discount factor, ψ equals the inter-temporal elasticity of substitution, and the parameter θ is defined by

$$\theta \equiv \frac{1-\gamma}{1-\frac{1}{\psi}}, \tag{3.5}$$

where γ refers to the coefficient of risk aversion. The expression in equation (3.4) is naturally interpreted as the continuous-time version of the widely used and celebrated discrete-time equilibrium relationship derived in [48]. In the following we will

[8] We also assume that $\mu_q\kappa_q > 0.50\varphi_q^2$, which ensures positivity of q_t, and that μ_σ is sufficiently large relative to κ_σ, so that negativity of $\sigma_{g,t}^2$ is highly unlikely and the subsequent approximations reasonable.

maintain the assumptions that $\gamma > 1$ and $\psi > 1$, which readily implies that $\theta < 0$.[9] Consistent with the empirical regularities discussed in the introductory section, these specific parameter restrictions ensure, among other things, that asset prices fall on news of positive volatility shocks and that volatility carries a positive risk premium.

3.2.2 Basic Model Solution

Let Ψ_t denote the price-dividend ratio, or equivalently the price-consumption or the wealth-consumption ratio, of the asset that pay the consumption endowment $\{C_{t+s}\}_{s\in[0,\infty)}$. It is convenient to express $\Psi_t = \Psi(x_t)$ a function of the state vector $x_t = (\sigma_{g,t}^2\,,\, q_t)$. The conventional linear method of solving for rational expectations models is complicated by the fact that the stochastic differential equation for $\log(\Psi_t)$ involves the reciprocal of $\Psi(x_t)$. To circumvent this difficulty, we approximate $\Psi(x)^{-1} = \exp(-\log\Psi(x))$ by the following first-order expansion,

$$\frac{1}{\Psi(x)} \;\approx\; \exp(-\overline{\log\Psi}) - \exp(-\overline{\log\Psi})(\log\Psi(x) - \overline{\log\Psi}) \;=\; \kappa_0 - \kappa_1 \log\Psi(x),$$

where $\kappa_1 > 0$. This approximation plays a similar role to that of the standard Campbell-Shiller discrete-time approximation to the returns in terms of the log price-dividend ratio. Similar expressions have also previously been used in the continuous-time setting by, e.g., [33].

Now conjecturing a solution for $\log(\Psi_t)$ as an affine function of the state variables, $\sigma_{g,t}^2$ and q_t,

$$\log(\Psi_t) = A_0 + A_\sigma \sigma_{g,t}^2 + A_q q_t, \tag{3.6}$$

[9] The assumption that $\gamma > 1$ is generally agreed upon, but the assumption that $\psi > 1$ is a matter of some debate; see, e.g., the discussion in [18].

and solving for the coefficients A_0, A_σ, and A_q, it follows from Appendix B that

$$A_0 = \frac{A_\sigma \kappa_\sigma \mu_\sigma + A_q \kappa_q \mu_q - \kappa_0 + (1 - 1/\psi)\mu_g - \rho}{\kappa_1}, \tag{3.7}$$

$$A_\sigma = \frac{-\gamma}{2} \frac{1 - \frac{1}{\psi}}{\kappa_\sigma + \kappa_1}, \tag{3.8}$$

$$A_q = \frac{\kappa_q + \kappa_1 - \sqrt{(\kappa_q + \kappa_1)^2 - \theta^2 A_\sigma^2 \varphi_q^2}}{\theta \varphi_q^2}. \tag{3.9}$$

The restrictions that $\gamma > 1$ and $\psi > 1$, readily imply that the impact coefficient associated with both of the volatility state variables are negative; i.e., $A_\sigma < 0$ and $A_q < 0$.[10] Or put differently, that the market falls on positive volatility "news." From these explicit solutions for the three coefficients it is now also relatively straightforward to deduce the reduced form expressions for other variables of interest.

In particular, as shown in equation (C.7) in Appendix B, the equilibrium dynamics of the logarithmic cumulative return process, R_t, is given by

$$\begin{aligned} d\log(R_t) = \left[\frac{\mu_g}{\psi} + \rho - \frac{\sigma_{g,t}^2}{2} + \frac{\gamma}{2}(1 - \frac{1}{\psi})\sigma_{g,t}^2 - A_q q_t (\kappa_1 + \kappa_q) \right] dt + \\ \sigma_{g,t} dW_t^c + A_\sigma \sqrt{q_t} dW_t^\sigma + A_q \varphi_q \sqrt{q_t} dW_t^q. \end{aligned} \tag{3.10}$$

The directional effects of increases in the endowment volatility, $\sigma_{g,t}^2$, on the local expected return are generally ambiguous. However, for sufficiently high levels of risk-aversion γ and inter-temporal substitution ψ, endowment volatility positively affects the local expected return. Meanwhile, increases in the volatility-of-volatility, q_t, unambiguously, increase the local expected return, reflecting the compensation for bearing volatility risk. On the other hand, diffusive-type innovations in the volatility and the volatility-of-volatility, dW_t^σ and dW_t^q, both have a negative impact on the local returns, consistent with a leverage type effect.

[10] The solution for A_q in equation (3.9) represents one of a pair of roots to a quadratic equation, but it is the economically meaningful root for reasons discussed below.

To further appreciate the implications of the model and how it might help explain the empirical regularities, it is instructive to consider the model-implied equity premium, $\pi_{r,t}$, as derived in equation (C.9),

$$\pi_{r,t} \equiv -\frac{1}{dt}\frac{d\left[R,M\right]_t}{R_t M_t} = \gamma\sigma_{g,t}^2 + (1-\theta)(A_\sigma^2 + A_q^2\varphi_q^2)q_t. \tag{3.11}$$

The first term, $\gamma\sigma_{g,t}^2$, is akin to a classic risk-return tradeoff relationship. It does not really represent a volatility risk premium *per se*, however, but rather changing prices of consumption risk induced by the presence of stochastic volatility. Instead, the second term, $(1-\theta)\kappa_1^2(A_q^2\varphi_q^2 + A_\sigma^2)q_t$, has the interpretation of a true volatility risk premium, representing the confounding impact of the two diffusive-type innovations, dW_t^σ and dW_t^q.[11] The existence of this true volatility risk premium depends crucially on the dual assumptions of recursive utility, or $\theta \neq 1$, as volatility would not otherwise be a priced factor, and time varying volatility-of-volatility, in the form of the q_t process. The restrictions that $\gamma > 1$ and $\psi > 1$ imply that the volatility risk premium is positive.

The expression for the instantaneous risk premium in equation (3.11) discussed above closely mirrors the expression for the premium in the corresponding discrete-time model of [29]. The specific formulations used in deriving these results both involve somewhat restrictive assumptions about the underlying volatility dynamics. We next discuss how the continuous-time setup and the developed model solution may be adapted to allow for more flexible and richer dynamic dependencies in the volatility, including longer-memory type effects.

[11] The specific root in equation (3.9) implies that $A_q^2\varphi_q^2 \to 0$ for $\varphi_q \to 0$, which guarantees that the premium disappears when q_t is constant, as would be required by the lack of arbitrage.

3.2.3 General Model Solution

Numerous more flexible continuous-time stochastic volatility models have been proposed in the literature. We build on the framework of [39] in assuming that the evolution in $\sigma_{g,t}^2$ may be described by

$$\sigma_{g,t}^2 = \sigma^2 + \int_{-\infty}^{t} a(t-s)\sqrt{q_s}dW_s^{\sigma}. \tag{3.12}$$

This representation is quite general and flexible. By appropriate choice of the moving average weights $\{a(s)\}_{s\in[0,\infty)}$ it obviously includes the affine process in equation (3.2) as a special case. Importantly, by suitable choice of the mapping $s \rightarrow a(s)$, the process for $\sigma_{g,t}^2$ may also exhibit various forms of long range dependence. In particular, setting

$$a(s) = \frac{\sigma}{\Gamma(1+\alpha)}\left(s^{\alpha} - ke^{-ks}\int_0^s e^{ku}u^{\alpha}du\right) \tag{3.13}$$

results in the classic fractionally integrated process, where α denotes the long-memory parameter.

To complete the specification of the model and still allow for closed form solutions, we will maintain the identical laws of motions for the consumption growth rate and the volatility-of-volatility given by equations (3.1) and (3.3), respectively. The actual solution strategy differs somewhat from that for the basic model. The full details are given in Appendix C; a precis follows.

In parallel to the solution for the short-memory model discussion above, we start by conjecturing a solution for the logarithmic price-consumption ratio now of the form

$$\log(\Psi_t) = A_0 + A_q q_t + \int_{-\infty}^{t} A(t-s)\sqrt{q_s}dW_s^{\sigma} \tag{3.14}$$

where A_0, A_q, and $\{A(s)\}_{s\in[0,\infty)}$ are to be determined. Some care is needed because

of subtleties related to possible arbitrage opportunities under long-memory type dependencies [92]. The strategy that we use relies on the fact that in the absence of arbitrage the return on a traded security must follow a semi-martingale. This allows us to split up the returns into a drift and a local martingale component. This decomposition is possible when $A(t)$ exists and is differentiable at zero. Substituting the conjectured solution into the pricing equation (3.4) yields the following ordinary differential equation for $t > s$,

$$A'(t-s) - \kappa_1 A(t-s) \;=\; \frac{\gamma(1-\frac{1}{\psi})}{2} a(t-s), \tag{3.15}$$

and two regular equations,

$$\frac{\theta}{2}\varphi_q^2 A_q^2 - (\kappa_q + \kappa_1)A_q + \frac{\theta A(0)^2}{2} \;=\; 0 \tag{3.16}$$

$$A_0 \;=\; \frac{A_q\kappa_q\theta_q - \kappa_0 + (1-\frac{1}{\psi})\mu_g - \rho - \frac{\gamma}{2}(1-\frac{1}{\psi})\sigma^2}{\kappa_1}. \tag{3.17}$$

From the Appendix, the solutions to this system of equations are

$$A(s) \;=\; -\int_s^{+\infty} \frac{\gamma(1-\frac{1}{\psi})}{2} e^{\kappa_1(t-\tau)} a(\tau)\, d\tau, \tag{3.18}$$

$$A_q \;=\; \frac{\kappa_q + \kappa_1 - \sqrt{(\kappa_q+\kappa_1)^2 - \theta^2\varphi_q^2 A(0)^2}}{\theta\varphi_q^2}, \tag{3.19}$$

$$A_0 \;=\; \frac{A_q\kappa_q\theta_q - \kappa_0 + (1-\frac{1}{\psi})\mu_g - \rho - \frac{\gamma}{2}(1-\frac{1}{\psi})\sigma^2}{\kappa_1}, \tag{3.20}$$

which exists and is well defined subject to a terminal condition ruling out explosive bubble solutions and other mild regularity conditions. As before, from this set of solutions it is possible to deduce the reduced form expressions for all other variables of interest.

In particular, in parallel to the expression for the returns in the short-memory model in equation (3.10) above, it follows from Appendix C that the reduced form expression for the returns in the general may be expressed as,

$$d\,\log(R_t) = \mu_{R,t}\,dt + \sigma_{g,t}dW_t^c + A_q\varphi_q\sqrt{q_t}dW_t^q + A(0)\sqrt{q_t}dW_t^\sigma, \tag{3.21}$$

where the drift is defined by,

$$\mu_{R,t} = \rho + \frac{\mu_g}{\psi} + [-\frac{1}{2} + \frac{\gamma}{2}(1-\frac{1}{\psi})]\sigma_{g,t}^2 - (\kappa_q+\kappa_1)A_q q_t. \tag{3.22}$$

Similarly, from equation (C.21) in Appendix C the equilibrium equity premium is determined by,

$$\pi_{r,t} \;=\; \gamma\sigma_t^2 + (1-\theta)[A_q^2\varphi_q^2 + A(0)^2]q_t \;=\; \gamma\sigma_t^2 + 2\left(\frac{1}{\theta}-1\right)(\kappa_q+\kappa_1)A_q q_t. \tag{3.23}$$

Under the previously discussed parameter restrictions $\gamma > 1$ and $\psi > 1$, implying that $\theta < 0$, the equity premium remains positive. So as long as $\gamma \neq \frac{1}{\psi}$, or $\theta \neq 1$, it also remains the case that stochastic volatility carries a positive risk premium. Note also that the instantaneous equity premium only depends on the $\{a(s)\}_{s\in[0,\infty)}$ weights and the possible long-run dependencies in the volatility through the cumulative impact determined by the integral solution for $A(0)$ in equation (3.18).

We next turn to a more specific discussion of the model's implications vis-a-vis the volatility risk premium and dynamic return-volatility dependencies.

3.3 Dynamic Equilibrium Dependencies

The equilibrium expressions discussed in the previous section characterize how the equity premium depends on the instantaneous volatility, and how the instantaneous return responds to contemporaneous volatility innovations within the model. This section further details the model's implications in regards to the dynamic dependencies in the volatility and the volatility risk premium, and how these volatility

measures co-vary with leads and lags of the returns at different horizons. We will subsequently confront these theoretical predictions with the key empirical regularities discussed in the introduction.

3.3.1 VIX and the Volatility Risk Premium

One of the key features of the general version of the model is that consumption volatility, $\sigma_{g,t}^2$, may exhibit long-range dependence while the volatility of the volatility, q_t, is a short-memory process. This in turn has important implications for the serial correlation properties of the equivalent to the VIX volatility index implied by the model and the corresponding volatility risk premium, and how these measures correlate with the returns.

To begin, consider the (forward) integrated variance, or quadratic variation, of the asset price S_t,

$$IV_{t,t+N} \equiv \int_{\tau=t}^{t+N} d\left[\log S, \log S\right]_\tau , \tag{3.24}$$

where the "brackets" $[\,]$ represents the standard quadratic variation process. From equation (C.22) in Appendix C the reduced form expression for the integrated variance may be conveniently written as,

$$IV_{t,t+N} = \int_t^{t+N} \sigma_{g,\tau}^2 d\tau + (A_q^2\varphi_q^2 + A(0)^2) \int_t^{t+N} q_\tau d\tau. \tag{3.25}$$

The integrated variance is, of course, random and not observed until time $t+N$.

The corresponding variance swap rate is defined as the time t risk-neutralized expectation of $IV_{t,t+N}$, say $E_t^{\mathbb{Q}}\left(IV_{t,t+N}\right)$. This risk-neutral expectation may in theory be calculated in a completely model-free fashion from a cross-section of option prices; see, e.g. [36, 32, 72]. This way of calculating the variance swap rate directly mirrors the definition of the (squared) VIX volatility index for the S&P 500,

$$\text{VIX}_t^2 \equiv E_t^{\mathbb{Q}}\left(IV_{t,t+N}\right), \tag{3.26}$$

where the horizon N is set to one month, or 22-days.[12]

This same risk-neutral expected variation may alternatively be calculated within the specific equilibrium model setting. In particular, it follows from equation (C.23) in Appendix C that

$$\text{VIX}_t^2 \;=\; \beta_{vx,0} \;+\; \int_{-\infty}^{t} h_{vx}(t-s)\sqrt{q_s}dW_s^{\sigma} \;+\; \beta_{vx,q}\, q_t, \tag{3.27}$$

where the dependence on N has been suppressed for notational convenience. The $\{h_{vx}(s)\}_{s\in[0,\infty)}$ weights depend on the $\{a(s)\}_{s\in[0,\infty)}$ moving average coefficients, and importantly inherit any long-memory decay in those coefficients. As such, an eventual slow hyperbolic decay in the autocorrelations for the VIX_t^2 would therefore be entirely consistent with the implications from the general theoretical model; i.e.,

$$\text{Corr}(\text{VIX}_t^2, \text{VIX}_{t+s}^2) \;=\; c_h s^{b_h} \qquad s > S, \tag{3.28}$$

where $c_h > 0$ and $b_h < 0$ are constants, and S denotes a sufficiently long lag so that the short-run dependencies have dissipated.

Next, consider the variance risk premium, as formally defined by the difference between the risk-neutral and objective expectation of $IV_{t,t+N}$,

$$vp_t \;\equiv\; \text{E}_t^{\mathbb{Q}}\left(IV_{t,t+N}\right) - \text{E}_t^{\mathbb{P}}\left(IV_{t,t+N}\right). \tag{3.29}$$

Whereas $E_t^Q(IV_{t,t+N})$ and $E_t^P(IV_{t,t+N})$ both depend on the consumption growth volatility and the volatility-of-volatility of that process, the variance risk premium is simply an affine function of the volatility-of-volatility, q_t. Specifically, from equation (C.24) in Appendix C,

$$vp_t \;=\; b_{vp,0} + b_{vp,1} q_t, \tag{3.30}$$

[12] A more detailed description of the mechanical calculation of the VIX index is available in the *white paper* on the CBOE website; see also the discussion in [73].

where $b_{vp,0} >$ and $b_{vp,1} > 0$, reflecting the positive premium for bearing volatility risk provided that $\theta > 0$. Consequently, the vp_t process simply inherits the short-memory dynamic dependencies in the q_t process, and should exhibit a relatively fast exponential decay in its autocorrelation structure,

$$\text{Corr}(vp_t, vp_{t+s}) \;=\; c_q e^{-\kappa_q s} \qquad s > 0, \tag{3.31}$$

where c_q denotes a positive constant.

The equilibrium expressions for the the variance swap rate and the premium in equations (3.27) and (3.30), respectively, also have unique implications for the dynamic cross-correlations with the returns. We review these next.

3.3.2 Return-Volatility Correlations

We are interested in calculating the dynamic cross-correlations between the two variance measures, VIX_t^2 and vp_t, and leads and lags of the returns. To help elucidate the economic mechanisms, it is instructive to first review the predictions under short-memory dynamics, which allow for closed form solutions for the cross-correlations. We subsequently discuss the general case explicitly allowing for long-memory dependencies in the underlying volatility process. The cross-correlations between the variance premium and the returns are easier to calculate than those for the VIX, and we begin by considering these.

Let $r_t \equiv d\log(R_t)$ denote the instantaneous return. We will refer to the cross-correlations between the time t premium vp_t and the future returns, r_{t+s} for $s > 0$, as the forward correlations. The forward correlations represent the extent to which the premium is able to forecast the returns. The correlations between the premium vp_t and the lagged returns, r_{t+s} for $s < 0$, on the other hand, represent the impact of movements in the past returns on the current variance premium. Given the well-known near unpredictability of returns, we would expect the forward correlations to

be positive, reflecting the premium for bearing volatility risk, but small and quickly declining to zero for longer interdaily return horizons. We would expect the lagged correlations to be negative, but increasing to zero for longer daily lags, consistent with the existence of a dynamic leverage type effect. The formal theoretical predictions from the model confirm these ideas.

Specifically, from the results for the short-memory model derived in Appendix B, it follows that for $s > 0$,

$$\text{Corr}(\, vp_t,\, r_{t+s}\,) \;=\; \beta_{R,q}\,\text{Var}(q_t)\,K_q\,e^{-\kappa_q\,s},$$

where $\beta_{R,q}$ represents the sensitivity of the instantaneous returns to the q_t process. Since $\beta_{R,q} > 0$ and $K_q > 0$, the forward correlations should all be positive. Similarly, it follows from the appendix that the cross-correlations for $s < 0$ satisfy,

$$\text{Corr}(\, vp_t,\, r_{t-s}\,) \;=\; (\beta_{R,q}\,\text{Var}(q_t) + A_q\phi_q^2\mu_q)\,K_q\,e^{-\kappa_q\,s}.$$

Since the high-frequency returns are close to unpredictable, the value for $\beta_{R,q}$ is likely to be small. Hence, we would expect the second term involving $A_q < 0$ to dominate the expression in parenthesis, and consequently all of the backward correlations to be negative. In summary, the model predicts,

$$\text{Corr}(vp_t,\, r_{t+s}\,) = \begin{cases} a_-\,e^{-\kappa_q\,|s|} & s < 0, \\ \\ a_+\,e^{-\kappa_q\,s} & s \geq 0, \end{cases} \tag{3.32}$$

where $a_- < 0$ and $a_+ > 0$. As discussed further below, this prediction does indeed adhere very closely with the relevant empirical shapes.

The dynamic cross-correlations between the VIX^2 and the return are a bit more involved than those for the variance premium. However, the basic intuition is essentially the same, except that the actual formulas now also depend on the volatility process $\sigma_{g,t}^2$ itself and its correlation with the returns. In particular, referring to

Appendix B the forward correlations for $s > 0$ takes the form,

$$\text{Corr}(\text{VIX}_t^2, r_{t+s}) \;=\; \beta_{R,\sigma}\,\text{Var}(\sigma_{g,t}^2)K_\sigma\,e^{-\kappa_\sigma\,s} + \beta_{R,q}\,\text{Var}(q_t)\,K_q\,e^{-\kappa_q\,s}.$$

The sign of $\beta_{R,\sigma}$ will depend upon the preference parameters ψ and γ. However, it may reasonably be expected to be positive,[13] so that the forward cross-correlations will again be positive, with the decay toward zero ultimately determined by the dominant value of κ_σ or κ_q. As for the premium, the backward correlations for $s < 0$ are slightly more complicated, taking the form,

$$\text{Corr}(\text{VIX}_t^2, r_{t-s}) = (\beta_{R,\sigma}\,\text{Var}(\sigma_{g,t}^2)+A_\sigma\mu_\sigma)K_\sigma\,e^{-\kappa_\sigma\,s}+(\beta_{R,q}\,\text{Var}(q_t)+A_q\phi_q^2\mu_q)\,K_q\,e^{-\kappa_q\,s}.$$

As discussed above, given the difficulties in predicting returns, we would expect the $\beta_{R,\sigma}$ and $\beta_{R,q}$ terms to be relatively small and dominated by the terms involving the $A_\sigma < 0$ and $A_q < 0$ coefficients determining the instantaneous response of the returns to volatility innovations. Consequently, the backward correlations are naturally expected to be all negative. In summary,

$$\text{Corr}(\text{VIX}_t^2,\, r_{t+s}\,) = \begin{cases} a_{q,-}\,e^{-\kappa_q\,|s|} \;+\; a_{\sigma,-}\,e^{-\kappa_\sigma\,|s|} & s < 0, \\ \\ a_{q,+}\,e^{-\kappa_q\,s} \;+\; a_{\sigma,+}\,e^{-\kappa_\sigma\,s} & s \geq 0, \end{cases} \tag{3.33}$$

where $a_{q,-}, a_{\sigma,-} < 0$ and $a_{q,+}, a_{\sigma,+} < 0$. Again, these theoretical model predictions closely match what we see in the data.

The general model allowing for long-memory in the volatility essentially give rise to the same basic patterns and predictions. However, the formal derivations are somewhat more complicated and the actual values of the cross-correlations will ultimately depend upon the specific process for $\sigma_{g,t}^2$ and the corresponding moving average coefficients $\{a(s)\}_{s\in[0,\infty)}$. We briefly sketch the relevant tools and ideas needed to determine the correlations.

[13] The prototypical values $\psi = 1.5$ and $\gamma = 10$ from [18] implies that $\beta_{R,\sigma} = 7/6$.

The economics of the problem remain exactly the same. The main interactions between the return and volatility are twofold: one consists in the forward effect of volatility innovations on future expected returns, the other involves the feedback effect of lagged return innovations, or the diffusive part of the returns, on current volatility. To elucidate these separate effects within the general model setting, it is useful to define the auxiliary variable

$$r_{t,s} \equiv \begin{cases} \sigma_{g,t+s} dW^c_{t+s} + A_q \varphi_q \sqrt{q_{t+s}} dW^q_{t+s} + A(0) \sqrt{q_{t+s}} dW^\sigma_{t+s} & s < 0, \\ \mu_{R,t+s} & s \geq 0. \end{cases}$$

The variable $r_{t,s}$ is equal to the local diffusive part of the equilibrium return process for $s < 0$, and it equals the local mean of the equilibrium return process for $s \geq 0$, as defined in equation (3.22). As such, the basic shape of the cross-covariances between the variance risk premium vp_t and $r_{t,s}$ directly mirrors that of the cross-covariances with the returns, r_{t+s}. In particular, it follows directly from the expression for vp_t in equation (3.30) that the forward correlations with $r_{t,s}$ must be proportional to the autocovariances of the q_t process. That is for $s > 0$,

$$\text{Cov}(vp_t, r_{t,s}) = K_r e^{-\kappa_q s},$$

where $K_r > 0$ denotes a positive constant of proportionality. To derive the backward correlations, write q_t in integral form,

$$q_t = \varphi_q \int_{u=-\infty}^{t} e^{\kappa_q (u-t)} \sqrt{q_u} dW^q_u.$$

From this expression it follows that for $s < 0$,

$$\text{Cov}(\, vp_t \,,\, r_{t,s}) = e^{-\kappa_q |s|} \text{E}\left(\varphi_q b_{vp,1} \sqrt{q_{t+s}} dW^q_{t+s}, A_q \varphi_q \sqrt{q_{t+s}} dW^q_{t+s}\right) = A_q b_{vp,1} \varphi_q^2 \,\text{E}\,(q_t)\; e^{-\kappa_q |s|}\,,$$

so that all of the backward autocovariances are again negative decaying at an exponential rate under the maintained assumption that the preference parameters $\gamma > 1$

and $\psi > 1$ and therefore $A_q < 0$. The dynamic patterns in the cross-correlations for $r_{t,s}$, of course, mimic those of the cross-covariances. While those for the returns r_{t-s} will be given by an attenuated version of the cross-correlations for $r_{t,s}$, consistent with the implications from the short-memory model summarized in equation (3.32) above.

The theoretical predictions for the dynamic cross-correlations between the VIX and the returns within the general model setting are not quite as clear-cut as those for the variance premium. The q_t process determining the variance risk premium essentially gets confounded with the classical consumption risk premium, and there are also potential side effects from long-range dependence. However, the underlying economic mechanisms remain the same as for the short-memory model, and we would expect a similar pattern of mixed negative backward correlations and positive forward correlations to hold true.

3.4 Empirical Results

The equilibrium framework developed above completely characterizes the dynamic dependencies in the returns and the volatility. Of course, the specific solution of the model will invariably depend upon the choice of preference parameters and the values of the parameters for the underlying consumption growth rate and volatility dynamics. However, the model is obviously somewhat stylized and direct estimation based on actual consumption data would be challenging at best. Instead, we will evaluate the model's direct implications in regards to the autocorrelations and cross-correlations derived in the previous section, and in particular how well the basic patterns implied by the model match those of the actual data depicted in Figures 3.1 and 3.2. We begin in the next subsection with a discussion of the data and pertinent summary statistics underlying the figures.

3.4.1 Data Description

Our tick-by-tick data for the S&P 500 futures contract was obtained from Tick Data Inc. To alleviate the impact of market microstructure "noise" in the calculation of the cross-correlations, and the realized volatility measures discussed further below, we follow the dominant approach in the literature and convert the tick-by-tick prices to equally spaced five-minute observations.[14] With 77 five-minute intervals per trading day and one overnight return, this leaves us with a total of 78 "high-frequency" return observations per day. Standard summary statistics for the corresponding daily returns over the January 2, 1990, through October 31, 2007, sample period and the five-minute returns over the shorter September 23, 2003, to August 31, 2007, sample are reported in the first column in Table 3.1.

The autocorrelations for the VIX in the top panel in Figure 3.1 are based on daily data from January 2, 1990, through October 31, 2007. These data are freely available from the Chicago Board Options Exchange (CBOE).[15] From the summary statistics reported in Table 3.1, the average value of the VIX^2 over the sample equals 32.81, or 16.41 in standard deviation form, but it varies quite considerably over the sample, as indicated by the standard deviation of 23.70. This variation is quite persistent, however, as suggested by the slow decay in the aforementioned autocorrelations depicted in Figure 3.1. This strong persistence is also immediately evident from the actual time series plot of the data in Figure C.1 in the appendix.

The tick-by-tick data for the VIX used in the construction of the cross-correlations shown in the top panel in Figure 3.2 was again obtained from Tick Data Inc. High-

[14] The specific choice of a five-minute smapling frequency strike a reasonable balance between confounding market microstructure effects when sampling too frequently and the loss of important information concerning fundamental price movements when sampling more coarsely; see, e.g., the discussion and references in [12], where the same futures data and five-minute sampling frequency have been used from a different perspective.

[15] The VIX index is reported in annualized units by the CBOE. We convert the series to monthly units using the transformation $VIX_t^2 = 30/365\ VIX_{CBOE,t}^2$.

Table 3.1: Summary Statistics

	r_t	VIX_t^2	$RV_{t,t+22}$	$\hat{v}p_t$
Daily Sampling (1990-2007)				
Mean	11.14	32.81	23.74	8.96
Standard Deviation	15.67	23.70	24.12	12.62
Skewness	-0.05	1.88	2.59	-1.86
Excess Kurtosis	3.75	4.66	8.01	17.53
5-Minute Sampling (2003-2007)				
Mean	8.96	17.48	11.28	6.17
Standard Deviation	11.31	8.40	7.68	5.25
Skewness	0.70	2.90	3.81	1.94
Excess Kurtosis	42.12	15.12	17.90	9.40

Note: The table reports summary statistics for continuously-compounded returns r_t, implied variances VIX_t^2, monthly realized variances $RV_{t,t+22}$, and the variance risk premium $\hat{v}p_t = VIX_t^2 - \widehat{E}RV_{t,t+22}$. The realized variances are constructed from the summation of high-frequency five-minute squared returns. The expectations for the future variances $\hat{E}_t RV_{t,t+22}$ are based on the HAR-RV forecasting model discussed in the text. All the variables are in percentage form. The daily data extend from from January 2, 1990 to October 31, 2007. The five-minute sample spans September 22, 2003 to August 31, 2007.

frequency data for the VIX has only been available since the introduction of the "new" model-free VIX index on September 22, 2003. The relevant summary statistics for the VIX^2 over the shorter September 22, 2003, to August, 2007, high-frequency sample, reported in the bottom part of Table 3.1, are broadly consistent with those over the longer daily sample, and the time series plots in Figures C.1 and C.2 in the appendix also reveal the same basic features. Most notably, the average value is somewhat lower over the more recent sample, and not surprisingly, the kurtosis is substantially higher when the data is sampled at the five-minute frequency.

The integrated variance $IV_{t,t+N}$ defined within the theoretical model is, of course, not directly observable. However, it may be consistently estimated in a completely model-free manner by the corresponding realized variation based on an increasing number of observations over the fixed time-interval $[t, t+N]$; see, e.g.,[10]. As previously noted, to guard against the adverse impact of market microstructure effects when sampling too frequently, we follow the common approach in the literature and

rely on the summation of equally spaced five-minute squared returns.[16] With 77 five-minute intervals per trading day and the overnight return, the N-day-ahead realized variation is then simply given by,

$$RV_{t,t+N} = \sum_{i=1}^{78N} (\log S_{t+i/78} - \log S_{t+(i-1)/78})^2.$$

With the notable exception of a lower mean, the summary statistics for the resulting one-month realized volatility measures $RV_{t,t+22}$ reported in Table 3.1 are generally close to those for the VIX. The corresponding time series plots in Figures C.1 and C.2 are also very similar to those for the VIX.

The higher average value of the VIX compared to the actual realized volatility reflects a premium for bearing volatility risk. Meanwhile, the variance risk premium is formally defined as the difference between the objective and risk-neutralized expectation of the forward integrated variance. While the risk-neutral expectation and the actually observed values of $IV_{t,t+N}$ may both be estimated in a completely model-free fashion by the VIX and the realized volatilities, respectively, the calculation of the objective expectation $E_t(IV_{t,t+N})$ necessitates some mild auxiliary modeling assumptions. Motivated by the results in [11] that simple autoregressive type models estimated directly for the realized volatility typically perform on par with, and often better, than specific parametric modeling approaches designed to forecast the integrated volatility,[17] we will here rely on the HAR-RV model structure first proposed by [40], and subsequently used by [9] among many others, in approximating the objective expectation. Specifically, define the one-day-ahead expectation by the linear

[16] Recent studies, e.g., [101] and [19], have proposed more efficient and complicated ways by which to annihilated the impact of market microstructure effects. However, the simple-to-implement estimator that we use here remains dominant, and importantly allows for easy verification and replication of the results.

[17] [14] have formally shown that for the stochastic volatility models most commonly applied, the loss in efficiency from the use of reduced form autoregressive models for the realized volatility is typically small; see also [94].

projection of the realized volatility on the lagged daily, weekly and monthly realized volatilities,

$$E_t(RV_{t,t+1}) = \beta_{rv,0} + \beta_{rv,1}RV_{t-1,t} + \beta_{rv,2}RV_{t-5,t} + \beta_{rv,2}RV_{t-22,t}.$$

The one-month expectation $E_t(IV_{t,t+22}) = E_t(RV_{t,t+22})$ is then simply obtained by iterating the projection forward.[18]

The summary statistics for the resulting variance risk premium $\hat{v}p_t = VIX_t^2 - \widehat{E}RV_{t,t+22}$, reported in the last column in Table 3.1, confirm the positive expected return for selling volatility, but also show that the magnitude of the premium varies substantially over time. At the same time, the plots in Figures C.1 and C.2 indicate much less persistent dependencies in the premium than for the VIX and the realized volatilities. Of course, the previously discussed autocorrelations in Figure 3.1 already confirm this important difference between the premium and the VIX. We next turn to our discussion of the model's ability to match this basic feature along with the other key dynamic dependencies observed in the data.

3.4.2 Model Implied Auto- and Cross-Correlations

A full characterization of the model-implied autocorrelations for the integrated volatility would require that all of the moving average weights $\{a(s)\}_{s\in[0,\infty)}$ in equation (3.12) be completely specified. Importantly, however, as discussed in Section 3.3.1, any long-run dependencies in these coefficients are directly translated to similar long-run dependencies in the moving average weights $\{h_{vx}(s)\}_{s\in[0,\infty)}$ describing the equilibrium process for the VIX_t^2 in equation (3.27). The top panel in Figure 3.3 shows the best fitting model-implied autocorrelations from estimating the slowly hyperbolic decaying autocorrelation structure in (3.28) to the actual daily sample autocorrela-

[18] The actual estimates for the β's are directly in line with the results reported in the extant literature and available upon request.

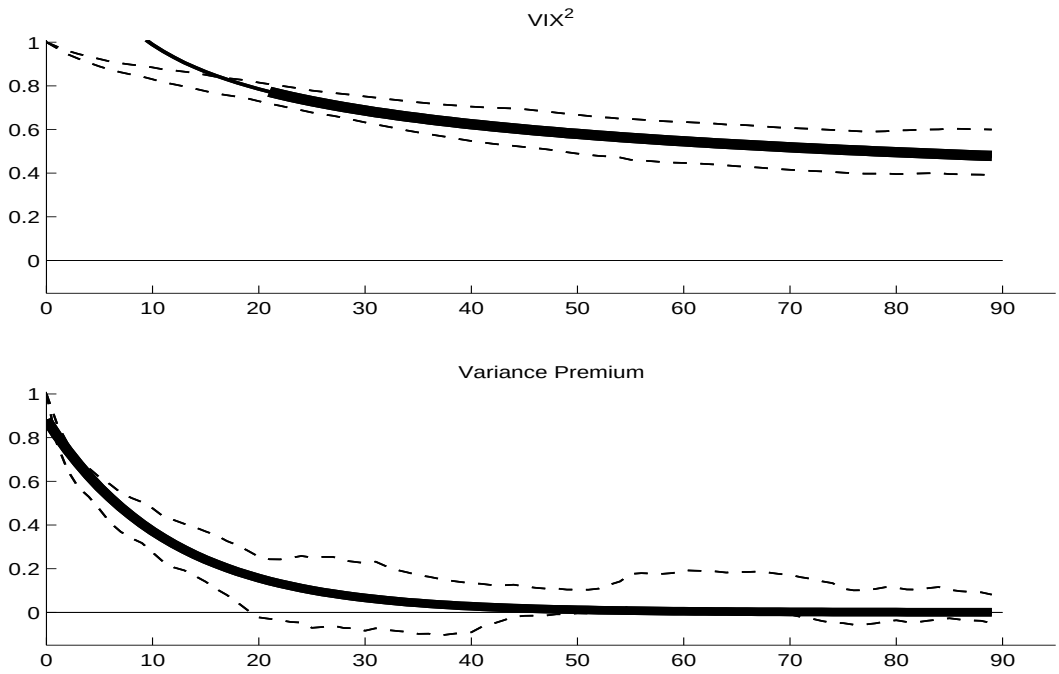

FIGURE 3.3: Model Implied Autocorrelations.The top panel shows the autocorrelations for the VIX^2 volatility index to a lag length of 90 days. The solid line gives the model implied autocorrelations under the assumption of long-memory in the underlying fundamental volatility process. The bottom panel shows the autocorrelations for the variance risk premium *vp*. The solid line gives the model implied autocorrelations. The pair of dashed lines included in both panels represent ninety-five percent confidence intervals for the corresponding sample autocorrelations based on daily data from 1990 through 2007.

tions for the VIX starting at a lag length of $S = 22$.[19] The figure also shows the conventional ninety-five percent confidence intervals for the sample autocorrelations. The model-implied autocorrelations do a remarkable job at describing the long-run dependencies inherent in the VIX, always falling well inside the confidence bands. Of course, the fit does *not* match at all well if extrapolated to the 1-22 day interval, which is to be expected.

One of the key predictions of the theoretical model is that the equilibrium volatil-

[19] The R^2 from estimating the functional relationship is an impressive 0.993, although this value should be carefully interpreted because of the strong serial correlation in the residuals from the fit.

ity risk premium is an affine function of the volatility-of-volatility, and thereby is short memory, despite the fact that the integrated volatility and its risk neutral expectation may both display long-memory dependencies. Intuitively, as discussed above, everything except for the volatility-of-volatility gets risk neutralized out in equation (3.30). This, of course, is consistent with the shape of the autocorrelation function displayed in Figure 3.1, which in sharp contrast to the one for the VIX dies out relatively quickly. As a more formal verification of this distinct implication form the model, we fitted the functional form in equation (3.31) to the actual sample autocorrelations for the premium. The fit shown in the bottom panel of Figure 3.3 is again excellent, and the model-implied autocorrelations easily fall within the ninety-five percent confidence intervals over the entire 1-90 day range.[20]

The observed volatility feedback and leverage effects evident in the dynamic cross-correlations in Figure 3.2 arguably present the more challenging and difficult to explain empirical dependencies. Consider first the cross-correlations for the variance risk premium. The negative backward correlations start out at a slightly larger absolute value than the forward correlations, and both decay toward zero at what appears to be an exponential rate. This apparent pattern in the cross-correlations between the premium and leads and lags of the returns is entirely consistent with the model-implied correlations summarized in (3.32). The bottom panel in Figure 3.4 shows the resulting fit along with the ninety-five percent confidence intervals for the sample cross-correlations.[21] The theoretical model obviously delivers very accurate predictions for the actually observed dynamic dependencies between the returns and the variance risk premium.

The theoretical predictions for the VIX-return cross-correlations are not quite as

[20] The fitted R^2 equals 0.934.

[21] The fitted R^2's for the backward and forward correlations equal 0.907 and 0.678, respectively. Of course, some of the sample cross-correlations used in the fit are not statistically different from zero.

clear-cut as those for the premium. As discussed in Section 3.3.2 above, the volatility risk premium in effect gets confounded with the classical consumption risk premium, and within the general theoretical model setting there may also be potential side effects from long-range dependence. In parallel to the model-implied autocorrelations for the VIX, a complete characterization of these separate effects would require that the underlying fundamental consumption growth rate volatility process $\sigma_{g,t}^2$ and the corresponding moving average weights $\{a(s)\}_{s\in[0,\infty)}$ be fully specified. Short of such a specification, the basic pattern and decay in the cross-correlations may naturally be expected to adhere to the functional form in (3.33). The top panel in Figure 3.4 shows the resulting fit to the sample cross-correlations.

Comparing the observed backward correlations for the VIX in the left part of Figure 3.2 to those for the variance premium in the bottom panel, the dynamic leverage effect is clearly more prolonged for the VIX. This slower decay is very well described by the mixed exponential functions shown in Figure 3.4. At the same time, the differences between the forward correlations for the VIX and the premium, and in turn the impact on future returns attributable to the classical consumption risk premium and mean-variance tradeoff, appear less pronounced. In fact, the relatively fast decay rates in the empirically observed forward correlations are well described by a single exponential function for both the premium and the VIX.[22].

Summing up the empirical results, the qualitative implications from the new theoretical model do an admirable job in terms of matching the key dynamic dependencies in the aggregate market returns and volatilities. The previously documented autocorrelations for the volatility and volatility risk premium and the puzzling high-frequency based cross-correlation patterns in Figures 3.1 and 3.2, may all be explained by the model, with the model predictions well within conventional statistical

[22] The fitted R^2's for a double exponential for the VIX backward correlations and a single exponential for the VIX forward correlations equal 0.941 and 0.709, respectively.

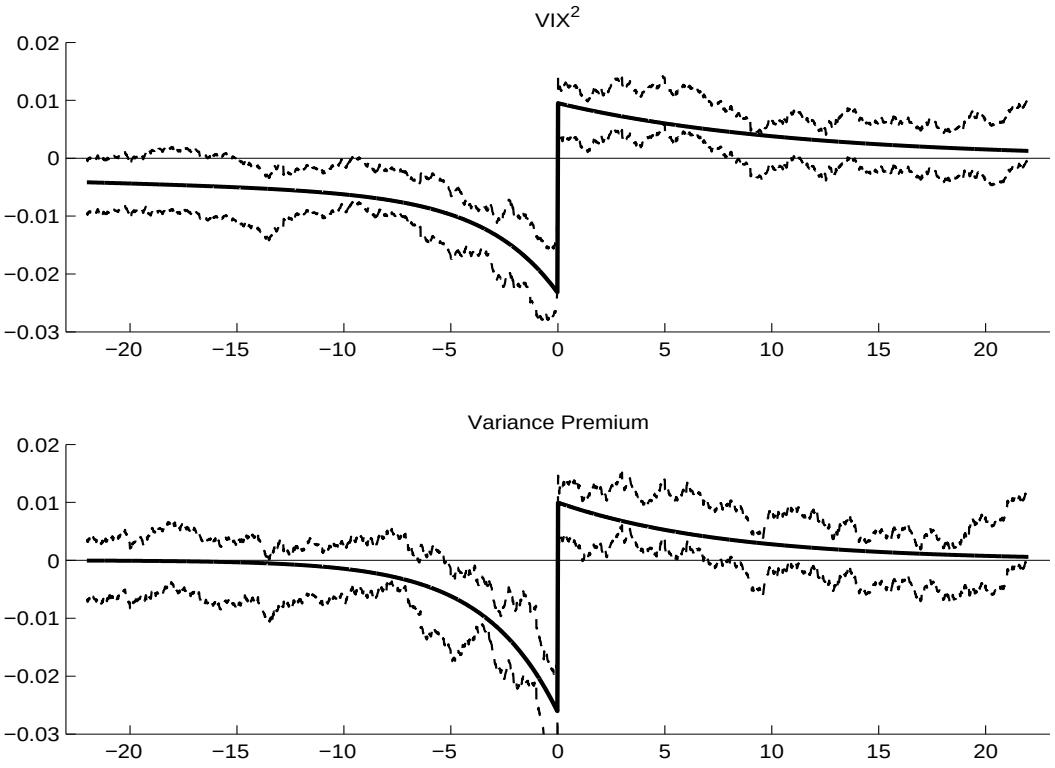

FIGURE 3.4: Model Implied Cross-Correlations. The top panel shows the sample cross-autocorrelations between the VIX^2 volatility index and lags and leads of the returns ranging from -22 to 22 days. The bottom panel shows the cross-correlations between the variance risk premium vp and the returns. The solid lines give the cross-correlations implied by the theoretical model. The pair of dashed lines represent ninety-five percent confidence intervals for the corresponding sample cross-correlations based on high-frequency 5-minute observations from 2003 through 2007.

confidence intervals.

3.5 Conclusion

The aggregate stock market volatility exhibits long-range dependence, while the variance risk premium, defined as the difference between the objective and risk-neutral expectation of the forward variance, shows much less persistence. Consistent with the well documented leverage effect, there is also a distinct and prolonged asymmetry in the relationship between volatility and past and future returns. We provide

a first self-contained equilibrium based explanation for these empirical facts. The return on the aggregate market defined within the new model depends not only on the prospects of future economy growth, but also on the current uncertainty about the future economic conditions, and importantly explains the presence of a separate premium for bearing variance risk.

The dynamic return-volatility dependencies that motivate our theoretical model were all well-established prior to the advent of the current financial crises. Although the same basic dependencies have been shown to hold true during other market downturns and recessions, the scope of the current crises is different. We look forward to the opportunity of analyzing more recent data and formally test the model's ability to explain the dependencies observed therein.

Our explanation of the empirical facts is entirely risk-based, and depends critically on the temporal variation in the variance risk premium defined within the model. The wedge between the objective and risk-neutral expectation of the forward variance may alternatively be interpreted as a proxy for the aggregate degree of risk aversion in the market, and any temporal variation in the empirically observed variance risk premium thus indicative of changes in the way in which systematic risk is valued; see, e.g.,[1, 27, 64]. Although it might be difficult to contemplate systematic changes in the level of risk aversion at the frequencies emphasized here, time-varying volatility risk *and* time-varying attitudes toward risk likely both play a role in explaining the temporal variation in expected returns and risk premia; e.g., [23]. It would be interesting to extend the model developed here to somehow allow for changes in the underlying preference parameters and risk-attitudes to more clearly delineate these effects. We conjecture that this may be especially important in fully understanding the return-volatility dependencies observed during the current financial crises.

Appendix A

Appendix to Chapter 1

A.1 Moments of Volatility Measures

- To keep formulas in this appendix parsimonious, we use the following notation. For multi-factor models of the form (1.3) we denote a weighted average of an arbitrary function $f(k_i), i = \overline{1,p}$ over all factors with a bar operator:

$$\overline{f}(k) = \frac{\sum_{i=1}^{p} a_i^2 f(k_i)}{\sum_{i=1}^{p} a_i^2},$$

  where $k_i, i = \overline{1,p}$ are persistence parameters and $a_i, i = \overline{1,p}$ are factor weights. Notably any function of this type depends only on persistence parameters k_i, and ratios $\frac{a_i}{a_1}$, $i = \overline{2,p}$.

- The following moments for the integrated variance were derived by [14]:

$$\mathrm{Cov}(IV_t^{t+1}, \sigma_t^2) = \overline{\frac{1-e^{-k}}{k}} \mathrm{Var}\sigma_t^2, \tag{A.1}$$

$$\mathrm{Var}IV_t^{t+1} = 2\mathrm{Var}\sigma_t^2 \overline{\left[\frac{1}{k} - \frac{1-e^{-k}}{k^2}\right]}, \tag{A.2}$$

$$\mathrm{Cov}(IV_t^{t+1}, IV_{t+l}^{t+l+1} | l > 0) = \mathrm{Var}\sigma_t^2 \overline{\left(\frac{1-e^{-k}}{k}\right)^2 e^{-k(l-1)}}.$$

Hence, the correlation structure of the integrated variance depends only on the persistence parameters k_i and ratios $\frac{a_i}{a_1}$, $i = \overline{2,p}$.

- For the no-leverage case, [82] obtained the following moments of the realized variance:

$$\mathrm{Cov}(RV_t^{t+1}, RV_{t-j}^{t-j+1}) = \mathrm{Cov}(IV_t^{t+1}, IV_{t-j}^{t-j+1}), \quad j \neq 0,$$
$$\mathrm{Var}RV_t^{t+1} = \mathrm{Var}IV_t^{t+1} + \mathrm{Var}u_{t+1}, \tag{A.3}$$

where the last component is the variance of the noise u_{t+1} in RV_t^{t+1}. Its variance has been derived by [21]:

$$\mathrm{Var}u_{t+1} = 2h\left[E^2\sigma^2 + \frac{Var(\sigma^2)}{h^2}\int_0^h\int_0^h \phi(s-\tau)dsd\tau\right] =$$
$$= 2h\mathrm{Var}(\sigma_t^2)\left[\frac{1-\lambda}{\lambda} + \frac{2}{h^2}\overline{(\frac{h}{k} - \frac{1-e^{-kh}}{k^2})}\right]. \tag{A.4}$$

- We also will need to find the second and fourth moments of the returns. Denote the h-period demeaned return by $\xi_{t+h} = h^{-1/2}\int_t^{t+h}\sigma_\tau dW_\tau^p$. From Itô–isometry for square-integrable prices, it follows that

$$E\xi_{t+h}^2 = \frac{1}{h}E\int_t^{t+h}\sigma_\tau^2 d\tau = E\sigma_t^2.$$

Similar to the proof of Itô–isometry, we can show that for the square-integrable spot variance under the no-leverage condition, the forth moment of the return

equals

$$E\xi_{t+h}^4 = \frac{3}{h^2}E\left[\int_t^{t+h}\int_t^{t+h}\sigma_\tau^2\sigma_s^2 d\tau ds\right] = 3E^2\sigma_t^2 + 3\mathrm{Var}\sigma_t^2\frac{\int_t^{t+h}\int_t^{t+h}\phi(\tau,s)d\tau ds}{h^2},$$

$$E[\xi_{t+h-jh}^2\xi_{t+h-ih}^2]_{i\neq j} = \frac{1}{h^2}E\left[\int_{t-(j+1)h}^{t-jh}\int_{t-(i+1)h}^{t-ih}\sigma_\tau^2\sigma_s^2 d\tau ds\right] =$$

$$= E^2\sigma_t^2 + \mathrm{Var}\sigma_t^2\frac{\int_{t-(j+1)h}^{t-jh}\int_{t-(i+1)h}^{t-ih}\phi(\tau,s)d\tau ds}{h^2}.$$

A variant of the above is proved by [81] for ESV-models. From the ESV-representation, it follows that the correlation function of the square-integrable process can be decomposed in the form $\mathrm{corr}(\sigma_{t+h}^2,\sigma_t^2) = \frac{\sum_{i=1}^p a_i^2 e^{-k_i h}}{\sum_{i=1}^p a_i^2}$. Therefore, for ESV-models $\phi(t,s) = \overline{e^{-k|t-s|}}$ and

$$E\xi_{t+h}^4 = 3E^2\sigma_t^2 + 6\frac{\mathrm{Var}\sigma_t^2}{h^2}\overline{\frac{h}{k} - \frac{1-e^{-kh}}{k^2}},$$

$$E[\xi_{t+h-jh}^2\xi_{t+h-ih}^2]_{i\neq j} = E^2\sigma_t^2 + \frac{\mathrm{Var}\sigma_t^2}{h^2}\overline{\left[\frac{1-e^{-kh}}{k}\right]^2 e^{-k(|j-i|-1)h}}. \qquad \text{(A.5)}$$

- Finally, we will derive the covariance structure between factors $P_i(x_t)$ and past h-period squared returns ξ_{t-jh}^2. Similarly to Itô–isometry, since covariances are preserved under L^2 convergence, we can show that

$$\mathrm{cov}(P_i(x_t),\xi_{t-jh}^2) = \frac{1}{h}\int_{t-(j+1)h}^{t-jh}\mathrm{cov}[P_i(x_t),\sigma_\tau^2]d\tau =$$

$$= \frac{1}{h}\int_{t-(j+1)h}^{t-jh}\mathrm{cov}[P_i(x_t),P_i(x_\tau)]d\tau = \frac{a_i^2}{h}\int_{t-(j+1)h}^{t-jh}e^{-k_i|t-\tau|}d\tau = \qquad \text{(A.6)}$$

$$= \frac{a_i^2}{h}e^{-k_i jh}\frac{1-e^{-k_i h}}{k_i}.$$

Summing up over the factors, we obtain the formula for the covariance of the

spot variance and past squared returns:

$$\text{cov}(\sigma_t^2, \xi_{t-jh}^2) = \text{Var}\sigma_t^2 \overline{e^{-kjh} \frac{1 - e^{-kh}}{kh}}. \tag{A.7}$$

A.2 Proof of Proposition 1

In this section we identify a complete set of parameters that affect the comparison of the model-based forecast and the reduced-form forecast. The model-based forecast employs an arbitrary stochastic volatility (SV) model with a variance dynamics of the form (1.19). The GMM approach is used to estimate the parameters, and the states are extracted by the efficient ARCH-filter. The reduced-form forecast employs the ARMA(p,p) model for realized variances RV to predict IV.

For the efficient ARCH-filter, $\hat{\sigma}_t^2$ follows the discrete time GARCH(1,1):

$$\hat{\sigma}_{t+h}^2 = \phi_h + a_h \sigma_t^2 + b_h \xi_{t+h}, \tag{A.8}$$

where $\xi_{t+h} = \frac{p_{t+h} - p_t - \mu h}{\sqrt{h}}$. The parameters of the model are chosen optimally to minimize asymptotic MSE.

Proposition 1. *Let σ_t^2 be square integrable with the correlation function: $corr(\sigma_{t+h}^2, \sigma_t^2) = \frac{\sum_{i=1}^p a_i^2 e^{-k_i h}}{\sum_{i=1}^p a_i^2}$. Suppose we apply the efficient ARCH-filter of the form (1.27) to extract the spot variance. The log price process is described by (2.27) with no leverage effect and zero drift. Then the comparison of the reduced-form forecast and the forecast based on a one-factor model depends only on the following set of parameters Ξ:*

- *Volatility-of-volatility – $\lambda = \frac{Var\sigma_t^2}{E\sigma_t^4}$;*
- *Relative weights of factors: $\frac{a_i}{a_1}, i = \overline{2, p}$;*
- *Persistence of factors: $k_i, i = \overline{1, p}$;*

- *Sampling frequency – h.*

Proof. The proof is organized as follows. First, we find the total MSPE of the "no-bias" model-based forecast. Second, we find the total MSPE of the reduced-form forecast. Finally, we will compare the errors and consider a few extensions, model-based forecast with a bias and multi-period forecasts.

A.2.1 MSPE of the Model-Based Forecast

As was shown in Section 1.3, the total MSPE for the "no-bias" forecast based on a one-factor model is the sum of the "genuine" forecast error:

$$\text{GFE}^{model} = E\left[IV_t^{t+1} - E\sigma_t^2 - \frac{\text{cov}(IV_t^{t+1}, \sigma_t^2)}{\text{Var}\sigma_t^2}(\sigma_t^2 - E\sigma_t^2)\right]^2, \tag{A.9}$$

the part that is due to the error in spot variance:

$$F(\hat{\sigma}_t^2 - \sigma_t^2) = E\left[E\sigma_t^2 + \frac{\text{cov}(IV_t^{t+1}, \sigma_t^2)}{\text{Var}\sigma_t^2}(\sigma_t^2 - E\sigma_t^2) - \tilde{\theta} - \frac{1 - e^{-\tilde{k}}}{\tilde{k}}(\hat{\sigma}_t^2 - \tilde{\theta})\right]^2, \tag{A.10}$$

and twice the covariance between the above two components:

$$\text{Cov}(\text{GFE}^{model}, F(\hat{\sigma}_t^2 - \sigma_t^2)) = \text{Cov}(IV_t^{t+1} - \frac{\text{cov}(IV_t^{t+1}, \sigma_t^2)}{\text{Var}\sigma_t^2}\sigma_t^2, \frac{\text{cov}(IV_t^{t+1}, \sigma_t^2)}{\text{Var}\sigma_t^2}\sigma_t^2 - \frac{1 - e^{-\tilde{k}}}{\tilde{k}}\hat{\sigma}_t^2). \tag{A.11}$$

The latter part is zero if the true model is one-factor. In this case, the "genuine" error is independent of all the past information and, therefore, of the error in the spot variance. However, under model misspecification the last term is no longer a null.

Since, by assumption, the model may be misspecified, the limits for parameter estimates $\tilde{\theta}, \tilde{k}$, and therefore part (A.10), will depend on the estimation method. For

now, we restrict our consideration to the "no-bias" case. Therefore, the limits for the parameters are given by the set of moments(1.23) and (1.24). And using formula (A.1), those moments imply that

$$\frac{1-e^{-\tilde{k}}}{\tilde{k}} = \overline{\frac{1-e^{-k}}{k}}. \tag{A.12}$$

Therefore, $\tilde{k}$ is a function of Ξ only. Moreover, part (A.10) simplifies to

$$F(\hat{\sigma}_t^2 - \sigma_t^2) = \left[\frac{\mathrm{cov}(IV_t^{t+1}, \sigma_t^2)}{\mathrm{Var}\sigma_t^2}\right]^2 E\left(\sigma_t^2 - \hat{\sigma}_t^2\right)^2.$$

Further analysis of the model-based forecast error will be organized into three steps. First, we will devise the formula for the "genuine" error (A.9). Second, we will find $F(\hat{\sigma}_t^2 - \sigma_t^2)$ from (A.10). Third, we will provide the formula for the covariance term (A.11).

Step 1: "Genuine" forecast error GFE^{model}

The expression for R^2 from the Mincer-Zarnowitz regression obtained by [14] can be readily converted into the expression for the corresponding mean squared error:

$$\mathrm{GFE}^{model} = \mathrm{Var}IV_t^{t+1}(1-R^2) = \mathrm{Var}IV_t^{t+1} - \frac{1}{\mathrm{Var}\sigma_t^2}\left[\sum_{i=1}^{p} a_i^2 \frac{1-e^{-k_i}}{k_i}\right]^2.$$

Substituting the moment (A.2) into the above expression yields the formula for the "genuine" error:

$$\mathrm{GFE}^{model} = \mathrm{Var}\sigma_t^2 \left(2\overline{\left[\frac{1}{k} - \frac{1-e^{-k}}{k^2}\right]} - \overline{\left[\frac{1-e^{-k}}{k}\right]}^2\right). \tag{A.13}$$

For comparison, we derive the "genuine" error of a forecast based on a multi-factor

model. It also follows from the expression for R^2 from the Mincer-Zarnowitz regression obtained by [14] and equals

$$\text{GFE}^{\text{p-F}model} = \text{Var}\sigma_t^2 \left(2\overline{\left[\frac{1}{k} - \frac{1-e^{-k}}{k^2}\right]} - \left[\frac{1-e^{-k}}{k}\right]^2 \right). \tag{A.14}$$

Step 2: Error due to volatility estimation

Rewrite the filter for the spot volatility in the following form:

$$\hat{\sigma}_t^2 = b_h \sum_{j=0}^{\infty} a_h^j \xi_{t-jh}^2 + \frac{\phi_h}{1-a_h}, \tag{A.15}$$

where the innovation is equal to $\xi_{t+h} = \frac{\int_t^{t+h} \sigma_\tau dW_\tau^p}{\sqrt{h}}$. That is, our estimation of the spot volatility is a weighted average of past squared demeaned returns and a constant that later will be defined to converge to zero as $h \to 0$.

After squaring the estimate $\hat{\sigma}_t^2$ and taking the expectation, we find that the second moment of the estimator $\hat{\sigma}_t^2$ equals

$$E\hat{\sigma}_t^4 = \frac{\phi_h^2}{(1-a_h)^2} + 2\frac{\phi_h}{1-a_h} b_h \sum_{j=0}^{\infty} a_h^j E(\xi_{t-j}^2) + b_h^2 E(\sum_{j=0}^{\infty} a_h^j \xi_{t-j}^2)^2.$$

After substituting moments from (A.5) we find that

$$E(\hat{\sigma}_t^4) = \frac{\phi_h^2}{(1-a_h)^2} + \frac{2\phi_h b_h}{(1-a_h)^2} E\sigma_t^2 + 3\frac{b_h^2}{1-a_h^2} \left[E^2\sigma_t^2 + 2\text{Var}\sigma_t^2 \overline{\frac{\frac{h}{k} - \frac{1-e^{-kh}}{k^2}}{h^2}} \right]$$
$$+2\frac{b_h^2}{1-a_h^2} \left[\frac{E^2\sigma_t^2 \, a_h}{1-a_h} + \text{Var}\sigma_t^2 \overline{\frac{(1-e^{-kh})^2}{k^2h^2} \frac{a_h}{1-a_h e^{-kh}}} \right]. \tag{A.16}$$

Analogously, the corresponding cross product equals

$$E(\hat{\sigma}_t^2 \sigma_t^2) = \frac{\phi_h}{1-a_h} E\sigma_t^2 + b_h \sum_{j=0}^{\infty} a_h^j E(\xi_{t-jh}^2 \sigma_t^2).$$

Substituting the correlation term from equation (A.7) yields that

$$E(\hat{\sigma}_t^2\sigma_t^2) = \frac{\phi_h}{1-a_h}E\sigma_t^2 + \frac{b_h}{1-a_h}E^2\sigma_t^2 + b_h\mathrm{Var}\sigma_t^2\overline{\frac{1-e^{-kh}}{kh}\frac{1}{1-a_he^{-kh}}}. \tag{A.17}$$

Thus, the mean-squared error in the spot variance estimate equals

$$E(\hat{\sigma}_t^2 - \sigma_t^2)^2 = E(\hat{\sigma}_t^4) + E\sigma^4 - 2E(\hat{\sigma}_t^2\sigma_t^2). \tag{A.18}$$

Under what conditions is the estimate $\hat{\sigma}_t^2$ consistent? Suppose, $\phi(h) = O(h)$, $b_h = O(\sqrt{h})$, $a_h + b_h = 1 - O(h)$. Then, taking limits of the expressions (A.16),(A.17) as $h \to 0$ yields that:

$$\lim_{h\to 0} E(\hat{\sigma}_t^4|k) = E\sigma^4,$$

$$\lim_{h\to 0} E(\hat{\sigma}_t^2\sigma_t^2|k) = E\sigma^4.$$

Hence, $\lim_{h\to 0} E(\hat{\sigma}_t^2 - \sigma_t^2|k)^2 = 0$. Therefore, any GARCH filter satisfying the assumptions outlined above is consistent, irrespective of the model. This result was first proved by Nelson(1992) for a general class of ARCH-filters.

What set of coefficients $a_h, b_h, \phi(h)$ renders the most efficient estimate $\hat{\sigma}_t^2$? To answer this question, we minimize the error (A.18) approximated around $h = 0$. Asymptotically, it holds that

$$E(\hat{\sigma}_t^2 - \sigma_t^2|k)^2 \approx hVar\sigma^2\frac{\bar{k}}{b_h} + b_hE\sigma^4,$$

$$\bar{k} = \frac{\sum_{i=1}^p a_i^2k_i}{\sum_{i=1}^p a_i^2}.$$

The above error is minimized at $b_h = \sqrt{\frac{Var\sigma^2}{E\sigma^4}\bar{k}h} = \sqrt{\lambda\bar{k}h}$. For any one-factor model, the efficient coefficients are $b_h = \sqrt{\lambda kh}$, $a_h = 1 - b_h - kh$,and $\phi_h = k\theta h$. After plugging in the limit for the estimate of k given by (A.12), the "optimal" coefficients

take values

$$\begin{aligned} b_h &= \sqrt{\lambda \tilde{k} h}, \\ a_h &= 1 - b_h - \tilde{k} h, \\ \phi_h &= \tilde{k} \theta h. \end{aligned} \tag{A.19}$$

Step 3: Covariance term.

Since the covariance between the genuine forecast error and the spot variance is zero for the unbiased forecast, the covariance term (A.11) simplifies to:

$$\mathrm{Cov}(\mathrm{GFE}^{model}, F(\hat{\sigma}_t^2 - \sigma_t^2)) = -\frac{1 - e^{-\tilde{k}}}{\tilde{k}} \mathrm{Cov}(IV_t^{t+1} - \frac{\mathrm{cov}(IV_t^{t+1}, \sigma_t^2)}{\mathrm{Var}\sigma_t^2} \sigma_t^2, \hat{\sigma}_t^2).$$

The covariance can be simplified further, since the integrated variance inside the covariance can be replaced by its conditional expectation $E_t(IV_t^{t+1}) = \theta +$
$+ \sum_{i=1}^{p} \frac{1-e^{-k_i}}{k_i} P_i(x_t)$.

$$\mathrm{Cov}(\mathrm{GFE}^{model}, F(\hat{\sigma}_t^2 - \sigma_t^2)) = -\frac{\overline{1 - e^{-k}}}{k} \mathrm{Cov}\left[\sum_{i=1}^{p} [\frac{1 - e^{-k_i}}{k_i} - \frac{\overline{1 - e^{-k}}}{k}] P_i(x_t), \hat{\sigma}_t^2 \right].$$

Combining the formula (A.6) with the definition of the filter (A.15) yields the expression for the above covariance:

$$\mathrm{Cov}(P_i(x_t), \hat{\sigma}_t^2) = b_h \frac{a_i^2}{h} \frac{1 - e^{-k_i h}}{k_i} \sum_{j=0}^{\infty} a_h^j e^{-k_i j h} = b_h a_i^2 \frac{1 - e^{-k_i h}}{k_i h} \frac{1}{1 - a_h e^{-k_i h}}.$$

Hence, the covariance between the genuine error and the error in $\hat{\sigma}_t^2$ is equal to

$$\begin{aligned} &\mathrm{Cov}(\mathrm{GFE}^{model}, F(\hat{\sigma}_t^2 - \sigma_t^2)) = \\ &\quad = \mathrm{Var}\sigma_t^2 \frac{\overline{1 - e^{-k}}}{k} \overline{\left(\left[\frac{\overline{1 - e^{-k}}}{k} - \frac{1 - e^{-k}}{k} \right] \frac{1 - e^{-kh}}{kh} \frac{b_h}{1 - a_h e^{-kh}} \right)}. \end{aligned} \tag{A.20}$$

Total MSPE of the model-based forecast.

This is the final step, in which we assemble together three parts of the total error of the unbiased model forecast and substitute for the efficient parameters a_h, b_h, ϕ_h from (A.19). Also, the expectation of the spot variance which enters formulas (A.16, A.17) is replaced by $\sqrt{\frac{1-\lambda}{\lambda}\text{Var}\sigma_t^2}$.

$$\text{Total MSPE} = \text{GFE}^{model} + F(\hat{\sigma}_t^2 - \sigma_t^2) + 2\text{Cov}(\text{GFE}^{model}, F(\hat{\sigma}_t^2 - \sigma_t^2))$$

$$\text{GFE}^{model} = \text{Var}\sigma_t^2 \left(2\overline{\left[\frac{1}{k} - \frac{1-e^{-k}}{k^2}\right]} - \overline{\left[\frac{1-e^{-k}}{k}\right]^2} \right)$$

$$F(\hat{\sigma}_t^2 - \sigma_t^2) = \left[\frac{1-e^{-\tilde{k}}}{\tilde{k}}\right]^2 E(\hat{\sigma}_t^2 - \sigma_t^2)^2 \qquad \text{(A.21)}$$

$$\text{Cov}(\text{GFE}^{model}, F(\hat{\sigma}_t^2 - \sigma_t^2)) =$$

$$= \text{Var}\sigma_t^2 \overline{\frac{1-e^{-k}}{k}} \overline{\left([\overline{\frac{1-e^{-k}}{k}} - \frac{1-e^{-k}}{k}] \frac{1-e^{-kh}}{kh} \frac{b_h}{1-a_h e^{-kh}} \right)}$$

where:

$$\frac{E(\hat{\sigma}_t^2 - \sigma_t^2)^2}{\text{Var}\sigma_t^2} =$$

$$= \frac{1-\lambda}{\lambda} \left(\frac{\tilde{k}^2 h^2}{1-a_h^2} + \frac{2\tilde{k}hb_h a_h}{(1-a_h)^2} - 2\frac{b_h}{1-a_h} \right) - 2b_h \overline{\frac{1-e^{-kh}}{kh} \frac{1}{1-a_h e^{-kh}}} + \frac{1}{\lambda} +$$

$$+ \frac{b_h^2}{1-a_h^2} \left[\frac{1-\lambda}{\lambda} \frac{3-a_h}{1-a_h} + 6\overline{\frac{\frac{h}{k} - \frac{1-e^{-kh}}{k^2}}{h^2}} + 2\overline{\frac{(1-e^{-kh})^2}{k^2h^2} \frac{a_h}{1-a_h e^{-kh}}} \right], \quad \text{(A.22)}$$

$$
\begin{aligned}
b_h &= \sqrt{\lambda \tilde{k} h}, && \text{(A.23)} \\
a_h &= 1 - b_h - \tilde{k} h, && \text{(A.24)} \\
\frac{1 - e^{-\tilde{k}}}{\tilde{k}} &= \overline{\frac{1 - e^{-k}}{k}}.
\end{aligned}
$$

From the expression above, it follows that the total MSPE of the model-based forecast is proportional to the variance of the spot variance and the coefficient of proportionality is a function of parameters Ξ only.

A.2.2 MSPE of the Reduced-Based Forecast

In this part, we will estimate the error from reduced-form forecasting. As follows from (1.15) the total Mean-Squared Prediction Error for the no-leverage case is decomposed in the following way:

$$
\text{Total MSPE} = \text{Var}\eta_{t+1}(h) - Var u_{t+1}^2 + E^2 u_{t+1}
$$

where the first component is an innovation in the ARMA representation for realized variance (1.14). The second component u_{t+1} is the difference between the realized variance and the integrated variance. Its variance is expressed by (A.4) and expectation is typically negligible, and is zero for the case of no-drift in returns.

To obtain the variance of the innovation in the ARMA-representation for RV, convert ARMA into an infinite AR-representation:

$$
\left[\frac{\prod_{i=1}^{p} (1 - e^{-k_i} L)}{1 - \sum_{i=1}^{p} \beta_i L^i} \right] (RV_t^{t+1} - \theta) = \eta_{t+1}(h). \qquad \text{(A.25)}
$$

Slope coefficients in the ARMA representation β_i are naturally functions of the correlation structure of RV. On the other hand, from formulas for the moments of IV and RV (A.1 - A.4) the correlation structure of the realized variance depends only on the parameters Ξ. Hence, the parameters of the ARMA representation are the functions

of only Ξ. In particular, [82] derived coefficients for the ARMA representation of realized variance for the case of one-factor and two-factor models.

Then it follows that the variance of the innovation is equal to

$$\mathrm{Var}\eta_{t+1}(h) = \mathrm{Var}\left(\left[\frac{\prod_{i=1}^{p}(1-e^{-k_i}L)}{1-\sum_{i=1}^{p}\beta_i L^i}\right] RV_t^{t+1}\right) = \mathrm{Var}\sigma_t^2\Psi(\Xi). \tag{A.26}$$

The exact formulas for the above variance in the case of one-factor and two-factor models are also derived by [82].

Summing up the parts of the error and substituting the variance of the noise (A.4), the total error of the reduced-form forecast equals:

$$E(IV_t^{t+1} - P(IV_t^{t+1}|RV))^2 = \mathrm{Var}\sigma_t^2\left[\Psi(\Xi) - 2h\frac{1-\lambda}{\lambda} - \frac{2}{h^2}\overline{(\frac{h}{k} - \frac{1-e^{-kh}}{k^2})}\right]. \tag{A.27}$$

A.2.3 Forecast Comparison

Hence, the total MSPE of both forecasts is proportional to the variance of the spot $\mathrm{Var}\sigma_t^2$. The corresponding coefficients of proportionality derived in (A.21) and (A.27) are functions of the coefficients Ξ. Hence, the comparison of the forecasts depends only on the parameters from the set Ξ . □

A.2.4 Extension 1: Bias in the Model-Based Forecast

In the above proof, we defined the parameter $\tilde{k}$ in a way that ensures that the resulting model-based forecast is unbiased. In general, we may assume that the estimated persistence parameter $\tilde{k}$ satisfies another moment condition:

$$f(\tilde{k}) = \overline{f(k)}.$$

For example, the method that matches n-period covariances of the spot variance results in $\tilde{k}$, defined by the condition:

$$e^{-n\tilde{k}} = \overline{e^{-nk}}.$$

If we allow for a bias and follow the same steps from the previous subsection, we find that the total MSPE of the model-based forecast is equal to:

$$\text{Total MSPE} = \text{GFE}^{model} + F(\hat{\sigma}_t^2, Bias) + 2\ \text{Cov}(\text{GFE}^{model}, F(\hat{\sigma}_t^2, Bias)), \quad \text{(A.28)}$$

$$F(\hat{\sigma}_t^2, Bias) =$$

$$= \left[\frac{1-e^{-\tilde{k}}}{\tilde{k}}\right]^2 E\hat{\sigma}_t^4 + \left[\overline{\frac{1-e^{-k}}{k}}\right]^2 E\sigma_t^4 - 2\frac{1-e^{-\tilde{k}}}{\tilde{k}}\overline{\frac{1-e^{-k}}{k}}E[\hat{\sigma}_t^2\sigma_t^2], \quad \text{(A.29)}$$

$$\text{Cov}(\text{GFE}^{model}, F(\hat{\sigma}_t^2, Bias)) =$$

$$= \text{Var}\sigma_t^2\frac{1-e^{-\tilde{k}}}{\tilde{k}}\overline{\left([\overline{\frac{1-e^{-k}}{k}} - \frac{1-e^{-k}}{k}]\frac{1-e^{-kh}}{kh}\frac{b_h}{1-a_he^{-kh}}\right)}, \quad \text{(A.30)}$$

where

$$\frac{E(\hat{\sigma}_t^4)}{\text{Var}\sigma_t^2} = [\frac{\tilde{k}^2h^2}{1-a_h^2} + \frac{2\tilde{k}hb_h}{(1-a_h)^2}]\frac{1-\lambda}{\lambda} + 3\frac{b_h^2}{1-a_h^2}\left[\frac{1-\lambda}{\lambda} + 2\overline{\frac{\frac{h}{k} - \frac{1-e^{-kh}}{k^2}}{h^2}}\right]$$

$$+2\frac{b_h^2}{1-a_h^2}\left[\frac{1-\lambda}{\lambda}\frac{a_h}{1-a_h} + \overline{\frac{(1-e^{-kh})^2}{k^2h^2}\frac{a_h}{1-a_he^{-kh}}}\right],$$

$$\frac{E(\hat{\sigma}_t^2\sigma_t^2)}{\text{Var}\sigma_t^2} = [\frac{\tilde{k}h}{1-a_h} + \frac{b_h}{1-a_h}]\frac{1-\lambda}{\lambda} + b_h\overline{\frac{1-e^{-kh}}{kh}\frac{1}{1-a_he^{-kh}}}.$$

Since the above error takes the form $\text{Var}\sigma_t^2 f(\Xi)$, then Proposition 1 is still valid for this more general case.

A.2.5 Extension 2 : Drift in Returns

Suppose there is a constant drift in returns μdt. This generalization of the base case will not change the formula for the errors in the model-based forecast, since we demeaned returns before extracting spot variances. However, the total prediction error of the reduced-form forecast will now be equal to:

$$\frac{\text{Total MSPE "Reduced-Form"}}{\text{Var}\sigma_t^2} = \Psi(\Xi) - 2h\frac{1-\lambda}{\lambda} - \frac{2}{h^2}\overline{(\frac{h}{k} - \frac{1-e^{-kh}}{k^2})} + \frac{E^2 u_{t+1}}{Var\sigma^2},$$

where the expectation of the noise is

$$Eu_{t+1} = \mu^2 h.$$

The last term also affects the comparison. However, it is normally small for intra-day data and can be omitted from consideration. In general, constructing RV as a sum of the squared demeaned returns will result in the same forecast comparison as in the case of zero drift.

A.2.6 Extension 3 : Multi-Period Forecast

For the model-based forecast (1.38), we decompose the error in the sum of the genuine part and the error coming from the error in spot variance. The genuine part follows from the prediction error of regressing IV on the last observable spot and can be found from R^2 of the Mincer-Zarnowitz regression derived by [14]:

$$\text{GFE}^{model} = \text{Var}IV_t^{t+T} - \text{Var}\sigma_t^2 \left(\overline{\frac{1-e^{-kT}}{k}}\right)^2 =$$

$$= \text{Var}\sigma_t^2 \left[\overline{\frac{2}{k}[T + \frac{1-e^{-kT}}{k}]} - \left(\overline{\frac{1-e^{-kT}}{k}}\right)^2\right].$$

The total error equals:

$$\text{Total MSPE} = \text{GFE}^{model} + F(\hat{\sigma}_t^2, Bias) + 2\ \text{Cov}(\text{GFE}^{model}, F(\hat{\sigma}_t^2, Bias)),$$

$$\text{GFE}^{model} = \text{Var}\sigma_t^2 \left[\overline{\frac{2}{k}[T + \frac{1-e^{-kT}}{k}]} - \left(\overline{\frac{1-e^{-kT}}{k}} \right)^2 \right],$$

$$F(\hat{\sigma}_t^2, Bias) = \left[\frac{1-e^{-\tilde{k}T}}{\tilde{k}} \right]^2 E\hat{\sigma}_t^4 + \left[\overline{\frac{1-e^{-kT}}{k}} \right]^2 E\sigma_t^4 - 2\frac{1-e^{-\tilde{k}T}}{\tilde{k}} \overline{\frac{1-e^{-kT}}{k}} E[\hat{\sigma}_t^2 \sigma_t^2],$$

$$\text{Cov}(\text{GFE}^{model}, F(\hat{\sigma}_t^2, Bias)) =$$

$$= \text{Var}\sigma_t^2 \frac{1-e^{-\tilde{k}T}}{\tilde{k}} \overline{\left([\overline{\frac{1-e^{-kT}}{k}} - \frac{1-e^{-kT}}{k}] \frac{1-e^{-kh}}{kh} \frac{b_h}{1-a_h e^{-kh}} \right)}.$$

For the reduced-form forecast based on the ARMA, the total MSPE is constructed in the same manner as for the one-period case:

$$\text{Total MSPE} =$$

$$= E^2(IV_t^{t+T} - RV_t^{t+T}) + \text{Var}(RV_t^{t+T} - P(RV_t^{t+T} | RV_{t-j}^{t+T-j})) - \text{Var}(IV_t^{t+T} - RV_t^{t+T}).$$

For zero drifts in returns, the difference between the integrated variance and the realized variance is unpredictable, and therefore serially uncorrelated. Hence, the variance of the noise is equal to

$$\text{Var}(IV_t^{t+T} - RV_t^{t+T}) = T\text{Var}u_t^{t+1}.$$

The innovation in the realized variance can be found by iterating the ARMA for the RV:

$$RV_t^{t+T} - P(RV_t^{t+T} | RV_{t-j}^{t+T-j}) = \left[(1 + F + ... + F^{T-1}) \frac{1 - \sum_{i=1}^{p} \beta_i(h) L^p}{\prod_{i=1}^{p} (1 - e^{-k_i} L)} \right]_+ \eta_{t+1}(h).$$

Since parameters of the ARMA-representation are functions of Ξ and the variance of the innovation η_t takes the form (A.26), the variance of the above expression is

equal to:

$$\text{Var}(RV_t^{t+T} - P(RV_t^{t+T}|RV_{t-j}^{t+T-j})) = \Psi_T(\Xi)\text{Var}\sigma_t^2.$$

The resulting total MSPE equals

$$\frac{\text{Total MSPE "Reduced-Form"}}{\text{Var}\sigma_t^2} = \Psi_T(\Xi) - 2hT\frac{1-\lambda}{\lambda} - \frac{2T}{h^2}(\frac{h}{k} - \frac{1-e^{-kh}}{k^2}).$$

Thus, the comparison for multi-period forecasts is similar to the case of one-period forecasts. As before, this comparison depends solely on the parameters in Ξ.

A.3 Tables and Figures

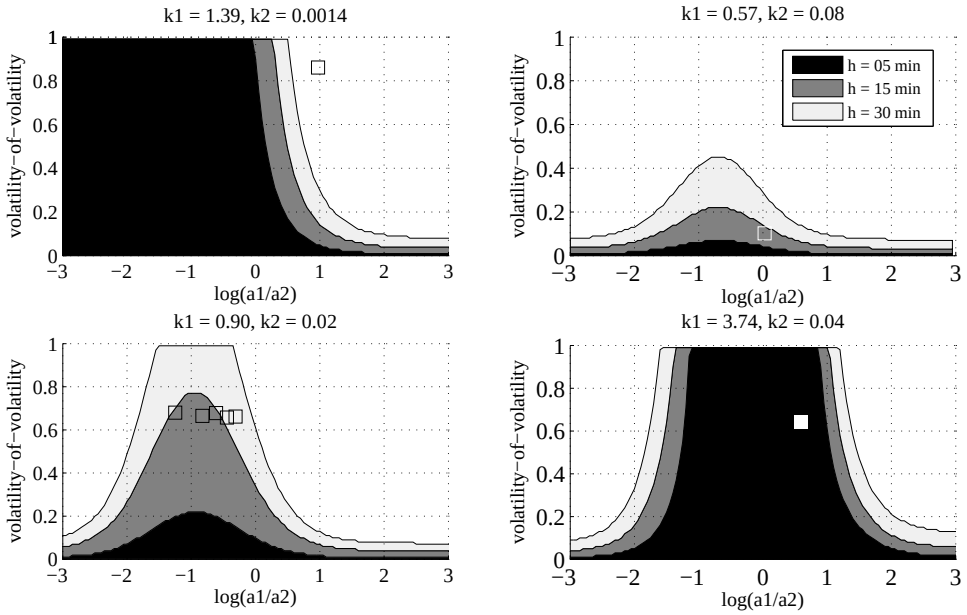

FIGURE A.1: Comparison of Day-Ahead Forecasts. Colored areas denote parameter sets for which the MSPE of the reduced-form forecast is lower than the MSPE of the model-based forecast. Different intensities of gray correspond to different sampling frequencies h. The X axis is the ratio of factor loadings $\ln(\frac{a_1}{a_2})$. The Y axis is the volatility-of-volatility $\frac{\text{Var}_{\sigma_t^2}}{E\sigma_t^4}$. The squares correspond to the parameter values from Table A.2.

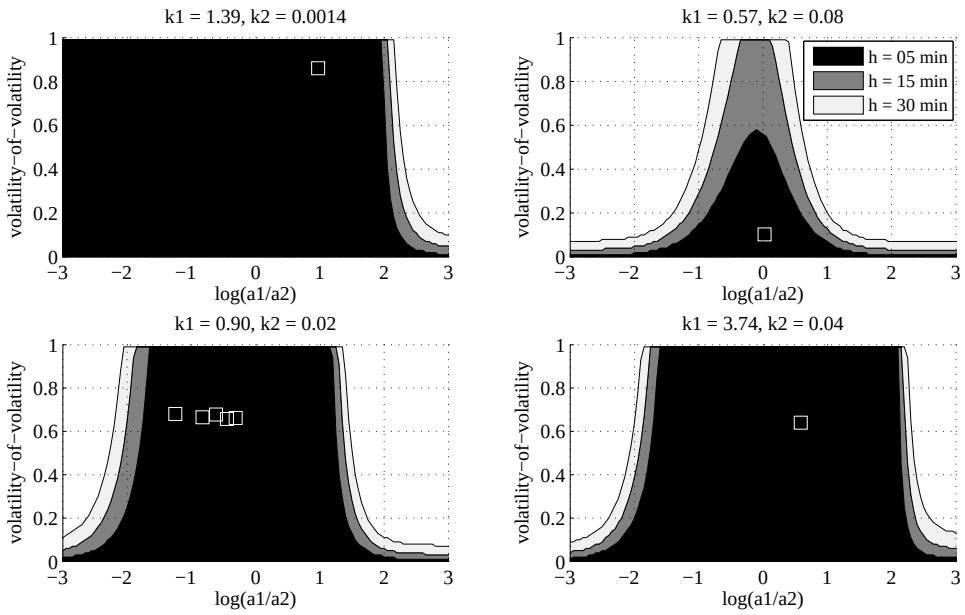

FIGURE A.2: Comparison of Week-Ahead Forecasts. Colored areas denote parameter sets for which the MSPE of the reduced-form forecast is lower than the MSPE of the model-based forecast. Different intensities of gray correspond to different sampling frequencies h. The X axis is the ratio of factor loadings $\ln(\frac{a_1}{a_2})$. The Y axis is the volatility-of-volatility $\frac{\mathrm{Var}_{\sigma_t^2}}{E\sigma_t^4}$. The squares correspond to the parameter values from Table A.2.

Table A.1: Models with One Component in the Variance

Model Name	Specification of the Variance Process
Square-Root SV	$d\sigma_t^2 = k(\theta - \sigma_t^2)dt + \eta\sigma_t dW_t$
Log-Volatility	$d\ln\sigma_t^2 = k(\theta - \ln\sigma_t^2)dt + \eta dW_t$
GARCH SV	$d\sigma_t^2 = k(\theta - \sigma_t^2)dt + \eta\sigma_t^2 dW_t$
Ornstein-Uhlenbeck	$d\sigma_t^2 = k(\theta - \sigma_t^2)dt + \eta dW_t$

Table A.2: Parameters for Two-Factor Models.

Forecast:	Model Type	k_1	k_2	$LN(a_1/a_2)$	λ
Bollerslev, Zhou (2002) Data: 5-min DM/USD spot	Affine	0.57	0.07	0.025	0.103
Huang, Tauchen (2005) Data: stock indices	Log-normal	1.386	0.00137	0.976	0.860
Alizadeh, Brandt, Diebold (2002) Data: daily exchange rates	Log-normal	0.81/0.95	0.02/0.03	-1.3/-0.3	0.66/0.68
Barndorff-Nielsen, Shephard (2002) Data: 5-min DM/USD spot	CEV	3.74	0.04	0.59	0.64

Note: The table reports parameters of the ESV representation for multi-factor models estimated on observed data sets; mean-reversions (k_1, k_2), model misspecification ($\ln(a_1/a_2)$), and volatility-of-volatility ($\lambda = \frac{\mathrm{Var}_{\sigma^2}}{E\sigma^4}$). All the parameters are in daily units. For log-normal models with $\ln\sigma_t^2 = x_{1t} + x_{2t}$, we define $a_1^2 = \mathrm{Var}(\sigma_t^2|x_{t2} = Ex_{t2})$, $a_2^2 = \mathrm{Var}(\sigma_t^2|x_{t1} = Ex_{t1})$.

Table A.3: GARCH-SV example.

Forecasting Model	True Model	Model-Based Infeasible	Model-Based Feasible	Reduced-Form Infeasible	Reduced-Form Feasible
1F-GARCH	1F-GARCH	0.977(0.023)	0.938(0.062)	0.958(0.043)	0.910(0.090)
1F-GARCH	2F-SR(I)	0.819(0.181)	0.674(0.393)	0.699(0.301)	0.646(0.354)
1F-GARCH	2F-SR(II)	0.328(0.672)	0.295(1.156)	0.315(0.686)	0.312(0.688)

Note: The table reports R^2 from the Mincer-Zarnowitz regressions and the normalized MSPE in parenthesis. Feasible forecasts are based on 5-minute returns. 2F-SR-SV(I) is calibrated from [14] and 2F-SR-SV(II) is calibrated from [21]. The results are calculated analytically from formulas presented in the Appendix.

Table A.4: Parameter Estimates of the One-Factor Model for DM/USD data.

	$\hat{k}$	$\hat{\lambda}$	$\hat{\theta}$
GMM based on: :			
- daily RV :	0.250	0.459	0.508
- weekly RV :	0.047	0.325	0.508
- monthly RV :	0.033	0.257	0.508
- quarterly RV :	0.018	0.216	0.509

Note: The table reports the GMM parameter estimates for model (1.40). Time is measured in daily units and returns are in percentage form.

Table A.5: Mincer-Zarnowitz R^2(MSPE) for Day-Ahead Forecasts.

Forecast	$R^2(MSPE)$
Model-Based(1-F):	
Daily $\hat{k} = 0.250$	0.236 (1.233)
Weekly $\hat{k} = 0.047$	0.324 (0.976)
Monthly $\hat{k} = 0.033$	0.341 (0.866)
Quarterly $\hat{k} = 0.018$	0.356 (0.765)
Model-Based(2-F):	
2-Factor(Particle)	0.370 (0.698)
2-Factor(Kalman)	0.356 (0.662)
Reduced-Form	0.357 (0.646)

Note: "Model-Based (1-F)" is a forecast based on the one-factor model (1.40). "Model-Based (2-F)" is a forecast based on the two-factor SR-SV model (1.33). "Reduced-Form" is a linear projection of the realized variance on its past values, with the number of lags chosen by the BIC-criterion. $\hat{k}$ are estimated mean-reversions of volatility from Table A.4. Data: DM/USD spot rates. The parameter estimation period for the one-factor model is December 2, 1986, through June 30, 1999; for the two-factor model it is December 2, 1986, through December 1, 1996. The forecast evaluation period is December 2, 1987, through June 30, 1999.

Table A.6: Mincer-Zarnowitz R^2(MSPE) for Multi-Period Forecasts.

Forecast:	Day Ahead	Week Ahead	Month Ahead
Model-Based 2-F (Particle)	0.370(0.698)	0.393(0.718)	0.316(0.822)
Model-Based 2-F (Kalman)	0.356(0.662)	0.345(0.666)	0.311(0.711)
Model-Based 1-F	0.356(0.765)	0.340(1.031)	0.277(1.307)
Reduced-Form	0.357(0.646)	0.377(0.623)	0.323(0.681)

Note: "Model-Based (1-F)" is a forecast based on the one-factor model (1.40). "Model-Based (2-F)" is a forecast based on the two-factor SR-SV model (1.33). "Reduced-Form" is a linear projection of the realized variance on its past values, with the number of lags chosen by the BIC-criterion. Data: DM/USD spot rates. The parameter estimation period for the one-factor model is December 2, 1986, through June 30, 1999; for the two-factor model it is December 2, 1986, through December 1, 1996. The forecast evaluation period is December 2, 1987, through June 30, 1999.

Appendix B

Appendix to Chapter 2

B.1 SV-CJ Model-Based Forecast

The SV-CJ forecast is based on the following model:

$$\begin{bmatrix} ds_t \\ d\sigma_t^2 \end{bmatrix} = \begin{bmatrix} \mu \\ \kappa(\theta - \sigma_t^2) \end{bmatrix} dt + \begin{bmatrix} \sigma_t dW_t^s \\ \sigma_v \sigma_t dW_t^v \end{bmatrix} + \begin{bmatrix} \xi_t^s \\ \xi_t^v \end{bmatrix} dJ_t,$$

where dW_t^v and dW_t^s are increments of standard Brownian motions with the correlation ρdt, J_t is a Poisson process with intensity λ, ξ_t^s and ξ_t^v are jump sizes: ξ_t^v is exponential with a mean $1/\mu_v$, and ξ_t^s is conditionally normal with a mean $\mu_j + \rho_j \xi_t^v$ and variance σ_j^2.

For each day t, we estimate the above model on daily close-to-close data (up to the day t) using the MCMC method described by [50]. To form the forecast we, first, generate draws of parameters, jumps and variances from their conditional distributions given the observed data. Then, at each draw we also generate the realized variance of the future return RV_t^{t+22}. The average across draws yields the

expectation of the future realized variance given the observed daily returns up to day t. Figure B.8 reports the estimated parameter values as they evolve with time.

B.2 One/Two-Factor Model-Based Forecasts

[80] demonstrated that each SV-model for log-prices s_t with square-integrable variances $\sigma_t^2 = \sigma_t^2(x_t)$ can be represented as an ESV-model:

$$\begin{aligned} ds_t &= \mu_t dt + \sigma(x_t) dW_t^s, \\ \sigma^2(x_t) &= a_0 + \sum_{i=1}^{p} a_i x_{it}, \\ dx_{it} &= -k_i x_{it} dt + \phi_i(x_{it}) dW_{it}, i = \overline{1,p}, \\ E(x_{it} x_{jt}) &= I(i=j), \end{aligned} \tag{B.1}$$

where $x_t = \{x_{1t}, ..., x_{pt}\}$ are referred to as factors. The above definition implies that the spot variance $\sigma^2(x_t)$ can be represented by a sum of unconditionally uncorrelated autoregressive processes. After small corrections for usually negligible drifts in intraday returns, the forecast of the future realized variance can be obtained for ESV-models in the following form:

$$E_t RV_t^{t+H} \approx a_0 H + \sum_{i=1}^{p} \frac{1 - e^{-k_i H}}{k_i} x_{i,t}. \tag{B.2}$$

Although factors x_t are not observable, they can be estimated. The same is true for parameters a_0 and k_i, $i \in \overline{1,p}$. It appears that we can form the forecast (B.2) using methods that do not rely on the specification of volatility-in-volatility terms $\phi_i()$ and the leverage effects, i.e. $\text{corr}(dW_{it}, dW_t^s)$. Thus, we employ the method that is to some extent robust to misspecification. The only parameter that is crucial to define is the number of factors p and we will fix it to $p = 1$ for one-factor forecasts, and $p = 2$ for two-factor forecasts. On the negative side, since the suggested method does not employ all the information about the model, it is not efficient.

The first step for the suggested method is the GMM estimation based on the correlation structure of bi-power variations BV similar to [97]. At the second step, the states x_t are extracted by Kalman filtering. Bi-power variations are defined as the sum:

$$BV_t^{t+1} = \sum_{j=2}^{1/h} \frac{\pi}{2} \left| s_{t+jh} - s_{t+jh-h} \right| \left| s_{t+jh-h} - s_{t+jh-2h} \right|.$$

Their limits – integrated variances $IV_t^{t+1} \equiv \int_t^{t+1} \sigma_s^2 ds$ – have the covariance structure that can be derived from the model (B.1):

$$\operatorname{cov}[IV_t^{t+1}, IV_{t+l}^{t+l+1}] = \sum_{i=1}^{p} a_i^2 \left[\frac{1 - e^{k_i}}{k_i} \right]^2 e^{-k_i(l-1)}, \forall l \in \overline{1, 40}. \tag{B.3}$$

As sampling frequency increases to infinity, the analytical expressions for the moments of bi-power variations in GMM can be replaced by those of integrated variances; see [98]. For the weighting matrix in GMM, we use the Newey-West variance-covariance matrix of sample moments with $T^{1/4}$ lags where T is a number of days in the sample.

Kalman-filter for extraction of states is based on the joint dynamics of IV and states:

$$IV_t^{t+\Delta} = a_0 \Delta + \Gamma' x_t + \varepsilon_{t+\Delta}^{IV}, \tag{B.4}$$

$$x_{t+\Delta} = A x_t + \varepsilon_{t+\Delta}^{x}, \tag{B.5}$$

where

$$\Gamma' = \begin{bmatrix} \frac{1-e^{-k_1\Delta}}{k_1} & \frac{1-e^{-k_2\Delta}}{k_2} \end{bmatrix}, \tag{B.6}$$

$$A = \operatorname{diag} \begin{bmatrix} e^{-k_1\Delta} & e^{-k_2\Delta} \end{bmatrix}, \tag{B.7}$$

$$\Omega \equiv \operatorname{Var} \begin{bmatrix} \varepsilon_t^{IV} \\ \varepsilon_t^{x} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{p} \frac{2a_i^2}{k_i} (\Delta - \frac{1-e^{-k_i\Delta}}{k_i}) - \Gamma' P_{0|0} \Gamma & a_i^2 k_i \left(\frac{1-e^{-k_i\Delta}}{k_i}\right)^2 \\ a_i^2 k_i \left(\frac{1-e^{-k_i\Delta}}{k_i}\right)^2 & (I - AA') P_{0|0} \end{bmatrix}, \tag{B.8}$$

$$P_{0|0} = \operatorname{diag} \begin{bmatrix} a_1^2 & a_2^2 \end{bmatrix}. \tag{B.9}$$

To extract states we followed standard steps as described by [66], starting with zero initial values for the states $\hat{x}_0 = \text{diag}[0\ 0]$ and corresponding error $E(x_0 - \hat{x}_0)^2 = P_{0|0}$.

There is a trade-off in the choice of the parameter Δ for the above Kalman filter. On one hand, equating Δ to the interval between observations ($\Delta = h$) minimizes the errors in $\hat{x}_t$ if $IV_t^{t+\Delta}$ is observed. On the other hand, if $\Delta = h$ the bi-power variations $BV_t^{t+\Delta}$ do not converge to $IV_t^{t+\Delta}$ for $h \to 0$. Therefore, we select optimal Δ based on simulations from SR-SV models, i.e. models described by (B.1) with diffusion terms $\phi_i(x) = \sqrt{x_i}$. For these simulations the errors in states were at their minimum for $\Delta = 1/2$ day if sampling frequency is $h = 5$ minutes.

The forecast that is based on the formula (B.2) with plugged-in estimates for parameters and states will be referred to as the two-factor model-based forecast:

$$\widehat{RV}_t^{t+H} = \hat{a}_0 T + \sum_{i=1}^{2} \frac{1 - e^{-\hat{k}_i H}}{\hat{k}_i} \hat{x}_{it}, \tag{B.10}$$

and the forecast:

$$\widehat{RV}_t^{t+H} = \hat{a}_0 T + \frac{1 - e^{-\hat{k} H}}{\hat{k}} (\hat{\sigma}_t^2 - \hat{a}_0), \tag{B.11}$$

where $\hat{\sigma}_t^2 = \hat{a}_0 + \hat{x}_{1,t} + \hat{x}_{2,t}$ will be referred to as the one-factor forecast. [1]

Forecasting Jumps.

For the SV model (B.1) we assumed that the dynamics of returns is continuous. This contradicts to the strong evidence presented in the literature that prices include both continuous and jump components; see e.g. [69]. If we allow for the jumps in prices, the forecast of the realized variance should include two parts: the forecast of the continuous part that is still equal to

$$a_0 H + \sum_{i=1}^{p} \frac{1 - e^{-k_i H}}{k_i} x_{i,t},$$

[1] The parameter $\hat{k}$ is estimated from the slope in the regression of RV_t^{t+H} on $\hat{\sigma}_t^2$.

and the forecast of the jump part. There is an evidence in the literature that the jump part of the variance is hard to predict; see e.g. [9]. In this case, the best prediction of the jump part is a constant. Several papers also argue that the intensity of jumps is positively related to the current level of the continuous part of the variance; see e.g. [17].

In this study we assume that the model yields the expectation of RV_t^{t+H} to be a linear combination of the states $x_{1,t}$ and $x_{2,t}$, and we find the intercept and the slopes by regressing RV_t^{t+H} on the estimated values of $\hat{x}_{1t}$ and $\hat{x}_{2t}$. This assumption incorporates both the case of unpredictable jumps and the case of jumps with the intensity proportional to the current level of variance.

Correction for Overnight Changes.

The five-minute S&P 500 data include one "large" interval per day, which is the overnight return. Ideally, for model-based forecasting overnight return ought to be modeled separately, as several studies showed that the properties of overnight returns are different from those of returns during trading hours; see e.g. [79]. However, all that is needed to form a forecast are the estimates of the states $\hat{x}_{1,t}$ and $\hat{x}_{2,t}$. Optimal extraction of the states depends on the dynamics of x_t through the night, rather then on the full model for overnight returns. We assume, that the model preserves a p-factor structure over the night with the similar dynamics for two continuous states and introduce a new parameter – the duration of the night Δ_{night}. That is, the overnight dynamics of the states is:

$$x_{t+\Delta_{night}} = A^{night} x_t + \varepsilon^{x\ night}_{t+\Delta_{night}}, \tag{B.12}$$

where

$$A^{night} = \text{diag}\left[\begin{array}{cc} e^{-k_1\Delta_{night}} & e^{-k_2\Delta_{night}} \end{array}\right], \tag{B.13}$$

$$\Omega^{x\ night} \equiv \text{Var}\varepsilon_t^{x\ night} = (I - A^{night}A^{night'})P_{0|0}. \tag{B.14}$$

To estimate the night duration, we add a new moment condition to the GMM estimation that measures the persistence of variance during regular trading hours:

$$\text{cov}[IV_t^{t+1/2}, IV_{t+1/2}^{t+1}] = \sum_{i=1}^{p} a_i^2 \left[\frac{1 - e^{k_i/2}}{k_i}\right]^2 e^{k_i}, \tag{B.15}$$

and modify the other moments to take into account overnight changes:

$$\text{cov}[IV_t^{t+1}, IV_{t+l(1+\Delta_{night})+1}^{t+l(1+\Delta_{night})+1}] = \sum_{i=1}^{p} a_i^2 \left[\frac{1 - e^{k_i}}{k_i}\right]^2 e^{-k_i(l(1+\Delta_{night})-1)}, \tag{B.16}$$

where $l = \overline{1, 40}$.

In Kalman filter, we update the states x_t from the closing quote to the next-day first quote in the following way:

$$\hat{x}_{t+\Delta_{night}} = A^{night}\hat{x}_t, \tag{B.17}$$

$$E(x_{t+\Delta_{night}} - \hat{x}_{t+\Delta_{night}})^2 = A^{night}E(x_t - \hat{x}_t)^2A^{night'} + \Omega^{x\ night}. \tag{B.18}$$

To form the two-factor model-based forecast at time t, we find slopes and the intercept from regressing $RV_\tau^{\tau+H}$ on the estimated vector of states $\hat{x}_\tau$ using the available data $\tau \leq t - H$.

B.3 Utility-Based Comparison: Robustness Checks

In the section that discusses the case of variable trader's premium, we demonstrated that the performances of the reduced-form forecasts and the two-factor model-based forecast are close, with the reduced-form forecast being slightly better for moderate risk-aversions of the trader.

To check the robustness of this result, here we, first, investigate if the result was affected by the simplifying assumption made in defining the ex-post utility (2.9). Figure B.9 reproduces the utility-based comparison for the quadratic utility in the "classic" form given by (2.8). As expected, the utility pattern seems more volatile, as the error in estimation of the average loss-function is increased due to the replacement of the smooth estimate $\widetilde{\text{Var}}_t RV_t^{t+H}$ by its realization $(RV_t^{t+H} - \tilde{E}_t RV_t^{t+H})^2$. Also, the average utilities in Figure B.9 are generally lower than the utilities in Figure B.4. The slump is due to the new source of uncertainty that effects the decision making of the trader: the uncertainty about the higher moments of RV.

Apart from these changes the general pattern of the utilities remained the same. Thus, in its highest point the economic value of the statistical forecasts reaches 7%, 6% and 2% of the contract price for the AR-7 reduced-form, two-factor model-based and SV-CJ forecasts respectively. Also, Figure B.9 shows that the ranking between these forecasts is preserved: reduced-form and two-factor model- based forecasts still lead with the reduced-form forecast being better for volatility premiums lower than 4%.

The second robustness check concerns the form of the trader's premium. For quadratic utility the premium is proportional to the conditional variance. However, traders may be less or more sensitive to uncertainty in the market, which may translate to the premiums proportional either to the standard deviation (for less uncertainty-sensitive traders) or to the higher moments of variance (for more uncertainty-sensitive traders). We check the robustness of our results with respect to different specifications of the premium. Figure B.10 plots the difference between AR-7 reduced-form and two-factor model-based forecasts as a function of the trader's risk aversion for different choices of m and n in the utility specification (2.10). We see the same pattern across all the graphs: reduced-form forecast is better for lower risk-aversions and the two-factor model-based forecast better for higher risk aversions.

The difference is within 0.5 % of the contract price for the region where traders are better off by using model-based forecasts and eventually vanishes for large risk aversions.

B.4 Reduced-Form Forecasts for RV (Simulations) and HAR-RV model for VIX (Actual Data)

The reduced-form forecast for the simulation study (2.35) is based on the following VAR for the realized variance and bi-power variation:

$$RV_{t+1|t} = \beta_0^{rv} + \sum_{i=1}^{6} \beta_i^{rv} BV_{t-i}^{t-i+1},$$

$$BV_{t+1|t} = \beta_0^{bv} + \sum_{i=1}^{6} \beta_i^{bv} BV_{t-i}^{t-i+1}.$$

The parameters estimated on 60000 days of 5-minute simulated observations are reported in Table B.1.

Table B.1: Parameters of the Reduced-Form Model for RV and BV.

Regressors	const	BV_{t-1}^{t}	BV_{t-2}^{t-1}	BV_{t-3}^{t-2}	BV_{t-4}^{t-3}	BV_{t-5}^{t-4}	BV_{t-6}^{t-5}	R^2
RV_t^{t+1}	0.116	0.428	0.252	0.128	0.068	0.034	0.041	27.7%
BV_t^{t+1}	0.043	0.450	0.242	0.124	0.065	0.045	0.031	84.9%

Table B.2: HAR-RV Model for VIX

β_1	β_2	β_3	β_4	β_5	β^{mon}	β_1^{qrt}	β_2^{qrt}	β_3^{qrt}	β_0
2.11	1.36	1.05	1.00	0.52	0.33	0.099	0.050	0.042	6.21
(0.33)	(0.29)	(0.31)	(0.26)	(0.24)	(0.03)	(0.011)	(0.005)	(0.004)	(0.22)

Note: HAR-RV Model for VIX. Table reports parameters for the following regression: $VIX_t^2 = \beta_0 + \sum_{i=1}^{5} \beta_i RV_{t-i}^{t-i+1} + \beta^{mon} RV_{t-20}^{t} + \sum_{i=1}^{3} \beta_i^{qrt} RV_{t-60i}^{t-60(i-1)}$
Parameters are estimated by OLS using the data on daily VIX and RV (constructed from 5-minute returns) from January, 1992, to October, 2007. Robust standard errors are reported in parenthesis.

B.5 Figures

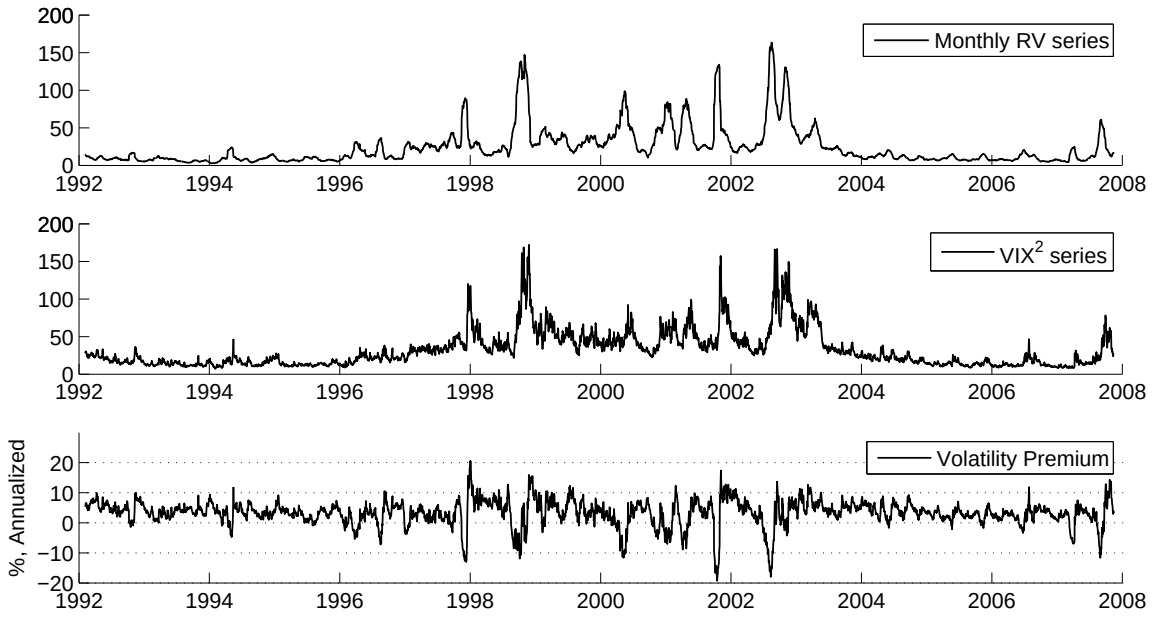

FIGURE B.1: Series of Monthly Realized Variances and VIX^2. The third panel shows the difference between VIX^2 and realized variance in the form $\sqrt{12VIX_t^2} - \sqrt{12RV_t^{t+22}}$.

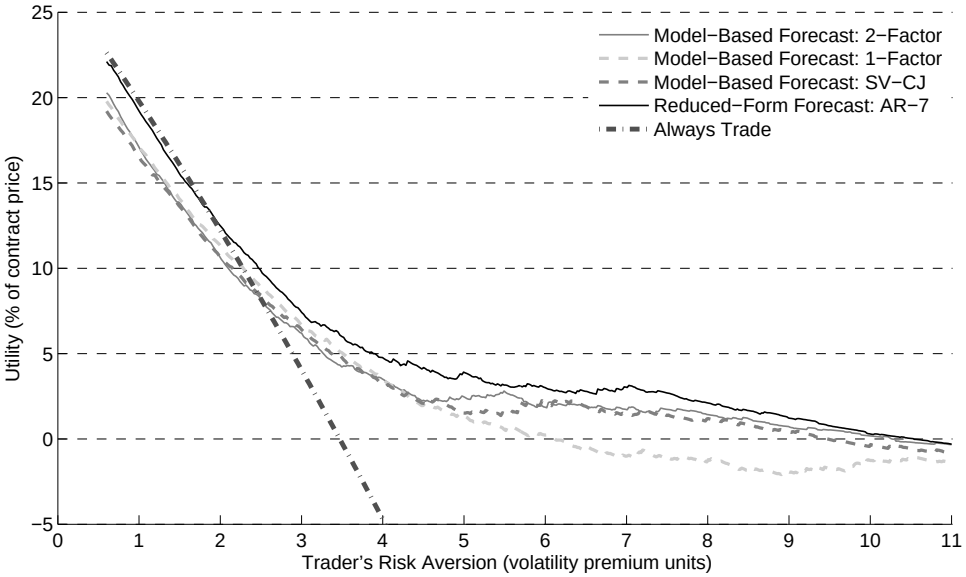

FIGURE B.2: Figure reports the utility of the variance trader (for different forecasts he employs to improve his trading strategy) as a function of the trader's risk aversion. Risk aversion on the axis X is measured in volatility premium units; see Section 2.3.1 for details. Utility on the axis Y is measured as a percentage of the contract price. The trader's premium (difference between his reservation value and his variance forecast) is constant. Forecasts are evaluated using the S&P 500 futures data from August, 2000, to October, 2007.

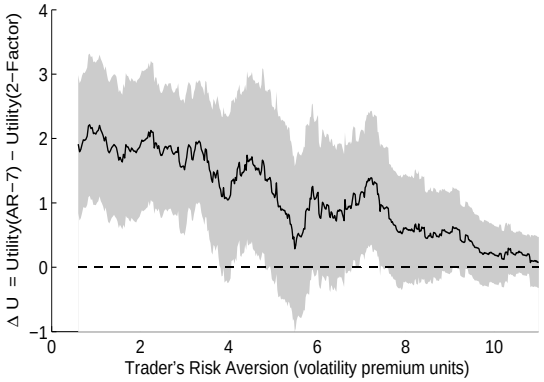

Figure B.3: Figure reports the difference between average utilities for the reduced-form AR-7 and the model-based two-factor forecasts. Grey area denotes the 95 % confidence interval. See Section 2.6.2 for details.

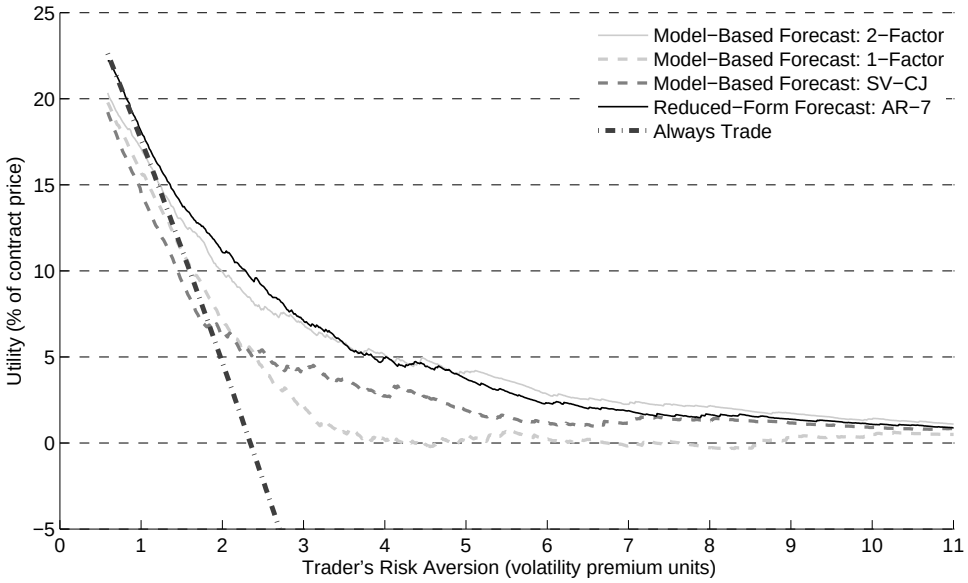

FIGURE B.4: Figure reports the utility of the variance trader (for different forecasts he employs to improve his trading strategy) as a function of the trader's risk aversion. Risk aversion on the axis X is measured in volatility premium units; see Section 2.3.1 for details. Utility on the axis Y is measured as a percentage of the contract price. The trader's premium (difference between his reservation value and his variance forecast) is proportional to the conditional variance $\text{Var}_t RV_t^{t+22}$. Forecasts are evaluated using the S&P 500 futures data from August, 2000, to October, 2007.

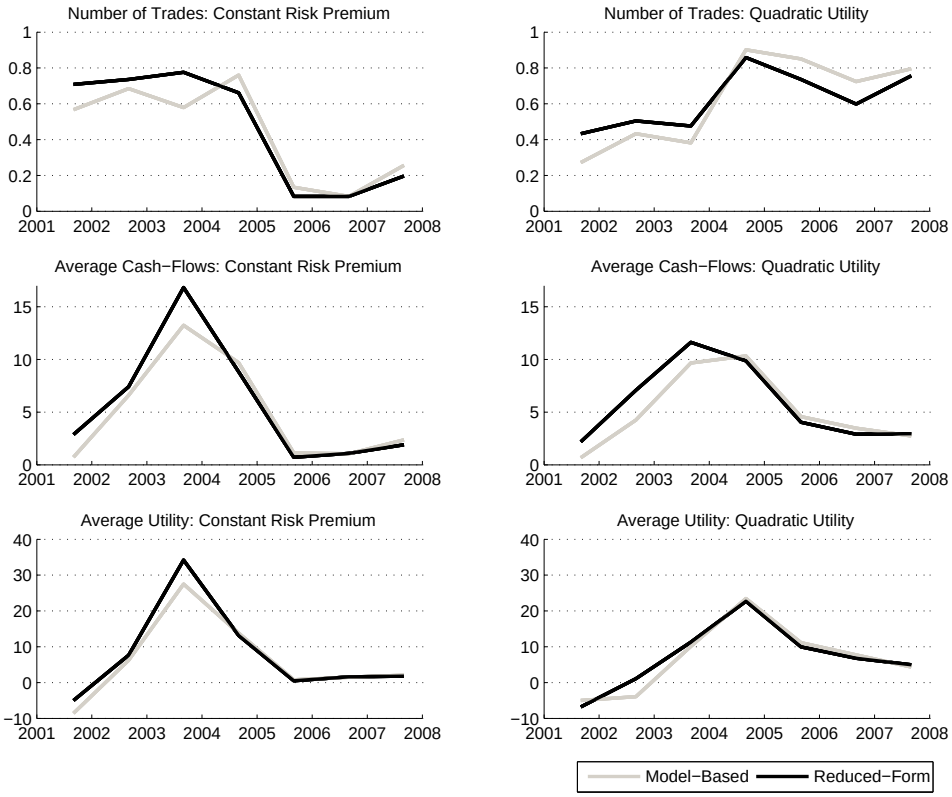

FIGURE B.5: Performance of Forecasts across Years. Panels in the first row report trading intensities $\frac{1}{T}\sum_{t=1}^{T} I_t$. Panels in the second row report average cash flows $\frac{1}{T}\sum_{t=1}^{T} C_t$. Panels in the third row report average utilities $\frac{1}{T}\sum_{t=1}^{T} u_t$ normalized by the contract price. Graphs show the averages across 7 periods starting August 2000 and ending in October 2007. Gray plots are for trading intensities, cash flows and utilities of the trader who employs the two-factor model-based forecast. Black plots are for the AR-7 reduced-form forecast.

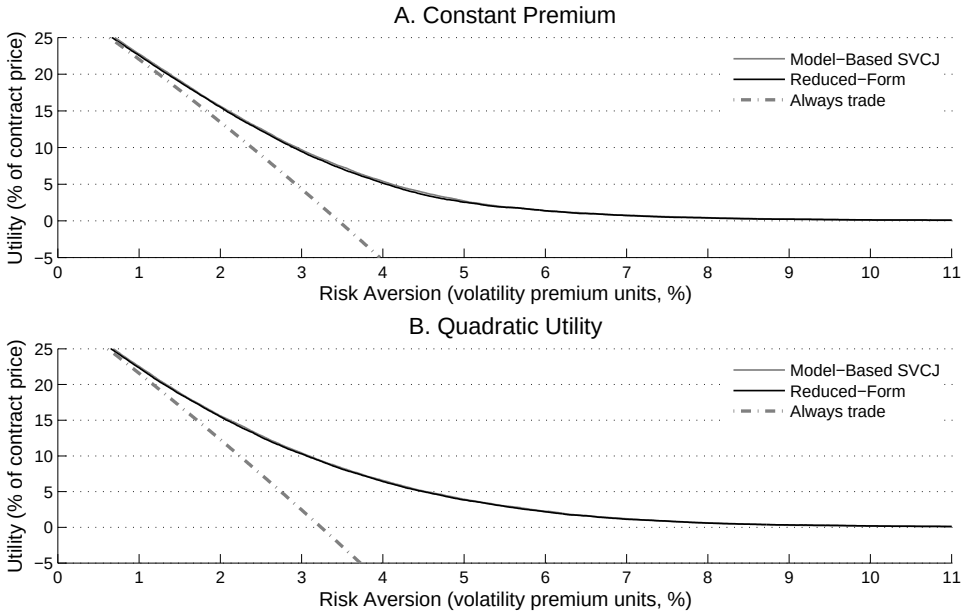

FIGURE B.6: Utility-Based Comparison of Variance Forecasts for Simulated Data: True Model. Figure reports the utility of the variance trader (for different forecasts he employs to improve his trading strategy) as a function of the trader's risk aversion. Risk aversion on the X axis is measured in volatility premium units. (See Section 2.3.1 for details.) Utility on the Y axis is measured as a percentage of the contract price. In Panel A, the trader's premium (the difference between his reservation value and his variance forecast) is constant; in Panel B, the premium is proportional to the conditional variance $\text{Var}_t IV_t^{t+T}$ that is calculated using the true model (2.35). The data for this figure was simulated from the SV-CJ model (2.35) and consists of 5 data sets of the length 4000 days.

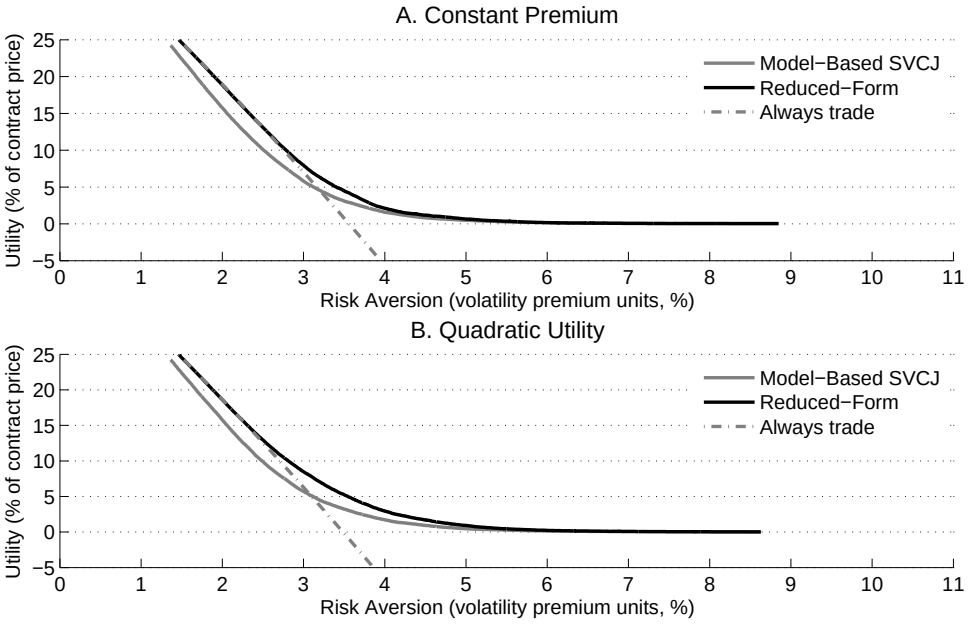

FIGURE B.7: Utility-Based Comparison of Variance Forecasts for Simulated Data: Misspecified Model. Figure reports the utility of the variance trader (for different forecasts he employs to improve his trading strategy) as a function of the trader's risk aversion. Risk aversion on the X axis is measured in volatility premium units. (See Section 2.3.1 for details.) Utility on the Y axis is measured as a percentage of the contract price. In Panel A, the trader's premium (the difference between his reservation value and his variance forecast) is constant; in Panel B, the premium is proportional to the conditional variance $\text{Var}_t IV_t^{t+T}$ that is calculated using the true model (2.37). The data consists of five data sets simulated from the model (2.37) of the length 4000 days. Model-based forecast is calculated using the SV-CJ model (2.35).

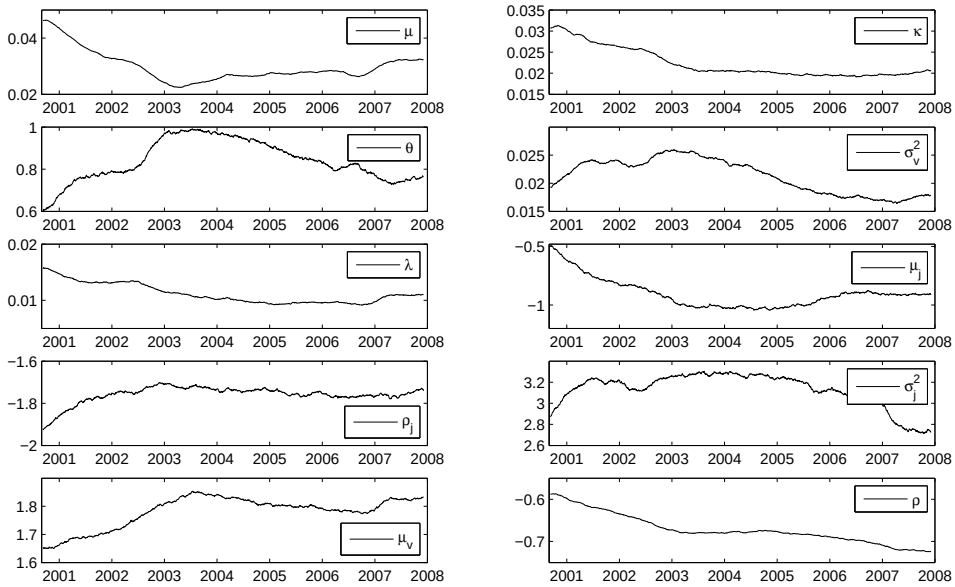

FIGURE B.8: Parameter Estimates for the SV-CJ model: Day-by-Day Estimation

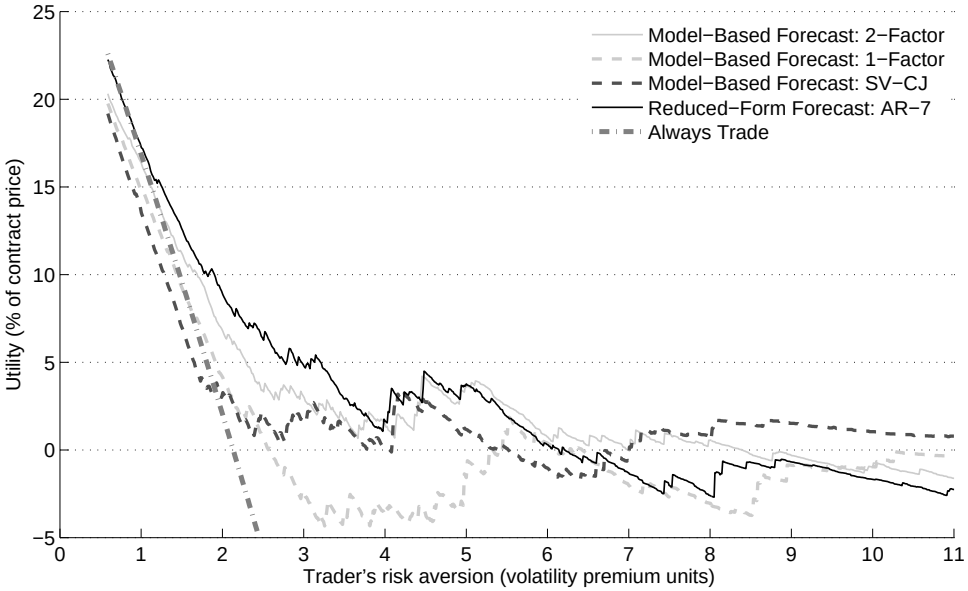

FIGURE B.9: Robustness Check: Comparison using Ex-post Utilities. Figure reports the utility of the variance trader (for different forecasts he employs to improve his trading strategy) as a function of the trader's risk aversion. Risk aversion on the axis X is measured in volatility premium units; see Section 2.3.1 for details. Utility on the axis Y is measured as a percentage of the contract price. Forecasts are evaluated using S&P 500 futures data from August 2000 to October 2007. The utility of the trader takes the form (2.7).

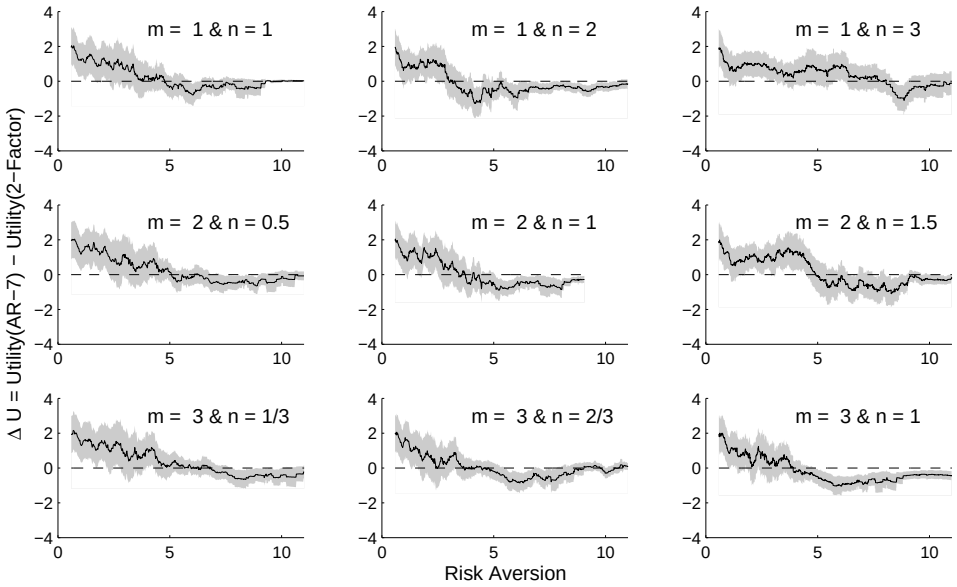

FIGURE B.10: Robustness Check. Difference in utilities between model-based and reduced-form forecasts for various specifications of the trader's risk premium with 95 % confidence intervals

Appendix C

Appendix to Chapter 3

C.1 Continuous-Time Equilibrium and SDF

The generalized preferences that we use are a continuous-time version of the Epstein-Zin-Weil discrete-time utility:

$$\tilde{V}_t^{1-\frac{1}{\psi}} = \left(1 - e^{-\rho h}\right) C_t^{1-\frac{1}{\psi}} + e^{-\rho h} \left[E_t \tilde{V}_{t+h}^{1-\gamma}\right]^{\frac{1-\frac{1}{\psi}}{1-\gamma}}. \tag{C.1}$$

Assuming that $\tilde{V}_t^{1-\gamma}$ is a semi-martingale, we approximate its conditional expectation over a short time-interval h by the linear function:

$$E_t \tilde{V}_{t+h}^{1-\gamma} \approx D\big(\tilde{V}_t^{1-\gamma}\big)h + \tilde{V}_t^{1-\gamma}, \tag{C.2}$$

where $D(.)$ denotes the drift of the argument. Plugging the conditional expectation in (C.2) into the definition (C.1), and taking limits around $h = 0$, the drift for the utility may be expressed as a function of consumption and utility:

$$D\big(\tilde{V}_t^{1-\gamma}\big) = \theta\rho \left[1 - \frac{C_t^{1-\frac{1}{\psi}}}{\tilde{V}_t^{1-\frac{1}{\psi}}}\right] \tilde{V}_t^{1-\gamma}.$$

The original utility $\tilde{V}_t$ in (C.1) can be replaced by any ordinally equivalent utility $V_t = \varphi(\tilde{V}_t)$, where the transformation $\varphi(.)$ is strictly increasing. Following [43], we apply the transformation $\varphi(.)$ that is linear in $\tilde{V}_t^{1-\gamma}$:

$$V_t = \frac{1}{1-\gamma}\tilde{V}_t^{1-\gamma}.$$

Given this choice of $\varphi(.)$, the preferences may be simply defined through the recursive condition:

$$DV_t + f(C_t, V_t) = 0, \tag{C.3}$$

where the normalized drift equals

$$f(c, v) = \frac{\rho}{1-\frac{1}{\psi}} \frac{c^{1-\frac{1}{\psi}} - [(1-\gamma)v]^{1/\theta}}{[(1-\gamma)v]^{1/\theta-1}}. \tag{C.4}$$

This in effect constitutes the formal definition of the Epstein-Zin-Weil preferences in continuous time; see also [44].

C.1.1 SDF in Continuous Time

In this section we derive the exact formula for the Stochastic Discount Factor (SDF) under the Epstein-Zin-Weil preferences in equation (C.4) as a function of the return on the consumption asset and the consumption growth rate.

For notational convenience, denote the logarithmic welfare-consumption ratio:

$$vc_t \equiv \log \frac{\tilde{V}_t}{C_t} = \frac{1}{1-\gamma} \log([1-\gamma]V_t) - \log C_t. \tag{C.5}$$

The normalized drift in equation (C.4) may then alternatively be represented as:

$$f(C_t, V_t) = \rho\theta V_t \left(e^{-(1-\frac{1}{\psi})vc_t} - 1 \right).$$

Duffie and Esptein(1992) have previously derive the SDF for recursive preferences with an arbitrary normalized drift $f(C_t, V_t)$:

$$M_t = \exp^{\int_0^t f_v(C_s,V_s)ds} f_c(C_t, V_t). \tag{C.6}$$

The dynamics for the SDF thus follows from the dynamics of the two partial derivative,

$$f_c \equiv \frac{\partial f}{\partial C_t} = (1-\gamma)\rho C_t^{-\gamma} e^{-\log(1-\gamma)+(\frac{1}{\psi}-\gamma)vc_t},$$

and

$$f_v \equiv \frac{\partial f}{\partial V_t} = \theta\rho(e^{-(1-\frac{1}{\psi})vc_t} - 1)(1-\frac{1}{\theta}) - \rho.$$

Now, note that

$$V_t^{\frac{1}{\theta}} M_t C_t^{\frac{1}{\psi}} = \rho e^{\frac{\log(1-\gamma)}{\theta}} V_t e^{\int_0^t f_v(C_s,V_s)ds}.$$

Extracting the drift of the process on the left-hand-side of the above equation, taking into account that the drift of the V_t process equals $-f(C_t, V_t)$, and rearranging the terms, we obtain the following relation:

$$D\left[\frac{e^{\frac{\log(1-\gamma)}{\theta}}}{\rho}\left(\frac{V_t^{\frac{1}{1-\gamma}}}{C_t}\right)^{1-\frac{1}{\psi}} M_t C_t\right] = -M_t C_t.$$

In other words,

$$\frac{e^{\frac{\log(1-\gamma)}{\theta}}}{\rho}\left(\frac{V_t^{\frac{1}{1-\gamma}}}{C_t}\right)^{1-\frac{1}{\psi}}$$

satisfies the Euler condition for the price-dividend ratio of the consumption asset. Together with appropriate terminal conditions, this implies that this expression must

be equal to the price-dividend ratio. In logarithm terms, there is a one-to-one correspondence between the price-dividend ratio of the consumption asset and the welfare-consumption ratio:

$$\log \Psi_t = -\log \rho + (1 - \frac{1}{\psi}) v c_t. \tag{C.7}$$

The formula for the price-dividend ratio implies that the total return satisfies:

$$\frac{dR_t}{R_t} \equiv \frac{dS_t + C_t dt}{S_t} = \frac{d\left[\Psi_t C_t\right]}{\Psi_t C_t} + \frac{dt}{\Psi_t}. \tag{C.8}$$

Combining the definitions of the return in (C.8), the SDF in (C.6), and the price-dividend ratio in (C.7), we obtain the following relation:

$$\begin{aligned} & d\log M_t + (1-\theta) d\log R_t + \frac{\theta}{\psi} d\log C_t \\ & = [f_v(C_t, V_t)dt + d\log f_c(C_t, V_t)] + (1-\theta)[d\log C_t + d\log \Psi_t + \frac{dt}{\Psi_t}] + \frac{\theta}{\psi} d\log C_t \\ & = \rho\theta dt, \end{aligned} \tag{C.9}$$

where the last equality follows from two identities:

$$f_v(C_t, V_t) + (1-\theta)\frac{1}{\Psi_t} = \rho\theta,$$

$$d\log f_c(C_t, V_t) + (1-\theta)[d\log C_t + d\log \Psi_t] + \frac{\theta}{\psi} d\log C_t = 0.$$

The expression for the SDF in equation (C.9) as a function of the return on aggregate consumption and consumption growth may naturally be seen as the continuous-time version of the similar discrete-time relationship in [18]. We next proceed to study the asset pricing implications of the model and this SDF, including expressions for the risk-free rate, the return on the consumption asset, and the variance risk premium.

C.2 Model Solution Under Short Memory Dynamics

C.2.1 Pricing of the Consumption Asset

Suppose, that the dynamics of the consumption growth $g_t \equiv \frac{dC_t}{C_t}$ is determined by the following system of equations:

$$\frac{dC_t}{C_t} = \mu_g dt + \sigma_{g,t} dW_t^c, \tag{C.1}$$

$$d\sigma_{g,t}^2 = \kappa_\sigma(\mu_\sigma - \sigma_{g,t}^2)dt + \sqrt{q_t} dW_t^\sigma, \tag{C.2}$$

$$dq_t = \kappa_q(\mu_q - q_t)dt + \varphi_q\sqrt{q_t} dW_t^q, \tag{C.3}$$

where all of the shocks are uncorrelated. Under the risk-neutral measure the asset return is a martingale with respect to information at time t, i.e.,

$$D\left(M_t R_t\right) = 0.$$

It follows from the formula for the SDF in (C.9) and the definition of the return in (C.8) that

$$d\log\left(M_t R_t\right) = \theta(d\log\Psi_t + d\log C_t + \frac{dt}{\Psi_t}) - \frac{\theta}{\psi} d\log C_t - \rho\theta dt$$

Substituting the price-dividend ratio $\Psi_t \equiv \Psi(x_t)$ into the above condition for the drift $D\left(M_t R_t\right)$ yields the following pricing relation:

$$\theta D\log\Psi(x_t) + \frac{\theta}{\Psi(x_t)} + (1-\gamma)(\mu_g - \frac{\sigma_{g,t}^2}{2}) - \rho\theta + \\ + \frac{\theta^2}{2}\frac{d\left[\log\Psi(x), \log\Psi(x)\right]_t}{dt} + \frac{1}{2}(1-\gamma)^2\sigma_{g,t}^2 = 0, \tag{C.4}$$

where $D\log\Psi(x_t)$ denotes the drift of $\log\Psi(x_t)$, and $\left[\log\Psi(x), \log\Psi(x)\right]_t$ refers to the quadratic variation, whose increment characterizes the variance of shocks to

$\log \Psi(x_t)$. The pricing relation in (C.4) may now be solved using the first-order approximations similar to [33],

$$\frac{1}{\Psi(x_t)} = \exp(-\log \Psi(x_t)) \approx \exp(-\overline{\log \Psi}) - \exp(-\overline{\log \Psi})(\log \Psi(x_t) - \overline{\log \Psi}).$$

In particular, under this linearization it is natural to conjecture that log price-dividend ratio is linear in the states:

$$\log \Psi(x_t) = A_0 + A_\sigma \sigma_{g,t}^2 + A_q q_t, \tag{C.5}$$

and therefore

$$\frac{1}{\Psi(x_t)} \approx -\kappa_0 - \kappa_1(A_0 + A_\sigma \sigma_{g,t}^2 + A_q q_t). \tag{C.6}$$

Substituting the conjectured solution for $\Psi(x_t)$ in (C.5) and its inverse value in (C.6) into the pricing condition (C.4), we find the coefficients:

$$\begin{aligned} A_0 &= \frac{A_\sigma \kappa_\sigma \mu_\sigma + A_q \kappa_q \mu_q - \kappa_0 + (1 - 1/\psi)\mu_g - \rho}{\kappa_1}, \\ A_\sigma &= \frac{-\gamma}{2} \frac{1 - \frac{1}{\psi}}{\kappa_\sigma + \kappa_1}, \\ A_q &= \frac{\kappa_q + \kappa_1 - \sqrt{(\kappa_q + \kappa_1)^2 - \theta^2 A_\sigma^2 \varphi_q^2}}{\theta \varphi_q^2}, \end{aligned}$$

where the value of A_q is the root of a quadratic equation that is bounded away from ∞ as ϕ_q goes to zero. Note that similar to the discrete-time case, the loadings A_σ and A_q are negative for values of the inter-temporal elasticity of substitution $\psi > 1$.

Combining the dynamics of dividends C_t and the dynamics of the price-dividend ratio, we obtain the dynamics for the total log-return under the objective measure:

$$\begin{aligned} d \log R_t = (\frac{\mu_g}{\psi} + \rho + \left(-\frac{1}{2} + \frac{\gamma}{2}(1 - \frac{1}{\psi})\right) \sigma_{g,t}^2 - A_q(\kappa_1 + \kappa_q) q_t) dt \\ + \sigma_{g,t} dW_t^c + A_\sigma \sqrt{q_t} dW_t^\sigma + A_q \varphi_q \sqrt{q_t} dW_t^q. \end{aligned} \tag{C.7}$$

The dynamics for the stochastic discount factor follows from (C.9):

$$\frac{dM_t}{M_t} = [\frac{\gamma}{2}(1+\frac{1}{\psi})\sigma_{g,t}^2 - (1-\frac{1}{\theta})(\kappa_q+\kappa_1)A_q q_t - \frac{1}{\psi}\mu_g - \rho]dt \\ - \gamma\sigma_{g,t}dW_t^g + (\theta-1)[A_\sigma\sqrt{q_t}dW_t^\sigma + A_q\varphi_q\sqrt{q_t}dW_t^q] \quad \text{(C.8)}$$

The dynamics of SDF readily defines the risk-free rate, the equity premium, and the risk-neutral probability measure. For example, the risk-free rate is simply given by the drift of the SDF:

$$r_{f,t} \equiv -E_t\frac{dM_t}{M_t} = \frac{1}{\psi}\mu_g + \rho - \frac{\gamma}{2}(1+\frac{1}{\psi})\sigma_{g,t}^2 + (1-\frac{1}{\theta})(\kappa_q+\kappa_1)A_q q_t.$$

The equity premium is obtained as the "covariance" between the total return and the SDF:

$$\pi_{r,t} \equiv \frac{1}{dt}\frac{d[R,M]_t}{R_t M_t} = \gamma\sigma_{g,t}^2 - (\theta-1)(A_\sigma^2 + A_q^2\varphi_q^2)q_t. \quad \text{(C.9)}$$

Lastly, the diffusion part of the SDF (C.8) defines the transition from the processes under the objective measure to the risk-neutral measure:

$$\frac{dC_t}{C_t} = (\mu_g - \gamma\sigma_{g,t}^2)dt + \sigma_{g,t}d\tilde{W}_t^c$$

$$d\sigma_{g,t}^2 = \left(\kappa_\sigma(\mu_\sigma - \sigma_{g,t}^2) + (\theta-1)A_\sigma q_t\right)dt + \sqrt{q_t}d\tilde{W}_t^\sigma \quad \text{(C.10)}$$

$$dq_t = \left(\kappa_q(\mu_q - q_t) + (\theta-1)A_q\varphi_q^2 q_t\right)dt + \varphi_q\sqrt{q_t}d\tilde{W}_t^q, \quad \text{(C.11)}$$

where $d\tilde{W}_t^c$, $d\tilde{W}_t^\sigma$, and $d\tilde{W}_t^q$ are all uncorrelated Brownian motions under the risk-neutral probability measure.

C.2.2 Variance Premium

The variability of the future asset price is determined by the integrated variance:

$$IV_{t,t+N} \equiv \int_{\tau=t}^{t+N} d[\log S, \log S]_\tau = \int_{\tau=t}^{t+N}\sigma_{g\tau}^2 d\tau + (A_\sigma^2 + A_q^2\varphi_q^2)\int_{\tau=t}^{t+N} q_\tau d\tau. \quad \text{(C.12)}$$

The variance risk premium by definition is given by the difference between the expected values of the integrated variance under the objective and risk-neutral measures:

$$vp_t \;\equiv\; E_t^Q IV_{t,t+N} \;-\; E_t^P IV_{t,t+N}.$$

Under the objective measure, the consumption variance $\sigma_{g,t}^2$ and the volatility-of-volatility q_t are both affine processes with expectations:

$$E_t^P q_{t+\Delta t} = [q_t - \mu_q]e^{-\kappa_q \Delta t} + \mu_q,$$

$$E_t^P \sigma_{g,t+\Delta t}^2 = [\sigma_{g,t}^2 - \mu_\sigma]e^{-\kappa_\sigma \Delta t} + \mu_\sigma.$$

Under the risk-neutral measure, the volatility-of-volatility q_t remains an affine process, with the mean-reversion $\tilde{\kappa}_q = \kappa_q - (\theta - 1)A_q\varphi_q^2$ and the mean $\tilde{\mu}_q = \frac{\kappa_q \mu_q}{\tilde{\kappa}_q}$ given by equation (C.11). Thus, the expectation of $q_{t+\Delta t}$ under the risk-neutral measure simply equals:

$$E_t^Q q_{t+\Delta t} = [q_t - \tilde{\mu}_q]e^{-\tilde{\kappa}_q \Delta t} + \tilde{\mu}_q.$$

The process for the variance $\sigma_{g,t}^2$ under the risk-neutral measure in equation (C.10) is qualitatively different. The conditional mean now depends not only on its own value, but also on the current realization of q_t:

$$E_t^Q \sigma_{gt+\Delta t} = \tilde{\mu}_\sigma + (\sigma_{g,t}^2 - \tilde{\mu}_\sigma - \Delta_q)e^{-\kappa_\sigma \Delta t} + \Delta_q e^{-\tilde{\kappa}_q \Delta t},$$

where,

$$\Delta_q = \frac{(\theta - 1)A_\sigma}{\kappa_\sigma - \tilde{\kappa}_q}[q_t - \tilde{\mu}_q],$$

and $\tilde{\mu}_\sigma = \mu_\sigma + \frac{(\theta-1)A_\sigma}{\kappa_\sigma}\tilde{\mu}_q$ is equal to the risk-neutral unconditional mean of the variance.

The differences in the conditional expectations of the state variables under the risk-neutral and the objective measures can be represented as:

$$E_t^Q q_{t+\Delta t} - E_t^P q_t = q_t \left[e^{-\tilde{\kappa}_q \Delta t} - e^{-\tilde{\kappa}_q \Delta t}\right] + \kappa_q \mu_q \left[\frac{1 - e^{-\tilde{\kappa}_q}}{\tilde{\kappa}_q} - \frac{1 - e^{-\kappa_q \Delta t}}{\kappa_q}\right],$$

$$\frac{E_t^Q \sigma_{g,t+\Delta t}^2 - E_t^P \sigma_{g,t+\Delta t}^2}{(\theta-1)A_\sigma} = q_t \left[\frac{e^{-\tilde{\kappa}_q \Delta t} - e^{-\kappa_\sigma \Delta t}}{\kappa_\sigma - \tilde{\kappa}_q}\right] + \tilde{\mu}_q \left[\frac{1-e^{-\kappa_\sigma \Delta t}}{\kappa_\sigma} - \frac{e^{-\tilde{\kappa}_q \Delta t} - e^{-\kappa_\sigma \Delta t}}{\kappa_\sigma - \tilde{\kappa}_q}\right].$$

Since $\exp(-x)$ and $(1-\exp(-x))/x$ are both decreasing functions in x, it follows that $E_t^Q q_{t+\Delta t} > E_t^P q_t$ for $\tilde{\kappa}_q < \kappa_q$ ($A_q < 0$). Similarly, it it is possible to show that for any positive $\tilde{\kappa}_q$, κ_σ, and Δt, the expressions in square brackets in the second equation above are both greater than zero. Thus, the variance and volatility-of-volatility are both expected to be higher under the risk-neutral measure. Since the integrated variance in (C.12) depends on future values of q_t and $\sigma_{g,t}^2$, the variance premium $vp_t \equiv E_t^Q IV_{t,t+N} - E_t^P IV_{t,t+N}$ must be positive.

Going one step further, the variance premium may be expressed as:

$$vp_t \;=\; \beta_{pr,0} + \beta_{pr,1} q_t, \tag{C.13}$$

where

$$\beta_{pr,0} = (\tilde{\mu}_\sigma - \mu_\sigma)\left[N - \frac{1-e^{-\kappa_\sigma N}}{\kappa_\sigma}\right] + (\tilde{\mu}_q - \mu_q)(A_\sigma^2 + A_q^2 \varphi_q^2)\left[N - \frac{1-e^{-\kappa_q N}}{\kappa_q}\right] - \beta_{pr,1}\tilde{\mu}_q,$$

$$\beta_{pr,1} = \left[\frac{1-e^{-\tilde{\kappa}_q N}}{\tilde{\kappa}_q} - \frac{1-e^{-\kappa_\sigma N}}{\kappa_\sigma}\right]\frac{(\theta-1)A_\sigma}{\kappa_\sigma - \tilde{\kappa}_q} + (A_\sigma^2 + A_q^2\varphi_q^2)\left[\frac{1-e^{-\tilde{\kappa}_q N}}{\tilde{\kappa}_q} - \frac{1-e^{-\kappa_q N}}{\kappa_q}\right].$$

The expression in (C.13) is obtained by taking the difference between the expectations of the integrated variance under the objective measure,

$$E_t^P IV_{t,t+N} = \mu_\sigma N + \frac{1-e^{-\kappa_\sigma N}}{\kappa_\sigma}(\sigma_{g,t}^2 - \mu_\sigma) + (A_\sigma^2 + A_q^2\varphi_q^2)\left(\mu_q N + \frac{1-e^{-\kappa_q N}}{\kappa_q}(q_t - \mu_q)\right),$$

and under the risk-neutral measure,

$$E_t^Q IV_{t,t+N} = \beta_{VIX,0} + \beta_{VIX,\sigma}\sigma_{g,t}^2 + \beta_{VIX,q} q_t^2, \tag{C.14}$$

where

$$\beta_{VIX,0} = \tilde{\mu}_\sigma \left[T - \frac{1-e^{-\kappa_\sigma T}}{\kappa_\sigma} \right] +$$
$$+ \left[\frac{(\theta-1)A_\sigma}{\kappa_\sigma - \tilde{\kappa}_q} \tilde{\mu}_q \right] \left[\frac{1-e^{-\kappa_\sigma T}}{\kappa_\sigma} - \frac{1-e^{-\tilde{\kappa}_q T}}{\tilde{\kappa}_q} \right] + (A_\sigma^2 + A_q^2 \varphi_q^2) \left[T - \frac{1-e^{-\tilde{\kappa}_q T}}{\tilde{\kappa}_q} \right] \tilde{\mu}_q,$$

$$\begin{aligned} \beta_{VIX,\sigma} &= \frac{1-e^{-\kappa_\sigma T}}{\kappa_\sigma}, \\ \beta_{VIX,q} &= \frac{(\theta-1)A_\sigma}{\kappa_\sigma - \tilde{\kappa}_q} \left[\frac{1-e^{-\tilde{\kappa}_q T}}{\tilde{\kappa}_q} - \frac{1-e^{-\kappa_\sigma T}}{\kappa_\sigma} \right] + (A_\sigma^2 + A_q^2 \varphi_q^2) \frac{1-e^{-\tilde{\kappa}_q T}}{\tilde{\kappa}_q}. \end{aligned}$$

As discussed further in the main text, the expectation under the risk-neutral measure corresponds directly to the VIX^2 volatility index, hence the subscript notation for the β's.

C.2.3 Return-Volatility Cross-Correlations

The return over a short time-interval Δt is approximately equal to:

$$\Delta \log(R_t) \approx \left(\frac{\mu_g}{\psi} + \rho - \frac{\sigma_{g,t}^2}{2} + \frac{\gamma}{2}(1 - \frac{1}{\psi})\sigma_{g,t}^2 - A_q(\kappa_1 + \kappa_q) q_t \right) \Delta t +$$
$$+ \sigma_{g,t} \Delta W_t^c + A_\sigma \sqrt{q_t} \Delta W_t^\sigma + A_q \varphi_q \sqrt{q_t} \Delta W_t^q,$$

where the operator Δ denotes the increment to the process over the $[t, t+\Delta]$ time-interval. The variance of the return equals:

$$\mathrm{Var}(\Delta \log R_t) = \left[\beta_{R,\sigma}^2 \mathrm{Var}\left(\sigma_{g,t}^2\right) + \beta_{R,q}^2 \mathrm{Var} q_t \right] (\Delta t)^2 + \mu_\sigma \Delta t + \left[A_\sigma^2 + A_q^2 \varphi_q^2 \right] \mu_q \Delta t,$$

where $\beta_{R,\sigma} = -0.5 + \frac{\gamma}{2}(1 - \frac{1}{\Psi})$ and $\beta_{R,q} = -A_q(\kappa_1 + \kappa_q)$.

From equation (C.13), the variance premium is directly proportional to q_t. Hence, the correlation between the premium and the return is solely determined by the correlation of the return with the q_t process. The covariance between q_t and a future

return, $\Delta \log R_{t+l}$ and $l > 0$, is equal to the covariance of q_t with the drift part of the return:

$$\text{cov}(q_t, \Delta \log R_{t+l}) = e^{-\kappa_q l} \text{Var} q_t \beta_{R,q} \Delta t, \quad l \geq 0.$$

The covariance of q_t with the past return, $\Delta \log R_{t-l}$ and $l < 0$, consists of two parts. The covariance with the drift $e^{-\kappa_q l} \text{Var} q_t \beta_{R,q} \Delta t$, and the covariance with the diffusive part $A_q \varphi_q \text{cov}(q_t, \sqrt{q_{t-l}} \Delta W^q_{t-l}) = A_q \varphi_q^2 \mu_q e^{-\kappa_q l} \Delta t$. Combining these effects, the cross-correlation function for the variance risk premium and the returns may be conveniently expressed as:

$$\text{corr}(q_t, \Delta \log R_{t+l}) = \frac{\left(\text{Var} q_t \beta_{R,q} + I_{l<0} A_q \varphi_q^2 \mu_q\right) e^{-\kappa_q |l|} \Delta t}{\sqrt{\text{Var} q_t} \sqrt{\text{Var}(\Delta \log R_t)}},$$

for any value of l.

The expression for the $\text{VIX}_t^2 \equiv E_t^Q IV_{t,t+N}$ in (C.14) involves a linear function of q_t and the variance $\sigma_{g,t}$, with loadings $\beta_{VIX,q}$ and $\beta_{VIX,\sigma}$, respectively. The covariance of the VIX_t^2 with any future return depends solely on the covariance with the drift of the return:

$$\text{cov}(\text{VIX}_t^2, \Delta \log R_{t+l}) = \left(\beta_{R,\sigma} \beta_{VIX,\sigma} \text{Var} \sigma_{g,t}^2 e^{-\kappa_\sigma l} + \beta_{R,q} \beta_{VIX,q} \text{Var} q_t e^{-\kappa_q l}\right) \Delta t.$$

The covariance of the VIX_t^2 with past returns includes the covariances with the drift and the diffusion:

$$\begin{aligned}\text{cov}(\text{VIX}_t^2, \Delta \log R_{t-l}) &= \beta_{VIX,\sigma}\ \beta_{R,\sigma} e^{-\kappa_\sigma l} \text{Var} \sigma_{g,t}^2 \Delta t + \beta_{VIX,\sigma}\ A_\sigma \text{cov}(\sigma_{g,t}^2, \sqrt{q_{t-l}} dW^\sigma_{t-l}) + \\ &\quad + \beta_{VIX,q}\ \beta_{R,q} e^{-\kappa_q l} \text{Var} q_t \Delta t + \beta_{VIX,q}\ A_q \varphi_q \text{cov}(q_t, \sqrt{q_{t-l}} dW^q_{t-l}) \\ &= \left([\beta_{R,\sigma} \text{Var} \sigma_{g,t}^2 + A_\sigma \mu_q] \beta_{VIX,\sigma} e^{-\kappa_\sigma l} + [\beta_{R,q} \text{Var} q_t + A_q \varphi_q^2 \mu_q] \beta_{VIX,q} e^{-\kappa_q l}\right) \Delta t.\end{aligned}$$

Combining these expressions, the cross-correlations between the VIX_t^2 and the returns

may be succinctly written:

$$\operatorname{corr}(\mathrm{VIX}_t^2, \Delta \log R_{t+l}) =$$

$$\frac{\left(\beta_{R,\sigma}\mathrm{Var}\sigma_{g,t}^2 + I_{l<0}A_\sigma\mu_q\right)\beta_{VIX,\sigma}e^{-\kappa_\sigma|l|}\Delta t + \left(\beta_{R,q}\mathrm{Var}q_t + I_{l<0}A_q\varphi_q^2\mu_q\right)\beta_{VIX,q}e^{-\kappa_q|l|}\Delta t}{\sqrt{\left(\beta_{VIX,\sigma}^2\mathrm{Var}\sigma_{g,t}^2 + \beta_{VIX,q}^2\mathrm{Var}q_t\right)\mathrm{Var}(\Delta \log R_t)}},$$

for any value of l.

C.3 General Model Solution

To allow for more flexible volatility dynamics, we assume that the process for the consumption variance has a general MA-representation:

$$\sigma_{g,t}^2 = \sigma^2 + \int_{-\infty}^{t} a(t-s)\sqrt{q_s}d.W_s^{\sigma} \tag{C.15}$$

This specification includes the short-memory model in equation (C.2) above as a special case, but importantly allows for much richer dynamic dependencies, including long-memory in which the $a(t-s)$ coefficients decrease at a slow hyperbolical rate. We maintain the identical short-memory process for the volatility-of-volatility in equation (C.3).

The pricing relation in (C.4) remains the same. In parallel to the solution method for the short-memory model used above, the linearization of the price-dividend ratio reduces the problem to a system of linear equations. In general, all the shocks in (C.15) need to be included in the conjectured solution for the dividend-price ratio:

$$\log \Psi_t = A_0 + A_q q_t + \int_{-\infty}^{t} A(t-s)\sqrt{q_s}dW_s^{\sigma}. \tag{C.16}$$

Following [92], if the price is a semi-martingale, as it must be to prevent arbitrage, and $A(t)$ exists and is differentiable at zero, the dynamics of the price-dividend ratio may be decomposed into separate drift and diffusion terms:

$$d\log \Psi_t = \left[A_q\kappa_q(\mu_q - q_t) + \int_{-\infty}^{t} A'(t-s)\sqrt{q_s}dW_s^{\sigma}\right]dt + A_q\varphi_q\sqrt{q_t}dW_t^q + A(0)\sqrt{q_t}dW_t^{\sigma}.$$

Now, substituting the conjectured solution into the pricing equation imply that the loadings on the variance shocks must satisfy:

$$A'(t-s) - \kappa_1 A(t-s) = \frac{\gamma(1-\frac{1}{\psi})}{2}a(t-s),$$

for all $t \geq s$. Solving this system along with the constant and the loading on q_t in equation (C.16) we obtain:

$$A(t) = -\int_t^{+\infty} \frac{\gamma(1-\frac{1}{\psi})}{2} e^{\kappa_1(t-\tau)} a(\tau) d\tau, \tag{C.17}$$

$$A_q = \frac{\kappa_q + \kappa_1 - \sqrt{(\kappa_q+\kappa_1)^2 - \theta^2\varphi_q^2 A(0)^2}}{\theta\varphi_q^2}, \tag{C.18}$$

$$A_0 = \frac{A_q\kappa_q\mu_q - \kappa_0 + (1-\frac{1}{\psi})\mu_g - \rho - \frac{\gamma}{2}(1-\frac{1}{\psi})\sigma^2}{\kappa_1}. \tag{C.19}$$

If absolute values of the coefficients $a(t)$ are decreasing in t (or grow at a rate less than exponential) and $|a(t)| < \infty$, then $A(t)$ is well-defined and $A'(0) = \kappa_1 A(0) + \frac{\gamma(1-\frac{1}{\psi})}{2} a(0)$ is finite.

Substituting the solution for the price-dividend ratio into the expression for the SDF, it follows that

$$\begin{aligned}\frac{dM_t}{M_t} = &\left[-\frac{\mu_g}{\psi} - \rho + \frac{\gamma}{2}(1+\frac{1}{\psi})\sigma_{g,t}^2 + (\frac{1}{\theta}-1)(\kappa_1+\kappa_q)A_q q_t\right] dt \\ &- \gamma\sigma_{g,t} dW_t^c + (\theta-1)A_q\varphi_q\sqrt{q_t} dW_t^q + (\theta-1)A(0)\sqrt{q_t} dW_t^\sigma.\end{aligned}$$

As before, the risk-free rate is simply defined by the drift of M_t:

$$r_t^{rf} \;=\; -E_t\frac{dM_t}{M_t dt} \;=\; \frac{\mu_g}{\psi} + \rho - \frac{\gamma}{2}(1+\frac{1}{\psi})\sigma_{g,t}^2 + (1-\frac{1}{\theta})(\kappa_1+\kappa_q)A_q q_t.$$

Since $d\log R_t = d\log C_t + d\log\Psi_t + \Psi_t^{-1}dt$, and therefore

$$d\log R_t \;=\; D\log R_t dt + \sigma_{g,t} dW_t^c + A_q\varphi_q\sqrt{q_t} dW_t^q + A(0)\sqrt{q_t} dW_t^\sigma, \tag{C.20}$$

the equity premium equals:

$$\pi_{r,t} \;=\; \gamma\sigma_{g,t}^2 + (1-\theta)[A_q^2\varphi_q^2 + A(0)^2]q_t = \gamma\sigma_{g,t}^2 + 2\left(\frac{1}{\theta}-1\right)(\kappa_q+\kappa_1)A_q q_t. \tag{C.21}$$

And, the dynamics of the return is determined by:

$$d\log R_t = \left[\rho + \frac{\mu_g}{\psi} + [-\frac{1}{2} + \frac{\gamma}{2}(1-\frac{1}{\psi})]\sigma^2_{g,t} - (\kappa_q + \kappa_1)A_q q_t\right]dt + \sigma_{g,t}dW^c_t + A_q\varphi_q\sqrt{q_t}dW^q_t + A(0)\sqrt{q_t}dW^\sigma_t.$$

The integrated variance may generally be expressed as:

$$IV_{t,t+N} = \int_t^{t+N}\sigma^2_{g,\tau}d\tau + (A^2_q\varphi^2_q + A(0)^2)\int_t^{t+N} q_\tau d\tau. \tag{C.22}$$

The expected value of the integrated variance under the objective measure equals:

$$E^P_t IV_{t,t+N} = \sigma^2 N + \int_{-\infty}^t\left[\int_t^{t+N} A(\tau-s)d\tau\right]\sqrt{q_s}dW^\sigma_s + (A^2_q\varphi^2_q + A(0)^2)\left[\mu_q N + \frac{1-e^{-\kappa_q N}}{\kappa_q}(q_t-\mu_q)\right].$$

The expectation under the risk-neutral measure is:

$$E^Q_t IV_{t,t+N} = \sigma^2 N + \int_{-\infty}^t\left[\int_t^{t+N} A(\tau-s)d\tau\right]\sqrt{q_s}dW^\sigma_s + E^Q_t\int_t^{t+N}\int_t^\tau A(\tau-s)\sqrt{q_s}dW^\sigma_s d\tau + (A^2_q\varphi^2_q + A(0)^2)\left[\tilde{\mu}_q N + \frac{1-e^{-\tilde{\kappa}_q N}}{\tilde{\kappa}_q}(q_t-\tilde{\mu}_q)\right], \tag{C.23}$$

where $\tilde{\kappa}_q = \kappa_q - (\theta-1)A_q\varphi^2_q$ refers to the mean-reversion of q_t under the risk-neutral probability, and $\tilde{\mu}_q = \kappa_q/\tilde{\kappa}_q\mu_q$ denotes the corresponding expectation. Moreover, under the risk-neutral measure:

$$E^Q_t\sqrt{q_s}dW^\sigma_s = E^Q_t[\sqrt{q_s}d\tilde{W}^\sigma_s + (\theta-1)A(0)q_s ds] = (\theta-1)A(0)(\tilde{\mu}_q + e^{-\tilde{\kappa}_q(s-t)}(q_t-\tilde{\mu}_q))ds.$$

Consequently, the variance risk premium defined by the difference between $E^Q_t IV_{t,t+N}$ and $E^P_t IV_{t,t+N}$ is again a linear function of q_t:

$$vp_t = \beta_{pr,0} + \beta_{pr,1}q_t, \tag{C.24}$$

with the two coefficients now defined by:

$$\begin{aligned}\beta_{pr,0} &= (\theta-1)A(0)\tilde{\mu}_q \int_t^{t+N}\int_t^{\tau} A(\tau-s)(1-e^{-\tilde{\kappa}_q(s-t)})dsd\tau + \\ &\quad + (A_q^2\varphi_q^2 + A(0)^2)\left[\tilde{\mu}_q(N-\frac{1-e^{-\tilde{\kappa}_q N}}{\tilde{\kappa}_q}) - \mu_q(N-\frac{1-e^{-\kappa_q N}}{\kappa_q})\right],\end{aligned}$$

$$\begin{aligned}\beta_{pr,1} &= (\theta-1)A(0) \int_t^{t+N}\int_t^{\tau} A(\tau-s)e^{-\tilde{\kappa}_q(s-t)}dsd\tau + \\ &\quad + (A_q^2\varphi_q^2 + A(0)^2)\left[\frac{1-e^{-\tilde{\kappa}_q N}}{\tilde{\kappa}_q} - \frac{1-e^{-\kappa_q N}}{\kappa_q}\right].\end{aligned}$$

C.4 Time Series Plots

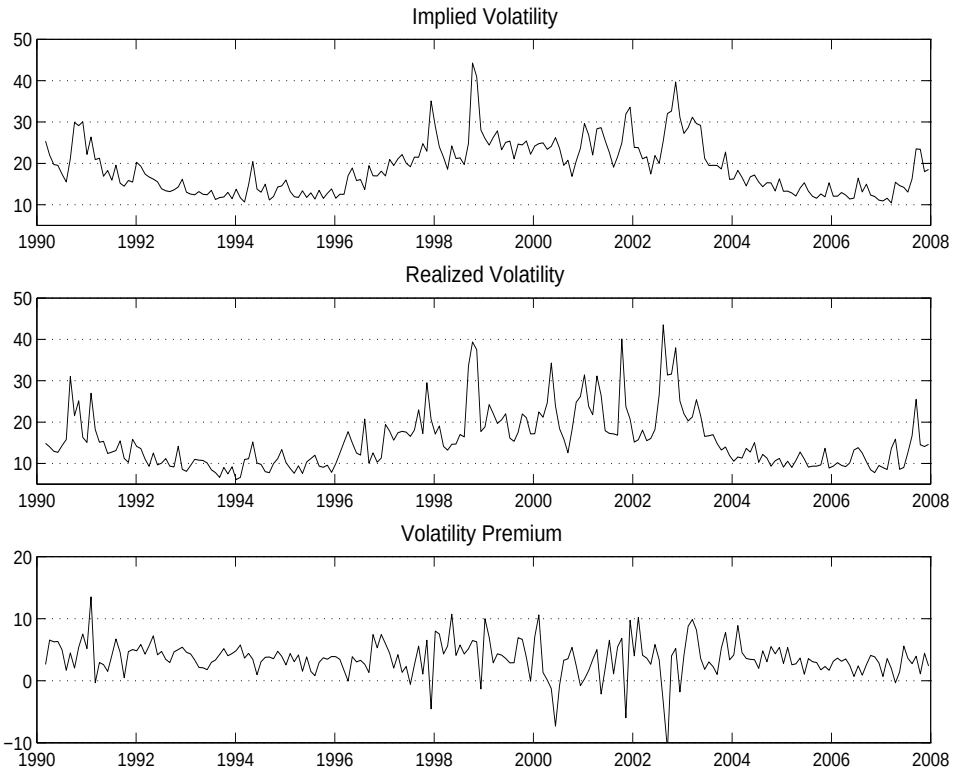

FIGURE C.1: The figure shows the VIX_t implied volatility index, the realized volatility $RV_{t,t+22}^{1/2}$, and the volatility risk premium $\hat{vp}_t = VIX_t - E_t(RV_{t,t+22})^{1/2}$ over the January 2, 1990 to October 31, 2007 sample period. All of the volatility measures are plotted at the monthly frequency in annualized percentage units. The realized volatilities are constructed from the summation of high-frequency five-minute squared returns. The expectations for the future variances $\hat{E}_t RV_{t,t+22}$ are based on the HAR-RV forecasting model discussed in the text.

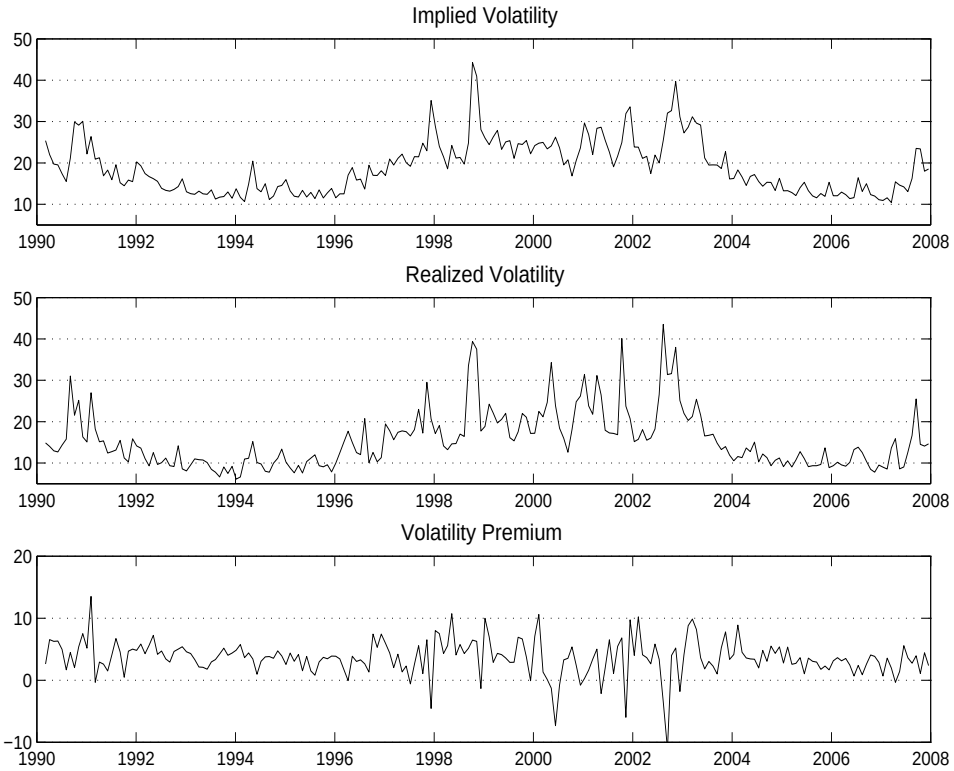

FIGURE C.2: The figure shows the VIX_t implied volatility index, the realized volatility $RV_{t,t+22}^{1/2}$, and the volatility risk premium $\hat{v}p_t = VIX_t - E_t(RV_{t,t+22})^{1/2}$ over the September 23, 2003 to August 31, 2007 sample period. All of the volatility measures are plotted at the daily frequency in annualized percentage units. The realized volatilities are constructed from the summation of high-frequency five-minute squared returns. The expectations for the future variances $\hat{E}_t RV_{t,t+22}$ are based on the HAR-RV forecasting model discussed in the text.

Bibliography

[1] Yacine Aït-Sahalia and Andrew W. Lo. Nonparametric risk management and implied risk aversion. *Journal of Econometrics*, 94:9–51, 2000.

[2] Yacine Aït-Sahalia, Per C. Mykland, and Lan Zhang. How often to sample a continuous-time process in the presence of market microstructure noise. *Review of Financial Studies*, 18(2):351–416, 2005.

[3] Sassan Alizadeh, Michael W. Brandt, and Francis X. Diebold. Range-based estimation of stochastic volatility models. *Journal of Finance*, 57(3):1047–1091, 2002.

[4] Torben G. Andersen, Luca Benzoni, Eric Ghysels, and Jesper Lund. An empirical investigation of continuous-time equity return models. *Journal of Finance*, 57(3):1239 – 1284, 2002.

[5] Torben G. Andersen and Tim Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4:115–158, 1997.

[6] Torben G. Andersen and Tim Bollerslev. Deutsche mark-dollar volatility: Intraday activity patterns, macroeconomic announcements, and longer run dependencies financial markets. *Journal of Finance*, 53(1):219–265, 1998.

[7] Torben G. Andersen, Tim Bollerslev, Peter F. Christoffersen, and Francis X. Diebold. Volatility forecasting. 2005. Working Paper, NBER.

[8] Torben G. Andersen, Tim Bollerslev, Peter F. Christoffersen, and Francis X. Diebold. *Risks of Financial Institutions*, chapter Practical Volatility and Correlation Modeling for Financial Market Risk Management, pages 513–548. University of Chicago Press for NBER, 2006.

[9] Torben G. Andersen, Tim Bollerslev, and Francis X. Diebold. Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *Review of Economics and Statistics*, 89:701–720, 2007.

[10] Torben G. Andersen, Tim Bollerslev, and Francis X. Diebold. *Handbook of Financial Econometrics*, chapter Parametric and Nonparametric Volatility Measurement. Elsevier Science B.V., Amsterdam, 2009. Forthcoming.

[11] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71:579–625, 2003.

[12] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Clara Vega. Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics*, 73:251–277, 2007.

[13] Torben G. Andersen, Tim Bollerslev, and Xin Huang. A semiparametric framework for modeling and forecasting jumps and volatility in speculative prices. 2008. Working paper, Duke University.

[14] Torben G. Andersen, Tim Bollerslev, and Nour Meddahi. Analytical evaluation of volatility forecasts. *International Economic Review*, 45, 2004. 1079-1110.

[15] Richard T. Baillie, Tim Bollerslev, and Hans-Ole Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74:3–30, 1996.

[16] Federico Bandi and Benoit Perron. Long memory and the relation between implied and realized volatility. *Journal of Financial Econometrics*, 4:636–670, 2006.

[17] Ravi Bansal and Ivan Shaliastovich. Confidence risk and asset prices. 2008. Working paper, Duke University.

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

[19] Ole E. Barndorff-Nielsen, Peter R. Hansen, Asger Lunde, and Neil Shephard. Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. 2007. Working Paper, Oxford University.

[20] Ole E. Barndorff-Nielsen, Peter R. Hansen, Asger Lunde, and Neil Shephard. Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise. *Econometrica*, 76:1481 – 1536, 2008.

[21] Ole E. Barndorff-Nielsen and Neil Shephard. Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *Journal of Royal Statistical Society, Series B*, 64:253–280, 2002.

[22] Ole E. Barndorff-Nielsen and Neil Shephard. *Identification and Inference for Econometric Models. Essays in Honor of Thomas Rothenberg*, chapter How Accurate is the Asymptotic Approximation to the Distribution of Realized Volatility, pages 513–548. Cambridge University Press, Cambridge, 2005.

[23] Geert Bekaert, Eric Engstrom, and Yuhang Xing. Risk, uncertainty and asset prices. *Journal of Financial Economics*, 2009. Forthcoming.

[24] Luca Benzoni, Pierre Collin-Dufresne, and Robert S. Goldstein. Can standard preferences explain the prices of out-of-the-money s&p 500 put options? 2005. Working Paper No.11861, National Bureau of Economic Research.

[25] Fischer Black. Studies in stock price volatility changes. *Proceedings from the American Statistical Association, Business and Economics Statistics Section*, pages 177–181, 1976.

[26] Tim Bollerslev and Robert Engle. Modeling the persistence of conditional variances. *Econometric Reviews*, 5:1 – 50, 1986.

[27] Tim Bollerslev, Mike Gibson, and Hao Zhou. Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities. 2006. Working Paper, Federal Reserve Board.

[28] Tim Bollerslev, Lulia Litvinova, and George Tauchen. Leverage and volatility feedback effects in high-frequency data. *Journal of Financial Econometrics*, 4:353–384, 2006.

[29] Tim Bollerslev, George Tauchen, and Hao Zhou. Expected stock returns and variance risk premia. *Review of Financial Studies*, 2009.

[30] Tim Bollerslev and Hao Zhou. Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *Journal of Econometrics*, 109(1):33 – 65, 2002.

[31] Tim Bollerslev and Hao Zhou. Volatility puzzles: A simple framework for gauging return-volatility regressions. *Journal of Econometrics*, 131:123–150, 2006.

[32] Mark Britten-Jones and Anthony Neuberger. Option prices, implied price processes, and stochastic volatility. *Journal of Finance*, 55:839–866, 2000.

[33] John M. Campbell and L. M. Viceira. *Strategic Asset Allocation: Portfolio Choice for Long-Term Investors*. Oxford University Press, Oxford, U.K., 2002.

[34] John Y. Campbell. Understanding risk and return. *Journal of Political Economy*, 104:298–345, 1996.

[35] John Y. Campbell and Ludger Hentschell. No news is good news: An asymmetric model of changing volatility in stock returns. *Journal of Financial Economics*, 31:281–318, 1992.

[36] Peter Carr and Dilip Madan. Towards a theory of volatility trading. *in Volatility: New Estimation Techniques for Pricing Derivatives*, 1998. chap.29, 417-427, Robert Jarrow (ed.). London: Risk Books.

[37] Peter Carr and Liuren Wu. Variance risk premia. *Review of Financial Studies*, 2007. Forthcoming.

[38] Mikhail Chernov, A. Ronald Gallant, Eric Ghysels, and George Tauchen. Alternative models for stock price dynamics. *Journal of Econometrics*, 116(1-2):225 – 257, 2003.

[39] Fabio Comte and Eric Renault. Long memory continuous time models. *Journal of Econometrics*, 73:101 – 149, 1996.

[40] Fulvio Corsi. A simple long memory model of realized volatility. 2004. Working Paper, University of Southern Switzerland.

[41] Z. Ding, Clive W.J. Grange, and Robert F. Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1:83–106, 1993.

[42] Itamar Drechsler and Amir Yaron. Whats vol got to do with it? 2008. Working Paper, University of Pennsylvania.

[43] Darrell Duffie and Larry Epstein. Stochastic differential utility. *Econometrica*, 60:353–394, 1992a.

[44] Darrell Duffie and Larry Epstein. Asset pricing with stochastic differential utility. *Review of Financial Studies*, 5:411–436, 1992b.

[45] Darrell Duffie, Jun Pan, and Kenneth Singleton. Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, 68(6):1343–1376, 2000.

[46] Garland B. Durham. Monte carlo methods for estimating, smoothing, and filtering one and two-factor stochastic volatility models. *Journal of Econometrics*, 133:273–305, 2006.

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

[48] Larry G. Epstein and Stanley E. Zin. Substitution, risk aversion, and the temporal behavior of consumption and asset returns: An empirical analysis. *Journal of Political Economy*, 99:263–286, 1991.

[49] Bjorn Eraker. The volatility premium. *Mathematical Finance*, 2008. Working Paper, Department of Finance, University of Wisconsin.

[50] Bjørn Eraker, Michael Johannes, and Nicholas Polson. The impact of jumps in volatility and returns. *Journal of Finance*, 58(3):1269 – 1300, 2003.

[51] Bjorn Eraker and Ivan Shaliastovich. An equilibrium guide to designing affine pricing models. *Mathematical Finance*, 18:519–543, 2008.

[52] Cheol S. Eun and Bruce G. Resnick. Estimating the correlation structure of international share prices. *Journal of Finance*, 39(5):1311–1324, 1984.

[53] Stephen Figlewski and Urich Thomas. Optimal aggregation of money supply forecasts: Accuracy, profitability, and market efficiency. *Journal of Finance*, 38:695–710, 1983.

[54] Stephen Figlewski and Xiaozu Wang. Is the 'leverage effect' a leverage effect? 2001. Working Paper, Stern School of Business, New York University.

[55] Jeff Fleming, Chris Kirby, and Barbara Ostdiek. The economic value of volatility timing. *Journal of Finance*, 56:329–352, 2001.

[56] Jeff Fleming, Chris Kirby, and Barbara Ostdiek. The economic value of volatility timing using "realized" volatility. *Journal of Financial Economics*, 67:473–509, 2003.

[57] Kenneth F. French, G. William Schwert, and Robert Stambaugh. Expected stock returns and volatility. *Journal of Financial Economics*, 19:3–29, 1987.

[58] Xavier Gabaix. Variable rare disasters: A tractable theory of ten puzzles in macro-finance. *American Economic Review*, 2008.

[59] A. Ronald Gallant and George Tauchen. Reprojecting partially observed systems with application to interest rate diffusions. *Journal of the American Statistical Association*, 93(441):10–24, 1998.

[60] Raffaella Giacomini and Halbert White. Test of conditional predictive ability. *Econometrica*, 74(6):1545 – 1578, 2006.

[61] Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48:1779–1801, 1993.

[62] A. Gloter and J. Jacod. Diffusions with measurement errors. i local asymptotic normality. *ESAIM: Probability and Statistics*, 5:225–242, 2001a.

[63] Arnoud Gloter and Jean Jacod. Diffusions with measurement errors. ii measurement errors. *ESAIM: Probability and Statistics*, 5:243260, 2001b.

[64] Stephen Gordon and Pascal St-Amour. Asset returns and state-dependent risk preferences. *Journal of Business and Economic Statistics*, 22:241–252, 2004.

[65] Hui Guo. On the out-of-sample predictability of stock market returns. *Journal of Business*, 79:645 – 670, 2006.

[66] James Douglas Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[67] Steven Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6:327–343, 1993.

[68] Chi-Fu Huang and Robert H. Litzenberger. *Foundations for financial economicss*. Prentice Hall, New York, 1998.

[69] Xin Huang and George Tauchen. The relative contribution of jumps to total price variance. *Journal of Financial Econometrics*, 3(4):456 – 499, 2005.

[70] Namwon Hyung, Ser-Huang Poon, and Clive W.J. Granger. *Forecasting in the Presence of Structural Breaks and Model Uncertaintys*, chapter The Source of Long Memory in Financial Market Volatility. Elsevier Series, Amsterdam, 2006. Forthcoming.

[71] Eric Jacquier, Nicholas G. Polson, and Peter Rossi. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, 12(4):371 – 389, 1994.

[72] George Jiang and Yisong Tian. Model-free implied volatility and its information content. *Review of Financial Studies*, 18:1305–1342, 2005b.

[73] George J. Jiang and Yisong S. Tian. Extracting model-free volatility from option prices: An examination of the vix index. *Journal of Derivatives*, 14:1–26, 2007.

[74] Michael Johannes, Arthur Korteweg, and Nicholas Polson. Sequential learning, predictive regressions, and optimal portfolio returns. 2008. Working paper, University of Chicago, Graduate School of Business.

[75] Michael Johannes and Nicholas Polson. *Handbook of Financial Econometrics*, chapter MCMC Methods for Continuous Time Financial Econometrics. University of Chicago, 2003.

[76] Michael Johannes, Nicholas Polson, and Jonathan Stroud. Sequential optimal portfolio performance: Market and volatility timing. 2003. Working paper, University of Chicago, Graduate School of Business.

[77] Gordon Leitch and J. Ernest Tanner. Economic forecast evaluation:profits versus the conventional error measures. *American Economic Review*, 81(3):580, 1991.

[78] Martin Lettau, Sydney C. Ludvigson, and Jessica A. Wachter. The declining equity premium: What role does macroeconomic risk play? *Review of Financial Studies*, 2009.

[79] Larry J. Lockwood and Scott C. Linn. Examination of Stock Market Return Volatility During Overnight and Intraday Periods,1964-1989. *Journal of Finance*, 45(2):591–601, 1990.

[80] Nour Meddahi. An eigenfunction approach for volatility modeling. 2001. Working paper, Centre de Recherche et Développement en Économique.

[81] Nour Meddahi. Theoretical comparison between integrated and realized volatility. *Journal of Applied Econometrics*, 17:479–508, 2002.

[82] Nour Meddahi. Arma representation of integrated and realized variances. *Econometrics Journal*, 6(2):335 – 356, 2003.

[83] Robert C. Merton. An intertemporal capital asset pricing model. *Econometrica*, 41:867–887, 1973.

[84] Daniel B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Ecometrica*, 59:347–370, 1991.

[85] Daniel B. Nelson. Filtering and forecasting with misspecified arch models i. *Journal of Econometrics*, 52:61–90, 1992.

[86] Daniel B. Nelson and Dean P. Foster. Asymptotic filtering theory for univariate arch models. *Econometrica*, 62(1):1 – 41, 1994.

[87] Morten O. Nielsen. Local whittle analysis of stationary fractional cointegration and the implied-realized volatility relation. *Journal of Business and Economic Statistics*, 25:427–446, 2007.

[88] Andrew J. Patton and Allan Timmermann. Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics*, 140(2):884 – 918, 2007a.

[89] Andrew J. Patton and Allan Timmermann. Testing forecast optimality under unknown loss. *Journal of the American Statistical Association*, 102(480):1172 – 1184, 2007b.

[90] Michael K. Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590 – 599, 1999.

[91] Peter Robinson. Testing for strong serial correlation and dynamic conditional heteroskedasticity in multiple regression. *Journal of Econometrics*, 47:67–84, 1991.

[92] Chris Rogers. Arbitrage with fractional brownian motion. *Mathematical Finance*, 7:95–105, 1997.

[93] Pedro Santa-Clara and Shu Yan. Crashes, volatility, and the equity premium: Lessons from s&p 500 options. *Review of Economics and Statistics*, 2009.

[94] Natalia Sizova. Integrated variance forecasting: Model-based versus reduced form. 2008. Working Paper, Department of Economics, Duke University.

[95] Jonathan R. Stroud, Peter Müller, and Nicholas G. Polson. Nonlinear state-space models with state-dependent variances. *Journal of the American Statistical Association*, 98(462):377 – 386, 2003.

[96] George Tauchen. Recent developments in stochastic volatilility: Statistical modelling and general equilibrium analysis. 2004. Working paper, Duke University.

[97] Viktor Todorov. Estimation of continuous-time stochastic volatility models with jumps using high-frequency data. 2007a. Working paper, Duke University.

[98] Viktor Todorov. Variance risk premium dynamics. 2007b. Working Paper, Northwestern University.

[99] Jessica A. Wachter. Can time-varying risk of rare disasters explain aggregate stock market volatility? 2008. Working Paper, Wharton School of Business, University of Pennsylvania.

[100] Kenneth D. West, Hali J. Edison, and Dongchul Cho. Utility-based comparison of some models of exchange rate volatility. *Journal of International Economics*, 35:23–45, 1993.

[101] Lan Zhang, Per A. Mykland, and Yacine Aït-Sahalia. A tale of two time scales: Determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, 100:1394–1411, 2005.

Biography

Natalia Sizova was born in Shikhany, Russia/USSR on February 6, 1981. She earned her Bachelor's degree (2002) and subsequent Master's degree (2004) from the Moscow Institute of Physics and Technology (Fiztech), Department of Applied Mathematics. She graduated with high honors ("Red Diploma") and was certified as an engineer-physicist. Concurrently, she received her M.A. in Economics from the New Economic School, Moscow. She expects to complete her Ph.D. in Economics at Duke University in 2009.