Blending Approach to Forecasting Stock Prices

Julius Caesar Ganaa

Department of Mathematical Sciences
Montana State University

May 9th, 2024

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Julius Caesar Ganaa

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date	Dr. John Smith Writing Project Advisor
Date	Dr. Katharine Banner Writing Projects Coordinator

Abstract

An effective forecasting approach is paramount in financial analysis and decision-making. However, the volatile nature of the stock market makes it challenging, if not impossible, to forecast. This study aims to understand the underlying dynamics of stock prices, to model the behavior of stock markets using advanced statistical techniques, and to forecast future movements of stock prices with a high degree of accuracy. To do this, we adopt a blending approach (Bröcker & Smith, 2007), which takes a weighted average of the forecast and historical climatology distribution to predict stock prices with high accuracy. We built our historical forecast distribution on the training data using kernel density estimation, followed by the development of a stochastic volatility model (SVM) in a Bayesian framework. Subsequently, we developed our blended model by taking a weighted average of the historical and forecast distributions, progressively decreasing the weight on the forecast distribution for predictions extending farther into the future. Our preliminary results based on logarithmic score find that this blended model provides robust forecasts compared to standalone historical or forecast-based models.

Contents

1 Introduction

1.1 Background

The stock market is a critical backbone of the global economy, offering a potentially lucrative platform for those who trade within its confines. Understanding its underlying dynamics is crucial for navigating its complex landscape. However, the market trends remain one of the most difficult things to predict.

In financial markets, accurate forecasting of stock prices is paramount in investment strategies and decision-making. These include but are not limited to hedging, taking advantage of arbitrage opportunities, and many more. The ability to anticipate or predict market movements can safeguard investment and guide economic policy. This transcends beyond individual benefits to society and the world economy at large. Stock markets serve as a conduit for money and liquidity, which are necessary for economic growth and stability (Chikwira & Mohammed, 2023).

It is undeniable the role the stock market plays in terms of advancing economic growth. It is one of the two main sources of capital for most companies (equity financing) – serving as a means of obtaining funds to facilitate the growth of the company which leads to the creation of jobs and economic growth at large.

This study seeks to understand the stock market behavior and how we can use this knowledge to predict stock market trends using stochastic volatility models and historical approaches (kernel density estimation). Volatility is central to many applied issues in finance and financial engineering, ranging from asset pricing and asset location to risk management (Joubert & Vencatasawmy, n.d.). Stochastic volatility models (SVMs) are non-linear state-space models that enjoy increasing popularity for fitting and predicting heteroskedastic time series (Hosszejni & Kastner, 2021). In recent financial econometrics research, stochastic volatility models have been extensively studied to better capture the underlying dynamics of market volatility. One notable advancement includes developments such as the continuous particle filtering technique, which offers a refined and robust alternative to traditional estimation methods. It simplifies the estimation process of complex stochastic volatility models, making it particularly suitable for handling the non-linearities and non-Gaussian features of financial data (Pitt, Malik, & Doucet, 2014). SVMs have been well-studied in statistics and have been used demonstrably in the development of new statistical methodology.

1.2 Motivation

This project is motivated by a dual desire to both broaden and deepen my analytical expertise, particularly in the realm of financial market dynamics, and to stay abreast with the demands of financial markets. At it's core, the aim of this project is to foster a comprehensive understanding of time series analysis – an important component for predicting financial trends. This includes improving practical abilities in handling and interpreting real-world finance data, alongside developing sophisticated statistical modeling techniques and pipelines. The importance of this endeavor extends beyond academic advancement, and is central to contemporary finance practices. In an era where data reinforces important segments such as algorithmic trading, risk evaluation, and portfolio strategy, mastering time series analysis is an indispensable skill.

1.3 Objectives

The aim of this study has three components. First, we seek to investigate the complex mechanisms that drive stock price fluctuations, shedding light on the patterns and behavior of the stock market. Second, we seek to capture and model stock market behaviors through the deployment of sophisticated statistical methods, constructing models that can accurately represent these complex systems. Finally, following from Bröcker and Smith (2007), we aim to leverage a blending approach to predict stock price movements with enhanced precision. This methodology, which combines forecasts from multiple models, is central to achieving our goal of delivering forecasts with a superior level of accuracy, thereby providing valuable insights into future market trends.

2 Methodology

This study aims to forecast future movements of stock prices with a high degree of accuracy. We intend to find a robust model that leverages the strengths of both predictive and historical insights. Following Bröcker and Smith (2007), we employ a blending technique that takes a weighted average of the forecast and historical climatology distribution to predict stock prices with high accuracy. We demonstrate this technique by applying it to Netflix and Apple stock data.

2.1 Data

Our analysis focuses on stock price movements of Apple and Netflix. To procure data, we use the `quantmod` (Ryan & Ulrich, 2024) package in `R` (R Core Team, 2021), a comprehensive tool tailored for financial data analysis and trading strategy development. Using tools from this package, we were able to download accurate historical stock data for both companies, spanning from January 1, 2018, to November 5, 2023. The dataset encompasses detailed daily trading information, including the date, opening prices, closing prices, lowest and highest prices of the day, trading volume, and adjusted prices.

Given the scope of our study, we chose to look at the adjusted prices from the dataset, as they provide a more accurate reflection of the stock's value over time, adjusting for factors like dividends, stock splits, and other corporate actions. We divided the dataset into training and test subsets:

- The first 1,440 data points of the dataset were designated for training our models. This selection allows the models to learn from the historical adjusted price movements of Apple and Netflix stocks, capturing essential trends and patterns.

- We reserved the subsequent 30 data points, immediately following the training data, as our test dataset. This strategic division enables an objective assessment of the models' forecasting capabilities, testing their predictions against actual market performance in a recent period.

2.2 Stochastic Volatility Model

We employed the stochastic volatility model to derive our predictive distribution. The key feature of the SV model is its stochastic and time-varying specification of the variance evolution (Hosszejni & Kastner, 2021). Stochastic volatility models (SVMs) are also crucial for capturing asymmetry and heavy tails in financial returns, providing a flexible framework for modeling such behavior (Vankov, Guindani, & Ensor, 2019), and we model this process as:

$$y_t = \epsilon_t \exp(h_t/2)$$

$$h_{t+1} = \mu + \phi(h_t - \mu) + \delta_t$$

$$h_1 \sim \text{Normal}\left(0, \frac{\sigma}{\sqrt{1-\phi^2}}\right)$$

$$\delta_t \sim \text{Normal}(0, 1)$$

$$\epsilon_t \sim \text{Normal}(0, 1)$$

$$y_t \sim \text{Normal}\left(0, \frac{h_t}{2}\right)$$

$$h_t \sim \text{Normal}(\mu + \phi(h_{t-1} - \mu), \sigma)$$

Table 1: Parameter Descriptions for the Stochastic Volatility Model (SVM)

Parameter	**Description**
y_t	Centered returns on the underlying asset: Apple and Netflix.
h_t	Log volatility, a latent parameter that evolves over time according to the model's dynamics.
μ	Mean log volatility, indicating the long-term average level of volatility in the model.
ϵ_t	Random shock term on the asset returns at time t, assumed to follow a normal distribution.
δ_t	Shock on the volatility term, representing random disturbances in the volatility process.

We estimate parameters using a Bayesian approach via Markov Chain Monte Carlo (Robert & Casella, 2004). MCMC is a flexible technique for parameter estimation, and is able to handle complex posterior distributions such as our SVM. As with any Bayesian analysis, we are required to choose prior distributions for our parameters. The prior distributions that we chose are :

- The autoregressive coefficient, ϕ, was assigned a Uniform prior distribution $\phi \sim Uniform(-1, 1)$. This choice is delibrate to ensure the stationarity of the autoregressive process by constraining ϕ within the bounds of -1 and 1. Stationarity is crucial for the stability and meaningful interpretation of the model.

- For the volatility parameter, σ, we opted for a Cauchy prior, $\sigma \sim Cauchy(0, 5)$. The Cauchy distribution, with its heavy tails, was selected to accommodate the potential for large variations in market volatility, a common characteristic in financial time series data.

- The mean return, μ, was also modeled with a Cauchy distribution, $\mu \sim Cauchy(0, 10)$, allowing for a wide range of possible values given the uncertainty surrounding average market returns over long periods.

We fit our SVM using the `R` package STAN (Stan Development Team, 2023)

2.3 Kernel Density Estimation

We employed kernel density estimation (KDE) to model the distribution of historical stock prices using the training data. Kernel density estimation is a non-parametric modeling approach that is used to estimate an empirical probability distribution for a dataset $\{x_1, ..., x_T\}$. Here, we use KDE to create an empirical probability distribution for the historical stock prices. For our KDE, we used a Gaussian kernel, and for the bandwidth (b) estimation we used Silverman's rule of thumb (Silverman, 1986). Silverman's rule of thumb balances the trade-off between smoothness and accuracy of the density estimate, and is well-suited for estimating unimodal probability distributions. Silverman's rule of thumb, alongside the Gaussian kernel, allow us to estimate our historical distribution $h(x)$ as follows:

$$h(x) \approx \frac{1}{bn} \sum_{i=1}^{n} \mathrm{N}(x_i, \sigma^2)$$
$$b = \left(\frac{4\hat{\sigma^5}}{3n} \right)^{\frac{1}{5}} \approx 1.06\hat{\sigma} n^{-\frac{1}{5}}$$

2.4 Blending

Following Bröcker and Smith (2007) proposition, we blend the forecasts from the stochastic volatility and KDE models using a weighted average approach to form a blended forecast model, $b_t(x)$. This allows us to leverage the strength between historical climatology and the forecast from the stochastic volatility model, therefore creating a more robust forecast that can adapt to different market conditions. For instance, during periods of high volatility, the stochastic volatility model might perform better and, therefore, receive a higher weight in the blend. Conversely, in more stable periods, the kernel density estimation model might perform more reliably.

$$b_t(x) = w_t f_t(x) + (1 - w_t) h(x), \; 0 \leq w_t \leq 1$$

Where w_t adjusts the weighting between the two distributions, f_t is the forecast distribution at time t, and $h(x)$ is the historical distribution

2.5 Schemes for Blending Parameters

In developing our blending model, $b_t(x)$, we employed two weighting schemes: linear and exponential decaying weights. Recognizing that stochastic volatility models tend to decrease in reliability for predictions extended further into the future, we strategically selected these weights to gradually shift emphasis between models. Specifically, as the forecast horizon extends, we increase the weight allocated to the historical model, simultaneously decreasing the weight on the predictive distribution derived from the stochastic volatility model. This approach aims to balance the immediate accuracy of stochastic models with the enduring stability offered by historical trends, optimizing our blended forecasts for both short-term precision and long-term consistency.

2.5.1 Exponential Decay Weights

We designed the exponential decay weights to give more importance to recent predictions, diminishing the weights exponentially as we predict farther into the future. The formula used to calculate the exponential decay weights is given below:

$$w_t = \exp(-\lambda t)$$

Where:

- w_t is the weight assigned to the tth prediction,
- λ is the decay rate, determining the rate at which the weights decrease.

For our analysis, we chose the decay rate, λ to 0.04621 with $t = 1, 2, \ldots, 30$

$$w_t = \exp(-0.04621t), \; t = 1, 2, \ldots, 30$$

2.5.2 Linear Decay Weights

For the linear decay weights, we designed an approach that reduces the weight on each subsequent prediction by a constant amount. The formula used to derived these weights is given below:

$$w_t = \frac{N - t}{N - 1}$$

Where:

N is the total number of predicted data points, t is the index of the predicted data point.

For our analysis, $N = 30$ and the weights are calculated as:

$$w_t = \frac{30 - t}{29},\ t = 1, 2, \ldots, 30$$

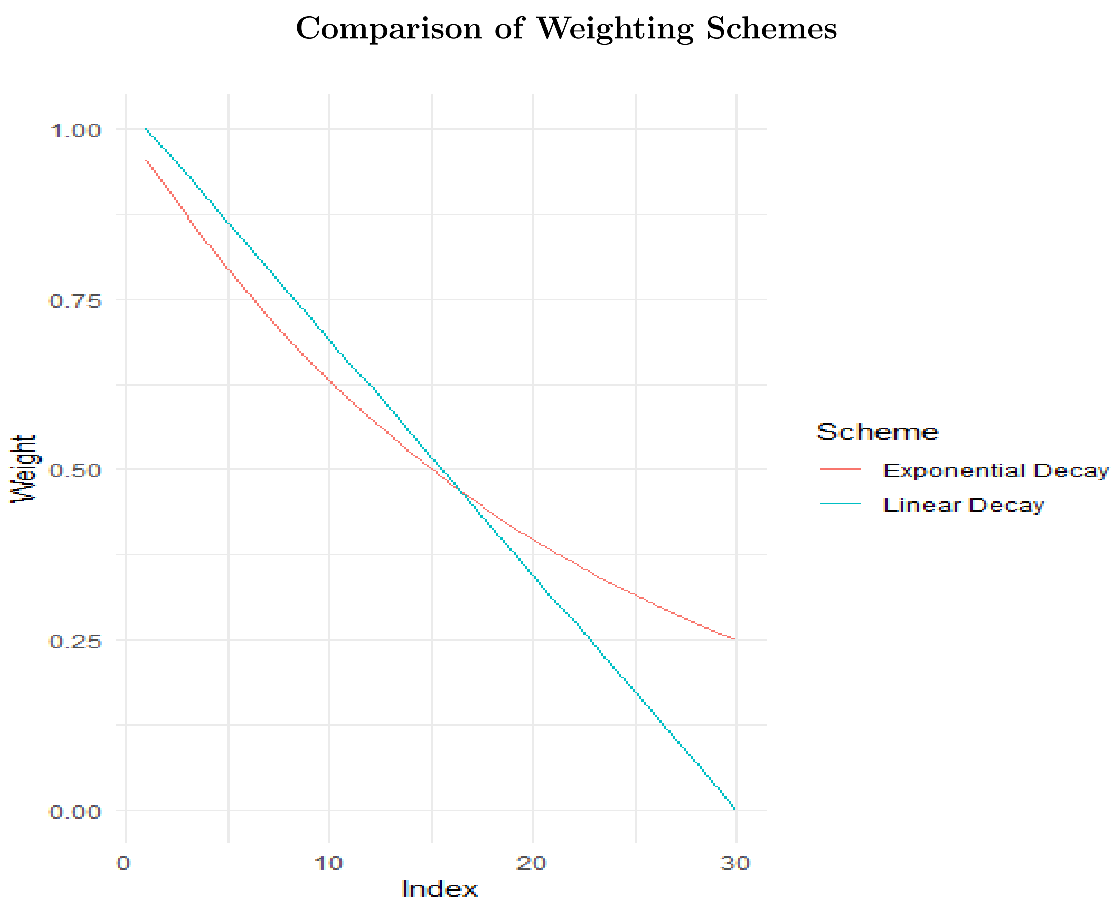

Figure 1: Weighting Schemes for the Blended Model: This plot illustrates the linear (green line) and exponential (red line) decay weighting schemes applied within the blended forecasting model for stock analysis.

2.6 Model Evaluation

2.6.1 Logarithmic Scoring

In assessing the performance of our predictive models, we applied the logarithmic scoring rule as introduced by Good (1952). This method calculates the score as

$$IGN(x_t) = -\log(f(x_t)),$$

where $f(x_t)$ is the predictive distribution of the stock returns at time t, and x_t is the materialization of the the stock return. A lower logarithmic score corresponds to a more accurate prediction, providing a quantitative measure of each model's performance. The logarithmic scores for each model applied to Apple and Netflix stock forecasts are graphically represented in Figure 1.

In Figure 2(a), due to the close proximity of the scores, distinguishing the superior model is challenging. However, for the Apple stock as shown in Figure 2(b), the logarithmic scores shed more light on model performance. The test data's low score indicates a close match with the predicted values, which is expected. Moreover, it is noteworthy that the blended model's score is consistently lower than that of the standalone stochastic volatility model, which lends credence to our hypothesis regarding its superior predictive power. This comparison underscores the potential benefits of the blended approach in achieving more accurate stock price forecasts.

(a) Netflix Stock (b) Apple Stock

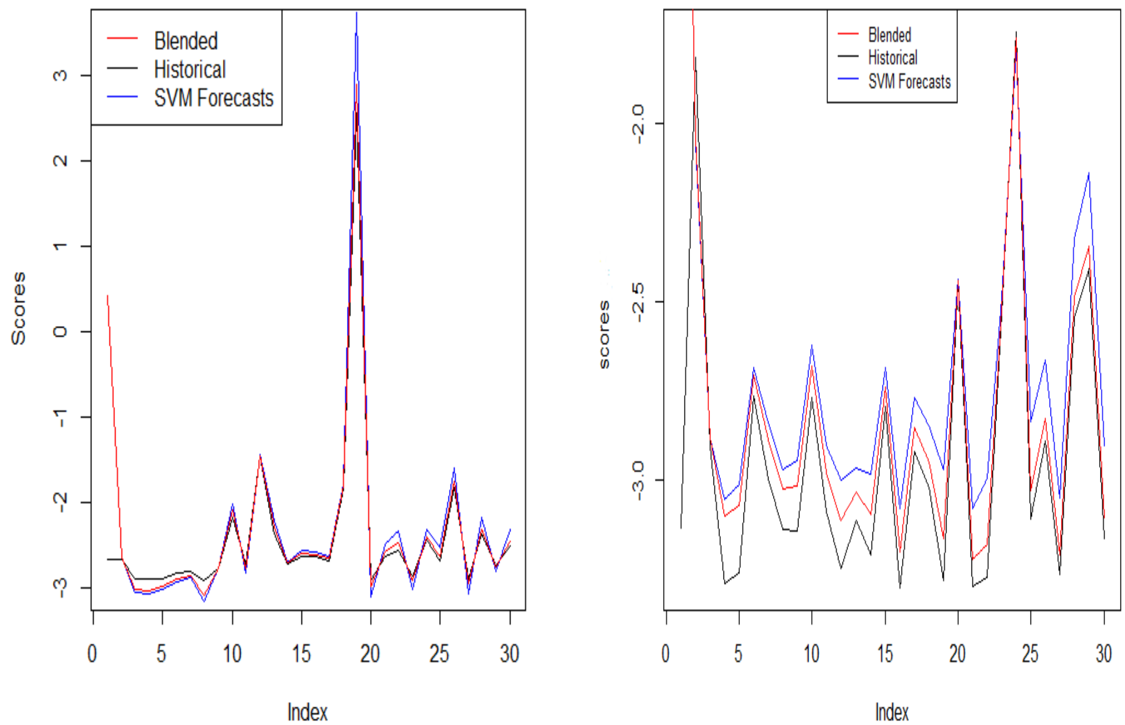

Figure 2: Comparative Logarithmic Scoring of Forecasting Models for Netflix (left) and Apple (right) Stocks. These plots demonstrate the logarithmic scores for each forecasting model, with lower scores representing more accurate predictions.

3 Results and Discussion

3.1 Results

The results presented in Figure 3 and 4 show the forecasting performance of the standalone stochastic volatility model (SVM) and the blended model relative to the historical data using the two weighting schemes: exponential and linear decaying weights. This comparative analysis is essential for evaluating the precision and reliability of each model in predicting the stock returns of Netflix and Apple.

3.1.1 Netflix Stock Performance

From the left plot in Figure 2 (a), we observe that the standalone stochastic volatility model occasionally underpredicts and overpredicts the stock returns, demonstrating a deviation from the pattern of the test data at certain points. However, the blended model, represented by the red line, follows more closely the actual stock return trajectory, especially during periods of heightened market volatility. This enhanced performance suggests the blended model's effectiveness in predicting stock returns for Netflix stock.

3.1.2 Apple Stock Performance

In the right plot of Figure 2 (b), the Apple stock analysis reflects a similar trend, with the blended forecasts often mirroring the actual returns more closely than the standalone SVM. This similarity between the blended forecasts and the actual data reinforces the robustness of the blended model compared to the standalone SVM.

The data visualization was enhanced using `ggplot2`, part of the `tidyverse` package (Wickham et al., 2019). The application of the linear decaying weights also mirrors these results, further reinforcing the robustness of the blended model in capturing the underlying dynamics of the stock market across the two weighing schemes.

(a) Netflix Stock (b) Apple Stock

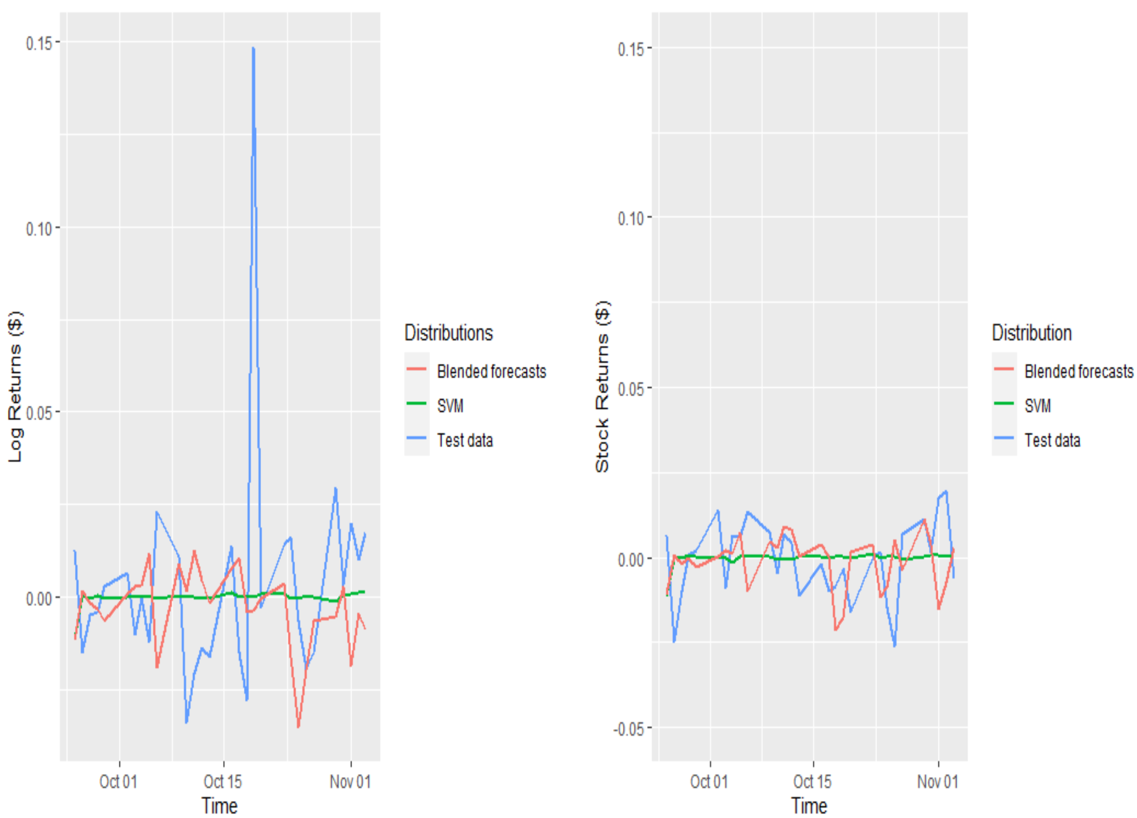

Figure 3: Comparative analysis of actual and predicted stock returns for Netflix (left) and Apple (right) using exponential decaying weights.

(a) Netflix Stock

(b) Apple Stock

Figure 4: Comparative analysis of actual and predicted stock returns for Netflix (left) and Apple (right) using linear decaying weights.

4 Conclusion and Future Directions

In conclusion, as evidenced by our results, the blending approach to forecasting, which combines the forecast and historical distributions through a weighted average, offers several significant advantages. These include the accuracy and reliability of our forecasts. Moreover, the blending approach allows for the flexibility of tuning the model according to the confidence in the forecast versus the historical data through changing the time dependent blending weights. In this manner, we may give more weight to the forecast distribution if it is more accurate, and vice versa.

We learned a lot of key lessons throughout this study. Most notably, the importance of model selection and the integration of diverse methodologies to address forecasting problems in financial time series data. This journey has reinforced the notion that a singular model often falls short in capturing the complexity of stock market behaviors, whereas a blended approach can yield more refined and robust predictions.

Moving forward, we intend to build upon the foundational work done in this research. Future improvements include the exploration of local climatologies via k-Nearest Neighbors algorithms to provide a more informed historical climatology distribution. Additionally, there is a plan to test different weighting schemes to improve the robustness of the forecasts, and to consider more advanced stochastic volatility models such as the SV-GARCH model proposed by Pitt et al. (2014), which is a hybrid model that attempts to bridge the elements of SV and GARCH specifications to better accommodate unexpected market movements. We also intend to employ the approximate Bayesian computation (ABC) based auxiliary particle filter proposed by Vankov et al. (2019) which enhances the performance of filtering and estimation in stochastic volatility models with intractable likelihoods. This approach proposed by Vankov et al. (2019) will also allow us to approximate the likelihood function through simulation, enabling effective estimation of the model parameters and latent space irrespective of how complex the underlying distribution may be.

References

Bröcker, J., & Smith, L. (2007). From ensemble forecasts to predictive distribution functions. Retrieved 2024-03-07, from `https://www.lse.ac.uk/CATS/Assets/PDFs/Publications/Papers/2008/74-Broecker-PDFs-tellus-2007.pdf`

Chikwira, C., & Mohammed, J. I. (2023, 06). The impact of the stock market on liquidity and economic growth: Evidence of volatile market. *Economies*, *11*, 155. Retrieved from `https://www.mdpi.com/2227-7099/11/6/155` doi: 10.3390/economies11060155

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*, 107–114. Retrieved 2022-08-14, from `https://www.jstor.org/stable/2984087`

Hosszejni, D., & Kastner, G. (2021). Modeling univariate and multivariate stochastic volatility in ¡i¿r¡/i¿ with ¡b¿stochvol¡/b¿ and ¡b¿factorstochvol¡/b¿. *Journal of Statistical Software*, *100*. Retrieved from `https://cran.r-project.org/web/packages/stochvol/vignettes/article2.pdf` doi: 10.18637/jss.v100.i12

Joubert, P., & Vencatasawmy, C. (n.d.). *Stochastic volatility models: Considerations for the lay actuary 1.* Retrieved from `https://www.actuaries.org.uk/system/files/documents/pdf/stochastic-volatiity-models-2005.pdf`

Pitt, M. K., Malik, S., & Doucet, A. (2014, 04). Simulated likelihood inference for stochastic volatility models using continuous particle filtering. *Annals of the*

Institute of Statistical Mathematics, *66*, 527-552. doi: 10.1007/s10463-014-0456-y

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Robert, C. P., & Casella, G. (2004). *Monte carlo statistical methods.* Springer New York. doi: 10.1007/978-1-4757-4145-2

Ryan, J. A., & Ulrich, J. M. (2024). quantmod: Quantitative financial modelling framework [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=quantmod` (R package version 0.4.26)

Silverman, B. W. (1986). Density estimation for statistics and data analysis.

Stan Development Team. (2023). *RStan: the R interface to Stan.* Retrieved from `https://mc-stan.org/` (R package version 2.26.23)

Vankov, E. R., Guindani, M., & Ensor, K. B. (2019, 03). Filtering and estimation for a class of stochastic volatility models with intractable likelihoods. *Bayesian Analysis*, *14*. doi: 10.1214/18-ba1099

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686