A Statistical Recurrent Stochastic Volatility Model for Stock Markets

Trong-Nghia Nguyen\textsuperscript{ab}, Minh-Ngoc Tran\textsuperscript{ab}, David Gunawan\textsuperscript{bc}, and Robert Kohn\textsuperscript{bd}

\textsuperscript{a}Discipline of Business Analytics, University of Sydney Business School, Sydney, Australia; \textsuperscript{b}Australian Research Council’s Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Melbourne, Australia; \textsuperscript{c}School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, Australia; \textsuperscript{d}School of Economics, UNSW Business School, Sydney, Australia

\begin{abstract}
The stochastic volatility (SV) model and its variants are widely used in the financial sector, while recurrent neural network (RNN) models are successfully used in many large-scale industrial applications of deep learning. We combine these two methods in a nontrivial way and propose a model, which we call the statistical recurrent stochastic volatility (SR-SV) model, to capture the dynamics of stochastic volatility. The proposed model is able to capture complex volatility effects, for example, nonlinearity and long-memory auto-dependence, overlooked by the conventional SV models, is statistically interpretable and has an impressive out-of-sample forecast performance. These properties are carefully discussed and illustrated through extensive simulation studies and applications to five international stock index datasets: the German stock index DAX30, the Hong Kong stock index HSI50, the France market index CAC40, the U.S. stock market index SP500 and the Canada market index TSX250. An user-friendly software package together with the examples reported in the article are available at \url{https://github.com/vbayeslab}.
\end{abstract}

\section{Introduction}

The volatility of a financial time series, such as stock returns, is defined as the variance of the returns and serves as a measure of the uncertainty about the returns. The volatility, which is of great interest to financial econometricians, is unobserved and therefore, often modeled statistically in order to estimate and predict it. The two most frequently used volatility model families are the generalized autoregressive conditional heteroscedastic (GARCH) models and the stochastic volatility (SV) models. The GARCH model (Bollerslev 1986) expresses the current volatility, conditional on the previous returns and volatilities, as a deterministic and linear function of the squared returns and the conditional volatilities in the previous time period. The SV model (Taylor 1982), on the other hand, uses a latent stochastic process to model the volatility, which is usually taken as a first-order autoregressive process. It is well documented that the GARCH and SV models can capture important effects exhibited in the variance of financial returns. For example, the volatilities in financial returns are observed to be highly autocorrelated in certain time periods and exhibit periods of both low and high volatility (Mandelbrot 1967). This so-called volatility clustering phenomenon can be modeled by the volatility processes introduced in the GARCH and SV models, leading to their extensive use in financial time series modeling.

Although the GARCH and SV models were independently and almost concurrently introduced, the GARCH models were initially more widely adopted as the likelihood of a GARCH model can be computed explicitly, while the likelihood of a SV model is intractable. However, the conditional variance process of GARCH models is deterministic and hence, GARCH models might not capture efficiently the random oscillatory behavior of financial volatility (Nelson 1991). SV models are considered as an attractive alternative to GARCH models because they overcome this limitation (Kim, Shephard, and Chib 1998; Yu 2002). Recent advances in Bayesian computation such as particle Markov chain Monte Carlo (PMCMC) (Andrieu, Doucet, and Holenstein 2010) allow straightforward estimation and inference for SV models.

Standard SV models still cannot appropriately capture some important features arising in financial volatility. For example, a large amount of both theoretical and empirical evidence indicates that there exists long-range persistence in the volatility process of many financial returns, see, for example, Lo (1991), Ding, Granger, and Engle (1993), Crato and de Lima (1994) and Bollerslev and Mikkelsen (1996). The long-memory property of a time series implies that the decay of the autocorrelations of the series is slower than exponential. The standard SV model of Taylor (1982) uses an AR(1) process to model the log of the volatility and hence, might fail to capture this type of persistence (Breidt, Crato, and de Lima 1998). Another line of the literature shows strong evidence of nonlinear auto-dependence in the volatility process of some stock and currency exchange returns (Kiliç 2011) and that the simple linear AR(1) process cannot effectively capture the underlying nonlinear volatility dynamics. Breidt, Crato, and de Lima (1998) proposed the long memory stochastic volatility (LMSV) model to overcome the short-memory limitation of the standard SV model. LMSV uses an ARFIMA process (Granger and Joyeux 1980) as an alternative.
to the AR(1) process to capture the long-memory dependence in the volatility. The empirical evidence in Breidt, Crato, and de Lima (1998) suggests that the LMSV model is able to capture the long-memory volatility behavior in some stock return datasets. However, the literature is unclear about whether the LMSV model can capture nonlinear dynamics within the volatility process, because the ARFIMA model is linear. Additionally, it is challenging to estimate the LMSV model as its likelihood is intractable. We are unaware of any available software package that implements the LMSV methodology. In another approach, Yu, Yang, and Zhang (2006) introduced a family of nonlinear SV (N-SV) models to capture the possible departure from the log transform commonly used in SV models. In the standard SV model, the logarithm of volatility is assumed to follow the AR(1) process; N-SV uses other nonlinear transformations, such as the Box-Cox power function, rather than the logarithm. The simulation studies and empirical results on currency exchange and option pricing data in Yu, Yang, and Zhang (2006) show that their N-SV model using the Box-Cox transformation is able to detect some interesting effects in the underlying volatility process. The general use of N-SV models requires the user to select an appropriate nonlinear transformation for the dataset under consideration, and this might lead to a challenging model selection problem. Neither Breidt, Crato, and de Lima (1998) nor Yu, Yang, and Zhang (2006) clearly discussed the out-of-sample forecast performance of their LMSV and N-SV models.

Recurrent neural networks (RNN) in the deep learning literature have impressive prediction performance and are successfully deployed in a large number of industrial-level applications (language translation, image captioning, speech synthesis, etc.). The RNN models are well-known for their ability to efficiently capture the long-range memory and nonlinear dependence existing within various types of sequential data, and are considered as the state-of-the-art models for many sequence learning problems (Lipton, Berkowitz, and Elkan 2015). Many researchers and practitioners use RNN for mean modeling (as opposed to variance modeling) in financial time series analysis, but the general consensus is that these machine learning models do not clearly outperform the traditional time series models such as ARMA and ARIMA; see, for example, Makridakis, Spiliotis, and Assimakopoulos (2018) and Zhang (2003). Makridakis, Spiliotis, and Assimakopoulos (2018) note that without careful modifications, machine learning models are usually less accurate than the statistical models that are extensively investigated in the financial time series literature. Recently, the idea of using RNN models to improve the predictive performance of GARCH-type models has also been proposed for volatility modeling. For example, Kim and Won (2018) use the volatility estimates from several GARCH-type models as inputs to a RNN model, which then nonlinearily transforms these inputs to output the final volatility estimate. The empirical results on the Korean stock market index show a significant improvement of forecast performance of the proposed hybrid model over several GARCH-type benchmark models. However, similar to many engineering-oriented machine learning models, Kim and Won’s model overlooks the interpretation aspect in volatility modeling, which is often of main interest to econometricians. One of our main motivations is to develop deep learning based volatility models that are not only able to produce accurate prediction, but also interpretable and have meaningful in-sample analysis. These models should not overlook the well-established features of traditional econometric models, that are motivated by the well-known stylized facts in financial time series such as volatility clustering and fat tails.

In the SV literature, there is still lack of research using RNN structures to model the stochastic volatility dynamics of financial time series, for two possible reasons. First, it is nontrivial to sensibly incorporate RNN into the statistical volatility models. Simple adaptations of RNN to volatility models easily overlook the important stylized facts exhibited in financial volatility, which are well captured by the AR(1) process in the SV model. It is important to select appropriate RNN structures that are not only able to produce accurate out-of-sample volatility forecast, but also explain properly the volatility dynamics. Second, a stochastic volatility model incorporating a RNN structure into its latent stochastic process is highly sophisticated and thus, challenging to estimate.

This article combines the SV and RNN models nontrivially, and proposes a new model, called the Statistical Recurrent Stochastic Volatility (SR-SV) model. In particular, we use the Statistical Recurrent Unit (SRU) structure of Oliva, Póczos, and Schneider (2017), which is a special type of RNN model, to capture complex volatility effects overlooked by an AR(1) process in the standard SV model but still retain the essential components of the SV model. This combination allows the SR-SV model to enjoy many advances from both worlds of deep learning, for example, flexibility and excellent predictive performance, and econometric volatility modeling, for example, excellent interpretability of volatility effects. The SR-SV model belongs to the class of parametric state space models whose Bayesian inference can be performed using recent advances in the Sequential Monte Carlo (SMC) and particle MCMC literatures (Andrieu, Doucet, and Holenstein 2010; Duan and Fulop 2015; Deligiannidis, Doucet, and Pitt 2018). The simulation studies and empirical results on the five stock index datasets demonstrate that the SR-SV model can efficiently capture the potential nonlinear and long-memory effects in the underlying volatility dynamics, and provide better out-of-sample forecasts than the standard SV, N-SV and LMSV models. We note that we have tested SR-SV on a wider range of stock returns but only report in this article the results for five of them, as we constantly observed a similar improvement of the model compared to the other three competitors. A Matlab software package implementing Bayesian estimation and inference for SR-SV together with the examples reported in this article are available at https://github.com/vbayeslab.

The rest of the article is organized as follows. Section 2 briefly reviews the SV and SRU models, and presents the SR-SV model. Section 3 discusses in detail Bayesian estimation and inference for the SR-SV model. Section 4 presents the simulation study and applies the SR-SV model to analyze the five stock index datasets. Section 5 concludes. The Appendix (online supplementary material) gives details of the implementation and further empirical results.
2. The SR-SV Model

2.1. The SV Model and its Possible Weaknesses

Let \( y = \{ y_t, t = 1, \ldots, T \} \) be a series of financial returns. We consider a basic version of the SV model (Taylor 1982)

\[
\begin{align*}
    z_t &= \mu + \phi(z_{t-1} - \mu) + \epsilon_t^z, \quad \epsilon_t^z \sim \mathcal{N}(0, \sigma^2), \quad t = 2, \ldots, T, \\
    z_1 &\sim \mathcal{N}(\mu, \frac{\sigma^2}{1 - \phi^2}), \quad (1) \\
    y_t &= e^{\frac{1}{2} z_t} \epsilon_t^y, \quad \epsilon_t^y \sim \mathcal{N}(0, 1), \quad t = 1, 2, \ldots, T. \quad (2)
\end{align*}
\]

The persistence parameter \( \phi \) is assumed to be in \((-1, 1)\) to enforce stationarity of both the \( z \) and \( y \) processes. The log volatility process \( z \) is assumed to follow an AR(1) model. It is well documented in the financial econometrics literature that financial time series data often exhibit a strong autocorrelation, which forces the persistence parameter \( \phi \) to be close to 1 (Jacquier, Polson, and Rossi 1994; Kim, Shephard, and Chib 1998). Write \( p(z|\theta) \) for the density of \( z \) given the model parameters \( \theta = (\mu, \phi, \sigma^2) \) and \( p(y|z) \) for the density of the data \( y \) conditional on \( z \). We can view \( p(z|\theta) \) as the prior with \( \theta \) being the hyperparameters and \( p(y|z) \) as the likelihood (Jacquier, Polson, and Rossi 1994).

Under this perspective, the SV model (1)–(2) puts nonzero prior mass on AR(1) stochastic processes, and zero or almost-zero mass on stochastic processes that are far from being well approximated by an AR(1). Thus, the SV model in (1)–(2) might not be able to capture more complex dynamics in the posterior behavior of the log volatility process \( z \), such as long-term memory or nonlinear auto-dependence, and that a more flexible prior distribution should be put on \( z \). We design such a flexible prior by combining the attractive features from both SV and RNN time series modeling techniques.

Yu, Yang, and Zhang (2006) propose a class of nonlinear SV (N-SV) models as a variant of SV which allows a more flexible link between the variance \( \text{var}(y_t|z_t) \) and the AR(1) process \( z_t \). Their N-SV model, using a Box-Cox like transformation for \( z_t \), is written as

\[
\begin{align*}
    z_t &= \mu + \phi(z_{t-1} - \mu) + \epsilon_t^z, \quad \epsilon_t^z \sim \mathcal{N}(0, \sigma^2), \quad t = 2, \ldots, T, \\
    z_1 &\sim \mathcal{N}(\mu, \frac{\sigma^2}{1 - \phi^2}), \quad (3) \\
    y_t &= (1 + \delta z_t)^{1/25} \epsilon_t^y, \quad \epsilon_t^y \sim \mathcal{N}(0, 1), \quad t = 1, 2, \ldots, T. \quad (4)
\end{align*}
\]

where \( \delta \) is the auxiliary parameter that measures the degree of nonlinearity rather than the log transform. As \( \delta \to 0 \), \( (1 + \delta z_t)^{1/25} \to e^{\frac{1}{25} z_t} \) and hence, the N-SV model includes the SV model as a special case. The term nonlinearity here might cause some confusion, as it does not refer to the nonlinear auto-dependence within the log volatility process \( z \), but the observation that \( \text{var}(y|z) \) is a nonlinear function of \( z \).

Breidt, Crato, and de Lima (1998) suggest to use an ARFIMA\((p,d,q)\) process for the log volatility \( z_t \) to capture the long-memory auto-dependence exhibited in financial time series. Their LMSV model is written as

\[
\begin{align*}
    (1 - B)^d \Phi(B)z_t &= \Theta(B)\eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2), \quad (5) \\
    y_t &= \sigma_t\epsilon_t, \quad \sigma_t = \kappa e^{\frac{1}{25} z_t}, \quad \epsilon_t \sim \mathcal{N}(0, 1), \quad t = 1, 2, \ldots, T. \quad (6)
\end{align*}
\]

where \( \Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q \), and \( B \) is the backshift operator, that is, \( B X_t := X_{t-1} \). To ensure the stationarity and invertibility of the log volatility process \( z_t \), the fractional integration parameter \( d \) is assumed to be in \((-0.5, 0.5)\) and the roots of \( \Phi(B) \) and \( \Theta(B) \) lie outside the unit circle.

We will use the standard SV, N-SV and LMSV models as the benchmarks to evaluate the SR-SV model. We will also compare the SR-SV against the engineering-oriented GP-Vol model of Wu, Hernández-Lobato, and Ghahramani (2014), who use a Gaussian process to capture nonlinearity in the volatility dynamics; see Section 4.2.3 for further details of this model.

2.2. The SRU Model

There are at least two approaches to modeling time series data. The first approach represents time effects explicitly via some simple function, often a linear function, of the lagged values of the time series. This is the mainstream time series data analysis approach in the statistics literature with the well-known models such as AR and ARMA. The second approach represents time effects implicitly via latent variables, which are designed to store the memory of the dynamics in the data. These latent variables, also called hidden states, are updated in a recurrent and deterministic manner using the information carried over by their values from the previous time steps and the information from the data at the current time step. Recurrent neural networks (RNN) belong to the second category, and were first developed in cognitive science and successfully used in computer science and other fields. A third class of models that represent time implicitly is state space models, albeit the recurrent update is stochastic, which are widely used in econometrics and statistics.

The SV model discussed in Section 2.1 is an example of state space models.

In this section, we denote the time series data as \( \{ D_t = \{ x_t, z_t \}, t = 1, 2, \ldots \} \) where \( x_t \) is the vector of inputs and \( z_t \) the scalar output. Our goal is to model the conditional distribution \( p(z_t|x_t, D_{1:t-1}) \). If the serial dependence structure is ignored, then a feedforward neural network (FNN) can be used to transform the raw input data \( x_t \) into a set of hidden units \( h_t \), also called learned features or summary statistics, for the purpose of explaining or predicting \( z_t \). However, this approach is unsuitable for time series data as the time effects, that is, the serial dependence, are totally ignored. The main idea behind RNN is to let the set of hidden units \( h_t \) to feed itself using its lagged value \( h_{t-1} \) from the previous \( (t - 1) \) time step. Hence, RNN can be best thought of as a FNN that allows a connection of the hidden units to their value from the previous time step, enabling the network to possess memory. Mathematically, the Simple RNN model (Elman 1990) is written as

\[
\begin{align*}
    h_t &= \Psi(w_x x_t + w_h h_{t-1} + b), \quad \eta_t = \beta_0 + \beta_1 h_t, \\
    z_t|\eta_t &\sim p(z_t|\eta_t). \quad (7)
\end{align*}
\]

The model parameters include \( w_x \), \( w_h \), \( b \), \( \beta_0 \) and \( \beta_1 \); \( \Psi(\cdot) \) is a nonlinear activation function, for example, common choices are the sigmoid \( \Psi(z) = 1/(1 + e^{-z}) \) and the tanh \( \Psi(z) = (e^z - e^{-z})/(e^z + e^{-z}) \). The density \( p(z_t|\eta_t) \) depends on the learning task. For example, if \( z_t \) is continuous, then typically \( p(z_t|\eta_t) \) is a
Gaussian density with mean $\eta_i$; if $z_t$ is binary, then $z_t|\eta_i$ follows a Bernoulli distribution with probability $\Psi(\eta_i) = 1/(1 + e^{-\eta_i})$. Usually, $h_t$ is set to 0, that is, the neural network initially does not have any memory.

Equation (7) suggests that the hidden state at time $t$ is the output of a composite function

$$ h_t = f\left(x_t, f(x_{t-1}, \ldots, f(x_1, h_0))\right), $$

where

$$ f(x_t, h_{t-1}) := \Psi(w_x h_{t-1} + b), $$

which somewhat resembles a multiplication structure in terms of the weight $w_x$. Consequently, the gradient of $h_t$ with respect to the model parameters might either explode or vanish if $t$ is sufficiently large and $w_x$ is not equal to 1, and hence, making it inefficient for the Simple RNN model to learn in long time series. See Goodfellow, Bengio, and Courville (2016) for further explanation.

Many sophisticated RNN structures have been proposed to overcome the problem above in the Simple RNN model; for example, the Long Short-term Memory model of Hochreiter and Schmidhuber (1997), the Gated Recurrent Unit of Cho et al. (2014) and the Statistical Recurrent Unit (SRU) of Oliva, Póczos, and Schneider (2017). The SRU allows the vector of summary statistics $h_t$ to traverse through the network using a moving average. We will use the SRU in this article as its structure and some of its main parameters carry statistical meaning; see Section 2.3. A general SRU structure is mathematically written as

$$ r_t = \Psi(w_r h_{t-1} + b_r), $$

$$ \varphi_t = \Psi(w_{\varphi} r_t + w_{\psi} x_t + b_{\psi}), $$

$$ h_t^{(j)} = \alpha_j h_{t-1}^{(j)} + (1 - \alpha_j) \varphi_t, \; j = 1, \ldots, m; $$

$$ h_t = (h_t^{(1)}, \ldots, h_t^{(m)})^T, $$

where $\alpha = (\alpha_1, \ldots, \alpha_m) \in (0, 1)^m$ is a vector of moving average weights, $r_t$ and $\varphi_t$ are auxiliary states to facilitate the computation of hidden state $h_t$, and $w_r, b_r, w_{\psi}, w_{\psi}$ are the model parameters. We denote the functional learning structure in (9a)–(9c) as $h_t = \text{SRU}(x_t, h_{t-1})$, which takes $x_t$—the input data at current time $t$—and $h_{t-1}$—the previous output of the SRU—as the input arguments. The moving average structure of the state $h_t$ allows the RNN network with SRU units to enjoy some advantages compared to other RNN models. The current state $h_t$ is related to the previous state $h_{t-1}$ both directly and indirectly, and thus, mitigates the problem of multiplying the same quantities multiple times as in the Simple RNN model. The novel architecture of the SRU allows the model to capture long term dependencies in data via simple moving averages.

### 2.3. The SR-SV Model

This section proposes the SR-SV model that combines SV and SRU for financial volatility modeling. The key idea is using the SRU structure to capture the complicated effects such as long-term memory and nonlinear auto-dependence, in the volatility dynamics that are overlooked by the basic SV models. This leads to a prior distribution for the log volatility process $z$ that is much more flexible than the AR(1) prior (see Section 2.1). The proposed SR-SV model is as follows

$$ r_t = \Psi(w_r h_{t-1} + b_r), \; t = 2, \ldots, T, $$

$$ \varphi_t = \Psi(w_{\varphi} r_t + w_{\psi} h_{t-1} + w_z z_{t-1} + b_{\psi}), \; t = 2, \ldots, T, $$

$$ h_t = ah_{t-1} + (1 - \alpha) \varphi_t, \; t = 2, \ldots, T, \; h_1 = 0, $$

$$ \eta_t = \beta_0 + \beta_1 h_t + \epsilon_t^\eta \sim \mathcal{N}(0, \sigma^2), \; t = 1, 2, \ldots, T, $$

$$ z_t = \eta_t + \phi z_{t-1}, \; t = 1, \ldots, T, $$

$$ y_t = e^{z_t^2} \epsilon_t^y \sim \mathcal{N}(0, 1), \; t = 1, 2, \ldots, T, $$

that is, we use a SRU to model the dynamics of the hidden states $h_t$. Here, $\eta_0$ is the initial value of the log volatility process and a convenient choice of $\eta_0$ is the log of the unconditional variance of the observed series $y_t$ that is, $\eta_0 = \log(\text{var}(y))$. Alternatively, one could treat $\eta_0$ as a model parameter. We follow the literature to initialize $h_1 = 0$ as the recurrent units initially have no memory. See Appendix A in the supplementary material for a graphical representation of the SR-SV model.

We note the following important properties of the SR-SV model. First, the SR-SV model in (10a)–(10f) retains the measurement Equation (10f) and the linear part $\phi z_{t-1}$ of the AR(1) process from the standard SV model, and captures the volatility effects not captured by the AR(1) process, for example, nonlinear and long-memory auto-dependence, via the latent state $h_t$ of the SRU structure. The log volatility at time $t$ in (10e) can be written as

$$ z_t = \beta_0 + \beta_1 \text{SRU}(\eta_{t-1}, z_{t-1}, h_{t-1}) + \phi z_{t-1} + \epsilon_t^\eta. $$

Therefore, the parameter $\beta_1$ characterizes all the effects in the underlying log volatility process $z$ rather than the short-term linear effect captured by the AR(1) process. We refer to $\beta_1$ as the nonlinearity long-memory coefficient. If $\beta_1 = 0$ and $\epsilon_t^\eta \sim \mathcal{N}(\beta_0/(1 - \phi), \sigma^2/(1 - \phi^2))$, the SR-SV model becomes the SV model (1)–(2) and hence, the SV model is a special case of the SR-SV model. We therefore, follow the SV literature and assume that $|\phi| < 1$. The $z$ process, and thus, the $y$ process of the SR-SV model, is not guaranteed to be stationary unless $\beta_1 = 0$ and $\epsilon_t^\eta \sim \mathcal{N}(\beta_0/(1 - \phi), \sigma^2/(1 - \phi^2))$. Nonstationarity for volatility is often argued to be more realistic in practice, for example, van Bellegem (2012), although it might be mathematically less appealing. The Equation (11) can be reexpressed as

$$ z_t = \beta_0 + \beta_1 \Psi(\eta_{t-1}, w_z z_{t-1}, h_{t-1}) + \phi z_{t-1} + \epsilon_t^\eta, $$

where $\Psi(\cdot)$ is a nonlinear function and $w_z$ is the weight corresponding to $z_{t-1}$. If $w_z = 0$ in (12), then $z_t$ only depends linearly on $z_{t-1}$, therefore, $w_z$ characterizes the serial dependence of the volatility $z_t$ on $z_{t-1}$, rather than just linear dependence. Section 4 analyzes $w_z$ in more detail.

Second, Oliva, Póczos, and Schneider (2017) set the scales $\alpha$ of the SRU model to prespecified values to obtain a vector of summary statistics $h_t$ at different moving average weights. We, however, treat $\alpha$ as a model parameter and learn it from the data. We note that a higher $\alpha$ puts more weight on the historical information while a smaller $\alpha$ puts more weight on the current information. We show later in the empirical study.
that this parameter \( \alpha \) is able to quantify the existence of the long-memory auto-dependence commonly exhibited in the volatility dynamics of the financial time series.

Third, neural networks are highly flexible and might suffer from overfitting, that is, having over-confident in-sample fit and bad out-of-sample forecasts. Regularization is often needed to avoid overfitting. Injecting noise into the layers of the network is an effective regularization approach in the Machine Learning literature, and seen as a form of data augmentation at multiple levels of abstraction (Sietsma and Dow 1991; Dieng et al. 2018). In the SR-SV model, by allowing \( z_{t-1} \) and \( \eta_{t-1} \) to be the inputs of the SRU structure at time \( t \), we inject the noise \( \epsilon_{t}^{z_{t-1}} \) of the volatility process to the input and hidden layers of the SRU. This noise-injecting regularization approach makes the SR-SV model perform well on both in-sample fitting and out-of-sample forecasting, even with the simplest specification of the SRU structure where all the \( r_{t}, \phi_{t} \) and \( h_{t} \) are scalars. Our SR-SV model can be categorized as a parametric model with the vector of model parameters \( \theta \) consisting of eleven parameters: four main parameters \( b_{0}, \beta_{1}, \phi, \sigma^{2} \) and the SRU parameters \( \alpha, w_{0}, b_{r}, w_{r}, w_{a}, w_{z} \) and \( b_{\epsilon} \).

Finally, \( \beta_{0} \) plays the role of the scale factor \( \tau = e^{\beta_{0}/2} \) for the variance of \( y_{1} \). One could set \( \beta_{0} = 0 \) and modify (10a) to \( y_{1} = \tau e^{\epsilon_{t}^{z_{t-1}}}/\epsilon_{t}^{\eta_{t-1}} \); however, this parameterization might be less statistically efficient in terms of Bayesian estimation, especially for the parameter \( \tau \) (see Kim, Shephard, and Chib 1998).

It is straightforward to extend the SR-SV model in (10a)–(10f) by incorporating other advances in the SV literature. For example, we can use a Student’s \( t \) distribution instead of a Gaussian for the measurement shock \( \epsilon_{t}^{\eta_{t-1}} \) and take into account the leverage effect by correlating \( \epsilon_{t}^{z_{t-1}} \) with the volatility shock \( \epsilon_{t}^{\eta_{t-1}} \). We do not consider these extensions here, however, because using the most basic version makes it easier to understand the strengths and weaknesses of the new model.

### 3. Bayesian Inference

This section discusses Bayesian estimation and inference for the SR-SV model. For a generic sequence \( \{x_{i}\} \) we use \( x_{ij}, \ i \leq j \), to denote the series \( \{x_{1}, \ldots, x_{t}\} \). The SR-SV model is a state-space model with the measurement equation \( y_{1}|z_{t} \sim \mathcal{N}(0,e^{\sigma^{2}}) \), and the state transition equation

\[
\begin{align*}
  z_{t}|z_{t-1}, h_{t} & \sim \mathcal{N}(\phi z_{t-1} + b_{0} + \beta_{1} h_{t}, \sigma^{2}), \ t \geq 2, \\
  z_{1} & \sim \mathcal{N}(\beta_{0}, \sigma^{2}).
\end{align*}
\]

We are interested in sampling from the posterior distribution \( \pi(\theta) = p(\theta|y_{1:T}) \propto p(y_{1:T}|\theta)p(\theta) \), where \( p(y_{1:T}|\theta) = \int p(y_{1:T}|z_{1:T}, \theta)p(z_{1:T}|\theta)dz_{1:T} \) is the likelihood function, \( p(\theta) \) is the prior and \( p(y_{1:T}) = \int p(y_{1:T}|\theta)p(\theta)d\theta \) is the marginal likelihood, \( \theta \) is the vector of 11 parameters of the SR-SV model discussed in Section 2.3.

The likelihood function \( p(y_{1:T}|\theta) \) is computationally intractable for nonlinear state space models like the SV and SR-SV models, but can be estimated unbiasedly by a particle filter (Del Moral 2004). Bayesian inference for SR-SV can be performed using recent advances in the Sequential Monte Carlo (SMC) literature presented next.

### 3.1. SMC for the SR-SV Model

Duan and Fulop (2015) propose the Density Tempered Sequential Monte Carlo (DT-SMC) approach to Bayesian inference for state space models with intractable likelihood; see also Tran et al. (2014). The DT-SMC sampler generalizes the SMC method of Neal (2001) and Del Moral, Doucet, and Jasra (2006) when the likelihood can be computed analytically. In order to sample from the posterior \( \pi(\theta) \), the DT-SMC method first samples a set of \( M \) weighted particles \( \{W_{0,j}, \theta_{0,j}^{M}\}_{j=1}^{M} \) from an easy-to-sample distribution \( \pi_{0}(\theta) \), such as the prior \( p(\theta) \), and then traverses these particles through intermediate distributions \( \pi_{k}(\theta) \), \( k = 1, \ldots, K \), which target the posterior distribution \( \pi(\theta) \) eventually, that is, \( \pi_{K}(\theta) = \pi(\theta) \). The DT-SMC method uses the following intermediate distributions

\[
\pi_{k}(\theta) := \pi_{k}(\theta|y_{1:T}) \propto \tilde{p}(y_{1:T}|\theta, u)^{K}p(\theta), \quad (14)
\]

where the \( \gamma_{K} \) is referred to as the level temperature and \( 0 = \gamma_{0} < \gamma_{1} < \gamma_{2} < \cdots < \gamma_{K} = 1 \). \( \tilde{p}(y_{1:T}|\theta, u) \) is the unbiased estimator of the likelihood \( p(y_{1:T}|\theta) \) and \( u \) is the set of pseudo random numbers used within a particle filter to estimate the likelihood \( p(y_{1:T}|\theta) \). In this article where it is possible to sample from the prior \( p(\theta) \), we set \( \pi_{0}(\theta) = p(\theta) \). Algorithm B.1 in the supplementary material summarizes the DT-SMC method for the SR-SV model.

The DT-SMC method consists of three main steps: reweighting, resampling and Markov moves. At the beginning of SMC iteration \( k \), the set of weighted particles \( \{W_{k-1,j}, \theta_{k-1,j}^{M}\}_{j=1}^{M} \) that approximate the intermediate distribution \( \pi_{k-1}(\theta) \) is reweighted to approximate the target \( \pi_{k}(\theta) \). The efficiency of these weighted particles is often measured by the effective sample size (ESS) (Liu and Chen 1998) defined in (B.3) in the supplementary material. If the ESS is below a prespecified threshold, the particles are resampled; the resulting equally-weighted resamples, which are now approximate samples from \( \pi_{k}(\theta) \), are then refreshed by a Markov kernel whose invariant distribution is \( \pi_{k}(\theta) \). For example, Duan and Fulop (2015) uses the pseudo marginal Metropolis-Hastings (PMMH) kernel of Andrieu, Doucet, and Holenstein (2010) with the likelihood estimated unbiasedly by the particle filter in the Markov move step. Pitt et al. (2012) and Doucet et al. (2015) suggest that the PMMH approach works efficiently when the variance of the log of the estimated likelihood is in between 1 and 3. For some state space models like the SR-SV model, a large number of particles might be required to obtain a likelihood estimator with log variance to be in between 1 and 3, which is computationally inefficient. To tackle this problem, we incorporate the correlated pseudo marginal (CPM) approach of Deligiannidis, Doucet, and Pitt (2018) (see also Quiroz et al. 2021) into the Markov move step. The CPM method makes the current set of random numbers \( u \) and proposal \( u^{'} \) correlated, and helps reduce the variance of the ratio \( \tilde{p}(y_{1:T}|\theta^{*}, u^{'})/\tilde{p}(y_{1:T}|\theta, u) \) in (B.4), thus, leading to a better mixing Markov chain while using less number of particles in the particle filter. Similar to the SMC methods of Del Moral, Doucet, and Jasra (2006) and Neal (2001), the DT-SMC method is parallelizable as the particles move independently in the Markov move step, and provides an estimate of the marginal likelihood as a by-product.
3.2. Model Choice by Marginal Likelihood

The marginal likelihood is often used to choose between models via the Bayes factor. In order to compare the relative performance between two models $M_1$ and $M_2$ on a given data $y_{1:T}$, we use the Bayes factor $F_{M_1,M_2} = p(y_{1:T} | M_1) / p(y_{1:T} | M_2)$. The larger the Bayes factor $F_{M_1,M_2}$, the stronger $M_1$ is supported by the data than $M_2$. Jeffreys (1961) and Kass and Raftery (1995) suggest a scale of interpretation of the Bayes factor $F_{M_1,M_2}$ as listed in Table 1. We note that the DT-SMC sampler in the previous section provides an efficient way to compute the marginal likelihood.

4. Simulation Studies and Applications

This section evaluates the performance of the SR-SV model relative to the SV, N-SV and LMSV models using a simulation study and real data applications. The results for GARCH model are not reported as it performs similarly to SV. We use the DT-SMC sampler for Bayesian inference in the SV, N-SV and SR-SV models. As the LMSV model does not have an explicit state-space representation and its likelihood function is analytically intractable, we follow Breidt, Crato, and de Lima (1998) and estimate the LMSV model in the frequency domain. The implementation of Bayesian inference for the LMSV model is presented in Appendix C. Table B.1 in Appendix B, supplementary material lists our implementation details of the DT-SMC sampler. All empirical results from the RNN literature show that the RNN parameters are often small. Finally, we set a normal prior with a zero mean and a small variance for the SRU parameters, except $w_2$, because empirical results from the RNN literature show that the RNN parameters are often small. We use a normal prior with a zero mean and a small variance for the intercept $\mu$ and the persistence parameter $\phi$ of the three models SV, N-SV and SR-SV. We also use an inverse-Gamma prior for the parameters $\sigma^2$ in all models, but make it flatter than the priors used in Yu, Yang, and Zhang (2006) and Kim, Shephard, and Chib (1998). We follow Yu, Yang, and Zhang (2006) to use an informative but reasonably flat prior distribution for the intercept $\mu$ in the SV and N-SV models. For the SR-SV model, we found that the posterior distributions of $\beta_1$ and $w_2$ are unimodal under inverse-Gamma priors. We use a normal prior with a zero mean and a small variance for the SRU parameters, except $w_2$, because empirical results from the RNN literature show that the RNN parameters are often small. Finally, we set a normal prior with a zero mean and a small variance for the intercept $\beta_0$ in the SR-SV model as the empirical results often show small values of $\beta_0$.

Table 3 lists the predictive scores used to measure the out-of-sample performance. The smaller the predictive score, the better. We use the posterior means estimated from in-sample data to perform one-step-ahead forecasts for the out-of-sample data. Alternatively, one could use a data annealing approach, such as SMC² of Chopin, Jacob, and Papaspiliopoulos (2013), to keep updating the posterior each time a new observation arrives. Data annealing approaches are more computationally expensive and we do not consider them in this article.

4.1. Simulation Studies

We consider the three volatility models in Table 4. Model 1 is a GARCH(1,1). Model 2 is an extension of GARCH(1,1) by applying a Box-Cox power transformation (Box and Cox 1964) to both the conditional variance equation and the volatility dynamics. Model 2 is similar to the nonlinear ARCH models of Higgins and Bera (1992), but they use lagged innovations to construct the conditional variance. Model 3 is a FIGARCH(1,d,1) model of Baillie, Bollerslev, and Mikkelsen (1996) which uses a long-memory process ARFIMA(1,d,1) to simulate the long-memory auto-dependence.

We generate time series of $T = 3000$ observations each from these three models and refer to these simulation settings as SIM I, SIM II, SIM III, respectively. The parameters are set so that $y_t$ somewhat resembles real financial time series data exhibiting volatility clustering with nonlinearity (SIM II) and long-memory (SIM III) auto-dependence in the underlying volatility dynamics. For each dataset, the first $T_{in} = 2000$ observations are used for model estimation and the last $T_{out} = 1000$ for out-of-sample analysis. Table 5 shows the posterior estimates for the parameters of the SV and SR-SV models; for the SR-

### Table 1. Scale of interpretation of the Bayes factor $F_{M_1,M_2}$.

| Grade | $F_{M_1,M_2}$ | $\log_{10} F_{M_1,M_2}$ | $\ln F_{M_1,M_2}$ | Strength of evidence |
|-------|--------------|------------------------|--------------------|---------------------|
| 0     | $< 10^{10}$  | $< 0$                  | $< 0$              | Negative (supports $M_2$) |
| 1     | $10^{0} - 10^{1/2}$ | $0.0 - 0.5$         | $0.0 - 1.2$         | Barely worth mentioning |
| 2     | $10^{1/2} - 10^{1}$ | $0.5 - 1.0$          | $1.2 - 2.3$         | Substantial          |
| 3     | $10^{-1} - 10^{-2}$ | $1.0 - 2.0$          | $2.3 - 4.6$         | Strong               |
| 4     | $> 10^{-2}$  | $> 2.0$                | $> 4.6$             | Decisive             |

### Table 2. Prior distributions for the parameters in the SR-SV, SV and N-SV models.

| Parameter | SR-SV | SV | N-SV |
|-----------|-------|----|------|
| $\beta_0$ | $\mathcal{N}(0,0.1)$ | $\mathcal{N}(0,0.25)$ | $\mathcal{N}(0,0.25)$ |
| $\theta+1$ | Beta(20,1.5) | Beta(20,1.5) | Beta(20,1.5) |
| $\sigma^2$ | $\mathcal{IG}(2.5,0.25)$ | $\mathcal{IG}(2.5,0.25)$ | $\mathcal{IG}(2.5,0.25)$ |
| $\phi^2$ | $\mathcal{IG}(2.5,0.25)$ | $\mathcal{IG}(2.5,0.25)$ | $\mathcal{IG}(2.5,0.25)$ |
| $\alpha$ | Beta(2,2) | Beta(2,2) | Beta(2,2) |
| $w_1, w_2, \gamma_1$ | $\mathcal{N}(0,0.1)$ | $\mathcal{N}(0,0.1)$ | $\mathcal{N}(0,0.1)$ |
| $b_1, b_2$ | $\mathcal{N}(0,0.1)$ | $\mathcal{N}(0,0.1)$ | $\mathcal{N}(0,0.1)$ |
| $w_2$ | $\mathcal{IG}(2.5,1)$ | $\mathcal{IG}(2.5,1)$ | $\mathcal{IG}(2.5,1)$ |

Note: The notation $\mathcal{N}$, $\mathcal{IG}$ and Beta denote the Gaussian, inverse-Gamma and Beta distributions, respectively. $\mathcal{IG}(a,b)$ denotes an inverse-Gamma distribution with scale parameter $a$ and shape parameter $b$.

1. https://sydneyuni.atlassian.net/wiki/spaces/RC/overview
Table 4. Simulation: Data generating process.

| Data Model Parameters | \( \sigma_t^2 = \mu + \alpha \sigma_{t-1}^2 + \beta \sigma_{t-1}^2, \quad t = 2, \ldots, T \) |
|-----------------------|-----------------------------------------------------------------|
|                       | \( y_t = \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, 1), \quad t = 1, \ldots, T \) |
| \( \nu_t^2 \)         | \( \alpha = 0.07, \beta = 0.92 \) |
| \( \mu \)             | \( \rho_1 = 0.1, \mu = 0.1 \) |
| \( \phi \)            | \( \phi = 0.01, \beta = 0.5 \) |

Table 5. Simulation: Posterior means of the parameters with the posterior standard deviations in brackets.

| \( \nu \) | \( \phi \) | \( \sigma_t^2 \) | \( \alpha \) | \( \beta_0 \) | \( \beta_1 \) | \( w_2 \) | Mar.llh |
|-----------|-----------|----------------|-----------|-------------|-------------|---------|--------|
| SIM I     |           |                |           |             |             |         |        |
| \( \nu \) | 2.145     | 0.985          | 0.019     | 0.016       | 0.026       | 0.027   | −5100.8|
|           | (0.237)   | (0.004)        | (0.003)   |             |             |         | (0.131)|
| SR-SV     | 0.974     | 0.020          | 0.534     | 0.027       | 0.388       | 0.205   | −5099.9|
|           | (0.023)   | (0.005)        | (0.166)   | (0.031)     | (0.235)     | (0.261) | (0.300)|
| SIM II    |           |                |           |             |             |         |        |
| \( \nu \) | 1.050     | 0.967          | 0.032     | 0.016       | 0.026       | 0.027   | −4060.3|
|           | (0.125)   | (0.008)        | (0.006)   |             |             |         | (0.164)|
| SR-SV     | 0.792     | 0.041          | 0.515     | 0.043       | 0.423       | 0.530   | −4057.7|
|           | (0.106)   | (0.010)        | (0.156)   | (0.044)     | (0.207)     | (0.256) | (0.306)|
| SIM III   |           |                |           |             |             |         |        |
| \( \nu \) | 0.134     | 0.984          | 0.041     | 0.016       | 0.026       | 0.027   | −3146.9|
|           | (0.329)   | (0.005)        | (0.007)   |             |             |         | (0.195)|
| SR-SV     | 0.896     | 0.056          | 0.645     | 0.093       | 0.325       | 0.290   | −3144.2|
|           | (0.035)   | (0.013)        | (0.240)   | (0.045)     | (0.132)     | (0.115) | (0.316)|

NOTE: The last column shows the estimated log marginal likelihood with the Monte Carlo standard errors in brackets, averaged over 10 different runs of the DT-SMC sampler. The asterisks indicate the cases when the Bayes factors strongly support the SR-SV model over the SV model.

Figure 1. SIM I: the filtered values of \( \nu_t \) and \( h_t \) of the SR-SV model, together with the SIM I in-sample data. (This is better viewed in color.)

SV model we only show the results for the main parameters. The last column in the table shows the marginal likelihood estimates, averaged over 10 different runs of the DT-SMC sampler, together with the Monte Carlo standard errors in the brackets. Figures 1–3 plot the filtered values of the \( \nu_t \) and \( h_t \) components of the SR-SV model in SIM I, SIM II and SIM III datasets, respectively. Figure D.2 in Appendix D in the supplementary material plots the true volatility together with the filtered volatility produced by the SV and SR-SV models, for the three simulated datasets.

The estimation results suggest the following conclusions. First, for SIM I, the difference of marginal likelihood estimates in Table 5 between the SV and SR-SV models is insignificant, the coefficient \( \beta_1 \) is insignificant, and the filtered volatilities from these two models in Figure D.2 are identical and close to the true volatility. This implies that the SV and SR-SV models fit
equally well to the SIM I data and that the SR-SV model is close to the SV model if the true data-generating process, which is GARCH(1,1) in this example, exhibits no other effects rather than short-memory linear auto-dependence within the volatility dynamics.

Figure 2. SIM II: the filtered values of $\eta_t$ and $h_t$ of the SR-SV model, together with the SIM II in-sample data. (This is better viewed in color.)

Figure 3. SIM III: the filtered values of $\eta_t$ and $h_t$ of the SR-SV model, together with the SIM III in-sample data. (This is better viewed in color.)

Second, the estimation results for SIM II and SIM III show that the additional neural network structure of the SR-SV model is able to efficiently capture the volatility effects overlooked by the basic SV model. This is supported by the Bayes factor of the SR-SV model compared to the SV model of more than $e^{2.3}$, which, according to the interpretation in Table 1, strongly supports the SR-SV model. The plots of the $h_t$ and $\eta_t$ components of the SR-SV model in Figures 1–3 clearly show that $h_t$, and hence, $\eta_t$, is well responsive to volatility effects rather than the linear short-memory effects. For example, in SIM I where the volatility exhibits no nonlinear and long-memory effects, the $h_t$ shown in Figure 1 is small at all time steps and hence, the $\eta_t$ component simply fluctuates around $\beta_0$ during both low and high volatility periods. Figures 2 and 3, on the other hand, show that the $h_t$ response adaptively to the changes in the volatility dynamics. As the result, $\eta_t$ is small during the low volatility periods and large in the high volatility periods. The nonlinear (SIM II) and long-memory (SIM III) auto-dependence of the simulated volatility are well captured by the SRU structure of the SR-SV model. The plots of filtered volatility in Figure D.2, supplementary material show that the filtered volatility of the SR-SV model are generally closer to the true volatility than those of the SV model.

Third, the SR-SV parameters are able to characterize the existence of the various volatility effects in the simulated data. The estimated posterior means of $\beta_1$ are more than two standard deviations from zero in SIM II and SIM III, suggesting the existence of volatility effects other than linearity in the volatility dynamics; while $\beta_1$ is less than two standard deviations from zero in SIM I, suggests that only simple linear effects are detectable in the volatility dynamics of this dataset. Similarly, the nonlinear coefficient $w_z$ of the SR-SV model (see Equation (12)) is more than two standard deviations from zero in SIM II and SIM III, but not in SIM I, indicating that the SR-SV model is able to detect the serial dependence rather than the linear dependence that the past log volatility $z_{t-1}$ has on the current log volatility $z_t$. The estimated posterior mean of the moving average weight parameter $\alpha$ in SIM III is higher than those in
is worth noting that the persistence parameter is much closer to 1 than those in SIM I and SIM II. Finally, it model. Figure 4 shows that the posterior mode of the SR-SV model also outperforms the SV model for all scores with the in-sample analysis showing that the SR-SV model fits these simulation datasets better than the SV model. For SIM I, the rest and SIM III data, and hence, the historical information is adequately stored in the $\eta_t$ process.

Table 6 reports the predictive performance scores of the SR-SV and SV models with the Monte Carlo standard errors in brackets. For SIM II and SIM III, the SR-SV model outperforms the SV model across all the predictive scores, which is consistent with the in-sample analysis showing that the SR-SV model fits these simulation datasets better than the SV model. For SIM I, the SR-SV model also outperforms the SV model for all scores except the PPS score. These results illustrate the impressive out-of-sample forecast ability of the SR-SV model. The results for the real data applications in the next section further support this claim.

4.2. Applications

This section evaluates the SR-SV model using five popular daily stock indexes from different international markets: The German stock index DAX30 (DAX), the Hong Kong stock index HS50 (HSI), the France market index CAC40 (FCHI), the U.S. stock index market SP500 (SPX) and the Canada market index TSX250 (TSX).

4.2.1. The Datasets and Exploratory Data Analysis

The datasets were downloaded from the Realized Library.\textsuperscript{2} We use adjusted closing prices $\{P_t, t = 1, \ldots, T_P\}$ and calculate the demeaned return process as

$$y_t = 100 \left( \log \frac{P_{t+1}}{P_t} - \frac{1}{T_p - 1} \sum_{i=1}^{T_p-1} \log \frac{P_{i+1}}{P_i} \right),$$

$$t = 1, 2, \ldots, T_p - 1.$$  (15)

The first $T_{in} = 2000$ returns are used for in-sample analysis and the rest $T_{out} = 1000$ for the out-of-sample analysis. See Table D.1 in the supplementary material for the relevant aspects of the datasets.

Figure D.1 in the supplementary material plots the time series data and shows the existence of the volatility clustering effect commonly seen in financial data. Table 7 reports some descriptive statistics together with Lo's modified R/S test (Lo 1991) for long-range memory in the absolute and squared returns. Lo's modified R/S test is widely used in the financial time series literature; see, for example, Lo (1991), Giraitis et al. (2003), Breidt, Crato, and de Lima (1998). All the index data exhibit some negative skewness, high excess kurtosis and high variation. The result of Lo's modified R/S test for long-memory dependence with several different lags $q$ indicates that there is significant evidence of long-memory dependence in the stock indices.

The Realized Library provides different realized volatility measures that can be used in financial econometrics as a proxy to the volatility $\sigma^2_t$. We use the following four common realized measures including realized variance (RV), bipower variation (BV), median realized volatility (MedRV) and Realized Kernel Variance (RKV) to evaluate the forecast performance of the

\textsuperscript{2}The Oxford-Man Institute, https://realized.oxford-man.ox.ac.uk/
volatility models using the predictive scores in Table 3. See Shephard and Shephard (2010) for more details about the Realized Library.

### 4.2.2. In-Sample Analysis

Table 8 summarizes the estimation results of fitting the SV, N-SV and SR-SV models to these five datasets; the results for LMSV are in Table C.1 in the supplementary material. For the SR-SV model, we only show the results of the key parameters. Figure 4 shows the posterior densities of the moving average weight parameter \( \alpha \) of the SR-SV model in all simulation and real datasets. We draw some conclusions from Table 8 and the listed figures.

First, the marginal likelihood estimates show that the SR-SV model fits the five index data better than the SV and N-SV models. The Bayes factors of the SR-SV model compared to the SV and N-SV models are more than \( e^{2.3} \), which strongly support the SR-SV in all cases. There are no significant differences between the SV and N-SV models in terms of marginal likelihood estimates across the five datasets.

Second, the evidence of more general volatility effects rather than just linearity, for example, nonlinearity and long-memory auto-dependence, in the volatility dynamics of the index datasets is clear as the posterior means of the nonlinearity long-memory parameter \( \beta_1 \) of the SR-SV model are all more than two standard deviations from zero. The LMSV estimation results in Table C.1 show that the posterior means of the fractional integration parameter \( d \) are also more than two standard deviations from zero and close to 0.5 in all cases, suggesting strong evidence of the long-memory dependence in the volatility process of these five index datasets. The posterior

---

**Table 7.** Descriptive statistics for the demeaned returns of the DAX, HSI, FCHI, SPX, and TSX datasets.

|          | Min  | Max  | Std  | Skew | Kurtosis | \( V_9(10) \) | \( V_9(20) \) | \( V_9(30) \) |
|----------|------|------|------|------|----------|--------------|--------------|--------------|
| DAX      | −4.37| 9.93 | 1.27 | 0.11 | 10.96    | 2.326*       | 2.501*       | 2.146*       |
| HSI      | −11.61| 12.155| 1.186| 0.307| 17.551   | 3.934*       | 3.030*       | 2.587*       |
| FCHI     | −7.215| 6.663| 1.132| 0.320| 7.383    | 2.694*       | 2.068*       | 1.844*       |
| SPX      | −9.351| 10.220| 1.037| 0.256| 12.502   | 3.188*       | 2.412*       | 2.047*       |
| TSX      | −9.879| 9.194| 1.262| 0.272| 12.202   | 3.558*       | 2.692*       | 2.277*       |

**NOTE:** \( V_9(q) \), \( q = 10, 20 \) and 30, shows the test statistics of Lo's modified R/S test of long memory with lag \( q \). Upper and lower values of the three last columns are the Lo's test statistics for absolute and squared returns, respectively. The asterisks indicate significance at the 5% level.

**Table 8.** Applications: Posterior means of the parameters with the posterior standard deviations in brackets.

|          | \( \mu \) | \( \phi \) | \( \sigma^2 \) | \( \delta \) | \( \alpha \) | \( \beta_0 \) | \( \beta_1 \) | \( w_x \) | Mar. Lnh |
|----------|-----------|----------|------------|----------|----------|----------|----------|----------|----------|
| DAX      | SV        | −0.098   | 0.979      | 0.038    | −0.117   | 0.410    | 0.397    | 0.315    | −2871.3  |
|          | N-SV      | −0.138   | 0.977      | 0.037    | −0.198   | 0.401    | 0.397    | 0.315    | −2827.4  |
|          | SR-SV     | 0.863    | 0.064      | 0.605    | 0.204    | 0.137    | 0.262    | 0.139    | −2868.8* |
| HSI      | SV        | −0.205   | 0.987      | 0.022    | 0.784    | 0.196    | 0.536    | 0.387    | −2692.0  |
|          | N-SV      | −0.366   | 0.987      | 0.021    | 0.242    | 0.497    | 0.526    | 0.378    | −2867.8* |
|          | SR-SV     | 0.824    | 0.054      | 0.780    | 0.197    | 0.070    | 0.262    | 0.139    | −2864.2* |
| FCHI     | SV        | −0.213   | 0.977      | 0.047    | −0.179   | 0.449    | 0.363    | 0.326    | −2787.3  |
|          | N-SV      | −0.217   | 0.979      | 0.041    | −0.198   | 0.497    | 0.526    | 0.378    | −2867.8* |
|          | SR-SV     | 0.843    | 0.093      | 0.780    | 0.197    | 0.070    | 0.262    | 0.139    | −2864.2* |
| SPX      | SV        | −0.228   | 0.985      | 0.034    | 0.527    | −0.180   | 0.481    | 0.373    | −2748.3  |
|          | N-SV      | −0.267   | 0.9837     | 0.036    | −0.121   | 0.527    | 0.481    | 0.373    | −2745.6* |
|          | SR-SV     | 0.844    | 0.056      | 0.527    | 0.186    | 0.186    | 0.241    | 0.132    | (0.311)  |
| TSX      | SV        | −0.200   | 0.985      | 0.028    | 0.697    | −0.129   | 0.414    | 0.355    | −2770.1  |
|          | N-SV      | −0.249   | 0.984      | 0.029    | −0.141   | 0.697    | −0.129   | 0.414    | −2769.9  |
|          | SR-SV     | 0.868    | 0.051      | 0.697    | 0.195    | 0.071    | 0.201    | 0.141    | (0.347)  |

**NOTE:** The last column shows the estimated log marginal likelihood with the Monte Carlo standard errors in brackets, averaged over 10 different runs of the DT-SMC sampler. The asterisks indicate the cases when the Bayes factors strongly support the SR-SV model over the SV model. The marginal likelihoods are reported in the natural log scale.
means of the nonlinear parameter $w_z$ with respect to the past log volatility $z_{t-1}$ are more than two standard deviations from zero, indicating the existence of more general serial dependence rather than just the linear dependence of the current log volatility $z_t$ on the past log volatility $z_{t-1}$, and that the SR-SV model is able to detect this more general serial dependence. The posterior density plots of the moving average weight $\alpha$ in Figure 4 suggest the existence of the long-memory auto-dependence as the posterior densities of $\alpha$ are highly skewed with the modes close to 1, which is similar to the results in SIM III.

Third, it is worth noting that, in all cases, the persistence parameter $\phi$ in the SR-SV model is smaller than the persistence parameters in the SV and N-SV models as the parameter $w_z$ is significant and hence, the linear effect that $z_{t-1}$ has on $z_t$ is reduced. As illustrated in SIM I, if the volatility process exhibits no volatility effects rather than a short-memory linear auto-dependence, the persistence parameters $\phi$ in the SV and SR-SV model are similar and the parameter $w_z$ is insignificant.

Using the posterior mean estimates in Table 8, the filtered values of $z_t$ of the SR-SV, SV and N-SV models can be computed using the particle filter. Appendix C in the supplementary material discusses how to obtain the filtered values of $z_t$ of the LMSV model. Figure 5 plots the filtered log volatility of the SV and SR-SV models, together with the filtered values of the components $h_t$ and $\eta_t$ of the SR-SV model in all time steps, for the SPX data. Figures D.3–D.6 in Appendix D in the supplementary material are similar plots for the other datasets. Figure 5 shows that the component $h_t$, and hence, $\eta_t$, of the SR-SV model is responsive to changes in the volatility dynamics, for example, being small during low volatility periods and large in high volatility periods of the SPX data. The SRU structure of the SR-SV model can capture these distinct behaviors of financial time series. We observe the similar behaviors of the $h_t$ and $\eta_t$ components for the other datasets as shown in the Figures D.3–D.6.

Table 9 provides summary statistics on the in-sample filtered volatilities and residuals $\hat{\epsilon}_t^y$. We note that if the assumption about the normality of returns of the three models is justified then the residuals $\tilde{\epsilon}_t^y$ should have a zero skewness and a kurtosis of 3. Table 9 shows that all the residual distributions produced from the four models are close to the standard normal distribution, but still slightly skewed and leptokurtic. The $p$-values of the

---

**Figure 5.** SPX: (Top) the filtered log volatility of the SR-SV and SV models. (Middle) The filtered values of $\eta_t$ and $h_t$ of the SR-SV model. (Bottom) The SPX in-sample data. (This is better viewed in color.)
Ljung-Box (LB) autocorrelation test of the residuals are high in all index datasets, suggesting that there is no evidence of autocorrelation in the residuals. We conjecture that extending the SR-SV model, for example, by using a Student’s t distribution instead of a Gaussian for the measurement shock $\epsilon_t$, is likely to lead to better residual diagnostics. However, these extensions are not considered here.

### 4.2.3. Out-of-Sample Analysis

The marginal likelihood estimates in Table 8 suggest that the SR-SV model fits the in-sample data of the five index datasets better than the SV and N-SV models. We now examine if this in-sample performance is consistent with the out-of-sample performance. Table 10 provides summary statistics on the one-step-ahead forecasts of volatility and standardized residuals for all the models. We note two conclusions from Table 10.

First, the SR-SV model does not suffer from overfitting as often observed in neural network based volatility models (Pagan and Schwert 1990; Donaldson and Kamstra 1997), as the one-often observed in neural network based volatility models (Pagan and Schwert 1990; Donaldson and Kamstra 1997), as the one-step-ahead forecast volatilities and the forecast residuals appear to be well behaved, compared to those from the more parsimonious SV and N-SV models. We emphasize that, as discussed in Section 2.3, the use of noise-injecting regularization in the novel structure of the SR-SV model helps prevent it from the well-known overfitting problem.

Second, the means and standard deviations of the one-step-ahead forecast volatilities of the SR-SV model are smaller than those of the other models in all five index datasets. The SR-SV forecasts are generally more conservative in low volatility periods, in the sense that the forecast intervals often have a smaller band compared to the forecasts produced by the other models. The comparison of 99% one-step-ahead forecast intervals during the period September 2014–May 2015 of the SPX data in Figure 6 shows that the SR-SV model gives a safe buffer against abrupt changes in low volatility regions, for example, November–December 2014, because it maintains a wider forecast band, while it does not produce overly large forecast intervals in high volatility regions, for example, October 2014, December 2014–January 2015. Therefore, the SR-SV model is less sensitive to the data values in the shorter time periods, and maintains a good tradeoff between the information in recent observations and the information in the long-term memory. The SV, N-SV and LMSV models, compared to the SR-SV model, produce a smaller forecast volatility in low volatility regions and a higher volatility forecast in high volatility regions. The figure also shows that the SV and N-SV forecasts depend mainly on the return at the previous step, as the persistence parameter $\phi$ in the SV and N-SV models are larger than the persistence parameter of the SR-SV model. The SR-SV intervals are closer to the intervals in high volatility regions, for example, October 2014, December 2014–January 2015, than the intervals made by the realized variance and hence, seem to track the out-of-sample returns better than the benchmark models.

Table 11 shows the out-of-sample performance of the SR-SV and benchmark models on the SPX data. We also report predictive scores of the GP-Vol model of Wu, Hernández-Lobato, and Ghahramani (2014), which is an engineering-oriented SV model using a Gaussian process to capture the nonlinearity in the volatility dynamics. As the GP-Vol model focuses mainly on prediction, we use it as another benchmark model to assess the predictive performance of the SR-SV model. We use the software package from Wu, Hernández-Lobato, and Ghahramani (2014)
to perform Bayesian inference and prediction for the GP-Vol model, with all settings at their default values. Tables D.2–D.5 in the supplementary material show the results for the other four datasets. Each table provides predictive scores separately in four panels, corresponding to the four realized measures discussed Section 4.2.1. In each panel, we count the number of times a particular model has lowest (best) predictive scores and list these numbers in the last column; the model with the highest count is preferred. We note that the PPS predictive score is independent of the realized measures of volatility, and that the PPS predictive score is not available for the LMSV and GP-Vol models.
Table 11 and D.2–D.5 show that the SR-SV model consistently has the best out-of-sample performance in all the five datasets. The superior predictive performance of the SR-SV model is consistent with its in-sample fitting discussed earlier and provides further evidence to support the conclusion that the SR-SV model does not overfit the data. We note that the forecast performances of the SV and N-SV models are mixed with no model consistently outperforming the other across the predictive scores and the datasets. The LMSV and GP-Vol model consistently make the least accurate forecasts, except for the HSI data where the GP-Vol model has a comparable forecast performance with the SR-SV model. We note that the GP-Vol model uses a Gaussian process with its covariance matrix expanding over time and hence, it becomes computationally expensive in applications with long time series.

5. Conclusions
This article proposes a statistical recurrent stochastic volatility (SR-SV) model, by combining the statistical recurrent unit architecture from Machine Learning and the stochastic volatility model from Financial Econometrics. These two techniques are combined in a principled and nontrivial way to form a new approach that is, as carefully illustrated through the extensive simulation and empirical studies, highly efficient for volatility modeling and forecasting. It is easy to carry out Bayesian inference in the SR-SV model using standard Bayesian computation methods, such as the Density Tempered Sequential Monte Carlo method as used in this article. The simulation and empirical studies suggest that the SR-SV model is able to capture various volatility effects overlooked by the SV benchmark models, and is able to produce highly accurate forecast volatilities.

Extending the SR-SV model by incorporating features such as the leverage effect is an interesting research question. Another interesting research question is extending the present SR-SV model to multivariate financial time series. We conjecture that the RNN architectures will be even more powerful in this case as they can naturally capture the interaction between the inputs. This research is in progress.

Supplementary Materials
The online supplementary materials contain the implementation details of the numerical examples and further empirical results.

Acknowledgments
The authors would like to thank the Associate Editor and two referees for their comments which helped improve both the methodology and presentation of the article.

Funding
The research of Tran was partially supported by the Australian Research Council’s grant DP200103015.

References
Andreieu, C., Doucet, A., and Holenstein, R. (2010), “Particle Markov Chain Monte Carlo Methods,” Journal of the Royal Statistical Society, Series B, 72, 1–33. [414,415,418]
Baillie, R. T., Bollerslev, T., and Mikkelsen, H. O. (1996), “Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity,” Journal of Econometrics, 74, 3–30. [419]
Bollerslev, T. (1986), “Generalized Autoregressive Conditional Heteroskedasticity,” Journal of Econometrics, 31, 307–327. [414]
Bollerslev, T., and Mikkelsen, H. O. (1996), “Modeling and Pricing Long Memory in Stock Market Volatility,” Journal of Econometrics, 73, 151–184. [414]
Box, G. E. P., and Cox, D. R. (1964), “An Analysis of Transformations,” Journal of the Royal Statistical Society, Series B, 26, 211–243. [419]
Breidt, F., Crato, N., and de Lima, P. (1998), “The Detection and Estimation of Long Memory in Stochastic Volatility,” Journal of Econometrics, 83, 325–348. [414,415,416,419,422]
Choi, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014), “Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation,” in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1724–1734, Doha, Qatar: Association for Computational Linguistics. [417]
Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013), “Smc2: An Efficient Algorithm for Sequential Analysis of State Space Models,” Journal of the Royal Statistical Society, Series B, 75, 397–426. [419]
Crato, N., and de Lima, P. J. (1994), “Long-Range Dependence in the Conditional Variance of Stock Returns,” Economics Letters, 45, 281–285. [414]
Del Moral, P. (2004), Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications, New York: Springer. [418]
Del Moral, P., Doucet, A., and Jasra, A. (2006), “Sequential Monte Carlo Samplers,” Journal of the Royal Statistical Society, Series B, 68, 411–436. [418]
Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018), “The Correlated Pseudo Marginal Method,” Journal of the Royal Statistical Society, Series B, 80, 839–870. [415,418]
Dieng, A. B., Ranganath, R., Altosaar, J., and Blei, D. M. (2018), “Noising: Unbiased Regularization for Recurrent Neural Networks,” in International Conference on Machine Learning (pp. 1252-1261), PMLR. [418]
Ding, Z., Granger, C. W. J., and Engle, R. F. (1993), “A Long Memory Property of Stock Market Returns and a New Model,” Journal of Empirical Finance, 1, 83–106. [414]
Donaldson, R. G., and Kamstra, M. (1997), “An Artificial Neural Network-Garch Model for International Stock Return Volatility,” Journal of Empirical Finance, 4, 17–46. [425]
Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015), “Efficient Implementation of Markov Chain Monte Carlo When Using an Unbiased Likelihood Estimator,” Biometrika, 102, 295–313. [418]
Duan, J.-C., and Fulop, A. (2015), “Density-Tempered Marginalized Sequential Monte Carlo Samplers,” Journal of Business & Economic Statistics, 33, 192–202. [415,418]
Elman, J. L. (1990), “Finding Structure in Time,” Cognitive Science, 14, 179–211. [416]
Giraitis, L., Kokoszka, P., Leipus, R., and Teyssière, G. (2003), “Rescaled Variance and Related Tests for Long Memory in Volatility and Levels,” Journal of Econometrics, 112, 265–294. [422]
Goodfellow, I., Bengio, Y., and Courville, A. (2016), Deep Learning, Cambridge, MA: MIT Press. [417]
Granger, C. W. J., and Joyeux, R. (1980), “An Introduction to Long-Memory Time Series Models and Fractional Differencing,” Journal of Time Series Analysis, 1, 15–29. [414]
Higgins, M. L., and Bera, A. K. (1992), “A Class of Nonlinear ARCH Models,” International Economic Review, 33, 137–158. [419]
Hochreiter, S., and Schmidhuber, J. (1997), “Long-Short-Term-Memory,” Neural Computation, 9, 1735–1780. [417]
Jacquier, E., Polson, N. G., and Rossi, P. E. (1994), “Bayesian Analysis of Stochastic Volatility Models” (with discussion), Journal of Business and Economic Statistics, 12, 371–417. [416]
Jeffreys, H. (1961), Theory of Probability (3rd ed.), Oxford: Clarendon Press. [419]
Kass, R. E., and Raftery, A. E. (1995), “Bayes Factors,” Journal of the American Statistical Association, 90, 773–795. [419]
Kılıç, R. (2011), “Long Memory and Nonlinearity in Conditional Variances: A Smooth Transition FIGARCH Model,” *Journal of Empirical Finance*, 18, 368–378. [414]

Kim, H. Y., and Won, C. H. (2018), “Forecasting the Volatility of Stock Price Index: A Hybrid Model Integrating LSTM with Multiple Garch-Type Models,” *Expert Systems with Applications*, 103, 25–37. [415]

Kim, S., Shephard, N., and Chib, S. (1998), “Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models,” *Review of Economic Studies*, 65, 361–393. [414, 416, 418, 419]

Lipton, Z., Berkowitz, J., and Elkan, C. (2015), “A Critical Review of Recurrent Neural Networks for Sequence Learning,” arXiv:1804.04359. [415]

Liu, J. S., and Chen, R. (1998), “Sequential Monte Carlo Methods for Dynamic Systems,” *Journal of the American Statistical Association*, 93, 1032–1044. [418]

Lo, A. W. (1991), “Long-Term Memory in Stock Market Prices,” *Econometrica*, 59, 1279–1313. [414, 422]

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018), “Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward,” *PLOS ONE*, 13, 1–26. [415]

Mandelbrot, B. (1967), “The Variation of Some Other Speculative Prices,” *The Journal of Business*, 40, 393–413. [414]

Neal, R. (2001), “Annealed Importance Sampling,” *Statistics and Computing*, 11, 125–139. [418]

Nelson, D. B. (1991), “Conditional Heteroskedasticity in Asset Returns: A New Approach,” *Econometrica*, 59, 347–370. [414]

Oliva, J. B., Póczos, B., and Schneider, J. G. (2017), “The Statistical Recurrent Unit,” in *ICML2017*. [415, 417]

Pagan, A. R., and Schwert, G. (1990), “Alternative Models for Conditional Stock Volatility,” *Journal of Econometrics*, 45, 267–290. [425]

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012), “On Some Properties of Markov Chain Monte Carlo Simulation Methods Based on the Particle Filter,” *Journal of Econometrics*, 171, 134–151. [418]

Quiroz, M., Tran, M.-N., Villani, M., Kohn, R., and Dang, K.-D. (2021), “The Block-Poisson Estimator for Optimally Tuned Exact Subsampling MCMC,” *Journal of Computational and Graphical Statistics*, 30, 877–888. [418]

Shephard, N., and Sheppard, K. (2010), “Realising the Future: Forecasting with High-Frequency-Based Volatility (Heavy) Models,” *Journal of Applied Econometrics*, 25, 197–231. [423]

Sietsma, J., and Dow, R. J. (1991), “Creating Artificial Neural Networks that Generalize,” *Neural Networks*, 4, 67–79. [418]

Taylor, S. J. (1982), “Financial Returns Modelled by the Product of Two Stochastic Processes—A Study of Daily Sugar Prices 1961–79,” in *Time Series Analysis: Theory and Practice*, eds. O. D. Anderson, pp. 203–226, Amsterdam: North-Holland. [414, 416]

Tran, M.-N., Strickland, C., Pitt, M. K., and Kohn, R. (2014), “Annealed Important Sampling for Models with Latent Variables,” Technical Report, University of New South Wales, Australia. arXiv:1402.6035. [418]

van Bellegem, S. (2012), “Locally Stationary Volatility Modeling,” in *Volatility Models and Their Applications*, eds. L. Bauwens, C. Hafner, and S. Laurent, pp. 249–262, New York: Wiley. [417]

Wu, Y., Hernández-Lobato, J. M., and Ghahramani, Z. (2014), “Gaussian Process Volatility Model,” in *Advances in Neural Information Processing Systems* (Vol. 27), eds. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Red Hook, NY: Curran Associates, Inc. [416, 425]

Yu, J. (2002), “Forecasting Volatility in the New Zealand Stock Market,” *Applied Financial Economics*, 12, 193–202. [414]

Yu, J., Yang, Z., and Zhang, X. (2006), “A Class of Nonlinear Stochastic Volatility Models and its Implications for Pricing Currency Options,” *Computational Statistics and Data Analysis*, 51, 2218–2231. [415, 416, 419]

Zhang, G. (2003), “Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model,” *Neurocomputing*, 50, 159–175. [415]