Erasmus University Rotterdam

Erasmus School of Economics

Master Thesis

Bivariate Stochastic Volatility Model with General Leverage and Cross-Sectional Correlations Specification

Author: Elwin Kardux (467905)

University Supervisor: dr.(Rutger-Jan) R-J Lange
Second assessor: dr.(Annika) A Camehl

January 16, 2022

A thesis submitted in partial fulfillment for the degree of
MSc Econometrics and Management Science

Bivariate Stochastic Volatility Model with General Leverage and Cross-Sectional Correlations Specification

Abstract

In this paper we introduce a bivariate stochastic volatility model that allows for contemporaneous, intertemporal and cross-sectional correlations between return shocks and volatility shocks. The model is reconstructed as a two-dimensional nonlinear Gaussian state-space model and estimated through Bellman filtering. An empirical analysis illustrates that the cross-sectional dependencies improve model fit and volatility predictions when a significant return correlation is present. Lastly, we find that cross-sectional dependencies substantiate the leverage effect through a skew term that is implied by our model.

Key words: bivariate stochastic volatility, leverage effect, bellman filter, multidimensional nonlinear state-space model.

Contents

1 Introduction

Financial time-series often exhibit time-varying and clustering volatility; they contain periods of swings interspersed with periods of relatively low and stable volatility. Over the years, researchers have found many ways to model the relation between return and volatility processes. This has led to many specifications that allow for different correlation structures between the processes, among which the timing of the correlation is of high importance. This paper adds to that strand of literature by introducing a bivariate stochastic volatility model that combines three different sources of correlation: contemporaneous (Jacquier et al., 2004), intertemporal (Harvey and Shephard, 1996) and cross-sectional correlations (Chan et al., 2006).

Historically, (G)ARCH type models (Bollerslev, 1986) were one of the only practical type of models to deal with the time-varying and clustering volatility features due to their simplicity. However, as more research is devoted to developing stochastic volatility (SV) models to deal with these volatility phenomena, they are progressively becoming a more suitable alternative. One of the earliest univariate SV models was proposed by Taylor (1982). He assumed that the volatility process of returns follow a stochastic process instead of a deterministic process assumed by GARCH-type models. This assumption makes the SV model more flexible than standard GARCH-type models in exchange for model complexity.

Although SV models have an intuitive appeal, their empirical application is often limited mainly due to difficulties involved in their estimation, with the main issue being that the likelihood function is hard to evaluate. A survey by Broto and Ruiz (2004) revised different estimation procedures for SV models and compared their efficiency. Simulation based approaches (e.g. Markov chain Monte Carlo by Jacquier et al. 2002); importance sampling methods (Richard and Zhang, 2007; Koopman et al., 2015) are found to be statistically more efficient than basic filtering methods (e.g. quasi maximum likelihood Harvey et al.1994), with the downside being computational intensity. This downside plays a critical role when modelling higher dimensional state space models, as simulation methods are generally subject to the curse of dimensionality.

We implement the Bellman filter (BF) of Lange (2020) for state estimation. This recent filtering method is based on Bellman and Dreyfus (1959)'s dynamic programming principles and performs comparably with simulation-based techniques like the particle filter of Malik and Pitt (2011) and the numerical accelerated importance sampling method of Koopman et al. (2015). Crucially, his findings demonstrate that the BF requires relatively little computational power. Therefore, the BF is preferred over simulation methods when working with higher dimensional state-space models.

A wide range of research has also been devoted to modelling the 'leverage effect' of financial

returns, which describes the asymmetric relation between the sign (positive or negative) of past returns and its volatility. Especially in equity markets, we often see falls in prices lead to an increase in volatility. This can cause negative skewness effects in returns. Basic SV models allow for leptokurtic distributions and volatility clustering, but do not immediately incorporate such asymmetric relations. Leverage effects are most commonly introduced in SV models through a correlation parameter between the return equation and (lagged) volatility equation. The timing, whether it is contemporaneous or intertemporal, presents substantial differences in theoretical model implications (Yu, 2005). Catania (2020) studied the propagation of the leverage effect and demonstrated that having a proper specification of this leverage effect over time is imperative for the model diagnostics and volatility forecasting. They followed up on their research in Catania (2021) by introducing a univariate SV model that combines contemporaneous and intertemporal correlation structures between the shocks of financial returns and volatilities. Their postulated general correlation specification leads to improved model fit and volatility predictions compared to the more restricted temporaneous or intertemporal specifications. We investigate whether cross-sectional dependencies between return shocks and volatility shocks impact the leverage effect results of Catania (2021).

Multivariate SV models are mainly different from univariate SV models in their cross-sectional correlation structure. Having a less restricted cross-sectional correlation structure can be beneficial in many financial applications, as such structures are known to be present in areas such as financial derivatives and fixed income products. Besides being more applicable, they can also lead to better model diagnostics. Asai et al. (2006) conducted a summary of existing multivariate SV models and emphasized on the economic value of using multivariate SV models. Omori et al. (2007) performed Bayesian analysis on bivariate SV models with leverage, where mixtures of bivariate normals and Student t-densities, and intertemporal correlations are used to incorporate leverage effects. Asai and McAleer (2009) propose another multivariate SV model with leverage using Bayesian methods, where they also use heavy-tailed return distributions and only allow for contemporaneous correlations between the return an volatility processes. Overall, several multivariate SV models are proposed in the literature where leverage effects are mainly introduced through either contemporary or intertemporal correlations between return and volatility shocks and are mainly based on Bayesian methods.

The focus of this research is on further generalizing the model of Catania (2021) in a bivariate SV framework. Cross-sectional correlations are introduced through four control parameters: one controlling the unconditional correlation between the return series, one controlling the instantaneous transmission of volatility shocks and two parameters controlling the transmission of one's lagged volatility to the other current volatility process.

The research question, then, is how effective are the cross-sectional correlations as an

addition to the contemporaneous and intertemporal specification. That is, to what extent does the model fit and out-of-sample performance benefit from the additional correlation parameters. We answer this question through an empirical analysis on three index data pairs: (S&P500, FTSE100), (S&P500, Nikkei225) and (S&P500, NASDAQ 100).

Our simulation studies have shown that our estimation procedure using the BF contains some biases and unstable estimates, especially for the instantaneous volatility transmission parameter. However, volatility predictions obtained using the biased parameters are found to be statistically indifferent from those obtained using the actual parameters.

In our empirical analysis, we find that allowing for a non-zero cross-sectional correlation between returns can be a promising improvement to the model fit and the quality of volatility predictions, given that the correlation is significant enough. The effect of lagged transmission between volatility shocks is shown to be insignificant for the data pairs, whereas the instantaneous transmission is positive and significant for some data pairs. Both lagged and instantaneous volatility transmission effects do not improve volatility predictions.

Lastly, we find that the leverage effect through the contemporaneous and intertemporal correlations remain present when cross-sectional dependencies are added. Cross-sectional correlations are shown to impact the model implied mean and variance of both the return and volatility processes.

Restricting the cross-sectional correlation parameters to 0 leads to an independent bivariate model of Catania (2021). The model of Catania (2021) encompasses the models of Jacquier et al. (2004), Yu (2005), Taylor (1994) and Harvey and Shephard (1996), and thus are also encompassed by our model under the right parameter constraints. This implies that our bivariate model framework is well suited for a wide variety of hypothesis tests to compare specific models.

The remainder of this paper is structured as follows. First, the bivariate model is specified and its state-space representation is constructed in Section 2. Then the Bellman filtering approach is described in Section 3. The empirical results of the model are discussed in Section 4. Finally, concluding remarks are given in Section 5.

2 The model

This section first presents the model of Catania (2021) in Section 2.1, then introduces a bivariate extension to the model and discuss model implications. Next, we derive a state-space format for the bivariate model in Section 2.3. In this paper, the diag(.) operator is used both as the typical vector-to-matrix operator when the argument is a vector, and the matrix-to-vector operator when the argument is a matrix resulting in a vector of its diagonal entries. The

vec(.) operator performs a linear transformation which converts a matrix argument into a column vector.

2.1 Univariate SV model with leverage propagation

The univariate SV model of Catania (2021) (SV-C) is unique in its generalized leverage specification. It extends the basic SV model by incorporating correlations of shocks between the measurement and log volatility equations at different temporal lags. The model, extended with constants in both the return (y_t) and volatility (x_t) equations, is specified as follows

$$y_t = \tilde{y} + \beta \exp\left\{\frac{x_t}{2}\right\} \varepsilon_t, \quad \varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \tag{1}$$

$$x_t = \phi x_{t-1} + \sigma_\eta \eta_t, \tag{2}$$

$$\eta_t = \sum_{j=0}^{k} \rho_j \varepsilon_{t-j} + \sigma_\xi \xi_t, \quad \xi_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{3}$$

The parameter space is assumed to ensure stationarity in y_t and x_t: $\theta_u = (\tilde{y}, \beta, \phi, \sigma_\eta, \boldsymbol{\rho}) \in \Theta_u$, with $\boldsymbol{\rho} = (\rho_0, \rho_1, ..., \rho_k)'$ and $\Theta_u = (0,\infty) \times (-1,1) \times (0,\infty) \times \mathcal{S}$, and $\mathcal{S} = \{(\rho_0, ..., \rho_k) \in (-1,1)^{k+1} | \sum_{j=0}^{k} \rho_j^2 < 1\}$. The constraint in $\mathcal{S}$ is for identification purposes of the σ_η parameter, as it is required that $\sigma_\xi = \sqrt{1 - \sum_{k=0}^{k} \rho_j^2}$, with $\sum_{j=0}^{k} \rho_j^2 \leq 1$, to ensure that the unconditional variance of η_t is 1.

In this model the volatility shocks η_t are linked in equation (3) through a linear function to the contemporaneous return shock ε_t and intertemporal return shocks $\varepsilon_{t-1}, \ldots, \varepsilon_{t-k}$. Parameters β, κ and ϕ control the level, standard deviation and persistence of the volatility process, respectively. As the number of temporal lags k increases, the model might get a better representation of the leverage propagation, with the drawback of less parsimony. Lange (2020) used a similar model specification but instead of the β parameter in equation (1), he added a constant mean term to the volatility equation in (2) as the level control parameter of the volatility equation. Furthermore, he argues that $\tilde{y}$ represents the median of process y_t due to the model implications on the skewness of the distribution of y_t. This skewness effect becomes more transparent in the state-space representation of the model in Section 2.3.

2.2 Bivariate extension

The following notation is used for the variables at time t for assets a and b:

- $\mathbf{y}_t = (y_{at}, y_{bt})'$: observed returns process;
- $\mathbf{x}_t = (x_{at}, x_{bt})'$: latent log-volatility process;

- $\boldsymbol{\varepsilon}_t = (\varepsilon_{at}, \varepsilon_{bt})'$: return shock;
- $\boldsymbol{\eta}_t = (\eta_{at}, \eta_{bt})'$: volatility shock;
- $\boldsymbol{\xi}_t = (\xi_{at}, \xi_{bt})'$: normally distributed error term.

The bivariate generalization of the stochastic volatility model of Catania (2021) can be modelled as follows

$$\mathbf{y}_t = \tilde{\mathbf{y}} + \mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}_2\left(\vec{\mathbf{0}}_2, \mathbf{R}_\varepsilon\right), \tag{4}$$

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t, \tag{5}$$

$$\boldsymbol{\eta}_t = \sum_{j=0}^{k} \mathbf{R}_j \boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\Sigma}_\xi^{1/2}\boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \overset{iid}{\sim} \mathcal{N}_2\left(\vec{\mathbf{0}}_2, \mathbf{1}_2\right) \tag{6}$$

where $\boldsymbol{\Omega}_t^{1/2}$, $\boldsymbol{\Sigma}_\eta^{1/2}$ and $\boldsymbol{\Sigma}_\xi^{1/2}$ are Cholesky decompositions of $\boldsymbol{\Omega}_t$, $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{\Sigma}_\xi$ respectively, and

$$\begin{gathered}
\tilde{\mathbf{y}} = \begin{bmatrix} \tilde{y}_a \\ \tilde{y}_b \end{bmatrix}, \quad \mathbf{B} = \operatorname{diag}(\beta_a, \beta_b), \quad \boldsymbol{\Omega}_t = \exp(\operatorname{diag}(\mathbf{x}_t)), \quad \mathbf{R}_\varepsilon = \begin{bmatrix} 1 & \rho_\varepsilon \\ \rho_\varepsilon & 1 \end{bmatrix}, \\
\boldsymbol{\mu} = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} \phi_{aa} & \phi_{ab} \\ \phi_{ba} & \phi_{bb} \end{bmatrix}, \quad \boldsymbol{\Sigma}_\eta = \operatorname{diag}(\sigma_{\eta a}^2, \sigma_{\eta b}^2), \\
\mathbf{R}_j = \operatorname{diag}(\rho_{aj}, \rho_{bj}) \text{ for } j = 0, 1, \ldots, k, \quad \boldsymbol{\Sigma}_\xi = \begin{bmatrix} \sigma_{\xi,a}^2 & \sigma_{\xi,ab} \\ \sigma_{\xi,ba} & \sigma_{\xi,b}^2 \end{bmatrix}.
\end{gathered} \tag{7}$$

The volatility equation could be written equivalently with a diagonal structure in $\boldsymbol{\Sigma}_\xi$ instead of $\boldsymbol{\Sigma}_\eta$, or with diagonal $\boldsymbol{\Sigma}_\varepsilon$ and non-diagonal $\boldsymbol{\Omega}_t$ and $\mathbf{R}_j$ for $j = 0, 1, \ldots, k$. However, notational wise, the presented model specification is preferred to let the correlation parameters be present in the covariances of the error terms $(\boldsymbol{\varepsilon}_t, \boldsymbol{\eta}_t)$ rather than in $\boldsymbol{\Omega}_t, \boldsymbol{\Sigma}_\eta$, such that the conditional structure in the next section can be derived more neatly.[1] Hence the imposed diagonal structure in $\boldsymbol{\Sigma}_\eta$, $\boldsymbol{\Omega}_t$ and $\mathbf{R}_j$ for $j = 0, 1, \ldots, k$.

In this model, both $\boldsymbol{\mu}$ and $\mathbf{B}$ are included as level parameters of the volatility process such that the derivations can be used for both specifications for further research. For identification purposes, during the parameter estimation either $\boldsymbol{\mu}$ should be restricted to $\vec{\mathbf{0}}_2$ or $\mathbf{B}$ to $\mathbf{1}_2$, where we denote $\vec{\mathbf{0}}_n$ as the zero vector of size n and $\mathbf{1}_n$ as the identity matrix of size n. Furthermore we propose the following restriction on $\boldsymbol{\Sigma}_\xi$:

[1] This is more convenient for the joint error structure in equation (15).

$$\boldsymbol{\Sigma}_\xi = \begin{bmatrix} 1 - \sum_{j=0}^{k} \rho_{aj}^2 & \rho_\eta - \sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon \\ \rho_\eta - \sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon & 1 - \sum_{j=0}^{k} \rho_{bj}^2 \end{bmatrix}. \tag{8}$$

This restriction causes the unconditional covariance of $\boldsymbol{\eta}_t$, derived in Appendix A, to be equal to the correlation matrix $\mathbf{R}_\eta$,

$$\mathbf{R}_\eta = \begin{bmatrix} 1 & \rho_\eta \\ \rho_\eta & 1 \end{bmatrix}, \tag{9}$$

such that the parameters in $\boldsymbol{\Sigma}_\eta$ and the correlation parameter ρ_η can be identified. $\boldsymbol{\Sigma}_\xi$ has to be positive definite in order for the Cholesky decomposition to work, $\mathbf{z}'\boldsymbol{\Sigma}_\xi\mathbf{z} > 0$ for all $\mathbf{z} \in \mathbb{R}^2$. This gives us the following two parameter restrictions:

- $\sum_{j=0}^{k} \rho_{ij}^2 < 1, \quad i \in \{a, b\}$;
- $1 - \sum_{j=0}^{k} \rho_{ij}^2 > \left|\rho_\eta - \sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon\right|, \quad i \in \{a, b\}$.

The first restriction ensures positive diagonal elements by limiting the effects of $\mathbf{R}_0, ..., \mathbf{R}_k$. The second restriction ensures $\boldsymbol{\Sigma}_\xi$ is diagonally dominant and entails that ρ_η may not be too distant from $\sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon$ when $\sum_{j=0}^{k} \rho_{ij}^2$ is large. In practice we find that neither constraints cause prominent issues.

The parameter set is given by $\boldsymbol{\theta} = \{\tilde{\mathbf{y}}, \text{diag}(\mathbf{B}), \boldsymbol{\mu}, \text{vec}(\boldsymbol{\Phi}), \text{diag}(\mathbf{R}_0), \ldots, \text{diag}(\mathbf{R}_k), \rho_\varepsilon, \rho_\eta\}$. The parameter space is assumed to ensure stationarity in the returns $\mathbf{y}_t$ and volatility $\mathbf{x}_t$ processes: $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta} = (-\infty, \infty)^2 \times (0, \infty)^2 \times (-\infty, \infty)^2 \times \mathcal{P} \times \mathcal{R}$, with $\mathcal{P}$ and $\mathcal{R}$ defined as

$$\mathcal{P} = \{\text{vec}(\boldsymbol{\Phi}) \in (-1, 1)^4 | \det(\mathbf{1}_2 - \boldsymbol{\Phi} z) \neq 0 \ \forall \ |z| \leq 1\}, \tag{10}$$

$$\mathcal{R} = \left\{(\text{diag}(\mathbf{R}_0), ..., \text{diag}(\mathbf{R}_k), \rho_\eta, \rho_\varepsilon) \in (-1, 1)^{2k+4} | \mathbf{z}'\boldsymbol{\Sigma}_\xi\mathbf{z} > 0 \ \forall \ \mathbf{z} \in \mathbb{R}^2\right\}. \tag{11}$$

The constraint in equation (10) represents the stationarity condition for a VAR(1) process (Lütkepohl, 2005) and the constraint in equation (11) ensures $\boldsymbol{\Sigma}_\xi$ is positive definite, similar to the parameter restrictions mentioned above.

The covariance matrix of $\mathbf{x}_t$ conditional on $\mathbf{x}_{t-1}$ is equal to

$$\text{Var}(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \text{Var}(\boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_\eta^{1/2}\mathbf{R}_\eta\boldsymbol{\Sigma}_\eta^{1/2} = \begin{bmatrix} \sigma_{\eta a}^2 & \sigma_{\eta a}\sigma_{\eta b}\rho_\eta \\ \sigma_{\eta a}\sigma_{\eta b}\rho_\eta & \sigma_{\eta b}^2 \end{bmatrix}. \tag{12}$$

Parameter ϕ_{ab} controls the transmission of volatility of process a at time $t-1$ to process b at time t, and vice versa for ϕ_{ba}. *Granger causality* (Granger, 1969) could play a role in the event that $\phi_{ab} = 0$ while $\phi_{ba} \neq 0$, or $\phi_{ba} = 0$ while $\phi_{ab} \neq 0$. In the first case the

volatility process of a is Granger caused by the volatility process of b, and conversely for the latter. The volatility noise correlation parameter ρ_η controls the instantaneous transmission of volatility shocks between the processes of a and b. The correlation parameter ρ_ε controls the unconditional correlation between y_{at} and y_{bt}. When $\rho_\varepsilon = \rho_\eta = \phi_{ab} = \phi_{ba} = 0$ holds, the model would be equivalent to stacking two independent univariate SV-Cs of equations (1-3).

2.3 State-space format

This section provides a shortened version of the derivation with more details on the model implications. The full derivation is given in Appendix C. In our derivations we allow for one intertemporal lag ($k = 1$ in equation (3)), so only $\mathbf{R}_0$ and $\mathbf{R}_1$ are incorporated in the model) for the sake of simplicity and parsimony. We define the latent state as $\boldsymbol{\alpha}_t = (\mathbf{x}_t, \mathbf{x}_{t-1})'$. The model in equations (4 - 6) can be written as a general state-space model with Gaussian nonlinear measurement and state-transition equations:

$$\mathbf{y}_t \sim p_{\mathbf{y}}\left(\mathbf{y}_t \mid \boldsymbol{\alpha}_t\right), \quad \boldsymbol{\alpha}_t \sim p_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{t-1}\right), \quad \boldsymbol{\alpha}_1 \sim p_{\boldsymbol{\alpha}_1}\left(\boldsymbol{\alpha}_1\right), \tag{13}$$

Due to the overlapping elements in $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{t-1}$, the state-transition density can be equivalently written for our model as

$$\boldsymbol{\alpha}_t \sim p_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{t-1}\right) = p_{\boldsymbol{\alpha}}\left((\mathbf{x}_t, \mathbf{x}_{t-1})' \mid (\mathbf{x}_{t-1}, \mathbf{x}_{t-2})'\right) = p_{\boldsymbol{\alpha}}\left(\mathbf{x}_t' \mid (\mathbf{x}_{t-1}, \mathbf{x}_{t-2})'\right) = p_{\mathbf{x}}\left(\mathbf{x}_t \mid \boldsymbol{\alpha}_{t-1}\right). \tag{14}$$

The error structure given the information set at time $t-2$, defined as $\mathcal{F}_{t-2} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{t-2}\}$, for the bivariate model has the following joint distribution:

$$\begin{bmatrix} \eta_{a,t} \\ \varepsilon_{a,t} \\ \varepsilon_{a,t-1} \\ \eta_{b,t} \\ \varepsilon_{b,t} \\ \varepsilon_{b,t-1} \end{bmatrix} \mid \mathcal{F}_{t-2} \sim \mathcal{N}_6 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{a0} & \rho_{a1} & \rho_\eta & 0 & 0 \\ \rho_{a0} & 1 & 0 & 0 & \rho_\varepsilon & 0 \\ \rho_{a1} & 0 & 1 & 0 & 0 & \rho_\varepsilon \\ \rho_\eta & 0 & 0 & 1 & \rho_{b0} & \rho_{b1} \\ 0 & \rho_\varepsilon & 0 & \rho_{b0} & 1 & 0 \\ 0 & 0 & \rho_\varepsilon & \rho_{b1} & 0 & 1 \end{bmatrix} \right). \tag{15}$$

The conditional normal lemma given in lemma 1 in the Appendix is used to find the distribution conditional on extra information. Similarly as in Lange (2020), we use the information that the past return shocks $\boldsymbol{\varepsilon}_{t-1}$ are non-random conditional on the information set and $\boldsymbol{\alpha}_{t-1}$. As we are looking for a state-space format as in equation (13), we can condition the state-transition equation, or equivalently $\boldsymbol{\eta}_t$, on $\boldsymbol{\alpha}_{t-1}$. Conditional on $\boldsymbol{\alpha}_{t-1}$, $\boldsymbol{\varepsilon}_{t-1}$ can be constructed using equation (4): $\boldsymbol{\varepsilon}_{t-1} = \boldsymbol{\Omega}_t^{-1/2}\mathbf{B}^{-1}(\mathbf{y}_t - \tilde{\mathbf{y}})$. Using this information, the

distribution of the volatility shock $\boldsymbol{\eta}_t$ conditional on $\mathcal{F}_{t-1}, \boldsymbol{\alpha}_{t-1}$ is equal to

$$\begin{gathered}\boldsymbol{\eta}_t \mid \mathcal{F}_{t-1}, \boldsymbol{\alpha}_{t-1} \sim \mathcal{N}_2\left(\boldsymbol{\mu}_{\eta|\alpha,t}, \boldsymbol{\Sigma}_{\eta|\alpha}\right), \\ \boldsymbol{\mu}_{\eta|\alpha,t} = \frac{1}{1-\rho_\varepsilon^2}\begin{bmatrix} \rho_{a1}\varepsilon_{a,t-1} - \rho_{a1}\rho_\varepsilon\varepsilon_{b,t-1} \\ -\rho_{b1}\rho_\varepsilon\varepsilon_{a,t-1} + \rho_{b1}\varepsilon_{b,t-1} \end{bmatrix}, \boldsymbol{\Sigma}_{\eta|\alpha} = \begin{bmatrix} 1-\frac{\rho_{a1}^2}{1-\rho_\varepsilon^2} & \rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} \\ \rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} & 1-\frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} \end{bmatrix}.\end{gathered} \tag{16}$$

The distribution of the current return shock $\boldsymbol{\varepsilon}_t$ conditional on $\boldsymbol{\eta}_t, \boldsymbol{\alpha}_t, \mathcal{F}_{t-1}$[2] is given as

$$\begin{aligned}\boldsymbol{\varepsilon}_t|\boldsymbol{\eta}_t, \boldsymbol{\alpha}_t, \mathcal{F}_{t-1} &\sim \mathcal{N}_2\left(\boldsymbol{\mu}_{\varepsilon|\eta,t}, \boldsymbol{\Sigma}_{\varepsilon|\eta}\right), \\ \boldsymbol{\mu}_{\varepsilon|\eta,t} &= \mathbf{R}_0\boldsymbol{\Sigma}_{\eta|\alpha}^{-1}\left(\boldsymbol{\eta}_t - \boldsymbol{\mu}_{\eta|\alpha,t}\right), \\ \boldsymbol{\Sigma}_{\varepsilon|\eta} &= \mathbf{R}_\varepsilon - \mathbf{R}_0\boldsymbol{\Sigma}_{\eta|\alpha}^{-1}\mathbf{R}_0\end{aligned} \tag{17}$$

The conditional density of $\mathbf{y}_t|\boldsymbol{\alpha}_{t-1}, \mathcal{F}_{t-1}$ follows from that of $\boldsymbol{\varepsilon}_t$:

$$\begin{aligned}p_{\mathbf{y}}(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1}) &= \frac{1}{\sqrt{(2\pi)^2\det(\boldsymbol{\Sigma}_{\mathbf{y},t})}}\exp\left(-\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t})'\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t})\right), \\ \boldsymbol{\mu}_{\mathbf{y},t} &= \tilde{\mathbf{y}} + \mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\mu}_{\varepsilon|\eta,t}, \\ \boldsymbol{\Sigma}_{\mathbf{y},t} &= \mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Sigma}_{\varepsilon|\eta}\boldsymbol{\Omega}_t^{1/2\prime}\mathbf{B}'.\end{aligned} \tag{18}$$

The conditional mean of the returns process, $\boldsymbol{\mu}_{\mathbf{y},t}$ in equation (18), consists of the constant vector $\tilde{\mathbf{y}}$ and a second term that effects the mean and variance of the conditional distribution of $\mathbf{y}_t$ over time, $\mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\varepsilon}_t$. From here on, we refer to this second term as the model-implied skew. This model-implied skew has some similarities with the nonparametric skewness (NPS), defined as

$$\text{NPS} = \frac{\text{mean}(\mathbf{y}) - \text{median}(\mathbf{y})}{\text{std. deviation}(\mathbf{y})}, \tag{19}$$

but conflating these terms would be spurious. Specifically, the model-implied skew and the NPS have the same sign, are both standardized by a standard deviation term and it reveals either left- or right-skewness equally, but the model-implied skew affects the variance of $\mathbf{y}_t$ whereas the NPS does not. The latter causes affine transformations to $\mathbf{y}_t$ to provide different results for the model-implied skew whereas the NPS would not change, which violates a main property of the NPS.

The sign and magnitude of the model-implied skew term are determined by the magnitudes of $\mathbf{R}_0, \mathbf{R}_1, \rho_\varepsilon, \rho_\eta$ and $\boldsymbol{\eta}_t$. Often asset returns exhibit a negative skewness, meaning large negative returns occur more often than large positive ones. In case $\mathbf{y}_t$ is negatively skewed, it is expected that mode($\mathbf{y}$) > median($\mathbf{y}$) > mean($\mathbf{y}$) (Von Hippel, 2005)[3], hence the model-

[2] $\boldsymbol{\eta}_t$ can be left out of the conditioning set as it is implied by $\boldsymbol{\alpha}_t$.

[3] This relation generally holds for close-to-Gaussian unimodal distributions.

implied skew term would also be expected to be more negative over time. The model-implied skew also implies that the conditional correlation between $y_{a,t}$ and $y_{b,t}$ are affected by $\rho_\eta, \mathbf{R}_0$ and $\mathbf{R}_1$.

The conditional (degenerate) density of the latent state $\mathbf{x}_t|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1}$ follows from the conditional distribution $\boldsymbol{\eta}_t \mid \mathcal{F}_{t-1}$ in equation (16):

$$
\begin{aligned}
p_{\mathbf{x}}(\mathbf{x}_t|\boldsymbol{\alpha}_{t-1}, \mathcal{F}_{t-1}) = \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma}_{\mathbf{x}})}} \exp\left(-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x},t})' {\boldsymbol{\Sigma}_{\mathbf{x}}}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x},t})\right) \times \delta_t \\
\boldsymbol{\mu}_{\mathbf{x},t} = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \boldsymbol{\Sigma}_\eta^{1/2} \boldsymbol{\mu}_{\eta|\alpha,t}, \\
\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Sigma}_\eta^{1/2} \boldsymbol{\Sigma}_{\eta|\alpha} \boldsymbol{\Sigma}_\eta^{1/2}, \\
\delta_t = \prod_{j=1}^{2} \delta_d(\alpha_{t,j+2} - \alpha_{t-1,j}).
\end{aligned}
\tag{20}
$$

The density follows from equation (16), where a product of Dirac delta[4] functions are added similarly as in Lange (2020). The Dirac deltas restrict the elements of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{t-1}$, in a way such that $\boldsymbol{\alpha}_{t,3} = \mathbf{x}_{a,t-1} = \boldsymbol{\alpha}_{t-1,1}$ and $\boldsymbol{\alpha}_{t,4} = \mathbf{x}_{b,t-1} = \boldsymbol{\alpha}_{t-1,2}$. This classical optimization technique forces the constraints onto the objective function without influencing the data generating process, but simplify the optimization problem as it reduces the number of state variables to optimize. The optimization problem is introduced in Section 3.2.

Intriguingly, return correlation parameter ρ_ε appears (indirectly) in every equation from (15- 20), often more than once. Parameter ρ_η is present in $\boldsymbol{\Sigma}_{\eta|\alpha}$, and thus also in every equation from (15- 20) except $\boldsymbol{\mu}_{\eta|\alpha,t}$. The parameters in $\mathbf{R}_0$ mainly impact the return shock's conditional distribution (thus also in the parameters of $p_{\mathbf{y}}(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1})$), whereas $\mathbf{R}_1$ mainly impacts the conditional volatility shock distribution, and thus also the conditional return shock distribution through $\boldsymbol{\mu}_{\boldsymbol{\eta}|\boldsymbol{\alpha},\boldsymbol{t}}$ and $\boldsymbol{\Sigma}_{\eta|\alpha}$.

3 State-space estimation

This section explains how we estimate the latent state $\boldsymbol{\alpha}_t$ for the given state-space model in equations (18-20). We first derive necessary derivatives for the state-space model in Section 3.1 and Appendix D. Then using these derivatives, the BF illustrated in Section 3.2 can be used for state filtering. Section 3.3 explains the parameter estimation process and Section 3.4 shows forecast accuracy measures.

[4] $\delta_d(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$

3.1 Gradient vector and Fisher information matrix

In this section, the gradient and the Fisher information matrix (FIM) are derived for the log likelihood functions of the observation density given in equation (18) and state-density given in equations (20). For notational convenience we define $\mathbf{z}_t := (\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t})$, $f(\mathbf{y}_t) := \log p_{\mathbf{y}}(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1})$ and $g(\mathbf{x}_t) := \log p_{\mathbf{x}}(\mathbf{x}_t|\boldsymbol{\alpha}_{t-1}, \mathcal{F}_{t-1})$. The colon operator (:) indicates the Frobenius inner product, which is analogous to the vector inner product but for matrices. The operation is a component-wise inner product of two vectorized matrices.[5] Trace(A) is defined as the sum of the diagonal elements of a square matrix A.

Furthermore, we define the gradient vector at time t ($\boldsymbol{\nabla}_t$) as the first derivative of a log likelihood with respect to the state vector $\boldsymbol{\alpha}_t = (x_{a,t}, x_{b,t}, x_{a,t-1}, x_{b,t-1})'$. Derivations for the gradient vector of the observation and state densities are given in Appendix D. For the observation density the gradient is defined as

$$\boldsymbol{\nabla}_t^{\mathbf{y}} := \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_t} f(\mathbf{y}_t) = \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathbf{z}_t : \frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}}{\mathrm{d}\boldsymbol{\alpha}_t} + \frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathbf{z}_t\mathbf{z}_t'\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} - \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\right) : \frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}\boldsymbol{\alpha}_t}. \tag{21}$$

In the state-transition density the volatility process $\mathbf{x}_t$ and its mean $\boldsymbol{\mu}_{\mathbf{x},t}$ are dependent on $\boldsymbol{\alpha}_t$, whereas the variance is constant. The gradient vector for the log state-transition density is given by

$$\boldsymbol{\nabla}_t^{\mathbf{x}} := \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_t} g(\mathbf{x}_t) = \boldsymbol{\Sigma}_x^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x},t}) : \frac{\mathrm{d}(\boldsymbol{\mu}_{\mathbf{x},t} - \mathbf{x}_t)}{\mathrm{d}\boldsymbol{\alpha}_t}. \tag{22}$$

The FIM is defined as the variance of the observation and state density gradient vectors:[6]

$$\mathcal{I}_{t,m,n}^{\mathbf{y}} = \mathrm{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_{t,m}} f(\mathbf{y}_t)\right)\left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_{t,n}} f(\mathbf{y}_t)\right)' \mid \boldsymbol{\alpha}_t\right], \tag{23}$$

$$\mathcal{I}_{t,m,n}^{\boldsymbol{\alpha}} = \mathrm{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_{t,m}} g(\mathbf{x}_t)\right)\left(\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\alpha}_{t,n}} g(\mathbf{x}_t)\right)' \mid \boldsymbol{\alpha}_{t-1}\right]. \tag{24}$$

Malagò and Pistone (2015) derived the FIM for the multivariate normal density, which is, applied to the observation and state densities, given as

$$\mathcal{I}_{t,m,n}^{\mathbf{y}} = \frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}'}{\mathrm{d}\alpha_{t,m}}\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}}{\mathrm{d}\alpha_{t,n}} + \frac{1}{2}\operatorname{trace}\left(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}\alpha_{t,m}}\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}\alpha_{t,n}}\right), \tag{25}$$

$$\mathcal{I}_{m,n}^{\boldsymbol{\alpha}} = \frac{\mathrm{d}(\boldsymbol{\mu}_{\mathbf{x},t} - \mathbf{x}_t)'}{\mathrm{d}\boldsymbol{\alpha}_{t,m}}\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\frac{\mathrm{d}(\boldsymbol{\mu}_{\mathbf{x},t} - \mathbf{x}_t)}{\mathrm{d}\boldsymbol{\alpha}_{t,n}}, \tag{26}$$

for $m, n \in \{1, 2, 3, 4\}$. Derivations for $\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}, \mathrm{d}\boldsymbol{\mu}_{\mathbf{x},t}, \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}$ and $\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{x}}$ with respect to $\mathrm{d}\boldsymbol{\alpha}_t$ are

[5] For $n \times m$ matrices A and B, $A : B = \mathrm{vec}(A)'\,\mathrm{vec}(B) = \mathrm{vec}(A) \cdot \mathrm{vec}(B)$.

[6] Equivalent to the expected negative Hessian under some regularity conditions, see Lehmann et al. (2005).

given in Appendix D equations (D.17-D.28).

3.2 Bellman filter

We use the BF to estimate the latent state vector in the state-space model. This section gives a short description of Lange (2020)'s filtering method and introduces small modifications to reduce computational burden. The method assumes the latent state $\boldsymbol{\alpha}_t$ has a unique mode for every time t, which holds for our model as it assumes normally distributed observation and state dynamics.

The method evaluates the latent states based on mode estimation, where the mode is defined as the maximiser of the complete data log likelihood:

$$(\tilde{\boldsymbol{\alpha}}_{1|t}, \tilde{\boldsymbol{\alpha}}_{2|t}, \ldots, \tilde{\boldsymbol{\alpha}}_{t|t}) = \underset{\boldsymbol{\alpha}_{1|t}, \boldsymbol{\alpha}_{2|t}, \ldots, \boldsymbol{\alpha}_{t|t}}{\arg\max} \log p(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_t, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t). \tag{27}$$

To circumvent the ever-increasing number of optimization variables over time, Lange (2020) defines a value function $V_t(\boldsymbol{\alpha}_t)$ as the maximum of the complete data log likelihood with respect to all states prior to the current state $\boldsymbol{\alpha}_t$:

$$V_t(\boldsymbol{\alpha}_t) = \max_{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_{t-1}} \log p(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_t, \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_t), \tag{28}$$

which satisfies the recursive Bellman's equation (Bellman and Dreyfus, 1959):

$$V_t(\boldsymbol{\alpha}_t) = \log p_{\mathbf{y}}(\mathbf{y}_t \mid \boldsymbol{\alpha}_t) + \max_{\boldsymbol{\alpha}_{t-1}} \{\log p_{\mathbf{x}}(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{t-1}) + V_{t-1}(\boldsymbol{\alpha}_{t-1})\}. \tag{29}$$

The Bellman equation allows us to use dynamic programming principles to filter the latent states $\tilde{\boldsymbol{\alpha}}_{t|t}$ and $\tilde{\boldsymbol{\alpha}}_{t-1|t}$ instead of all modes $(\tilde{\boldsymbol{\alpha}}_{1|t}, \tilde{\boldsymbol{\alpha}}_{2|t}, \ldots, \tilde{\boldsymbol{\alpha}}_{t|t})$ in equation (27). The optimization problem for the filtered states becomes

$$\begin{bmatrix} \tilde{\boldsymbol{\alpha}}_{t|t} \\ \tilde{\boldsymbol{\alpha}}_{t-1|t} \end{bmatrix} := \underset{\begin{bmatrix} \boldsymbol{\alpha}_t \\ \boldsymbol{\alpha}_{t-1} \end{bmatrix}}{\arg\max} \{\log p_{\mathbf{y}}(\mathbf{y}_t \mid \boldsymbol{\alpha}_t) + \log p_{\mathbf{x}}(\boldsymbol{a}_t \mid \boldsymbol{\alpha}_{t-1}) + V_{t-1}(\boldsymbol{\alpha}_{t-1})\}. \tag{30}$$

Lange (2020) approximates $V_{t-1}(\boldsymbol{\alpha}_{t-1})$ by

$$V_{t-1}(\boldsymbol{\alpha}_{t-1}) = -\frac{1}{2}\left(\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_{t-1|t-1}\right)' \mathbf{I}_{t-1|t-1}\left(\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_{t-1|t-1}\right), \tag{31}$$

where $\mathbf{I}_{t-1|t-1}$ is the precision matrix at time $t-1$. Bellman's equation requires a maximisation over the state $\boldsymbol{\alpha}_t$ and the lagged state $\boldsymbol{\alpha}_{t-1}$. As both terms contain the lagged state $\mathbf{x}_{t-1}$, we

introduce $\boldsymbol{\zeta}_t$ as $\boldsymbol{\zeta}_t = (\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2})' = (\boldsymbol{\alpha}_t, \mathbf{x}_{t-2})'$ for notational convenience. Furthermore, the state-transition density is degenerate and requires modifications to drop the degenerate parts ($\mathbf{x}_{t-1}$ in $\boldsymbol{\alpha}_{t-1}$). These modifications are added in the gradient and FIM for $\boldsymbol{\zeta}_t$, which follow from Lange (2020):

$$\boldsymbol{\nabla}_t^{\boldsymbol{\zeta}} = \begin{bmatrix} \boldsymbol{\nabla}_t^{\boldsymbol{\alpha}} + \boldsymbol{\nabla}_t^{\mathbf{y}} \\ \vec{\mathbf{0}}_2 \end{bmatrix} - \begin{bmatrix} \vec{\mathbf{0}}_2 \\ \mathbf{I}_{t-1|t-1}\left(\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_{t-1|t-1}\right) \end{bmatrix}, \tag{32}$$

$$\mathcal{I}_t^{\boldsymbol{\zeta}} = \left[\begin{pmatrix} \mathcal{I}_t^{\mathbf{y}} + \mathcal{I}_t^{\boldsymbol{\alpha}} & \mathbf{0}_{4\times 2} \\ \mathbf{0}_{2\times 4} & \mathbf{0}_{2\times 2} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{2\times 2} & \mathbf{0}_{2\times 4} \\ \mathbf{0}_{4\times 2} & \mathbf{I}_{t-1|t-1} \end{pmatrix} \right]. \tag{33}$$

The state precision matrix $\mathbf{I}_{t|t}$, defined as the negative Hessian of the value function at its peak, can be obtained by taking the Schur complement[7] of the top-left $[4 \times 4]$ block of $\mathcal{I}_t^{\boldsymbol{\zeta}}$:

$$\text{Partition } \mathcal{I}_t^{\boldsymbol{\zeta}} \text{ as } \begin{bmatrix} \mathbf{A}_{[4\times 4]} & \mathbf{B}_{[4\times 2]} \\ \mathbf{C}_{[2\times 4]} & \mathbf{D}_{[2\times 2]} \end{bmatrix}, \text{ then } \mathbf{I}_{t|t} = \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}. \tag{34}$$

The predicted precision matrix, $\mathbf{I}_{t|t-1}$, can be computed similarly, but the $\mathcal{I}_t^{\mathbf{y}}$ term has to be omitted in equation (33) as this term is not yet available at time $t-1$.

We use a Fishers scoring algorithm with an added iteration-based learning rate parameter to optimize the value function of Bellman's equation (29) with respect to the state at every time step:

$$\boldsymbol{\zeta}_{t|t-1}^{(0)} = \begin{bmatrix} \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{x}_{t-1|t-1} + \boldsymbol{\Sigma}_{\eta}^{1/2}\boldsymbol{\mu}_{\eta|\alpha,t} \\ \boldsymbol{\alpha}_{t-1|t-1} \end{bmatrix} \tag{35}$$

$$\boldsymbol{\zeta}_t^{(i+1)} = \boldsymbol{\zeta}_t^{(i)} + \gamma^{(i)}(\mathcal{I}_t^{\boldsymbol{\zeta}})^{-1}\boldsymbol{\nabla}_t^{\boldsymbol{\zeta}}, \tag{36}$$

$$\gamma^{(i+1)} = \gamma^{(i)}/(1.025), \tag{37}$$

where $\mathcal{I}_t^{\boldsymbol{\zeta}}$ and $\boldsymbol{\nabla}_t^{\boldsymbol{\zeta}}$ denote the Fisher's information matrix and gradient of $\boldsymbol{\zeta}_t$ respectively, and $\gamma^{(0)} = 1$. The initial value for $\boldsymbol{\zeta}_{t|t-1}$ in equation (35) contains the state prediction $\mathbf{x}_{t|t-1}$ and the filtered state of the previous time step $\alpha_{t-1|t-1}$.

The decreasing iteration-based learning rate γ is used to achieve convergence in less iterations, to prevent oscillations in the updating process of $\boldsymbol{\zeta}_t$ and reduces the probability of getting stuck in local minima (Patterson and Gibson, 2017). The decreasing rate of 1.025 is chosen arbitrarily such that $\gamma^{(50)} \approx 0.30$.

The algorithm for the modified BF applied to the state-space model of equations (18-20)

[7] See Petersen et al. (2008) Section 9.1.5. The Schur complement is useful when solving for a partition of x in the linear system: $Ax = b$.

is illustrated in Algorithm 1. The state is initialized as a zero vector and the state precision matrix as an arbitrary multiple of the identity matrix. The unconditional expectation and variance of $\mathbf{x}_t$ are derived in Appendix B and can also be used for state and state precision matrix initialization.

3.3 Parameter estimation

Based on a combination of maximising the fit between $\boldsymbol{\alpha}_{t|t}$ and $\mathbf{y}_t$ and minimizing the Kullback-Leibler divergence (Kullback and Leibler, 1951) between $\boldsymbol{\alpha}_{t|t}$ and $\boldsymbol{\alpha}_{t|t-1}$, Lange (2020) provides an approximate likelihood function given by

$$\begin{aligned}\log p_{\mathbf{y}}\left(\mathbf{y}_t \mid \mathcal{F}_{t-1}\right) \approx & \log p_{\mathbf{y}}\left(\mathbf{y}_t \mid \boldsymbol{\alpha}_{t|t}\right)+\frac{1}{2} \log \det\left\{\mathbf{I}_{t|t-1}\right\}-\frac{1}{2} \log \det\left\{\mathbf{I}_{t|t}\right\} \\ & -\frac{1}{2}\left(\boldsymbol{\alpha}_{t|t}-\boldsymbol{\alpha}_{t|t-1}\right)^{\prime} \mathbf{I}_{t|t-1}\left(\boldsymbol{\alpha}_{t|t}-\boldsymbol{\alpha}_{t|t-1}\right).\end{aligned} \tag{38}$$

The control parameters can be estimated numerically by finding the optimal solution to

$$\widehat{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \sum_{t=1}^{T} \log p_{\mathbf{y}}\left(\mathbf{y}_t \mid \mathcal{F}_{t-1}\right), \tag{39}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Each term in equation (38) is indirectly dependent on parameters in $\boldsymbol{\theta}$. For the parameter optimization, we performed the following steps:

1. Perform a grid search for $\mathbf{R}_0, \mathbf{R}_1$: Optimize the model without cross-sectional dependencies for the standard SV parameters: $\{\tilde{\mathbf{y}}, \operatorname{diag}(\mathbf{B}), \boldsymbol{\mu}, \operatorname{diag}(\boldsymbol{\Phi})\}$ while plugging in grid values for $\mathbf{R}_0$ and $\mathbf{R}_1$. The grid consists of values between -0.95 and 0.95 in steps of 0.05.[8]

2. Take the optimized parameter values with the highest likelihood as initial values. Optimize the model with cross-sectional parameters:
   $\boldsymbol{\theta} = \{\tilde{\mathbf{y}}, \operatorname{diag}(\mathbf{B}), \boldsymbol{\mu}, \operatorname{vec}(\boldsymbol{\Phi}), \operatorname{diag}(\mathbf{R}_0), \operatorname{diag}(\mathbf{R}_1), \rho_\varepsilon, \rho_\eta\}$.

The optimization is done through the Quasi-Newton algorithm (Broyden, 1967). The grid search is used to find promising initial values for $\mathbf{R}_0$ and $\mathbf{R}_1$.

3.4 Out-of-sample forecast accuracy measures

For comparison of forecasts we use the average quasi likelihood (QLIKE) and the mean squared error (MSE) of the volatility forecasts, using the squared daily return of the forecasted

[8] Some grid values are skipped as they are not a subset of the parameter space in equation (11).

Algorithm 1 Modified Bellman filter

Consider the state-space model of equations (18) & (20):

$$\mathbf{y}_t \sim p_{\mathbf{y}}(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1}), \qquad \boldsymbol{\alpha}_t \sim p_{\mathbf{x}}(\mathbf{x}_t|\boldsymbol{\alpha}_{t-1}, \mathcal{F}_{t-1})$$

Initialize $\mathbf{x}_{0|0} = \vec{\mathbf{0}}_2, \boldsymbol{\alpha}_{0|0} = \vec{\mathbf{0}}_4, \boldsymbol{\zeta}_{0|0} = \vec{\mathbf{0}}_6, \mathcal{I}^{\boldsymbol{\alpha}}_{0|0} = \mathbf{1}_{4\times4} * 10, \mathcal{I}^{\boldsymbol{\zeta}}_{0|0} = \mathbf{1}_{6\times6} * 10.$

for $t = 1, ..., T$ **do**

$\mathbf{x}_{t|t-1} \leftarrow \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{x}_{t-1|t-1} + \boldsymbol{\Sigma}_{\eta}^{1/2}\boldsymbol{\mu}_{\eta|\alpha,t}$ ▷ Prediction Steps

$\boldsymbol{\zeta}_{t|t-1} \leftarrow \begin{bmatrix} \mathbf{x}_{t|t-1} \\ \boldsymbol{\alpha}_{t-1|t-1} \end{bmatrix}$, $\boldsymbol{\alpha}_{t|t-1} \leftarrow \begin{bmatrix} \mathbf{x}_{t|t-1} \\ \mathbf{x}_{t-1|t-1} \end{bmatrix}$

$\mathcal{I}^{\boldsymbol{\alpha}}_{t|t-1,m,n} \leftarrow \frac{\mathrm{d}(\boldsymbol{\mu}_{\mathbf{x},t} - \mathbf{x}_t)'}{\mathrm{d}\alpha_{t,m}}\boldsymbol{\Sigma}_x^{-1}\frac{\mathrm{d}(\boldsymbol{\mu}_{\mathbf{x},t} - \mathbf{x}_t)}{\mathrm{d}\alpha_{t,n}}\Big|_{\boldsymbol{\alpha}_t=\boldsymbol{\alpha}_{t|t-1}}, m, n \in \{1,2,3,4\}$ ▷ FIM equation (26)

$\mathcal{I}^{\boldsymbol{\zeta}}_{t|t-1} \leftarrow \left[\begin{pmatrix} \mathcal{I}^{\boldsymbol{\alpha}}_{t|t-1} & \mathbf{0}_{4\times2} \\ \mathbf{0}_{2\times4} & \mathbf{0}_{2\times2} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{2\times2} & \mathbf{0}_{2\times4} \\ \mathbf{0}_{4\times2} & \mathcal{I}^{\boldsymbol{\alpha}}_{t-1|t-1} \end{pmatrix}\right]$ ▷ FIM equation (33)

Partition $\mathcal{I}^{\boldsymbol{\zeta}}_{t|t-1}$ as $\begin{bmatrix} \mathbf{A}_{[4\times4]} & \mathbf{B}_{[4\times2]} \\ \mathbf{C}_{[2\times4]} & \mathbf{D}_{[2\times2]} \end{bmatrix}$,

$\mathbf{I}_{t|t-1} \leftarrow \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ ▷ State precision prediction step (34)

Set $\boldsymbol{\zeta}^{(0)}_{t|t} = \boldsymbol{\zeta}_{t|t-1}$, $\boldsymbol{\alpha}^{(0)}_{t|t} = \boldsymbol{\alpha}_{t|t-1}$, $i = 0$ and $\gamma^{(0)} = 1$.

while $i \leq i_{max}$ and $\max\{|\boldsymbol{\alpha}^{(i+1)}_{t|t} - \boldsymbol{\alpha}^{(i)}_{t|t}|\} \geq 10^{-3}$ **do** ▷ Stopping Criteria

$\boldsymbol{\alpha}_t \leftarrow \boldsymbol{\alpha}^{(i)}_{t|t}$. ▷ Evaluate $\boldsymbol{\alpha}_t$ at its last iterate.

for $m, n \in \{1,2,3,4\}$ **do**

$\mathcal{I}^{\mathbf{y}}_{t,m,n} \leftarrow \frac{\mathrm{d}\boldsymbol{\mu}'_{\mathbf{y},t}}{\mathrm{d}\alpha_{t,m}}\boldsymbol{\Sigma}^{-1}_{\mathbf{y},t}\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}}{\mathrm{d}\alpha_{t,n}} + \frac{1}{2}\mathrm{trace}\left(\boldsymbol{\Sigma}^{-1}_{\mathbf{y},t}\frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}\alpha_{t,m}}\boldsymbol{\Sigma}^{-1}_{y}\frac{\mathrm{d}\boldsymbol{\Sigma}_{\boldsymbol{y}}}{\mathrm{d}\alpha_{t,n}}\right)$ ▷ FIM equation (25)

$\mathcal{I}^{\boldsymbol{\alpha}}_{t,m,n} \leftarrow \frac{\mathrm{d}(\boldsymbol{\mu}_{\mathbf{x},t} - \mathbf{x}_t)'}{\mathrm{d}\alpha_{t,m}}\boldsymbol{\Sigma}_x^{-1}\frac{\mathrm{d}(\boldsymbol{\mu}_{\mathbf{x},t} - \mathbf{x}_t)}{\mathrm{d}\alpha_{t,n}}$ ▷ FIM equation (26)

end for

$\boldsymbol{\nabla}^{\boldsymbol{\zeta}}_t \leftarrow \begin{bmatrix} \boldsymbol{\nabla}^{\mathbf{y}}_t + \boldsymbol{\nabla}^{\boldsymbol{\alpha}}_t \\ \vec{\mathbf{0}}_2 \end{bmatrix} - \begin{bmatrix} \vec{\mathbf{0}}_2 \\ \mathcal{I}^{\boldsymbol{\alpha}}_{t-1|t-1}\left(\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_{t-1|t-1}\right) \end{bmatrix}$ ▷ Equations (32)

$\mathcal{I}^{\boldsymbol{\zeta}}_t \leftarrow \left[\begin{pmatrix} \mathcal{I}^{\mathbf{y}}_t + \mathcal{I}^{\boldsymbol{\alpha}}_t & \mathbf{0}_{4\times2} \\ \mathbf{0}_{2\times4} & \mathbf{0}_{2\times2} \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{2\times2} & \mathbf{0}_{2\times4} \\ \mathbf{0}_{4\times2} & \mathbf{I}_{t-1|t-1} \end{pmatrix}\right]$ ▷ FIM equation (33)

$\boldsymbol{\zeta}^{(i+1)}_{t|t} \leftarrow \boldsymbol{\zeta}^{(i)}_{t|t} + \gamma^{(i)}(\mathcal{I}^{\boldsymbol{\zeta}}_t)^{-1}\boldsymbol{\nabla}^{\boldsymbol{\zeta}}_t$ ▷ Fishers Scoring update

$\gamma^{(i+1)} \leftarrow \gamma^{(i)}/1.025$ ▷ Learning rate update

$\boldsymbol{\alpha}^{(i+1)}_{t|t} \leftarrow \left(\boldsymbol{\zeta}^{(i+1)}_{t|t}\right)_{1:4}$

$i \leftarrow i + 1$

end while

$\boldsymbol{\alpha}_{t|t} \leftarrow \boldsymbol{\alpha}^{(i+1)}_{t|t}$ ▷ Updating steps

$\mathbf{x}_{t|t} \leftarrow \left(\boldsymbol{\zeta}^{(i+1)}_{t|t}\right)_{1:2}$

Partition $\mathcal{I}^{\boldsymbol{\zeta}}_{t|t}$ as $\begin{bmatrix} \mathbf{A}_{[4\times4]} & \mathbf{B}_{[4\times2]} \\ \mathbf{C}_{[2\times4]} & \mathbf{D}_{[2\times2]} \end{bmatrix}$

$\mathbf{I}_{t|t} \leftarrow \mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ ▷ State precision updating step (34)

end for

date as the volatility proxy.

$$\text{QLIKE}_i = \frac{1}{T-t_0}\sum_{t=t_0}^{T} x_{i,t+1|t} + y^2_{i,t+1|t}\exp\big(-x_{i,t+1|t}\big) \tag{40}$$

$$\text{MSE}_i = \frac{1}{T-t_0}\sum_{t=t_0}^{T}\big(\exp\big(x_{i,t+1|t}\big) - y_{i,t+1|t}\big)^2, \tag{41}$$

for $i \in \{a, b\}$, where t_0 and T are the first and final out-of-sample dates and we used $\hat{\sigma}^2_{i,t+1|t} = \text{Var}(y_{i,t+1}|\mathcal{F}_t) = \exp\big(x_{i,t+1|t}\big)$. The QLIKE (Patton, 2011) is found to be more robust than the MSE for outliers, which occur frequently in our out-of-sample period (e.g. during The Great Recession). The squared daily return is known to be a noisy but viable volatility proxy. Realized variance based proxies are shown to be more robust proxies (Patton and Sheppard, 2009), but often require additional intraday data.

To test whether the volatility forecasts with cross-sectional effects are significantly different from the cross-sectional independent predictions, we use the Diebold-Mariano (DM) test statistic (Diebold and Mariano, 1991) given by

$$\text{DM} = \frac{\bar{d}}{\sqrt{\text{Var}(d)/T}} \sim \mathcal{N}(0,1), \tag{42}$$

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T}(e^2_{1,t} - e^2_{2,t}), \quad \text{Var}(d) = (\gamma^d_0 + 2\sum_{k=1}^{h}\gamma^d_j) \tag{43}$$

where γ_j denotes the j-th autocovariance of $\bar{d}$, $e^2_{1,t}$ and $e^2_{2,t}$ denote the variance prediction errors using the squared return volatility proxy, and we choose forecast horizon h as 1.

4 Results

In this section, we present the in-sample estimation and out-of-sample volatility forecasting results for the bivariate model. The bivariate model is applied on the daily log returns of Standard & Poor's 500 (SP500) index paired with three other indices: FTSE 100 (FTSE), Nikkei 225 (N225) and the NASDAQ 100 (NASD). Log returns are collected from Yahoo finance from 3 Jan 1990 to 31 Dec 2019[9], and the intersection of the dates is taken in each pair to synchronize the time series.[10] The in-sample period runs from 3 Jan 1990 to 30 June 2005. The out-of sample period runs from 1 July 2005 to 31 Dec 2019. Lastly, we investigate the

[9]Data can be obtained through: https://finance.yahoo.com/quote/%5EGSPC/history/.

[10](SP500,FTSE): 7471 observations, (SP500, N225): 7130 observations, (SP500, NASD): 7559 observations.

model-implied skewness for the SP500 when taking into account the different cross-sectional relations between the SP500 and the respective three other datasets.

We consider four parameter settings:

- Setting 1: Unrestricted, estimate all 16 parameters;
- Setting 2: Restrict $\rho_\eta = 0$, estimate 15 parameters;
- Setting 3: Restrict $\phi_{ab} = \phi_{ba} = \rho_\varepsilon = 0$, estimate 13 parameters;
- Setting 4: Cross-sectional independence, restrict $\phi_{ab} = \phi_{ba} = \rho_\varepsilon = \rho_\eta = 0$, estimate 12 parameters.

The cross-sectional dependence is controlled by ϕ_{ab}, ϕ_{ba} and ρ_ε in setting 2, whereas setting 3 only allows cross-sectional dependence through ρ_η. Other possible parameter settings where $\phi_{ab} = \phi_{ba} = \rho_\eta = 0$ and $\rho_\varepsilon = \rho_\eta = 0$ are left out as these were dominated by either Setting 2 or 3.[11] The setting where we only restrict $\phi_{ab} = \phi_{ba} = 0$ is also omitted as its results closely resemble Setting 1. Setting 4 is equivalent to stacking two independent univariate SV-Cs in equation (1-3), and is used as a benchmark to evaluate the relative performance of the cross-sectional dependent other three settings.

4.1 Simulation results

In our simulation study, we compare the parameter estimates of the true parameters used for the simulation with those obtained through Bellman filtering and parameter optimization. 30 bivariate samples of length 7500 are simulated using the true parameter sets, where we use the first half for parameter estimation and the second half for comparing volatility predictions. The length of 7500 represents the length of our empirical data sets. Four sets of realistic parameter values are used where the sets follow the parameter restrictions as in Setting 1 to 4 as described above. Table 1 reports the simulation estimates for Setting 1 and Setting 2, Table 2 reports simulation estimates for Setting 3 and Setting 4.

[11] Dominated in an arbitrary in-sample preliminary analysis. Out-of-sample performance might contradict this finding.

Table 1: Setting 1 and 2 simulation based average parameter estimates.

	True values 1		Estimates 1		True values 2		Estimates 2	
	a	b	a	b	a	b	a	b
$\tilde{y}_i$	0.015	0.015	0.015	0.014	0.015	0.015	0.015	0.014
			(0.002)	(0.003)			(0.001)	(0.001)
μ_i	0.000	0.000	-0.001	0.005	0.000	0.000	0.001	0.002
			(0.003)	(0.004)			(0.003)	(0.003)
ϕ_i	0.980	0.970	0.981	0.968	0.980	0.980	0.984	0.977
			(0.002)	(0.003)			(0.002)	(0.002)
ϕ_{ij}	0.010	0.000	0.013	0.001	0.000	0.000	0.003	-0.002
			(0.002)	(0.002)			(0.002)	(0.002)
$\sigma_{\eta i}$	0.200	0.200	0.214	0.190	0.200	0.200	0.187	0.211
			(0.031)	(0.029)			(0.028)	(0.033)
ρ_{0i}	-0.300	-0.300	-0.306	-0.279	-0.300	-0.300	-0.326	-0.279
			(0.057)	(0.064)			(0.049)	(0.055)
ρ_{1i}	-0.300	-0.300	-0.269	-0.285	-0.300	-0.300	-0.276	-0.279
			(0.055)	(0.069)			(0.056)	(0.048)
ρ_η	0.500		0.364		0		0	
			(0.097)					
ρ_ε	0.500		0.476		0.500		0.473	
			(0.025)				(0.022)	
MAE	0.380	0.376	0.379	0.377	0.376	0.377	0.385	0.384
MSE	0.479	0.473	0.480	0.479	0.474	0.475	0.486	0.485
DM			0.086	0.012			1.152	0.895

Notes: We simulate 30 bivariate samples of length 7500 and, based on the first half, compute parameter estimates and numerical standard deviations (in parentheses) by taking the square root of the diagonal elements of the inverse Hessian, where the Hessian is computed through the Quasi-Newton algorithm (Broyden, 1967). Mean absolute errors (MAE) and Mean squared errors (MSE) are computed using the predicted state $x_{i,t|t-1}$ and true simulated state x_t for the second half ($t > 3750$). DM statistics compare the volatility forecast errors obtained through the difference of the volatility forecasts obtained with the true and estimated parameters, with the squared return proxy $y^2_{t|t-1}$.

Table 2: Setting 3 and 4 simulation based average parameter estimates.

	True values 3		Estimates 3		True values 4		Estimates 4	
	a	b	a	b	a	b	a	b
$\tilde{y}_i$	0.015	0.015	0.017	0.015	0.015	0.015	0.014	0.015
			(0.002)	(0.002)			(0.002)	(0.001)
μ_i	0.000	0.000	-0.004	-0.003	0.000	0.000	0.001	0.002
			(0.004)	(0.004)			(0.003)	(0.003)
ϕ_i	0.980	0.980	0.983	0.976	0.980	0.980	0.980	0.981
			(0.002)	(0.003)			(0.002)	(0.002)
ϕ_{ij}	0	0	0	0	0	0	0	0
$\sigma_{\eta i}$	0.200	0.200	0.195	0.209	0.200	0.200	0.187	0.211
			(0.029)	(0.028)			(0.026)	(0.03)
ρ_{0i}	-0.300	-0.300	-0.274	-0.267	-0.300	-0.300	-0.284	-0.301
			(0.068)	(0.066)			(0.039)	(0.046)
ρ_{1i}	-0.300	-0.300	-0.280	-0.278	-0.300	-0.300	-0.288	-0.274
			(0.049)	(0.053)			(0.045)	(0.042)
ρ_η	0.500		0.401		0		0	
			(0.071)					
ρ_ε	0		0		0		0	
MAE	0.384	0.383	0.383	0.385	0.382	0.379	0.382	0.377
MSE	0.476	0.487	0.485	0.487	0.482	0.480	0.484	0.482
DM			0.029	0.051			0.075	0.031

Notes: We simulate 30 bivariate samples of length 7500 and, based on the first half, compute parameter estimates and numerical standard deviations (in parentheses) by taking the square root of the diagonal elements of the inverse Hessian, where the Hessian is computed through the Quasi-Newton algorithm (Broyden, 1967). Mean absolute errors (MAE) and Mean squared errors (MSE) are computed using the predicted state $x_{i,t|t-1}$ and true simulated state x_t for the second half ($t > 3750$). DM statistics compare the volatility forecast errors obtained through the difference of the volatility forecasts obtained with the true and estimated parameters, with the squared return proxy $y^2_{t|t-1}$.

In line with Lange (2020) and Catania (2021), the method does not find difficulty in estimating the standard SV parameters $\tilde{\mathbf{y}}$, $\boldsymbol{\mu}$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}_\eta$. Estimates of $\mathbf{R}_0, \mathbf{R}_1, \rho_\varepsilon$ are slightly biased (mainly about 0.02 closer to 0), with $\mathbf{R}_0$ and $\mathbf{R}_1$ parameters relatively more unstable than the other

parameters with average numerical standard errors ranging between 0.046 and 0.069. ρ_η estimates are the most unstable with numerical standard deviation of 0.097 for Setting 1 and 0.071 for Setting 2. The biases indicate the estimation process is not flawless or we require a larger simulated sample size. The latter we did not perform due to time constraints, but could be an important factor due to the non-parsimonious volatility equation. Setting 4 estimates are slightly more accurate for the leverage parameters $\mathbf{R}_0$ and $\mathbf{R}_1$ but still find relatively larger numerical standard errors than the standard SV parameters. The latter result is in line with Lange (2020) and Catania (2021).

Mean absolute errors (MAE) and mean squared errors (MSE) are computed using the BF predicted state $\mathbf{x}_{t|t-1}$ and actual (simulated) state $\mathbf{x}_t$ on the second half of the simulated data set. We find minor differences between Setting 3 and 4, and slightly lower errors for Setting 1 and 2 than Setting 4. Overall, we find that the estimated parameters obtain MAE and MSE values close to the ones obtained with the true parameters. The null hypothesis of equal predictive ability between the volatility predictions obtained with the true parameters and the estimated parameters and the squared return proxy is tested using the DM test. No significant differences (on a 5% level) are found between the volatility forecasts of the true parameter and estimated parameters when using the squared returns as volatility proxy in all four parameter settings. Hence the biases do not significantly influence the out-of-sample volatility predictions.

4.2 In-sample results

Tables 3- 5 report the parameter estimates of the 4 parameter settings for the corresponding data sets with numerical standard errors shown in parentheses. The estimates for the SP500 in Setting 4 differ slightly in each pair due to the synchronization of the daily returns causing different missing values in the respective other datasets to be omitted. As the simulation studies has indicated, the parameter estimates of $\mathbf{R}_0$, $\mathbf{R}_1$, ρ_η and $\rho_{v\varepsilon}$ might contain a bias.

Table 3: Parameter Estimates S&P500, FTSE.

	Setting 1		Setting 2		Setting 3		Setting 4	
	a	b	a	b	a	b	a	b
$\tilde{y}_i$	0.059	0.041	0.063	0.043	0.062	0.039	0.044	0.018
	(0.013)	(0.014)	(0.013)	(0.014)	(0.013)	(0.01)	(0.013)	(0.014)
μ_i	-0.007	-0.007	-0.010	-0.005	-0.011	-0.009	-0.007	-0.004
	(0.003)	(0.002)	(0.003)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)
ϕ_i	0.984	0.987	0.971	0.986	0.983	0.985	0.980	0.988
	(0.003)	(0.003)	(0.007)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)
ϕ_{ij}	0.000	0.000	0.021	0.021	0	0	0	0
	(0.003)	(0.003)	(0.006)	(0.028)				
$\sigma_{\eta i}$	0.256	0.185	0.254	0.185	0.246	0.187	0.245	0.177
	(0.026)	(0.025)	(0.074)	(0.032)	(0.028)	(0.022)	(0.032)	(0.024)
ρ_{0i}	-0.298	-0.274	-0.352	-0.312	-0.377	-0.308	-0.382	-0.342
	(0.048)	(0.046)	(0.044)	(0.051)	(0.018)	(0.042)	(0.059)	(0.056)
ρ_{1i}	-0.213	-0.207	-0.286	-0.254	-0.320	-0.321	-0.409	-0.439
	(0.015)	(0.021)	(0.038)	(0.038)	(0.038)	(0.023)	(0.074)	(0.075)
ρ_η	0.818		0		0.537		0	
	(0.077)				(0.019)			
ρ_ε	0.399		0.387		0		0	
	(0.014)		(0.014)					
LogL	-0.948		-0.948		-0.978		-0.999	
AIC	1.899		1.900		1.958		2.000	
BIC	1.908		1.909		1.967		2.008	

Notes: This table shows the parameter estimates for the bivariate model of the pair: (a, b) = (S&P500, FTSE 100). The date-intersection of the daily log returns (multiplied by 100) of the pair are taken from 3 Jan 1990 to 30 June 2005. i denotes the corresponding column a or b and j denotes the corresponding other column, $i \neq j$. Numerical standard errors are shown in parentheses. Reported values for LogL, AIC and BIC are $\times 10^{-4}$.

Table 4: Parameter Estimates S&P500, Nikkei 225.

	Setting 1		Setting 2		Setting 3		Setting 4	
	a	b	a	b	a	b	a	b
$\tilde{y}_i$	0.049	-0.050	0.050	-0.040	0.045	-0.056	0.045	-0.057
	(0.013)	(0.023)	(0.013)	(0.024)	(0.014)	(0.023)	(0.014)	(0.022)
μ_i	-0.014	0.021	-0.013	0.019	-0.008	0.016	-0.008	0.016
	(0.004)	(0.005)	(0.004)	(0.005)	(0.003)	(0.004)	(0.003)	(0.004)
ϕ_i	0.980	0.967	0.981	0.968	0.983	0.971	0.983	0.972
	(0.004)	(0.006)	(0.004)	(0.006)	(0.004)	(0.005)	(0.004)	(0.005)
ϕ_{ij}	0.008	0.008	0.006	0.006	0	0	0	0
	(0.003)	(0.003)	(0.003)	(0.003)				
$\sigma_{\eta i}$	0.255	0.276	0.250	0.281	0.247	0.293	0.245	0.295
	(0.040)	(0.053)	(0.039)	(0.051)	(0.039)	(0.051)	(0.039)	(0.052)
ρ_{0i}	-0.322	-0.242	-0.326	-0.279	-0.380	-0.309	-0.382	-0.308
	(0.049)	(0.055)	(0.051)	(0.065)	(0.056)	(0.048)	(0.054)	(0.047)
ρ_{1i}	-0.441	-0.563	-0.439	-0.585	-0.459	-0.634	-0.409	-0.641
	(0.063)	(0.069)	(0.072)	(0.08)	(0.072)	(0.058)	(0.067)	(0.055)
ρ_η	-0.091		0		0.0264		0	
	(0.067)				(0.029)			
ρ_ε	0.101		0.091		0		0	
	(0.021)		(0.019)					
LogL	-1.095		-1.095		-1.097		-1.112	
AIC	2.223		2.222		2.225		2.225	
BIC	2.232		2.232		2.234		2.233	

Notes: This table shows the parameter estimates for the bivariate model of the pair: (a, b) = (S&P500, Nikkei 225). The date-intersection of the daily log returns (multiplied by 100) of the pair are taken from 3 Jan 1990 to 30 June 2005. i denotes the corresponding column a or b and j denotes the corresponding other column, $i \neq j$. Numerical standard errors are shown in parentheses. Reported values for LogL, AIC and BIC are $\times 10^{-4}$.

Table 5: Parameter Estimates S&P500, Nasdaq 100.

	Setting 1		Setting 2		Setting 3		Setting 4	
	a	b	a	b	a	b	a	b
$\tilde{y}_i$	0.056	0.097	0.059	0.103	0.066	0.112	0.043	0.082
	(0.013)	(0.023)	(0.013)	(0.023)	(0.012)	(0.022)	(0.013)	(0.024)
μ_i	-0.004	0.005	-0.016	0.012	-0.012	0.007	-0.008	0.003
	(0.003)	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
ϕ_i	0.990	0.993	0.983	0.989	0.983	0.987	0.986	0.993
	(0.002)	(0.002)	(0.004)	(0.003)	(0.004)	(0.003)	(0.003)	(0.005)
ϕ_{ij}	-0.001	-0.001	0.006	0.005	0	0	0	0
	(0.002)	(0.002)	(0.003)	(0.003)				
$\sigma_{\eta i}$	0.241	0.308	0.247	0.305	0.238	0.319	0.245	0.320
	(0.023)	(0.019)	(0.015)	(0.011)	(0.045)	(0.039)	(0.033)	(0.019)
ρ_{0i}	-0.248	-0.409	-0.289	-0.473	-0.275	-0.367	-0.377	-0.410
	(0.032)	(0.037)	(0.029)	(0.032)	(0.040)	(0.039)	(0.057)	(0.067)
ρ_{1i}	-0.207	-0.205	-0.210	-0.206	-0.247	-0.274	-0.409	-0.404
	(0.01)	(0.011)	(0.012)	(0.012)	(0.022)	(0.042)	(0.068)	(0.104)
ρ_η	0.647		0		0.733		0	
	(0.068)				(0.043)			
ρ_ε	0.817		0.816		0		0	
	(0.026)		(0.023)					
LogL	-1.012		-1.015		-1.207		-1.212	
AIC	2.028		2.132		2.317		2.426	
BIC	2.038		2.141		2.325		2.433	

Notes: This table shows the parameter estimates for the bivariate model of the pair: (a, b) = (S&P500, NASDAQ100). The date-intersection of the daily log returns (multiplied by 100) of the pair are taken from 3 Jan 1990 to 30 June 2005. i denotes the corresponding column a or b and j denotes the corresponding other column, $i \neq j$. Numerical standard errors are shown in parentheses. Reported values for LogL, AIC and BIC are $\times 10^{-4}$.

Overall, we find that pairing two US market indices (SP500 and NASD) finds great value in the additional cross-sectional dependencies in the in-sample fit. The UK market finds considerable value when pairing with the SP500, whereas the Japanese market (through N225) finds only minor improvement from the dependencies with SP500. As expected, ρ_ε has the largest effect on the (SP500, NASD) pair, as these indices are known to be highly

correlated due to the way the indices are constructed. Particularly, we observe that allowing ρ_ε to deviate from 0 improves the model fit considerably when the value is significant for the data pair.

The introduction of possible Granger causality through ϕ_{ab} and ϕ_{ba} only finds significant effects for Setting 2 (SP500, FTSE), but overall we find mostly insignificant effects of ϕ_{ab} and ϕ_{ba} for the data pairs. This is line with the (SP500,NASD) results of Chen et al. (2009). Moreover, the μ_i and σ_η estimates indicate that NASD and N225 have a higher average and more unstable volatility process than the SP500. When ρ_η and ρ_ε are large and significant, the estimates for $\mathbf{R}_0$ and $\mathbf{R}_1$ tend to be less significant, but a direct relation between the parameters among all data pairs remains unclear. For all three data pairs we find significant negative values for $\mathbf{R}_0$ and $\mathbf{R}_1$. This indicates that contemporaneous and intertemporal correlations between the return and volatility shocks both substantiate the leverage effect.

The Setting 1 and Setting 2 parameter estimates for ρ_ε and the empirical in-sample correlation for the data pairs is shown in Table 6. Overall, the ρ_ε estimates for (SP500, NASD) are close to the empirical correlation, whereas (SP500, FTSE) and (SP500, N225) finds sizeable differences. A possible explanation for the differences could be the larger estimates of $\mathbf{R}_0$ and $\mathbf{R}_1$ in (SP500, FTSE) and (SP500, N225) capturing some of the effects of ρ_ε.

Table 6: Return correlation ρ_ε comparison.

	Setting 1	Setting 2	$\hat{\rho}_\varepsilon$
SP500, FTSE	0.399	0.387	0.445
	(0.014)	(0.014)	
SP500, N225	0.101	0.091	0.152
	(0.016)	(0.014)	
SP500, NASD	0.817	0.816	0.830
	(0.006)	(0.006)	

Notes: This table shows the in-sample period estimates for the returns correlation parameter ρ_ε and the empirical correlation estimate $\hat{\rho}_\varepsilon$. The in-sample period runs from 3 Jan 1990 to 30 June 2005. The date-intersection of the daily log returns (multiplied by 100) is taken of the (S&P500, FTSE 100), (S&P500, Nikkei 225) and (S&P500, NASDAQ 100) pairs.

4.3 Out-of-sample performance

Does allowing for cross-sectional dependencies improve the quality of volatility predictions? To answer this question, we conduct an out-of-sample analysis on one-period variance predictions using the measures shown in Section 3.4. Table 7 reports the Setting 1 to Setting 3 average QLIKE and MSE measures relative to the Setting 4 predictions. Values above one indicate the cross-sectional independent Setting 4 outperformed the corresponding setting and vice versa. Asterisks indicate that the corresponding variance predictions are significantly better than the Setting 4 variance predictions according to a DM test on a 5% significance level.

Table 7: Relative quasi likelihood and squared error average losses.

	QLIKE			MSE		
Setting:	1	2	3	1	2	3
SP500	0.989	0.985	1.002	0.987	0.975	1.001
FTSE	0.999	0.987	1.003	0.996	0.987	1.009
SP500	0.998	0.988	1.001	1.001	1.004	1.002
N225	1.010	1.002	1.001	1.003	1.004	1.002
SP500	0.963*	0.941*	0.999	0.961*	0.949*	1.000
NASD	0.970	0.953*	1.000	0.975	0.937*	1.001

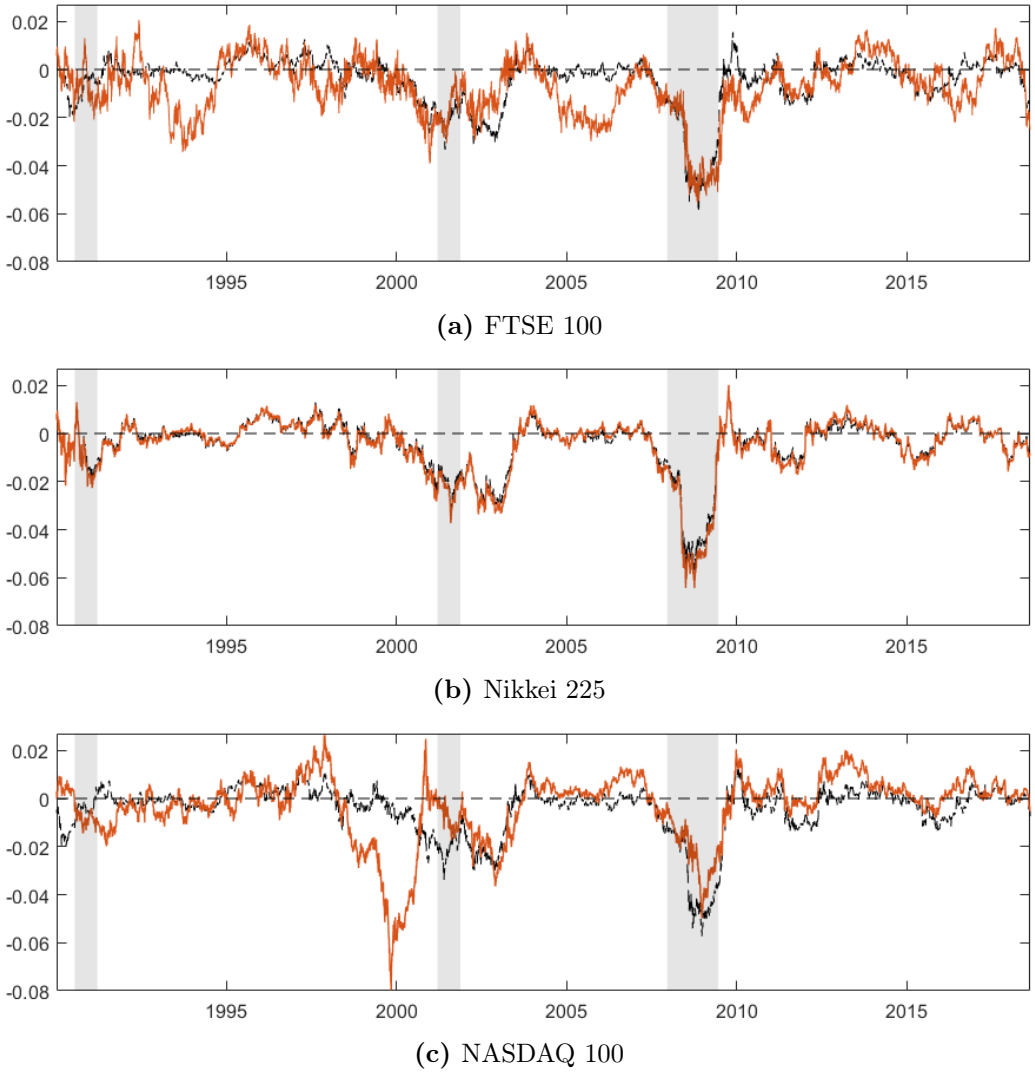
Notes: One period prediction are compared with the squared return variance proxy for the 3 data pairs: (S&P500, FTSE 100), (S&P500, Nikkei 225), (S&P500, NASDAQ100). Reported values are relative to Setting 4. The date-intersection of the daily log returns (multiplied by 100) of the pair are taken from 1 July 2005 to 31 Dec. 2019. Values lower than 1 indicate the corresponding setting outperforms Setting 4 and vice versa. Asterisks denote significance at 5% level, where significance is based on the null hypothesis of equal volatility predictions between the corresponding settings and Setting 4 computed with the DM test.

Setting 1 and Setting 2 predictions find significant improvements for the (SP500, NASD) pair except for the NASD predictions in Setting 1. Moreover, for these 2 pairs Setting 2 performs slightly better than Setting 1 in all cases. Setting 3 predictions seem inauspicious and are similar to the Setting 4 predictions. This insinuates the ρ_η estimate does not improve the out-of-sample fit, and restricting it to 0 provides volatility predictions of similar quality.

The bivariate model looks promising when ρ_ε is unrestricted and large for the data pair, suggesting that the ρ_ε estimates fit the out-of-sample data better than restricting it to 0 for highly correlated data pairs. At first sight, one would expect the return correlation parameter to have no impact on the volatility predictions. However, the term $\boldsymbol{\mu}_{\eta|\alpha,t}$ in the volatility prediction step (35) is dependent on ρ_ε as can be seen in equation (16). The parameter also

impacts the volatility prediction through $\mathbf{\Sigma}_{\eta|\alpha}$ in equation (16). Besides the direct effects, the improvement in volatility prediction could be indirectly caused by the differences in other parameter estimates for the different settings.

4.4 S&P500 model-implied skew effect



(a) FTSE 100

(b) Nikkei 225

(c) NASDAQ 100

Figure 1: model-implied skewness moving average of the S&P500 using cross-sectional dependencies. A moving average window of 126 (half a trading year) is used. The dashed black line indicates the no cross-sectional dependence case; the solid orange line uses (a) FTSE 100, (b) Nikkei 225, (c) NASDAQ 100 dependencies. Daily log returns are used from 3 Jan 1990 to 31 Dec 2019 using the unrestricted Setting 1 parameters in Tables (3-5). Shaded areas indicate US recession periods.

Figure 1 compares the model-implied skewness moving averages of the cross-sectional independent model (Setting 4) and the unrestricted model (Setting 1) for the SP500. A

moving window of 126 trading days is used to visualize the term over time and shaded areas indicate National Bureau of Economic Research (NBER) US Recession periods.[12] For Setting 2 and 3, the model-implied skews were mostly in between those of Setting 1 and 4 and are excluded for clarity.

Particularly during recession periods we find large drops at the start that revert back to 0, which is in line with the skewness results of Hou and Li (2020). During the Dot com crisis (1997-2001), the tech focused NASD suffered a substantial drawdown of 78%[13]. The crisis for the NASD appears to have a large negative impact for the SP500 during that period. FTSE shows similar patterns of lower magnitudes, and N225 finds only minor differences due to its relatively low cross-sectional dependence parameter estimates. Overall, we find a more pronounced negative than positive model-implied skew over time, further substantiating the presence of the leverage effect in the SP500.

5 Conclusion

In this paper, we studied a generalized bivariate SV model that allows for contemporaneous, intertemporal and cross-sectional correlations between returns and volatility shocks. The model encompasses the SV model of Catania (2021) and adds cross-sectional dependencies through four sources: a correlation parameter controlling the unconditional correlation between the return series, a correlation parameter controlling the instantaneous transmission of volatility shocks, and two parameters controlling the transmission of one's lagged volatility to the other current volatility process.

With my research, I tried to answer the question: "How effective are the cross-sectional correlations as an addition to the contemporaneous and intertemporal specification?". Our research has yielded several important findings that answer this question. First and foremost, we showed that allowing for a non-zero cross-sectional correlation between returns can be a promising improvement to the model fit and the quality of volatility predictions, given that the correlation is large enough. This was the case for the US market index pair (S&P500, NASDAQ100). The effect of lagged transmission between volatility shocks is shown to be insignificant for the data pairs, whereas the instantaneous transmission is positive and significant for (S&P500, FTSE100) and (S&P500, NASDAQ100). The latter does not find improvements in volatility predictions.

The BF of Lange (2020) provide us a method for state estimation and parameter opti-

[12] National Bureau of Economic Research (NBER), Business Cycle Expansions and Contractions, https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions.

[13] https://www.nasdaq.com/articles/3-lessons-investors-tech-bubble-2015-02-11

mization of our multivariate non-linear degenerate state-space model. The estimation method is relatively new and to our knowledge has no existing literature with an implementation on multivariate models yet. The simulation studies have shown that our implementation find slightly biased estimates for $\rho_\varepsilon, \mathbf{R}_0$ and $\mathbf{R}_1$, and we find relatively unstable instantaneous volatility transmission ρ_η estimates. However, the simulation studies also find no significant differences in volatility predictions between the actual parameters and estimated parameters.

Lastly, we examined the model-implied skewness of the S&P 500 returns. The findings showed a more pronounced negative than positive model-implied skew over time, particularly during recession periods. In line with existing literature, the leverage effect through the contemporaneous and intertemporal correlations remain present. Furthermore, the additional cross-sectional correlations impact the model-implied mean and variance of both the return and volatility processes, and thus affects the model-implied skewness for the S&P500 returns during recession periods.

We believe this new general specification is useful to understand how a pair of return and volatility shocks relate to each other with the addition of being able to differentiate contemporaneous and intertemporal correlations. It provides a framework that is well suited for model comparison and hypothesis testing as the model encompasses several other stochastic volatility models. The model is recommended when working with highly correlated data pairs, especially allowing for a return correlation ρ_ε could improve the model quality. The instantaneous volatility correlation parameter ρ_η also improves the model fit but appears unstable in the simulation study and inauspicious in the out-of-sample performance, thus the parameter might need additional research.

The difficult estimation process could be considered a major limitation of the model, as it mainly relies on correct manual derivations which make it less user friendly. Lange (2020) stated that black-box numerical optimisers can also be used in the optimization step in exchange for increased computational time. To our knowledge, existing literature does not provide a better state estimator for multidimensional nonlinear state-space models with degenerate state dynamics.

Due to the large number of parameters, the bivariate model does require a large sample size for estimation purposes and parameter restrictions are recommended to increase model simplicity and parsimony. On the contrary, we only considered 2 different temporal lags, which might lead to sub optimal performance of the models, as Catania (2021) empirically illustrated that for the same indices the optimal number of different temporal lags often range between two and five. Moreover, its unclear how the model-implied skewness as introduced in this paper exactly relates to the leverage effect of financial returns. Asai and McAleer (2009) emphasised on the assumption that there should be no connection in the leverage effects

between different returns, whereas our model inherently does have this connection.

The model could be extended in several ways. Different distributional assumptions can be specified and negative temporal lags can be added to the correlation specification. Applying the model framework on options data could give a better view of the skew effects of the model. Furthermore, as the model encompasses several existing SV models, it might be interesting to compare the bivariate model framework with other bivariate SV models and their leverage effects and perform robustness checks.

References

Asai, M. and McAleer, M. (2009). Multivariate stochastic volatility, leverage and news impact surfaces. *The Econometrics Journal*, 12(2):292–309.

Asai, M., McAleer, M., and Yu, J. (2006). Multivariate stochastic volatility: a review. *Econometric Reviews*, 25(2-3):145–175.

Bellman, R. and Dreyfus, S. (1959). Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, pages 247–251.

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

Broto, C. and Ruiz, E. (2004). Estimation methods for stochastic volatility models: A survey. *Journal of Economic Surveys*, 18(5):613–649.

Broyden, C. G. (1967). Quasi-newton methods and their application to function minimisation. *Mathematics of Computation*, 21(99):368–381.

Catania, L. (2020). The leverage effect and propagation. *Available at SSRN 3578656*.

Catania, L. (2021). A stochastic volatility model with a general leverage specification. *Journal of Business & Economic Statistics*, 0(0):1–23, Advance online publication.

Chan, D., Kohn, R., and Kirby, C. (2006). Multivariate stochastic volatility models with correlated errors. *Econometric Reviews*, 25(2-3):245–274.

Chen, A.-S., Cheng, L.-Y., and Cheng, K.-F. (2009). Intrinsic bubbles and granger causality: Evidence from long-term data. *Journal of Banking & Finance*, 33(12):2275–2281.

Diebold, F. X. and Mariano, R. S. (1991). Comparing predictive accuracy: An asymptotic test. Technical report, Federal Reserve Bank of Minneapolis.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264.

Harvey, A. C. and Shephard, N. (1996). Estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business & Economic Statistics*, 14(4):429–434.

Hou, Y. G. and Li, S. (2020). Volatility and skewness spillover between stock index and stock index futures markets during a crash period. *International Review of Economics & Finance*, 66:166–188.

Jacquier, E., Polson, N. G., and Rossi, P. E. (2002). Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics*, 20(1):69–87.

Jacquier, E., Polson, N. G., and Rossi, P. E. (2004). Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *Journal of Econometrics*, 122(1):185–212.

Koopman, S. J., Lucas, A., and Scharth, M. (2015). Numerically accelerated importance sampling for nonlinear non-gaussian state-space models. *Journal of Business & Economic Statistics*, 33(1):114–127.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lange, R.-J. (2020). Bellman filtering for state-space models. *arXiv preprint arXiv:2008.11477*.

Lehmann, E. L., Romano, J. P., and Casella, G. (2005). *Testing statistical hypotheses*, volume 3. Springer.

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.

Malagò, L. and Pistone, G. (2015). Information geometry of the gaussian distribution in view of stochastic optimization. In *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms XIII*, pages 150–162.

Malik, S. and Pitt, M. K. (2011). Particle filters for continuous likelihood evaluation and maximisation. *Journal of Econometrics*, 165(2):190–209.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2):425–449.

Page Jr, T. J. (1984). Multivariate statistics: A vector space approach. *Journal of Marketing Research*, 21(000002):236.

Patterson, J. and Gibson, A. (2017). *Deep learning: A practitioner's approach*. O'Reilly Media, Inc.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1):246–256.

Patton, A. J. and Sheppard, K. (2009). Evaluating volatility and correlation forecasts. In *Handbook of Financial Time Series*, pages 801–838. Springer.

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark*, 7(15):510.

Richard, J.-F. and Zhang, W. (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics*, 141(2):1385–1411.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes-a study of the daily sugar prices 1961-75. *Time series analysis: Theory and practice*, 1:203–226.

Taylor, S. J. (1994). Modeling stochastic volatility: A review and comparative study. *Mathematical Finance*, 4(2):183–204.

Von Hippel, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of Statistics Education*, 13(2).

Yu, J. (2005). On leverage in a stochastic volatility model. *Journal of Econometrics*, 127(2):165–178.

Appendices

A Unconditional variance of $\boldsymbol{\eta}_t$

The unconditional variance matrix of $\boldsymbol{\eta}_t$ can be derived from equation (6):

$$\boldsymbol{\eta}_t = \sum_{j=0}^{k} \mathbf{R}_j \boldsymbol{\varepsilon}_{t-j} + \boldsymbol{\Sigma}_\xi^{1/2} \boldsymbol{\xi}_t, \quad \boldsymbol{\xi}_t \overset{iid}{\sim} \mathcal{N}_2\left(\vec{\mathbf{0}}_2, \mathbf{1}_2\right), \tag{A.1}$$

$$\mathrm{Var}(\boldsymbol{\eta}_t) = \sum_{j=0}^{k} \mathbf{R}_j \mathrm{Var}(\boldsymbol{\varepsilon}_{t-j}) \mathbf{R}_j' + \boldsymbol{\Sigma}_\xi^{1/2} \mathbf{1}_2 \boldsymbol{\Sigma}_\xi^{1/2\prime} \tag{A.2}$$

$$= \sum_{j=0}^{k} \mathbf{R}_j \mathrm{Var}(\boldsymbol{\varepsilon}_{t-j}) \mathbf{R}_j' + \boldsymbol{\Sigma}_\xi \tag{A.3}$$

$$= \sum_{j=0}^{k} \mathrm{diag}(\rho_{aj}, \rho_{bj}) \begin{bmatrix} 1 & \rho_\varepsilon \\ \rho_\varepsilon & 1 \end{bmatrix} \mathrm{diag}(\rho_{aj}, \rho_{bj})' \tag{A.4}$$

$$+ \begin{bmatrix} 1 - \sum_{j=0}^{k} \rho_{aj}^2 & \rho_\eta - \sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon \\ \rho_\eta - \sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon & 1 - \sum_{j=0}^{k} \rho_{bj}^2 \end{bmatrix} \tag{A.5}$$

$$= \begin{bmatrix} \sum_{j=0}^{k} \rho_{aj}^2 & \rho_\varepsilon \sum_{j=0}^{k} \rho_{aj}\rho_{bj} \\ \rho_\varepsilon \sum_{j=0}^{k} \rho_{aj}\rho_{bj} & \sum_{j=0}^{k} \rho_{bj}^2 \end{bmatrix} + \begin{bmatrix} 1 - \sum_{j=0}^{k} \rho_{aj}^2 & \rho_\eta - \rho_\varepsilon \sum_{j=0}^{k} \rho_{aj}\rho_{bj} \\ \rho_\eta - \rho_\varepsilon \sum_{j=0}^{k} \rho_{aj}\rho_{bj} & 1 - \sum_{j=0}^{k} \rho_{bj}^2 \end{bmatrix} \tag{A.6}$$

$$= \begin{bmatrix} 1 & \rho_\eta \\ \rho_\eta & 1 \end{bmatrix} = \mathbf{R}_\eta. \tag{A.7}$$

Here we used the restriction in equation (8):

$$\boldsymbol{\Sigma}_\xi = \begin{bmatrix} 1 - \sum_{j=0}^{k} \rho_{aj}^2 & \rho_\eta - \sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon \\ \rho_\eta - \sum_{j=0}^{k} \rho_{aj}\rho_{bj}\rho_\varepsilon & 1 - \sum_{j=0}^{k} \rho_{bj}^2 \end{bmatrix}, \tag{A.8}$$

and the unconditional variance of $\boldsymbol{\varepsilon}_{t-j}$:

$$\mathrm{Var}(\boldsymbol{\varepsilon}_{t-j}) = \mathbf{R}_\varepsilon = \begin{bmatrix} 1 & \rho_\varepsilon \\ \rho_\varepsilon & 1 \end{bmatrix}, \text{ for } j = 0, 1, \ldots, k. \tag{A.9}$$

B Unconditional expectation and variance of $\mathbf{x}_t$

From equation (5):

$$\mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t, \tag{B.1}$$

$$\mathbb{E}(\mathbf{x}_t) = \mu + \boldsymbol{\Phi}\mathbb{E}(\mathbf{x}_{t-1}) + \vec{\mathbf{0}}_2, \tag{B.2}$$

$$\mathrm{Var}(\mathbf{x}_t) = \boldsymbol{\Phi}\mathrm{Var}(\mathbf{x}_{t-1})\boldsymbol{\Phi}' + \mathrm{Var}(\boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t). \tag{B.3}$$

Assuming stationarity for $\mathbf{x}_t$ implies $\mathbb{E}(\mathbf{x}_t) = \mathbb{E}(\mathbf{x}_{t-1}) = \boldsymbol{\Psi}_0$ and $\mathrm{Var}(\mathbf{x}_t) = \mathrm{Var}(\mathbf{x}_{t-1}) = \boldsymbol{\Gamma}_0$, hence we need to solve the following equations for $\boldsymbol{\Psi}_0$ and $\boldsymbol{\Gamma}_0$:

$$\boldsymbol{\Psi}_0 = \boldsymbol{\mu} + \boldsymbol{\Phi}\boldsymbol{\Psi}_0 \tag{B.4}$$

$$\boldsymbol{\Gamma}_0 = \boldsymbol{\Phi}\boldsymbol{\Gamma}_0\boldsymbol{\Phi}' + \mathrm{Var}(\boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t). \tag{B.5}$$

The first one can be solved directly:

$$\boldsymbol{\Psi}_0 = (\mathbf{1}_2 - \boldsymbol{\Phi})^{-1}\boldsymbol{\mu} \tag{B.6}$$

The latter, $\boldsymbol{\Gamma}_0$, can be solved by first applying the vec operator and rewrite equation (B.4) to

$$\mathrm{vec}(\boldsymbol{\Gamma}_0) = (\boldsymbol{\Phi} \otimes \boldsymbol{\Phi}')\,\mathrm{vec}(\boldsymbol{\Gamma}_0) + \mathrm{vec}\left(\mathrm{Var}(\boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t)\right) \tag{B.7}$$

$$= (\mathbf{1}_4 - \boldsymbol{\Phi} \otimes \boldsymbol{\Phi}')^{-1}\,\mathrm{vec}\left(\mathrm{Var}(\boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t)\right), \tag{B.8}$$

where $\otimes$ denotes the Kronecker product.[14] The last term in equation (B.3) can be further derived using the unconditional variance of $\boldsymbol{\eta}_t$ derived in Appendix A:

$$\mathrm{Var}(\boldsymbol{\Sigma}_\eta^{1/2}\boldsymbol{\eta}_t) = \boldsymbol{\Sigma}_\eta^{1/2}\mathbf{R}_\eta\boldsymbol{\Sigma}_\eta^{1/2} = \begin{bmatrix} \sigma_{\eta a}^2 & \sigma_{\eta a}\sigma_{\eta b}\rho_\eta \\ \sigma_{\eta a}\sigma_{\eta b}\rho_\eta & \sigma_{\eta b}^2 \end{bmatrix}. \tag{B.9}$$

The unconditional expectation and variance of $\mathbf{x}_t$ can be used as initialization for the state and the state variance respectively.

[14] See Petersen et al. (2008) equation (505).

C State-space construction

This section focuses on how the state-space representation can be obtained. We allow for two different temporal lags ($k = 1$ in equation 3) to prevent overfitting. The model in equations (4 -6) can be written as a general state-space model with Gaussian nonlinear measurement and state-transition equations:

$$\mathbf{y}_t \sim p_{\mathbf{y}}\left(\mathbf{y}_t \mid \boldsymbol{\alpha}_t\right), \quad \boldsymbol{\alpha}_t \sim p_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}_t \mid \boldsymbol{\alpha}_{t-1}\right), \quad \boldsymbol{\alpha}_1 \sim p_{\boldsymbol{\alpha}_1}\left(\boldsymbol{\alpha}_1\right), \tag{C.1}$$

where the state is defined as $\boldsymbol{\alpha}_t = (\mathbf{x}_t, \mathbf{x}_{t-1})'$.

First we look at the error structure given the information set at time $t-2$, defined as $\mathcal{F}_{t-2} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{t-2}\}$, for the bivariate model has the following joint distribution:

$$\begin{bmatrix} \eta_{a,t} \\ \varepsilon_{a,t} \\ \varepsilon_{a,t-1} \\ \eta_{b,t} \\ \varepsilon_{b,t} \\ \varepsilon_{b,t-1} \end{bmatrix} \mid \mathcal{F}_{t-2} \sim \mathcal{N}_6 \left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho_{a0} & \rho_{a1} & \rho_{\eta} & 0 & 0 \\ \rho_{a0} & 1 & 0 & 0 & \rho_{\varepsilon} & 0 \\ \rho_{a1} & 0 & 1 & 0 & 0 & \rho_{\varepsilon} \\ \rho_{\eta} & 0 & 0 & 1 & \rho_{b0} & \rho_{b1} \\ 0 & \rho_{\varepsilon} & 0 & \rho_{b0} & 1 & 0 \\ 0 & 0 & \rho_{\varepsilon} & \rho_{b1} & 0 & 1 \end{bmatrix} \right) \tag{C.2}$$

Following the same process as Lange (2020) in his Appendix, we condition the current errors $\boldsymbol{\varepsilon}_t, \boldsymbol{\eta}_t$ on the past errors $\boldsymbol{\varepsilon}_{t-1}, \boldsymbol{\eta}_{t-1}$ using the multivariate normal conditional distribution lemma of Page Jr (1984):

Lemma 1 (**Multivariate normal conditional distribution lemma**)
If for $\mathbf{x} \in \mathbb{R}^n$*:*

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}_n \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right), \tag{C.3}$$

with $\mathbf{x}_1 \in \mathbb{R}^q, \mathbf{x}_2 \in \mathbb{R}^{n-q}$*, then the conditional distribution of* $\mathbf{x}_1 \mid \mathbf{x}_2$ *is given by* $\mathcal{N}_q(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ *where*

$$\boldsymbol{\mu}_c = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}_2 - \boldsymbol{\mu}_2\right) \tag{C.4}$$

$$\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \tag{C.5}$$

Similarly as in Lange (2020), we use the information that the past return shocks $\boldsymbol{\varepsilon}_{t-1}$ are non-random conditional on the information set and state at time $t-1$. As we are looking for a state-space format as in equation (C.1), we can condition the state-transition equation, or equivalently $\boldsymbol{\eta}_t$, on $\boldsymbol{\alpha}_{t-1}$. Conditional on $\boldsymbol{\alpha}_{t-1}$, $\boldsymbol{\varepsilon}_{t-1}$ can be constructed using equation (4): $\boldsymbol{\varepsilon}_{t-1} = \boldsymbol{\Omega}_t^{-1/2}\mathbf{B}^{-1}(\mathbf{y}_t - \tilde{\mathbf{y}})$. Hence, by conditioning on these historical errors, we obtain the

following joint conditional Gaussian distribution for $\boldsymbol{\eta}_t, \boldsymbol{\varepsilon}_t | \mathcal{F}_{t-1}, \boldsymbol{\alpha}_{t-1}$:

$$\begin{bmatrix} \eta_{a,t} \\ \eta_{b,t} \\ \varepsilon_{a,t} \\ \varepsilon_{b,t} \end{bmatrix} \mid \mathcal{F}_{t-1}, \boldsymbol{\alpha}_{t-1} \sim \mathcal{N}_4 \left(\begin{bmatrix} \frac{\rho_{a1}\varepsilon_{a,t-1} - \rho_{a1}\rho_\varepsilon \varepsilon_{b,t-1}}{1-\rho_\varepsilon^2} \\ \frac{-\rho_{b1}\rho_\varepsilon \varepsilon_{a,t-1} + \rho_{b1}\varepsilon_{b,t-1}}{1-\rho_\varepsilon^2} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 - \frac{\rho_{a1}^2}{1-\rho_\varepsilon^2} & \rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} & \rho_{a0} & 0 \\ \rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} & 1 - \frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} & 0 & \rho_{b0} \\ \rho_{a0} & 0 & 1 & \rho_\varepsilon \\ 0 & \rho_{b0} & \rho_\varepsilon & 1 \end{bmatrix} \right). \tag{C.6}$$

The conditional distribution of $\boldsymbol{\eta}_t$, $\boldsymbol{\eta}_t | \mathcal{F}_{t-1}, \boldsymbol{\alpha}_{t-1} \sim \mathcal{N}_2 \left(\boldsymbol{\mu}_{\eta|\alpha,t}, \boldsymbol{\Sigma}_{\eta|\alpha} \right)$ is a subset of the multivariate normal random vector in equation (C.6), and hence $\boldsymbol{\mu}_{\eta|\alpha,t}, \boldsymbol{\Sigma}_{\eta|\alpha}$ can be read of the first two dimensions in equation (C.6). Next, we use the information that the observation equation in the general state-space structure in equation (C.1) conditions on $\boldsymbol{\alpha}_t$. The current volatility shock $\boldsymbol{\eta}_t$ becomes non-random after conditioning on $\boldsymbol{\alpha}_t$ and can be constructed through equation (5): $\boldsymbol{\eta}_t = \boldsymbol{\Sigma}_\eta^{-1/2} \left(\mathbf{x}_t - \boldsymbol{\mu} - \boldsymbol{\Phi} \mathbf{x}_{t-1} \right)$. Hence we can apply the conditional Normal lemma again and condition on $\boldsymbol{\eta}_t$ to obtain the conditional distribution of $\boldsymbol{\varepsilon}_t | \boldsymbol{\eta}_t, \boldsymbol{\alpha}_{t-1}, \mathcal{F}_{t-1}$:

$$\begin{aligned} \boldsymbol{\varepsilon}_t | \boldsymbol{\eta}_t, \boldsymbol{\alpha}_t, \mathcal{F}_{t-1} &\sim \mathcal{N}_2 \left(\boldsymbol{\mu}_{\varepsilon|\eta,t}, \boldsymbol{\Sigma}_{\varepsilon|\eta} \right), \\ \boldsymbol{\mu}_{\varepsilon|\eta,t} &= \begin{bmatrix} \rho_{a0} & 0 \\ 0 & \rho_{b0} \end{bmatrix} \boldsymbol{\Sigma}_{\eta|\alpha}^{-1} \left(\boldsymbol{\eta}_t - \boldsymbol{\mu}_{\eta|\alpha,t} \right), \\ \boldsymbol{\Sigma}_{\varepsilon|\eta} &= \begin{bmatrix} 1 & \rho_\varepsilon \\ \rho_\varepsilon & 1 \end{bmatrix} - \begin{bmatrix} \rho_{a0} & 0 \\ 0 & \rho_{b0} \end{bmatrix} \boldsymbol{\Sigma}_{\eta|\alpha}^{-1} \begin{bmatrix} \rho_{a0} & 0 \\ 0 & \rho_{b0} \end{bmatrix}, \end{aligned} \tag{C.7}$$

where $\boldsymbol{\Sigma}_{\eta|\alpha}^{-1}$ is equal to

$$\begin{aligned} \boldsymbol{\Sigma}_{\eta|\alpha}^{-1} &= \begin{bmatrix} 1 - \frac{\rho_{a1}^2}{1-\rho_\varepsilon^2} & \rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} \\ \rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} & 1 - \frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} \end{bmatrix}^{-1} = \frac{1}{\det\left(\boldsymbol{\Sigma}_{\eta|\alpha}\right)} \begin{bmatrix} 1 - \frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} & -\left(\rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} \right) \\ -\left(\rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} \right) & 1 - \frac{\rho_{a1}^2}{1-\rho_\varepsilon^2} \end{bmatrix}, \\ \det\left(\boldsymbol{\Sigma}_{\eta|\alpha}\right) &= \begin{vmatrix} 1 - \frac{\rho_{a1}^2}{1-\rho_\varepsilon^2} & \rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} \\ \rho_\eta - \frac{\rho_\varepsilon \rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2} & 1 - \frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} \end{vmatrix}. \end{aligned} \tag{C.8}$$

Conditional on time t, the values of $\boldsymbol{\eta}_t$ are known and can be constructed from equation (5): $\boldsymbol{\eta}_t = \boldsymbol{\Sigma}_\eta^{-1/2} \left(\mathbf{x}_t - \boldsymbol{\mu} - \boldsymbol{\Phi} \mathbf{x}_{t-1} \right)$.

Using the conditional distribution of $\boldsymbol{\varepsilon}_t$, the conditional density for $\mathbf{y}_t | \boldsymbol{\alpha}_{t-1}, \mathcal{F}_{t-1}$ follows directly as

$$\begin{aligned} p_{\mathbf{y}}(\mathbf{y}_t | \boldsymbol{\alpha}_t, \mathcal{F}_{t-1}) &= \frac{1}{\sqrt{(2\pi)^2 \det(\boldsymbol{\Sigma}_{\mathbf{y},t})}} \exp\left(-\frac{1}{2} (\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t})' \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t}) \right), \\ \boldsymbol{\mu}_{\mathbf{y},t} &= \tilde{\mathbf{y}} + \mathbf{B} \boldsymbol{\Omega}_t^{1/2} \boldsymbol{\mu}_{\varepsilon|\eta,t}, \quad \boldsymbol{\Sigma}_{\mathbf{y},t} = \mathbf{B} \boldsymbol{\Omega}_t^{1/2} \boldsymbol{\Sigma}_{\varepsilon|\eta} \boldsymbol{\Omega}_t^{1/2\prime} \mathbf{B}'. \end{aligned} \tag{C.9}$$

Finally, the conditional degenerate density of the latent state $\mathbf{x}|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1}$ is given in equation (C.10).

$$
\begin{aligned}
p_{\mathbf{x}}(\mathbf{x}_t|\boldsymbol{\alpha}_{t-1}, \mathcal{F}_{t-1}) &= \frac{1}{\sqrt{(2\pi)^n \det(\boldsymbol{\Sigma}_{\mathbf{x}})}} \exp\left(-\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x},t})' {\boldsymbol{\Sigma}_{\mathbf{x}}}^{-1} (\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{x},t})\right) \times \delta_t \\
\boldsymbol{\mu}_{\mathbf{x},t} &= \boldsymbol{\mu} + \boldsymbol{\Phi}\mathbf{x}_{t-1} + \boldsymbol{\Sigma}_{\eta}^{1/2} \boldsymbol{\mu}_{\eta|\alpha,t}, \\
\boldsymbol{\Sigma}_{\mathbf{x}} &= \boldsymbol{\Sigma}_{\eta}^{1/2} \boldsymbol{\Sigma}_{\eta|\alpha} \boldsymbol{\Sigma}_{\eta}^{1/2}, \\
\delta_t &= \prod_{j=1}^{2} \delta_d(\alpha_{t,j+2} - \alpha_{t-1,j}).
\end{aligned}
\tag{C.10}
$$

The density follows from equation (16), where a product of Dirac delta[15] functions are added similarly as in Lange (2020). The Dirac deltas restrict the elements of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_{t-1}$, in a way such that $\boldsymbol{\alpha}_{t,3} = \mathbf{x}_{a,t-1} = \boldsymbol{\alpha}_{t-1,1}$ and $\boldsymbol{\alpha}_{t,4} = \mathbf{x}_{b,t-1} = \boldsymbol{\alpha}_{t-1,2}$. This classical optimization technique forces the constraints onto the objective function without influencing the data generating process, but simplify the optimization problem as it reduces the number of state variables to optimize.

D State-space derivatives

As in Section 3.1, we define $\mathbf{z}_t = (\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t})$. $\mathbf{J}^{i,j}$ denotes the single entry matrix, $(\mathbf{J}^{i,j})_{k,l} = \delta_{ik}\delta_{jl}$, where δ_{ik} is the Kronecker's delta function. First we look at the total derivative of the quadratic part in the multivariate normal distribution inside the exponent. We rewrite the function in terms of Frobenius products with the colon operator (:) for notational convenience.

$$
\begin{aligned}
Q &= \mathbf{z}_t' \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathbf{z}_t && \text{(D.1)} \\
&= \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathbf{z}_t \mathbf{z}_t' && \text{(D.2)} \\
\mathrm{d}Q &= \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : (\mathrm{d}\mathbf{z}_t \mathbf{z}_t' + \mathbf{z}_t \mathrm{d}\mathbf{z}_t') + \mathbf{z}_t \mathbf{z}_t' : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} && \text{(D.3)} \\
&= \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathbf{z}_t : \mathrm{d}\mathbf{z}_t + \mathbf{z}_t' \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\mathbf{z}_t' - \mathbf{z}_t \mathbf{z}_t' : \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t} \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} && \text{(D.4)} \\
&= \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathbf{z}_t : \mathrm{d}\mathbf{z}_t + \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1\prime} \mathbf{z}_t : \mathrm{d}\mathbf{z}_t - \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathbf{z}_t \mathbf{z}_t' \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t} && \text{(D.5)} \\
&= 2\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathbf{z}_t : \mathrm{d}\mathbf{z}_t - \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathbf{z}_t \mathbf{z}_t' \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t} && \text{(D.6)} \\
&= -2\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} (\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t}) : \mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t} - \boldsymbol{\Sigma}_{\mathbf{y},t}' \mathbf{z}_t \mathbf{z}_t' \boldsymbol{\Sigma}_{\mathbf{y},t}' : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t} && \text{(D.7)}
\end{aligned}
$$

In D.3, we used rule (59) from Petersen et al. (2008): $\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} = \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t} \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}$, and rule (37): the product rule for matrix derivatives. In D.4 and D.5 we used the symmetric property of

[15] $\delta_d(x) = \begin{cases} +\infty, & x = 0 \\ 0, & x \neq 0 \end{cases}$

$\boldsymbol{\Sigma}_{\mathbf{y},t}$, and its inverse $\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1\prime} = \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}$.

Now we can further derive the total derivative of the log likelihood function as follows

$$\log f(\mathbf{y}_t|\boldsymbol{\alpha}_t) = \frac{1}{2}(\log\det(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}) - Q) \tag{D.8}$$

$$= \frac{1}{2}(\operatorname{trace}(\log(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1})) - Q) \tag{D.9}$$

$$\mathrm{d}\log f(\mathbf{y}_t|\boldsymbol{\alpha}_t) = \frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} - \mathrm{d}Q\right) \tag{D.10}$$

$$= \frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} - 2\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}(\boldsymbol{\mu}_{\mathbf{y},t} - \mathbf{y}_t) : \mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t} + \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathbf{z}_t\mathbf{z}_t'\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}\right) \tag{D.11}$$

$$= \frac{1}{2}\left(-\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} - 2\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}(\boldsymbol{\mu}_{\mathbf{y},t} - \mathbf{y}_t) : \mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t} + \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathbf{z}_t\mathbf{z}_t'\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}\right) \tag{D.12}$$

$$= \frac{1}{2}\left(-\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t} - 2\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}(\boldsymbol{\mu}_{\mathbf{y},t} - \mathbf{y}_t) : \mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t} + \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathbf{z}_t\mathbf{z}_t'\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}\right) \tag{D.13}$$

$$= \frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathbf{z}_t\mathbf{z}_t'\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} - \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\right) : \mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t} - \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}(\boldsymbol{\mu}_{\mathbf{y},t} - \mathbf{y}_t) : \mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t} \tag{D.14}$$

$$\frac{\partial \log f(\mathbf{y}_t|\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\mu}_{\mathbf{y},t}} = \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},t}) \tag{D.15}$$

$$\frac{\partial \log f(\mathbf{y}_t|\boldsymbol{\alpha}_t)}{\partial \boldsymbol{\Sigma}_{\mathbf{y},t}} = \frac{1}{2}\left(\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\mathbf{z}_t\mathbf{z}_t'\boldsymbol{\Sigma}_{\mathbf{y},t}^{-1} - \boldsymbol{\Sigma}_{\mathbf{y},t}^{-1}\right) \tag{D.16}$$

The gradient for the state density can be derived in a similar fashion with $\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{0}_2$ and note that $\mathrm{d}\mathbf{z}_t = \mathrm{d}\mathbf{x}_t - \mathrm{d}\boldsymbol{\mu}_{\mathbf{x}}$.

The derivatives of $\boldsymbol{\mu}_{\mathbf{y},t}$ and $\boldsymbol{\Sigma}_{\mathbf{y},t}$ w.r.t the elements in $\boldsymbol{\alpha}_t$ are given as

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}}{\mathrm{d}x_{a,t}} =& \mathbf{B}\frac{\mathrm{d}\boldsymbol{\Omega}_t^{1/2}}{\mathrm{d}x_{a,t}}\boldsymbol{\mu}_{\varepsilon|\eta,t} + \mathbf{B}\boldsymbol{\Omega}_t^{1/2}\frac{\mathrm{d}\boldsymbol{\mu}_{\varepsilon|\eta,t}}{\mathrm{d}x_{a,t}} = \mathbf{B}\frac{\exp(x_{a,t}/2)}{2}\mathbf{J}^{1,1}\boldsymbol{\mu}_{\varepsilon|\eta,t} + \mathbf{B}\boldsymbol{\Omega}_t^{1/2} \\ & \frac{1}{\det(\boldsymbol{\Sigma}_{\eta|\alpha})}\begin{bmatrix}\rho_{a0} & 0\\ 0 & \rho_{b0}\end{bmatrix}\begin{bmatrix}1-\frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} & -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right)\\ -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right) & 1-\frac{\rho_{a1}^2}{1-\rho_\varepsilon^2}\end{bmatrix}\boldsymbol{\Sigma}_\eta^{-1/2}\begin{bmatrix}1\\0\end{bmatrix},\end{aligned} \tag{D.17}$$

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}}{\mathrm{d}x_{b,t}} =& \mathbf{B}\frac{\mathrm{d}\boldsymbol{\Omega}_t^{1/2}}{\mathrm{d}x_{b,t}}\boldsymbol{\mu}_{\varepsilon|\eta,t} + \mathbf{B}\boldsymbol{\Omega}_t^{1/2}\frac{\mathrm{d}\boldsymbol{\mu}_{\varepsilon|\eta,t}}{\mathrm{d}x_{b,t}} = \mathbf{B}\frac{\exp(x_{b,t}/2)}{2}\mathbf{J}^{2,2}\boldsymbol{\mu}_{\varepsilon|\eta,t} + \mathbf{B}\boldsymbol{\Omega}_t^{1/2} \\ & \frac{1}{\det(\boldsymbol{\Sigma}_{\eta|\alpha})}\begin{bmatrix}\rho_{a0} & 0\\ 0 & \rho_{b0}\end{bmatrix}\begin{bmatrix}1-\frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} & -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right)\\ -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right) & 1-\frac{\rho_{a1}^2}{1-\rho_\varepsilon^2}\end{bmatrix}\boldsymbol{\Sigma}_\eta^{-1/2}\begin{bmatrix}0\\1\end{bmatrix},\end{aligned} \tag{D.18}$$

$$\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}}{\mathrm{d}x_{a,t-1}} = \mathbf{B}\boldsymbol{\Omega}_t^{1/2}\frac{1}{\det\left(\boldsymbol{\Sigma}_{\eta|\alpha}\right)}\begin{bmatrix}\rho_{a0} & 0\\ 0 & \rho_{b0}\end{bmatrix}\begin{bmatrix}1-\frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} & -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right)\\ -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right) & 1-\frac{\rho_{a1}^2}{1-\rho_\varepsilon^2}\end{bmatrix}\left(-\boldsymbol{\Sigma}_\eta^{-1/2}\Phi\begin{bmatrix}1\\0\end{bmatrix} + \frac{\exp(-x_{a,t-1}/2)y_{a,t-1}}{2\beta_a(1-\rho_\varepsilon^2)}\begin{bmatrix}\rho_{a1}\\ -\rho_{b1}\rho_\varepsilon\end{bmatrix}\right), \tag{D.19}$$

$$\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{y},t}}{\mathrm{d}x_{b,t-1}} = \mathbf{B}\boldsymbol{\Omega}_t^{1/2}\frac{1}{\det\left(\boldsymbol{\Sigma}_{\eta|\alpha}\right)}\begin{bmatrix}\rho_{a0} & 0\\ 0 & \rho_{b0}\end{bmatrix}\begin{bmatrix}1-\frac{\rho_{b1}^2}{1-\rho_\varepsilon^2} & -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right)\\ -\left(\rho_\eta - \frac{\rho_\varepsilon\rho_{a1}\rho_{b1}}{1-\rho_\varepsilon^2}\right) & 1-\frac{\rho_{a1}^2}{1-\rho_\varepsilon^2}\end{bmatrix}\left(-\boldsymbol{\Sigma}_\eta^{-1/2}\Phi\begin{bmatrix}0\\1\end{bmatrix} + \frac{\exp(-x_{b,t-1}/2)y_{b,t-1}}{2\beta_b(1-\rho_\varepsilon^2)}\begin{bmatrix}-\rho_{a1}\rho_\varepsilon\\ \rho_{b1}\end{bmatrix}\right), \tag{D.20}$$

$$\frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}x_{a,t}} = \frac{\mathrm{d}\mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Sigma}_{\varepsilon|\eta}\boldsymbol{\Omega}_t^{1/2}\mathbf{B}}{\mathrm{d}x_{a,t}} = \frac{1}{2}\mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Sigma}_{\varepsilon|\eta}\boldsymbol{\Omega}_t^{1/2}\mathbf{J}^{1,1}\mathbf{B} + \frac{1}{2}\mathbf{B}\mathbf{J}^{1,1}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Sigma}_{\varepsilon|\eta}\boldsymbol{\Omega}_t^{1/2}\mathbf{B}, \tag{D.21}$$

$$\frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}x_{b,t}} = \frac{\mathrm{d}\mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Sigma}_{\varepsilon|\eta}\boldsymbol{\Omega}_t^{1/2}\mathbf{B}}{\mathrm{d}x_{b,t}} = \frac{1}{2}\mathbf{B}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Sigma}_{\varepsilon|\eta}\boldsymbol{\Omega}_t^{1/2}\mathbf{J}^{2,2}\mathbf{B} + \frac{1}{2}\mathbf{B}\mathbf{J}^{2,2}\boldsymbol{\Omega}_t^{1/2}\boldsymbol{\Sigma}_{\varepsilon|\eta}\boldsymbol{\Omega}_t^{1/2}\mathbf{B}, \tag{D.22}$$

$$\frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}x_{a,t-1}} = \frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{y},t}}{\mathrm{d}x_{b,t-1}} = \mathbf{0}_{2\times 2}. \tag{D.23}$$

The individual expressions on the right side of equation D.21 and D.22 are not necessarily symmetric as $\boldsymbol{\Omega}_t^{1/2}$ and $\mathbf{J}^{i,i}$ for $i = 1, 2$, do not have to commute. Lastly, for the total derivative of the state density, we require the derivatives w.r.t $\boldsymbol{\mu}_{\mathbf{x},t}$, $\mathbf{x}_t$ and $\boldsymbol{\Sigma}_{\mathbf{x}}$. These are given as

$$\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{x},t}}{\mathrm{d}x_{a,t}} = \frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{x},t}}{\mathrm{d}x_{a,t}} = \vec{\mathbf{0}}_2, \tag{D.24}$$

$$\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{x},t}}{\mathrm{d}x_{a,t-1}} = \boldsymbol{\Phi}\begin{bmatrix}1\\0\end{bmatrix} - \frac{1}{2}\frac{1}{1-\rho_\varepsilon^2}\boldsymbol{\Sigma}_\eta^{1/2}\begin{bmatrix}\rho_{a1}\exp(-x_{a,t-1}/2)\beta_a^{-1}(y_{a,t-1}-\tilde{y}_a)\\ -\rho_{b1}\rho_\varepsilon\exp(-x_{a,t-1}/2)\beta_a^{-1}(y_{a,t-1}-\tilde{y}_a)\end{bmatrix}, \tag{D.25}$$

$$\frac{\mathrm{d}\boldsymbol{\mu}_{\mathbf{x},t}}{\mathrm{d}x_{b,t-1}} = \boldsymbol{\Phi}\begin{bmatrix}0\\1\end{bmatrix} - \frac{1}{2}\frac{1}{1-\rho_\varepsilon^2}\boldsymbol{\Sigma}_\eta^{1/2}\begin{bmatrix}-\rho_{a1}\rho_\varepsilon\exp(-x_{b,t-1}/2)\beta_b^{-1}(y_{b,t-1}-\tilde{y}_b)\\ \rho_{b1}\exp(-x_{b,t-1}/2)\beta_b^{-1}(y_{b,t-1}-\tilde{y}_b)\end{bmatrix}, \tag{D.26}$$

$$\frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{x}}}{\mathrm{d}x_{a,t}} = \frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{x}}}{\mathrm{d}x_{b,t}} = \frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{x}}}{\mathrm{d}x_{a,t-1}} = \frac{\mathrm{d}\boldsymbol{\Sigma}_{\mathbf{x}}}{\mathrm{d}x_{b,t-1}} = \mathbf{0}_{2\times 2}. \tag{D.27}$$

$$\frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}x_{a,t}} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}x_{b,t}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}x_{a,t-1}} = \frac{\mathrm{d}\mathbf{x}_t}{\mathrm{d}x_{a,t-1}} = \vec{\mathbf{0}}_2. \tag{D.28}$$