Multivariate Parameter- and Observation-Driven Models with Applications in Risk Management

Master Thesis in Quantitative Finance

Second Draft

Nando Vermeer

447763

Supervisor: dr. Lange, R-J.

Second assessor: prof. dr. Zhou, C.

Erasmus School of Economics

ERASMUS UNIVERSITY ROTTERDAM

November 15th, 2020

Contents

Abstract

In this paper we study how well multivariate parameter-driven models can be approximated by observation-driven models, since observation-driven models are less computationally intensive to estimate. To study this, we perform multiple simulation and empirical studies. For these studies, three different observation densities are chosen. The first one models the mean and variance of a univariate time series, and the other two model the covariance matrix of a two-dimensional time series. We find multivariate observation-driven models are able to approximate parameter-driven specifications quite well, although performance decreases in the presence of strong cross effects between the hidden states. Furthermore, in both Value at Risk analysis as well as minimum variance portfolio construction, observation-driven models perform equally well or better than parameter-driven models.

1 Introduction

Models where state variables are unobserved have been used for decades, and are often referred to as parameter-driven models. The most well-known models of this class are the linear state-space models with Gaussian innovations. For this specific case, analytic methods have been developed (Kalman, 1960) to estimate the unobserved states and perform one-step-ahead forecasts. However, when we depart from the assumptions of linear relations between the unobserved state variables and the observed dependent variables, these methods are no longer effective. These models are known as non-linear non-Gaussian state-space models, as the predictive one-step-ahead density is no longer Gaussian.

While analytic methods are often not available for these models, numerical methods have been developed. For example, efficient importance sampling introduced by Richard and Zhang (2007) allows for a numerical evaluation of the joint likelihood. With this method, parameter estimation, as well as one-step-ahead forecasts can be performed. Koopman (2015) proposes a faster variant of this algorithm, referred to as NAIS, which relies on the numerical evaluation of certain low-dimensional integrals. However, the approximation becomes less efficient when the approximated integrals are multidimensional. Recently Lange (2020) developed the Bellman filter, which approximates the likelihood and optimal state filtration by applying a generalized Kalman filter approach.

Another model type is known as observation-driven models. In contrast with the models described above, observation-driven models have perfect one-step-ahead predictions for the state variables. Well-known variants of these models are conditional volatility models such as ARCH and GARCH. Here, the volatility at time t depends only on the volatility at time $t-1$, as well as the dependent variable at time $t-1$. These models have been studied and used thoroughly, with many new variations still being published.

It was shown by Creal et al. (2013) and Harvey (2013) that these conditional volatility models, as well as other well-known observation-driven models (dynamic exponential family, multiplicative error models) can be seen as special cases of a more general model class, referred to as generalized autoregressive score (GAS) models. An advantage of these models is that the likelihood is often available in closed form. One of the more interesting aspects of GAS models is how the state variables are updated, which is (partly) done using the scaled score of the dependent variable. The choice of scaling matrix used influences the structure of the model. For a more in-depth overview of this class of models, we refer to Harvey (2013).

There is existing literature on the comparisons between parameter- and observation-driven models, for example Koopman, Lucas and Scharth (2016). They analyze how well parameter-driven models can be approximated by observation-driven models. The advantage of this approximation is that likelihood evaluation is often less cumbersome for observation-driven models, compared to parameter-driven models. They mainly study models with a single time-varying parameter. For all models considered, they use the inverse square root of the Fisher Information matrix as scaling for the score. However, they do not consider the performance of observation-driven models when there are multiple driving variables. Because of this, we propose to analyze the performance of both parameter- as well as observation-driven models under a multidimensional driving process. Here we investigate if the observation-driven models perform as well as similar parameter-driven models, both with underlying multivariate driving processes.

To answer this question, we consider multiple subquestions. One of the main questions answered in Koopman, Lucas and Scharth (2016) was how well parameter-driven models can be approximated by observation-driven models. We repeat this research, but with multidimensional driving processes. In order to study these effects, we investigate the performance of the approximation for three different observation densities. We note that stochastic means and variances are commonly seen in the literature, but are rarely used together. For stock returns, returns in periods of high volatility are on average more negative than in periods of lower volatility, as was shown by Giot (2005). Because of this, we propose a stochastic mean variance (SMV) specification, where both the real mean and variance are time-varying and unobserved. It is further investigated how well such a parameter-driven model can be approximated by a similar observation-driven model. It is especially interesting to see if the observation-driven models are able to pick up on the cross effects in the underlying states. Apart from the SMV model, we model a time-varying unobserved covariance matrix, for which we consider two different specifications. Both of these are further explained in Section 5.

We initially follow Koopman, Lucas and Scharth (2016), where the score is scaled by the inverse Cholesky decomposition of the Fisher Information matrix. We also consider different scaling matrices, such as the the inverse Fisher matrix itself and the identity. We examine if a full model specification is necessary, and compare this to the performance of a diagonal specification.

For our approximation study, we find similar results as found in Koopman, Lucas and Scharth (2016). When the hidden states follow a diagonal specification, the observation-driven models perform similarly or even better than the corresponding parameter-driven specification. We find that the intuitive reasoning that diagonal observation-driven models should be used for diagonal parameter-driven models holds. However, when the innovations in the underlying states are correlated, as was chosen for the SMV model, the observation-driven models perform noticeably worse. As for the scaling matrix, we are unable to find one general scaling choice that works best for all model specifications. Instead, we find that the best scaling choice depends on the choice of observation density.

While these theoretical properties are interesting in their own right, we also consider more practical applications by means of a small empirical study. The models we propose are fitted on stock index data of the AEX and S&P500. For the SMV model, we analyze how well it is able to give Value at Risk (VaR) predictions. For the two multivariate volatility models, minimum variance portfolio construction is performed. The performance of these models is then compared to other well-known models.

We find that the observation-driven models compete quite well with the parameter-driven models. They perform similar to their parameter-driven variant in the VaR study, and outperform the parameter-driven variant in the GMV study. The observation-driven model that performs especially well for the GMV study is the SCC model with a full specification. As full parameter-driven models of that size are not computationally feasible to estimate with EIS, this gives an advantage to the observation-driven models for this purpose.

2 Literature

In order to obtain a better grasp on the models we investigate, we consider the literature on the subject. Let us first discuss some of the relevant literature on parameter-driven models. The most relevant paper in our studies on this subject is Richard and Zhang (2007), where the EIS method for evaluating high-dimensional integrals is introduced. This is especially useful for evaluating the likelihood of parameter-driven models, which is usually not available in closed form. While this paper was published in 2007, multiple papers by the same author were written using a very similar method (see Richard and Zhang 1996, 1997, 1998). Based on these papers an empirical application of EIS was published in Liesenfeld and Richard (2003), where they apply the method to estimate multivariate stochastic volatility models. They also show how to apply the EIS algorithm to allow for filtered estimates of the hidden driving process. We apply this algorithm to evaluate the likelihood of parameter-driven specifications.

Later Koopman, Lucas and Scharth (2015) published a different variant of EIS, dubbed NAIS, which accelerates the estimation procedure by making use of numerical integration. However, we find that this algorithm is only efficient for small state-space models. This is mainly because it requires the repeated approximation of an integral of the same dimension as the hidden state vector. This method of approximation becomes challenging for models with a two or three dimensional state vector. This is because the numerical integration method they propose (Gauss–Hermite quadrature) requires an exponentially growing number of function evaluations in the dimension of the state vector. This is also referred to as the curse of dimensionality.

Another approach to the likelihood approximation problem is the Bellman filter from Lange (2020). An advantage of this method is that it requires no auxiliary regressions or integral approximations, but instead relies on estimating the mode of the posterior distribution of the hidden process. This is done by utilizing a generalized Kalman Filter. This approximation is significantly faster than NAIS even for univariate hidden processes. Lange (2020) also notes that due to the lack of regressions or other integral approximations this speedup scales well for larger hidden variable processes.

For observation-driven models, we find that a large number of models in this category also belong in the generalized autoregressive score model (or GAS) class, first introduced by Creal et al. (2013) and Harvey (2013). They show that a large number of observation-driven models depend on the scaled score. Furthermore, they show that for this class of models, the likelihood is typically available in closed form, which generally allows for shorter estimation times compared to similar parameter-driven models. Koopman, Lucas and Scharth (2016) also investigate if parameter-driven models could be approximated well by observation-driven models. They found that parameter- and observation-driven models have similar predictive power, even when the underlying process is parameter-driven. Since our observation-driven specifications are also in the GAS class,

we apply the GAS filter in order to evaluate the likelihood of these types of models.

3 Parameter-Driven Models

3.1 Model Introduction

The first model class that we consider is known as parameter-driven models. A defining feature of these models is that the driving variables are unobserved, and only the indirect effects can be seen. A commonly implemented subgroup of parameter-driven models is known as the linear state-space models. These are a class of models that can be written in the following form

$$\begin{aligned}\mathbf{y}_t|\boldsymbol{\alpha}_t &\sim g_N(\mathbf{y}_t; \boldsymbol{\mu} = \mathbf{H}'\boldsymbol{\alpha}_t, \boldsymbol{\Sigma} = \mathbf{R}),\\ \boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t &\sim g_k(\boldsymbol{\omega} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t, \mathbf{Q}),\end{aligned}$$

where $\mathbf{y}_t$ and $\boldsymbol{\alpha}_t$ are multivariate $N \times 1$ and $k \times 1$ observed and unobserved variables at time t respectively. Furthermore, we let $g_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the p-variate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. When the rest of the model parameters are constant, the conditional expectation and variance of the unobserved $\boldsymbol{\alpha}_t$ is known in closed form, and can be computed recursively for all $t \in \{1, \dots, T\}$ by applying the Kalman Filter, first introduced in Kalman, (1960). This allows for a exact evaluation of the likelihood in $\mathcal{O}(T)$ (linear time in T). Because of this, numerical maximization of the likelihood over the parameters is relatively fast for linear state-space models.

While the statistical properties of the linear state-space models are desirable, there are also parameter-driven models where these properties do not hold. The general models that we consider are referred to as non-linear non-Gaussian state-space models by Koopman, Lucas and Scharth (2016). These are models which can be written as

$$\begin{aligned}\mathbf{y}_t|\boldsymbol{\alpha}_t &\sim p(\mathbf{y}_t|\boldsymbol{\theta}(\boldsymbol{\alpha}_t), \boldsymbol{\Psi}),\\ \boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t &\sim g_k(\boldsymbol{\alpha}_{t+1}; \boldsymbol{\mu} = \boldsymbol{\omega} + \boldsymbol{\Phi}\boldsymbol{\alpha}_t, \boldsymbol{\Sigma} = \mathbf{Q}) \equiv g(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, \boldsymbol{\Psi}),\end{aligned}$$

where $\mathbf{y}_t$ and $\boldsymbol{\alpha}_t$ are defined the same as before, and p is the observation density (either discrete or continuous). In our paper, we will refer to $\boldsymbol{\alpha}_t$ as the driving process, as it drives the parameterization of $\mathbf{y}_t$. Furthermore, we define $\boldsymbol{\theta}$ as a vector function of $\boldsymbol{\alpha}_t$, such that the distribution p can be fully identified with $\boldsymbol{\theta}_t \equiv \boldsymbol{\theta}(\boldsymbol{\alpha}_t)$. This also implies that $p(\mathbf{y}_t|\boldsymbol{\theta}_t, \boldsymbol{\Psi}) = p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \boldsymbol{\Psi})$. For our observation densities $\boldsymbol{\theta}$ is invertible, however this is not necessary in general (for an example, see the stochastic volatility models in Liesenfeld and Richard (2003)). Also define the set of auxiliary parameters $\boldsymbol{\Psi} \equiv \{\boldsymbol{\omega}, \boldsymbol{\Phi}, \mathbf{Q}\}$. While p could also depend on auxiliary shape parameters, our implementations do not. Because of this, we omit $\boldsymbol{\Psi}$ in the observation densities.

In this formulation, we see that the conditional normality assumption for $\mathbf{y}_t$ is relaxed, and that its conditional mean is no longer (necessarily) a linear function of $\boldsymbol{\alpha}_t$. For these types of models, the likelihood is typically not available in closed form. Because of this, alternative algorithms have been designed in order to evaluate the likelihood, often by means of approximation.

Since exact likelihood evaluation is often not possible, we resort to numerical approximation. There are multiple relevant methods to perform this approximation, which all rest on the idea of importance sampling.

Here the method consists of applying random sampling and the law of large numbers in order to approximate an integral, but with some alterations in order to increase efficient sampling. Using the convention that $L(\mathbf{y}_{1:T}|\mathbf{\Psi})$ is the joint likelihood of observing $\mathbf{y}_{1:T}$ given the parameters $\mathbf{\Psi}$, then following Koopman, Lucas and Scharth (2015), the likelihood of observing $\{\mathbf{y}_1, \dots, \mathbf{y}_T\} \equiv \mathbf{y}_{1:T}$ can be rewritten as follows (assuming that $\boldsymbol{\alpha}_0$ is deterministic)

$$\begin{aligned} &L(\mathbf{y}_{1:T}|\mathbf{\Psi}) = \int_{\mathbb{R}^k \times T} L(\mathbf{y}_{1:T}, \boldsymbol{\alpha}_{1:T}|\mathbf{\Psi}) d\boldsymbol{\alpha}_{1:T} = \\ &\int_{\mathbb{R}^{k\times T}} \prod_{t=1}^{T} L(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathbf{\Psi}) L(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \mathbf{\Psi}) d\boldsymbol{\alpha}_{1:T} \equiv \int_{\mathbb{R}^{k\times T}} \prod_{t=1}^{T} p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathbf{\Psi}) g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \mathbf{\Psi}) d\boldsymbol{\alpha}_{1:T}, \end{aligned} \tag{1}$$

where we define $d\boldsymbol{\alpha}_{1:T} \equiv d\boldsymbol{\alpha}_1 \dots d\boldsymbol{\alpha}_T$, and p, g the probability distribution functions of $\mathbf{y}_t|\boldsymbol{\alpha}_t$ and $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \mathbf{\Psi}$ respectively.

While both p and g are available in closed form, the integral of their product is unlikely to be. Even though the integral can theoretically be evaluated by means of direct Monte Carlo simulation, convergence is very slow. Because of this, importance sampling can be used (Glynn, 1989). The idea here is to change the probability distribution of the stochastic variable in order to reduce the sampling variance. How this new distribution is chosen depends on the technique used. In Koopman, Lucas and Scharth (2015,2016), NAIS was used. However, it should be noted that this method is intended to be used when the dimensionality of the driving process $\boldsymbol{\alpha_t}$ is significantly smaller than the dimensionality of the parameter vector $\boldsymbol{\theta}_t$. Furthermore, one of the techniques used in NAIS is numerical integration, which is suitable when the signal is one dimensional, but quickly becomes computationally difficult when the dimensionality increases. For the models that we consider, the state vector has the same dimensionality as the signal vector, which is larger than 1. Because of these reasons, we instead opt to apply a multivariate variant of the EIS (Efficient Importance Sampling) algorithm from Richard and Zhang (2007), as well as the Bellman filter from Lange (2020).

3.2 Efficient Importance Sampling

3.2.1 Likelihood Estimation

The idea of efficient importance sampling, or EIS, is built upon standard importance sampling. Following the example given in Richard and Zhang (2007), suppose that we need to evaluate the following integral

$$G(\theta) = \int_{\mathbb{R}} \varphi(x|\theta) dx, \tag{2}$$

where we can decompose $\varphi(x|\theta) = p(x|\theta)g(x|\theta)$ such that $\int_{\mathbb{R}} g(x|\theta)dx = 1$. Note that this implies that g can be seen as a probability density function. Therefore, the integral in (2) can be seen as an expectation of $p(X|\theta)$, where X has the probability density function $g(x|\theta)$. Using the law of large numbers, we can then approximate this expectation by randomly sampling from X, generating $x_1, \dots, x_n$, and taking the average of $p(x_1|\theta), \dots, p(x_n|\theta)$

$$G(\theta) = \int_{\mathbb{R}} \varphi(x|\theta)dx = \int_{\mathbb{R}} p(x|\theta)g(x|\theta)dx = \mathbb{E}(p(X|\theta)) \approx \frac{1}{n}\sum_{i=1}^{n} p(x_i, \theta) \equiv \widetilde{G}_n(\theta).$$

This technique is also known as Monte Carlo integration. One of the problems with this method is that the variance of $\widetilde{G}(\theta)$ might be relatively large for small n. If it is possible to decrease the variance of $p(x_i, \theta)$, then $\widetilde{G}(\theta)$ would better approximate $G(\theta)$ in finite samples. This is the main idea of importance sampling. Note that the integral can also be rewritten in the following manner

$$G(\theta) = \int_{\mathbb{R}} \varphi(x|\theta)dx = \int_{\mathbb{R}} p(x|\theta)g(x|\theta)dx = \int_{\mathbb{R}} \frac{p(x|\theta)g(x|\theta)}{m(x|a)} m(x|a)dx =$$
$$\mathbb{E}(w(X, \theta, a)) \approx \frac{1}{n}\sum_{i=1}^{n} w(x_i, \theta, a) \equiv \widetilde{G}(\theta, a),$$

where we define $w(x, \theta, a) \equiv \frac{g(x|\theta)f(x|\theta)}{m(x|a)}$ and $x_1, \ldots, x_n$ are now sampled from the probability distribution $m(x|a)$. We define $m(x|a)$ to belong to some set of probability distributions, indexed by a. If a is chosen correctly, the variance of $\widetilde{G}$ decreases with respect to the original Monte Carlo estimator, and convergence to the true value of the integral will be faster. This is a very desirable property, since numerical optimization over the model parameters often requires us to evaluate such a integral many times. Our goal is therefore to choose a in such a way that the sample variance $\mathbb{V}(\widetilde{G}(\theta, a))$ is minimized. Richard and Zhang (2007) show that the optimal solution $\widehat{a}(\theta)$ can be well approximated as the solution to the following minimization problem

$$\begin{aligned}(\hat{a}(\theta), \hat{c}(\theta)) &= \arg\min_{a\in A, c\in\mathbf{R}} Q(a, c, \theta), \\ Q(a, c, \theta) &= \int_{\mathbb{R}} d^2(x, a, c, \theta) \cdot \varphi(x|\theta)dx, \\ d(x, a, c, \theta) &= \ln \varphi(x|\theta) - c - \ln k(x|a),\end{aligned} \tag{3}$$

where A is defined as the set of all values a is allowed to take, and $k(x|a)$ denotes the kernel of $m(x|a)$ with corresponding normalizing constant $\chi(a)$. This can be seen as a functional version of generalized least squares, where we approximate the dependent variable. Note that $Q(a, c|\theta)$ cannot be evaluated analytically in most cases, which prompts us to again approximate the integral by applying the same technique as was used in (3)

$$Q(a, c, \theta) = \int_{\mathbb{R}} d^2(x, a, c, \theta) \cdot \varphi(x|\theta)dx =$$
$$\int_{\mathbb{R}} d^2(x, a, c, \theta) \cdot w(x, \theta, a) \cdot m(x|a)dx =$$
$$\mathbb{E}(d^2(X, a, c, \theta) \cdot w(X, \theta, a)) \approx \frac{1}{n}\sum_{i=1}^{n} d^2(x_i, a, c, \theta) \cdot w(x_i, \theta, a) \equiv \widetilde{Q}(a, c, \theta),$$

where the x_i's are once again sampled from $m(x|a)$. Note that in order to approximate $Q(a, c|\theta)$ and find the optimal $\widehat{a}(\theta)$, we need to sample from $m(x|a)$, which requires an estimate for a. The essential EIS step is to iteratively optimize $Q(a, c, \theta)$ to obtain a new estimate for a, until convergence is reached. For an initial estimate of Q, $g(x|\theta)$ can be used to sample the x_i's.

This is how the problem of choosing a is solved for the case when the evaluated integral is low-dimensional. the problem can also be solved in a similar way using one auxiliary m function for higher-dimensional integrals. However, there is much efficiency to be gained when we split the integral up into multiple parts, and construct an m function for every part. Here we once more follows Richard and Zhang (2007), and use their

EIS algorithm adapted for non-linear state-space models. Define the following functions

$$\begin{aligned}
&\varphi_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\Psi}\right) \equiv p(\mathbf{y}_t|\boldsymbol{\alpha}_t)g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Psi}),\\
&m_t\left(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\Psi}\right) = \frac{k_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t, \boldsymbol{\Psi}\right)}{\chi_t\left(\boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t, \boldsymbol{\Psi}\right)},\\
&\chi_t\left(\boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t, \boldsymbol{\Psi}\right) = \int k_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t, \boldsymbol{\Psi}\right) d\boldsymbol{\alpha}_t.
\end{aligned}$$

Then the likelihood in (1) can be written in the following form (for a full derivation, we once again refer to Richard and Zhang (2007))

$$L(\mathbf{y}_{1:T}|\boldsymbol{\Psi}) = \chi_1\left(\boldsymbol{\beta}_1, \boldsymbol{\Psi}\right) \cdot \int \cdots \int \prod_{t=1}^{T} \left[\frac{\varphi_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\Psi}\right) \cdot \chi_{t+1}\left(\boldsymbol{\alpha}_t|\boldsymbol{\beta}_{t+1}, \boldsymbol{\Psi}\right)}{k_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t, \boldsymbol{\Psi}\right)}\right] m_t\left(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\Psi}\right) \cdot d\boldsymbol{\alpha}_T \cdots d\boldsymbol{\alpha}_1.$$

Since optimization over all $\boldsymbol{\beta}_t$ is infeasible for large T, a heuristic is chosen where the auxiliary minimization problem (3) is solved T times where the variance of the weights

$$\frac{\varphi_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\Psi}\right) \cdot \chi_{t+1}\left(\boldsymbol{\alpha}_t|\boldsymbol{\beta}_{t+1}, \boldsymbol{\Psi}\right)}{k_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t, \boldsymbol{\Psi}\right)}$$

is (approximately) minimized. It should be noted that choosing the functional class from which we pick m is non-trivial. We follow Liesenfeld and Richard (2003), and choose k_t to be the following k-variate Gaussian kernel

$$\begin{aligned}
&k_t(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}) \equiv g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Psi})\zeta(\boldsymbol{\alpha}_t),\\
&\zeta(\boldsymbol{\alpha}_t|\boldsymbol{\beta}_t) \equiv \exp\left(\boldsymbol{\alpha}_t'\boldsymbol{\lambda}_t + \boldsymbol{\alpha}_t'\boldsymbol{\Gamma}_t\boldsymbol{\alpha}_t\right),\\
&\boldsymbol{\beta}_t \equiv \{\boldsymbol{\lambda}_t, \boldsymbol{\Gamma}_t\}.
\end{aligned}$$

For this model, an analytic expression for the integration constant $\chi_{t+1}\left(\boldsymbol{\alpha}_t|\boldsymbol{\beta}_{t+1}, \boldsymbol{\Psi}\right)$ can also be derived, which is done in the appendix.

The optimization problem over $\boldsymbol{\beta}_t$ can then be expressed as solving the following weighted linear regression problem, where we sample $\{\tilde{\boldsymbol{\alpha}}_{t,i}, i : 1 \to S\}$ from $m_t\left(\tilde{\boldsymbol{\alpha}}_{t,i}|\tilde{\boldsymbol{\alpha}}_{t-1,i}, \boldsymbol{\beta}_t^{(j-1)}, \boldsymbol{\Psi}\right)$

$$\begin{aligned}
&\min \left\{\sum_{i=1}^{S} w_{t,i}^{(j-1)} \left(\log\left(p(\mathbf{y}_t|\tilde{\boldsymbol{\alpha}}_{t,i}, \boldsymbol{\Psi})\right) + \log\left(\chi_{t+1}\left(\tilde{\boldsymbol{\alpha}}_{t,i}|\boldsymbol{\beta}_{t+1}^{(j)}, \boldsymbol{\Psi}\right)\right) - c + \tilde{\boldsymbol{\alpha}}_{t,i}'\boldsymbol{\lambda}_t - \tilde{\boldsymbol{\alpha}}_{t,i}'\boldsymbol{\Gamma}_t\tilde{\boldsymbol{\alpha}}_{t,i}\right)^2\right\},\\
&w_{t,i}^{(j)} = \frac{\varphi_t\left(\tilde{\boldsymbol{\alpha}}_{t,i}, \tilde{\boldsymbol{\alpha}}_{t-1,i}|\boldsymbol{\Psi}\right)}{m_t\left(\tilde{\boldsymbol{\alpha}}_{t,i}|\tilde{\boldsymbol{\alpha}}_{t-1,i}, \boldsymbol{\beta}_t^{(j-1)}, \boldsymbol{\Psi}\right)},
\end{aligned} \tag{4}$$

where we define $\boldsymbol{\beta}_t^{(j)}$ to be the j^{th} iterative estimate of the auxiliary parameters. When the variance of the weights is relatively small, Richard and Zhang (2007) note that they can also be set to 1, which is equivalent to OLS. This is also done for the stochastic variance models of Liesenfeld and Richard (2003). More information about when this approximation can be justified is given in Section 3.2.2. Since this expression requires an estimate of $\boldsymbol{\beta}_{t+1}^{(j)}$, the auxiliary parameter set $\boldsymbol{\beta}_1^{(j)}, \ldots, \boldsymbol{\beta}_T^{(j)}$ is estimated using backwards recursion, beginning at time $t = T$ with initialization $\chi_{T+1}\left(\boldsymbol{\alpha}_t; \boldsymbol{\beta}_{t+1}^{(j)}\right) = 1$ and working backwards until $t = 1$. Similar to the univariate case described earlier, this procedure is then repeated multiple times until convergence of

$\boldsymbol{\beta}_1^{(j)}, \ldots, \boldsymbol{\beta}_T^{(j)}$ is reached. After that, the auxiliary samplers $m_1, \ldots, m_T$ are used to sample S realizations of $\boldsymbol{\alpha}_{1:T}$, given by $\{(\tilde{\boldsymbol{\alpha}}_{t,i}, t : 1 \to T) ; i : 1 \to S\}$ which are then plugged into the weight function. At the end, the average is calculated as an estimate of the likelihood function

$$\widehat{L}(\mathbf{y}_{1:T}|\boldsymbol{\Psi}) = \frac{1}{S}\sum_{i=1}^{S}\left[\prod_{t=1}^{T}\frac{\varphi_t\left(\boldsymbol{\alpha}_t, \boldsymbol{\alpha}_{t-1}|\boldsymbol{\Psi}\right)}{m_t\left(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\beta}_t, \boldsymbol{\Psi}\right)}\right]_{\{\boldsymbol{\alpha}_t = \tilde{\boldsymbol{\alpha}}_{t,i}\}}.$$

By applying the EIS algorithm, the likelihood of parameter-driven models can be approximated given the observations $\mathbf{y}_{1:T}$ and the model parameters $\boldsymbol{\Psi}$. Maximum likelihood can then be performed numerically over $\boldsymbol{\Psi}$.

3.2.2 Weights Simplification

As mentioned in the Section 3.2, some cases permit the weights in (4) to be set to 1, which reduces the problem from WLS to OLS. Since normally these weights are computed for every iteration of every EIS sample, we find that this simplification reduces the computational time significantly. However, in some scenarios this adaption leads to problems. For simplicity, consider diagonal variance $\mathbf{Q}$ and autoregressive parameter $\boldsymbol{\Phi}$ matrices. In this case, convergence is no longer guaranteed when the diagonal entries are chosen too large in absolute size (the exact numbers vary depending on the model of choice). This can be explained as follows. Note that the unconditional distribution for the hidden process is given by

$$\boldsymbol{\alpha}_t \sim g_k\left(\boldsymbol{\alpha}_t; \boldsymbol{\mu} = (\mathbf{I}_k - \boldsymbol{\Phi})^{-1}\boldsymbol{\omega}, \boldsymbol{\Sigma} = (\mathbf{I}_k - \boldsymbol{\Phi}^2)^{-1}\mathbf{Q}\right),$$

where we use that diagonal matrices commute, and that the time series is stationary. Here we see that an increase in the absolute values of the diagonal elements of both $\mathbf{Q}$ and $\boldsymbol{\Phi}$ leads to an increase in the unconditional variance.

During the initialization of the EIS algorithm, the initial sampler is set to the natural sampler, which implies that the weights during this initialization can be written as

$$w_{t,i}^1 = p(\mathbf{y}_t|\tilde{\boldsymbol{\alpha}}_{t,i}).$$

That is, the weights are equal to the conditional probability density of $\mathbf{y}_t$, which is a Gaussian distribution for our models. As you increase the variance of $\boldsymbol{\alpha}_t$, the probability that $\tilde{\boldsymbol{\alpha}}_{t,i}$ is a realistic parameterization of decreases, which leads to weights which are very close to zero, and should therefore have very little effect on the WLS auxiliary parameters. However, by setting all weights equal to 1, the importance shrinking does not happen. This eventually causes a large deviation in the auxiliary parameters, leading to very bad estimates of the likelihood. Because of this, keeping the weights as defined in (4) is often needed when the variance of the unconditional distribution of the hidden process is relatively large.

In some extreme cases, it could occur that the natural sampler for time t generates $\{\tilde{\boldsymbol{\alpha}}_{t,i}; i : 1 \to S\}$, such that all the weights are very close to 0. This can happen for a number of reasons, for example when $\mathbf{y}_t$ is an outlier, and none of the generated hidden variables fit the found parameter well. If this is the case, the auxiliary regression is likely to be close to singular, which will lead the EIS algorithm to fail. In order to solve this problem, we consider what values for $\boldsymbol{\alpha}_t$ are likely to occur, given the observed variable $\mathbf{y}_t$. To

answer this question, we study the posterior distribution of $\boldsymbol{\alpha}_t$ given $\mathbf{y}_t$

$$p(\boldsymbol{\alpha}_t|\mathbf{y}_t) = \frac{p(\mathbf{y}_t|\boldsymbol{\alpha}_t)p(\boldsymbol{\alpha}_t)}{p(\mathbf{y}_t)}.$$

We want to construct an auxiliary sampler $m_t^*(\boldsymbol{\alpha}_t; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, such that the mode of the posterior is in the same place as the mode of the auxiliary sampler. For a Gaussian sampler, the mean and the mode coincide, which gives us the following equation for its mean

$$\boldsymbol{\mu} = \arg\max_{\boldsymbol{\alpha}_t} \{p(\boldsymbol{\alpha}_t|\mathbf{y}_t)\} = \arg\max_{\boldsymbol{\alpha}_t} \{\log\left(p(\mathbf{y}_t|\boldsymbol{\alpha}_t)\right) + \log\left(p(\boldsymbol{\alpha}_t)\right)\},$$

where we use that the logarithm preserves order, and that $p(\mathbf{y}_t)$ does not depend on $\boldsymbol{\alpha}_t$. This can then be solved numerically. For the covariance matrix $\boldsymbol{\Sigma}$, the unconditional covariance matrix of $\boldsymbol{\alpha}_t$ suffices. If this does not hold, one can perform a similar calculation by equating the Hessian of the log posterior and log auxiliary sampler in order to obtain the covariance matrix for the sampler.

3.2.3 Filtering

In the case of the linear state-space model, the estimation of the hidden driving process $\mathbb{E}\left(\boldsymbol{\alpha}_t|\mathbf{y}_{1:t-1}\right)$ is obtainable analytically via the Kalman filter. However, this does not hold for the non-linear case. Because it is an expectation that needs to be approximated, a similar idea to EIS can be applied. Here, we follow Liesenfeld and Richard (2003), and write the filtered estimate in the following matter

$$\mathbb{E}\left(\boldsymbol{\theta}\left(\boldsymbol{\alpha}_t\right)|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi}\right) = \frac{\int \boldsymbol{\theta}\left(\boldsymbol{\alpha}_t\right) g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}, \boldsymbol{\Psi}) L(\mathbf{y}_{1:t-1}, \boldsymbol{\alpha}_{1:t-1}|\boldsymbol{\Psi}) d\boldsymbol{\alpha}_{1:t}}{\int L(\mathbf{y}_{1:t-1}, \boldsymbol{\alpha}_{1:t-1}|\boldsymbol{\Psi}) d\boldsymbol{\alpha}_{1:t-1}}. \tag{5}$$

Here the denominator is the likelihood of the first $t-1$ observations, and can therefore be approximated by means of EIS. As for the numerator, Liesenfeld and Richard (2003) show that it can be approximated by

$$\frac{1}{N}\sum_{i=1}^{N}\left\{\boldsymbol{\theta}\left(\tilde{\boldsymbol{\alpha}}_{t,i}\right)\prod_{\tau=1}^{t-1}\left[\frac{\varphi_\tau\left(\boldsymbol{\alpha}_\tau, \boldsymbol{\alpha}_{\tau-1}|\boldsymbol{\Psi}\right)}{m_\tau\left(\boldsymbol{\alpha}_\tau|\boldsymbol{\alpha}_{\tau-1}, \boldsymbol{\beta}_\tau\right)}\right]_{\{\boldsymbol{\alpha}_\tau = \tilde{\boldsymbol{\alpha}}_{\tau,i}\}}\right\},$$

where the driving variables $\{(\tilde{\boldsymbol{\alpha}}_{\tau,i}, \tau : 1 \to t-1); i : 1 \to S\}$ are generated using the auxiliary samplers $m_1, \ldots, m_{t-1}$, and $\{\tilde{\boldsymbol{\alpha}}_{t,i}; i : 1 \to S\}$ is generated from $\{g(\boldsymbol{\alpha}_t|\tilde{\boldsymbol{\alpha}}_{t-1,i}, \boldsymbol{\Psi}); i : 1 \to S\}$.

This formula holds not only for $\boldsymbol{\theta}$, but also for many other well-behaved deterministic functions $f(\boldsymbol{\alpha}_t)$ of the hidden process $\boldsymbol{\alpha}_t$. If one can show that a test statistic can be written in terms of $\boldsymbol{\alpha}_t$, then it can be calculated using the same approximation as was done above. A advantage of this method is that it gives a Bayesian estimate of the statistic (if we ignore parameter uncertainty in $\widehat{\boldsymbol{\Psi}}$). This leads to better estimates, compared to first estimating $\boldsymbol{\alpha}_t$ and plugging this in f (for non affine functions f). This is because the Bayesian estimate correctly handles the uncertainty in $\boldsymbol{\alpha}_t$. This idea is applied in Section 3.5, where we derive a method to perform Value at Risk estimation for parameter-driven models.

3.2.4 Value at Risk Estimation

If the parameter-driven model studied is used to model univariate returns $\mathrm{y}_1, \ldots, \mathrm{y}_T$, it may be interesting to estimate the conditional Value at Risk $\mathrm{VaR}_{p,t+1}$ implied by the model, which is defined as the solution to

the equation

$$\mathbb{P}(\mathrm{y}_t < \mathrm{VaR}_{p,t} \,|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) = p. \tag{6}$$

To our knowledge, the approximation of $\mathrm{VaR}_{p,t}$ has not been studied in the literature for parameter-driven models. We propose the following method of estimating this conditional Value at Risk. First of all, note that the probability in (6) can be written in the following matter

$$\begin{aligned}
&\mathbb{P}(\mathrm{y}_t < x|\boldsymbol{\Psi}, \mathrm{y}_{1:t-1}) \\
&= \int_{-\infty}^{x} p(\mathrm{y}_t|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) d\mathrm{y}_t \\
&= \int_{-\infty}^{x} \int_{\mathbb{R}^k} p(\mathrm{y}_t, \boldsymbol{\alpha}_t|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) d\boldsymbol{\alpha}_t d\mathrm{y}_t \\
&= \int_{-\infty}^{x} \int_{\mathbb{R}^k} p(\mathrm{y}_t|\boldsymbol{\alpha}_t, \boldsymbol{\Psi}) p(\boldsymbol{\alpha}_t|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) d\boldsymbol{\alpha}_t d\mathrm{y}_t \\
&= \int_{\mathbb{R}^k} p(\boldsymbol{\alpha}_t|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) \int_{-\infty}^{x} p(\mathrm{y}_t|\boldsymbol{\alpha}_t, \boldsymbol{\Psi}) d\mathrm{y}_t d\boldsymbol{\alpha}_t \\
&= \int_{\mathbb{R}^k} p(\boldsymbol{\alpha}_t|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) F_y(x|\boldsymbol{\theta}(\boldsymbol{\alpha}_t)) d\boldsymbol{\alpha}_t,
\end{aligned} \tag{7}$$

where we can interchange the integrals by Fubini's theorem, and $F_y(x|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))$ is the cumulative probability distribution for y_t with known parameters $\theta(\boldsymbol{\alpha}_t)$). For many probability distributions p, this function has a known closed form. For example, when $p(y_t|\theta(\boldsymbol{\alpha}_t))$ is a univariate Gaussian pdf, we have that

$$F_y(x|\boldsymbol{\theta}(\boldsymbol{\alpha}_t)) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x - \mu(\boldsymbol{\alpha}_t)}{\sigma(\boldsymbol{\alpha}_t)\sqrt{2}}\right)\right],$$

where erf(.) refers to the error function. We can use the filtering algorithm described in Section 3.2.3 to obtain an estimate for the integral given in (7), since

$$\mathbb{P}(\mathrm{y}_t < x|\boldsymbol{\Psi}, \mathrm{y}_{1:t-1}) = \int_{\mathbb{R}^k} p(\boldsymbol{\alpha}_t|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) F_y(x|\boldsymbol{\theta}(\boldsymbol{\alpha}_t)) d\boldsymbol{\alpha}_t = \mathbb{E}(F_y(x|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi}),$$

Using this formulation, the solution to the equation $\mathbb{P}(\mathrm{y}_t < x|\mathrm{y}_{1:t-1}, \boldsymbol{\Psi}) = p$ can be found, with solution $x = \mathrm{VaR}_{p,t}$. To solve this equation, we apply Newton's algorithm. If we approximate the cumulative distribution function $F_y(x|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))$ around the point x_0 and substitute this into the expectation above, we find

$$\mathbb{E}(F_y(x|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi}) \approx \mathbb{E}(F_y(x_0|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi}) + \mathbb{E}(p(x_0|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi})(x - x_0),$$

since the derivative of the cdf is the pdf, which exists if we assume that the cdf is strictly increasing in $\mathbb{R}$. If x_0 is close to the true value of $\mathrm{VaR}_{p,t}$, this linear approximation can then be used to iteratively solve the equation for x. If we denote x_i as the i^{th} iterative estimate, then we can obtain x_{i+1} as follows

$$\begin{aligned}
&\mathbb{E}(F_y(x_i|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi}) + \mathbb{E}(p(x_i|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi})(x_{i+1} - x_i) = p \implies \\
&x_{i+1} = x_i + \frac{p - \mathbb{E}(F_y(x_i|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi})}{\mathbb{E}(p(x_i|\boldsymbol{\theta}(\boldsymbol{\alpha}_t))|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi})},
\end{aligned}$$

where the algorithm stops when $|\mathbb{P}(\mathrm{y}_t < x_n|\boldsymbol{\Psi}, \mathrm{y}_{1:t-1}) - p| < \varepsilon$ for some previously chosen ε. After the algorithm has finished, we set $\widehat{\mathrm{VaR}}_{p,t} = x_n$. Since the pdf is positive on the support, the found value for the Value

of Risk is the unique solution. As both the probability and cumulative density functions are known functions of $\boldsymbol{\alpha}_t$, their expected values can be approximated according to filtering algorithm described in Section 3.2.3.

Due to the nature of the equation that we want to solve, a good initial estimate is essential. If the initial estimate x_0 is not close to the true Value at Risk, then $\mathbb{P}(\mathrm{y}_t < x_1 | \mathrm{y}_{1:t-1}, \boldsymbol{\Psi})$ is likely to be very close to either 0 or 1. While this theoretically is not a problem, the estimate for the derivative with respect to the Value at Risk is very close to zero, which leads to practical problems when using finite memory. As an initial choice for $x_0 \approx \mathrm{VaR}_{p,t}$, we note that the estimate probability given in (6) can be expanded as follows

$$
\begin{aligned}
&w_i \equiv \prod_{\tau=1}^{t-1} \left[\frac{\varphi_\tau \left(\boldsymbol{\alpha}_\tau, \boldsymbol{\alpha}_{\tau-1} | \boldsymbol{\Psi} \right)}{m_\tau \left(\boldsymbol{\alpha}_\tau | \boldsymbol{\alpha}_{\tau-1}, \boldsymbol{\beta}_\tau \right)} \right]_{\{\boldsymbol{\alpha}_\tau = \tilde{\boldsymbol{\alpha}}_{\tau,i}\}}, \quad \bar{w} \equiv \frac{1}{N} \sum_{i=1}^{N} w_i, \\
&p = \mathbb{P}(\mathrm{y}_t < \mathrm{VaR}_{p,t} | \boldsymbol{\Psi}, \mathrm{y}_{1:t-1}) \approx \frac{1}{N} \sum_{i=1}^{N} F_y(\mathrm{VaR}_{p,t} | \boldsymbol{\theta}((\tilde{\boldsymbol{\alpha}}_{t,i})) \cdot w_i \approx \\
&\frac{1}{N} \sum_{i=1}^{N} F_y \left(\mathrm{VaR}_{p,t} | \boldsymbol{\theta} \left(\frac{1}{N\bar{w}} \sum_{j=1}^{N} w_j \tilde{\boldsymbol{\alpha}}_{t,j} \right) \right) w_i = \bar{w} F_y \left(\mathrm{VaR}_{p,t} | \boldsymbol{\theta} \left(\frac{1}{N\bar{w}} \sum_{j=1}^{N} w_j \tilde{\boldsymbol{\alpha}}_{t,j} \right) \right) \implies \\
&\mathrm{VaR}_{p,t} \approx F_y^{-1} \left(\frac{p}{\bar{w}} \mid \boldsymbol{\theta} \left(\frac{1}{N\bar{w}} \sum_{j=1}^{N} w_j \tilde{\boldsymbol{\alpha}}_{t,j} \right) \right),
\end{aligned}
$$

with F_y^{-1} being the inverse cdf, which exists because the cdf is increasing on the support. Here the approximation is obtained by expanding the cdf such that the sum over the linear term becomes zero. We found that this approximation works very well in practice, only requiring a few iteration of Newton's algorithm until convergence was reached.

3.3 Bellman Filter

The EIS method is a suitable technique to approximate the likelihood of parameter-driven models. However, the regression step that needs to be performed at every time point for multiple EIS iterations are computationally expensive. Recent developments in the field by Lange (2020) have yielded a different method to approximate the likelihood, and filtrate the hidden process. This method is referred to as the Bellman filter, as it is based on techniques from Bellman dynamic programming (Bellman, 1957). In a nutshell, the method relies on finding a good filtration $\widetilde{\boldsymbol{\alpha}}_{1|1}, \ldots, \widetilde{\boldsymbol{\alpha}}_{T|T}$ to approximate the likelihood, which is found by generalizing the Kalman Filter (Kalman, 1960). We give a short description of the technique to capture its idea, but leave a full motivation to Lange (2020).

In order to describe how the filtered estimates are obtained, we first introduce the value function $V_t(\boldsymbol{\alpha}_t)$, which is defined as

$$
\begin{aligned}
V_t \left(\boldsymbol{\alpha}_t \right) &\equiv \max_{\boldsymbol{\alpha}_{1:t-1} \in \mathbb{R}^{k \times (t-1)}} \quad L(\mathbf{y}_{1:T}, \boldsymbol{\alpha}_{1:T} | \boldsymbol{\Psi}), \\
\widetilde{\boldsymbol{\alpha}}_{t|t} &\equiv \arg\max_{\boldsymbol{\alpha}_t \in \mathbb{R}^k} V_t \left(\boldsymbol{\alpha}_t \right),
\end{aligned}
$$

which can be shown to be equivalent to maximizing

$$V_t(\boldsymbol{\alpha}_t) = p(\mathbf{y}_t|\boldsymbol{\alpha}_t) + \max_{\boldsymbol{\alpha}_{t-1}\in\mathbb{R}^k} \{g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}) + V_{t-1}(\boldsymbol{\alpha}_{t-1})\}.$$

The values $\widetilde{\boldsymbol{\alpha}}_{1|1}, \ldots, \widetilde{\boldsymbol{\alpha}}_{T|T}$ can then be found by the following iterative process

$$\begin{aligned}
&\widetilde{\boldsymbol{\alpha}}_{t|t-1} = \boldsymbol{\omega} + \boldsymbol{\Phi}\widetilde{\boldsymbol{\alpha}}_{t-1|t-1},\\
&\boldsymbol{I}_{t|t-1} = \mathbf{Q}^{-1} - \mathbf{Q}^{-1}\boldsymbol{\Phi}\left(\boldsymbol{I}_{t-1|t-1} + \boldsymbol{\Phi}'\mathbf{Q}^{-1}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}'\mathbf{Q}^{-1},\\
&\widetilde{\boldsymbol{\alpha}}_{t|t}^{(k+1)} = \widetilde{\boldsymbol{\alpha}}_{t|t}^{(k)} + \left\{\boldsymbol{I}_{t|t-1} - \mathbb{E}\left[\frac{\partial^2 p(\mathbf{y}_t|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}|\boldsymbol{\alpha}\right]\right\}^{-1}\left\{\frac{\partial p(\mathbf{y}_t|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} - \boldsymbol{I}_{t|t-1}\left(\boldsymbol{\alpha} - \widetilde{\boldsymbol{\alpha}}_{t|t-1}\right)\right\}\Bigg|_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}_{t|t}^{(k)}},\\
&\widetilde{\boldsymbol{\alpha}}_{t|t} = \widetilde{\boldsymbol{\alpha}}_{t|t}^{(k_{\max})},\\
&\boldsymbol{I}_{t|t} = \boldsymbol{I}_{t|t-1} - \mathbb{E}\left[\frac{\partial^2 p(\mathbf{y}_t|\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}'}|\boldsymbol{\alpha}\right]_{\boldsymbol{\alpha}=\widetilde{\boldsymbol{\alpha}}_{t|t}},
\end{aligned}$$

which can be seen as a generalized version of the Kalman filter, where the updating step is done by maximizing $V_t(\boldsymbol{\alpha}_t)$ by performing gradient ascent. Note that minus the expectation of the hessian is equal to the information matrix, which is often available in closed form.

When the filtered estimates $\widetilde{\boldsymbol{\alpha}}_{1|1}, \ldots, \widetilde{\boldsymbol{\alpha}}_{T|T}$ are obtained, the likelihood can be approximated as follows

$$\begin{aligned}
&L(\mathbf{y}_{1:T}|\boldsymbol{\Psi}) = \sum_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \boldsymbol{\Psi}) \approx\\
&\sum_{t=1}^{T} p\left(\mathbf{y}_t|\widetilde{\boldsymbol{\alpha}}_{t|t}, \boldsymbol{\Psi}\right) - \frac{1}{2}\log\det\left(\boldsymbol{I}_{t|t}\right) + \frac{1}{2}\log\det\left(\boldsymbol{I}_{t|t-1}\right) - \frac{1}{2}\left(\widetilde{\boldsymbol{\alpha}}_{t|t} - \widetilde{\boldsymbol{\alpha}}_{t|t-1}\right)'\boldsymbol{I}_{t|t-1}\left(\widetilde{\boldsymbol{\alpha}}_{t|t} - \widetilde{\boldsymbol{\alpha}}_{t|t-1}\right).
\end{aligned}$$

For a motivation of the approximation, we once more refer to Lange (2020). This likelihood approximation can then be maximized by numerical methods to obtain a ML estimate for $\boldsymbol{\Psi}$. An advantage to this technique is that the gradient ascent iterations are often much faster than the regression step. Furthermore, only one filtration iteration needs to be performed, compared to multiple EIS iterations. Apart from the significant decrease in estimation time, there is another advantage to the Bellman filter. Testing on multiple multivariate models indicate that it is able to give a valid likelihood evaluation for every (stationary) parameter specification. This is in contrast with the EIS procedure, which has likelihood convergence problems when the unconditional variance of the driving process becomes too large (see also Section 3.2.2).

4 Observation-Driven Models

4.1 Model Introduction

The second model type that we investigate is the class of observation-driven models. These are very similar to the previously discussed parameter-driven models, with the main difference being that the driving process is no longer unobserved, but a deterministic function of the previous observations. Notable members of this class includes the ARCH/GARCH model families, as well as the family of dynamic exponential models. For more examples, we refer to Creal et al. (2013). In mathematical terms, define $\mathbf{y}_t$ as the observed dependent

variables, with $\boldsymbol{\alpha}_t$ as the (observed) driving process. Then observation-driven models can be written in the following form

$$\begin{aligned}&\mathbf{y}_t|\boldsymbol{\alpha}_t \sim p(\mathbf{y}_t|\boldsymbol{\alpha}_t),\\&\boldsymbol{\alpha}_t \text{ is known at time } t-1.\end{aligned}$$

4.2 Generalized Autoregressive Score Models

A more specific subclass of the model type described in the section before is denoted as generalized autoregressive score models, first introduced by Creal et al. (2013) and Harvey (2013). The main defining feature of these models is that the observed driving process $\boldsymbol{\alpha}_t$ depends on a linear transformation of the score, as well as on earlier values of $\boldsymbol{\alpha}_t$.

$$\begin{aligned}&\boldsymbol{\alpha}_{t+1} = \boldsymbol{\omega} + \sum_{i=1}^{p} \mathbf{A}_i \boldsymbol{s}_{t-i+1} + \sum_{i=1}^{q} \mathbf{B}_j \boldsymbol{\alpha}_{t-j+1},\\&\boldsymbol{s}_t = \boldsymbol{S}_t \cdot \boldsymbol{\nabla}_t, \quad \boldsymbol{\nabla}_t = \frac{\partial \log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t\right)\right)}{\partial \boldsymbol{\alpha}_t}, \quad \boldsymbol{S}_t = \boldsymbol{S}\left(t, \boldsymbol{\alpha}_t\right),\end{aligned} \tag{8}$$

where $\boldsymbol{S}_t$ is some matrix function known at time $t-1$. Depending on the choices of p and q, as well as the scaling matrix $\boldsymbol{S}_t$, different dynamics can be captured. Creal et al. (2013) and Harvey (2013) mention that natural choices for the scaling matrix depend on the Fisher information matrix $\boldsymbol{\mathcal{I}}_{t|t-1} = \mathbb{E}_{t-1}\left[\boldsymbol{\nabla}_t \boldsymbol{\nabla}_t'\right]$. For example, they note that when $\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t|t-1}^{-1}$, the model encompasses the GARCH model from Engle (1982) and Bollerslev (1986), as well as multiple others. For our applications, we only consider the case where $p = q = 1$, such that the specification can be written as

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\omega} + \mathbf{A}\boldsymbol{s}_t + \mathbf{B}\boldsymbol{\alpha}_t.$$

One of the more interesting choices for our applications is to set the scaling matrix equal to the inverse square root of the information matrix, e.g

$$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t|t-1}, \quad \boldsymbol{\mathcal{J}}_{t|t-1}' \boldsymbol{\mathcal{J}}_{t|t-1} \equiv \boldsymbol{\mathcal{I}}_{t|t-1}^{-1},$$

where $\boldsymbol{\mathcal{J}}_{t|t-1}$ can be seen as a Cholesky factor of $\boldsymbol{\mathcal{I}}_{t|t-1}^{-1}$. An advantage of this specification is that it gives the scaled score $\boldsymbol{s}_t$ an identity covariance matrix

$$\mathbb{V}_{t-1}(s_t) = \boldsymbol{\mathcal{J}}_{t|t-1} \mathbb{V}_{t-1}(\boldsymbol{\nabla}_t) \boldsymbol{\mathcal{J}}_{t|t-1}' = \boldsymbol{\mathcal{J}}_{t|t-1} \boldsymbol{\mathcal{I}}_{t|t-1} \boldsymbol{\mathcal{J}}_{t|t-1}' = \boldsymbol{\mathcal{J}}_{t|t-1} \boldsymbol{\mathcal{J}}_{t|t-1}^{-1} (\boldsymbol{\mathcal{J}}_{t|t-1}')^{-1} \boldsymbol{\mathcal{J}}_{t|t-1}' = \mathbf{I}_k,$$

where we use that the score has conditional expectation $\mathbf{0}_k$, and that the scaling matrix $\boldsymbol{\mathcal{J}}_{t|t-1}$ is invertible if and only if the information matrix is invertible.

4.3 Parameter Estimation

Parameter estimation for GAS models is straightforward. For ease of notation we assume that the observed driving variables $\boldsymbol{\alpha}_0$ are deterministic and known at time $t = 0$. This can be done by requiring it to be equal to the unconditional mean. In the cases that we study, we have that $\boldsymbol{\alpha}_{t+1}$ is a deterministic function of $y_1, \ldots, y_t$ (which can be proven by induction using the recursion given in (8)). We can then split up the log

likelihood in the following fashion

$$\log(L(\mathbf{y}_{1:T}|\mathbf{\Psi})) = \sum_{t=1}^{T} \log(L(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \mathbf{\Psi})) = \sum_{t=1}^{T} \log(p(\mathbf{y}_t|\boldsymbol{\alpha}_t)).$$

Using the GAS filter described in (9), $\boldsymbol{\alpha}_{1:T}$ can be found. After that, the likelihood can be evaluated analytically. Repeating these steps many times for different values of $\mathbf{\Psi}$, numerical maximization of the likelihood can be performed to find the approximate ML parameter estimate $\widehat{\mathbf{\Psi}}$.

The fact that the likelihood can be calculated analytically is an advantage observation-driven models have over parameter-driven models, where the likelihood is often not available in closed form. Furthermore, the analytic availability of the likelihood should make numerical optimization faster, especially for the case when the number of model parameters which need to be estimated is relatively large

4.4 Value at Risk Estimation

Now suppose that we are interested in some conditional $p \times 100\%$ Value at Risk estimate $\text{VaR}_{p,t}$, but this time y_t follows a known GAS specification. As is the case with maximum likelihood estimation, this estimation is more straightforward for observation-driven models compared to parameter-driven models. Noting that $\boldsymbol{\alpha}_t$ is deterministic in $\text{y}_{1:t-1}$, it holds that

$$\mathbb{P}(\text{y}_t < \text{VaR}_{p,t} \,|\text{y}_{1:t-1}, \mathbf{\Psi}) = \int_{-\infty}^{\text{VaR}_{p,t}} p(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1}, \mathbf{\Psi}) =$$
$$F_y(\text{VaR}_{p,t} \,|\boldsymbol{\alpha}_t, \mathcal{F}_{t-1}, \mathbf{\Psi}),$$

where we define F_y as the conditional CDF of y_t. As mentioned earlier, this is often available in closed form. A clear advantage of this type of model is that the Value at Risk can be calculated much faster, compared to the parameter-driven models. An application of both methods is given in Section 7.2.

5 Observation Density Choices

In this section we define the observation densities $p(\mathbf{y}_t|\boldsymbol{\alpha}_t)$ which are studied in the paper. Here the conditional densities of the dependent variables will be given. Note that for every observation density, two different general implementations are possible, a parameter-driven as well as an observation-driven implementation. While this choice does not influence the conditional density, it does affect the density of the driving process $\boldsymbol{\alpha}_{1:T}$. How these variables are then updated is given in Section 3.1 and Section 4.2 for parameter- and observation-driven specifications respectively.

5.1 Stochastic Mean Variance

The first model that we study is dubbed the stochastic mean variance (SMV) model. As the name implies, the main feature of this model is a time-varying mean and variance. Separate implementations are widely known, such as the random walk model for a time-varying mean, and the SV models for time-varying stochastic variance. However, combined specifications are rarely seen in the literature. We do believe that these models are relevant to study, as they are able to capture the proven correlation between high volatility

and negative returns (also see Giot (2005)).

Let us first define the time-varying components. Let μ_t and σ_t^2 respectively be the conditional mean and volatility of the dependent variable y_t at time t. Given these driving factors, we assume that y_t is normally distributed. In mathematical terms, this implies that

$$p(y_t|\mu_t, \sigma_t^2) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left\{-\frac{(y_t - \mu_t)^2}{2\sigma_t^2}\right\}.$$

For the parameter-driven variant, the dynamics of μ_t and σ_t^2 need to be described explicitly. Similar to the parameter-driven models described in Koopman, Lucas and Scharth (2016), in order to restrict σ_t^2 to always be positive, we model the joint dynamics of μ_t and $\log(\sigma_t^2)$ as the VAR(1) model

$$\begin{bmatrix} \mu_{t+1} \\ \log(\sigma_{t+1}^2) \end{bmatrix} = \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} \mu_t \\ \log(\sigma_t^2) \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}.$$

Here it is assumed that the VAR specification given above is stationary. The covariance matrix $\mathbf{Q}$ is also allowed to be non-diagonal. For the observation-driven variant, the dynamics depend on the choice of scaling matrix $\boldsymbol{S}_t$, as well as the transition matrices $\mathbf{A}$ and $\mathbf{B}$. Note that the score and corresponding information matrix are available in closed form (which are derived in the appendix)

$$\boldsymbol{\nabla}_t = \begin{bmatrix} (y_t - \mu_t)\exp\{-\log(\sigma_t^2)\} \\ -\frac{1}{2} + \frac{1}{2}(y_t - \mu_t)^2 \exp\{-\log(\sigma_t^2)\} \end{bmatrix}, \; \boldsymbol{\mathcal{I}}_{t|t-1} = \begin{bmatrix} \exp\{-\log(\sigma_t^2)\} & 0 \\ 0 & \frac{1}{2}. \end{bmatrix}. \tag{9}$$

Since this information matrix is diagonal and invertible, real powers p of this matrix are well defined, and are given by raising the diagonal elements to the power p. For choices $p \in \{-1, -\frac{1}{2}, 0\}$, we have the following expression for the scaled score

$$\boldsymbol{s}_t = \begin{bmatrix} (y_t - \mu_t)\exp\{-(p+1)\log(\sigma_t^2)\} \\ -\left(\frac{1}{2}\right)^{1+p} + \left(\frac{1}{2}\right)^{1+p}(y_t - \mu_t)^2 \exp\{-\log(\sigma_t^2)\} \end{bmatrix} \equiv \begin{bmatrix} h_1(\bar{d}_1) \\ h_2(\bar{d}_2) \end{bmatrix}.$$

Note that $\boldsymbol{s}_t$ depends only on $\bar{d}_1 = (y_t - \mu_t)$ and $\bar{d}_2 = (y_t - \mu_t)^2$, as well as μ_t and $\log(\sigma_t^2)$. Here $\bar{d}_1$ gives the total deviation from the mean, and $\bar{d}_2$ can be seen as the squared deviation of the mean. Both can be seen as a shock to the mean and volatility respectively. Furthermore, note that the first and second element of the score are both positive affine functions $h_i(\bar{d}_i)$ of these variables respectively. This implies that shock to the mean is fully contained in the first element, and the shock to the variance is fully contained in the second element.

In order to obtain the final effect that the observed variable y_t has on the new driving variables $\boldsymbol{\alpha}_{t+1}$, the scaled score is premultiplied with the parameter matrix $\mathbf{A}$. Writing the i^{th} element of the score as an affine function $h_i(\bar{d}_i)$, the effect of the scaled score on the updated driving variables is given by the following expression

$$\mathbf{A}\boldsymbol{s}_t = \begin{bmatrix} a_{11}h_1(\bar{d}_1) + a_{12}h_1(\bar{d}_1) \\ a_{21}h_2(\bar{d}_2) + a_{22}h_2(\bar{d}_2) \end{bmatrix}, \tag{10}$$

where a_{ij} is element i, j of $\mathbf{A}$. This shows that the values on the diagonal describe how a shock to the mean or variance updates the new mean or variance respectively. Furthermore, the off-diagonal elements describe the cross effects. For example, if we hypothesize that a large negative shock increases the volatility more than a large positive shock, then we would expect a_{21} to be negative.

5.2 Multivariate Stochastic Volatility Models

In this section we consider two different multivariate conditional volatility models. Both can be seen as multivariate extensions of the stochastic volatility model. For these models, we assume that $\mathbf{y}_t$ is a N dimensional vector of observed stochastic variables, with the following distribution

$$\mathbf{y}_t|\boldsymbol{\alpha}_t \sim g_N(\boldsymbol{\mu} = \mathbf{0}_N, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\alpha}_t)), \tag{11}$$

where g_N denotes the N-variate Gaussian distribution with mean $\mathbf{0}_N$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\alpha}_t)$. Note that $\boldsymbol{\Sigma}(\boldsymbol{\alpha}_t)$ is some positive semidefinite matrix function of the driving process $\boldsymbol{\alpha}_t$. Here some freedom exists in choosing the transformation $\boldsymbol{\Sigma}(\boldsymbol{\alpha}_t)$. The implementations that we study are based on existing literature on multivariate GARCH models, adapted to parameter-driven models.

5.2.1 Stochastic Conditional Correlation

The first model of the multivariate stochastic volatility class that we introduce is denoted as the stochastic conditional correlation, or SCC model. In this model the individual volatility and the correlations are modeled separately, which is done using a similar decomposition as is done in DCC Garch (also see Bauwens, (2006)). For our applications, we mainly consider a bivariate observed series $\mathrm{y}_{t,i}$, such that $N = 2$. Denote the time-varying volatility of the i^{th} observed variable $\mathrm{y}_{t,i}$ as $\sigma^2_{t,i}$. Furthermore, define ρ_t as the correlation between $\mathrm{y}_{t,1}$ and $\mathrm{y}_{t,2}$. Then the covariance matrix in (11) can be decomposed in the following manner

$$\boldsymbol{\Sigma}_t = \boldsymbol{H}_t \boldsymbol{R}_t \boldsymbol{H}_t,$$

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma^2_{t,1} & \sigma_{t,1}\sigma_{t,2}\rho_t \\ \sigma_{t,1}\sigma_{t,2}\rho_t & \sigma^2_{t,2} \end{bmatrix}, \boldsymbol{H}_t = \begin{bmatrix} \sigma_{t,1} & 0 \\ 0 & \sigma_{t,2} \end{bmatrix}, \boldsymbol{R}_t = \begin{bmatrix} 1 & \rho_t \\ \rho_t & 1 \end{bmatrix}.$$

We then model the volatilities and correlations as separate variables using transformed driving factors in order to ensure that the relevant variables are within bounds. Since we must have that $\sigma_{t,1,1}, \sigma_{t,2,2} > 0$, we apply the following transformation

$$\begin{bmatrix} \sigma_{t,1} \\ \sigma_{t,2} \end{bmatrix} = \begin{bmatrix} \exp\{\alpha_{t,1}\} \\ \exp\{\alpha_{t,3}\} \end{bmatrix}.$$

In order to ensure that these lie (strictly) between -1 and 1, we also apply the transformation $\rho = \tanh\{\alpha_{t,2}\}$.

For the parameter-driven variants, these models are usually too large to estimate without restrictions. Therefore we only consider diagonal implementations of $\boldsymbol{\Phi}$ and $\mathbf{Q}$. Because the likelihood of observation-driven models is available in closed form, this restriction does not apply to these specifications.

5.2.2 Stochastic Exponential Volatility

For the second stochastic volatility model, we define the Stochastic Exponential Volatility, or SEV model. This model is inspired by similar work done on matrix exponential GARCH models by Kawakatsu (2006). One of the main challenges in modeling stochastic volatility is keeping the time-series within its allowed bounds. When $N = 1$, we only require that the volatility is positive. However, for multivariate stochastic volatility models, we also require that the volatility process $\boldsymbol{\Sigma}(\boldsymbol{\alpha}_t)$ is positive definite for every t. In the univariate case, taking the exponent of $\boldsymbol{\alpha}_t$ is most often used to ensure the desired conditions. This idea can also be extended to the multivariate case by means of the matrix exponential, defined as follows

$$e^{\mathbf{A}} \equiv \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!}.$$

Similar to the univariate case, it can be shown that this sum converges for all square matrices $\mathbf{A}$ (in the sense that its norm is finite). Once more assume that the N-dimensional observed dependent variable $\mathbf{y}_t$ has a conditional covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\alpha}_t)$. Furthermore, define the vectorization operator $\text{vec}(\boldsymbol{\Theta}_t)$ such that it takes in an symmetric $N \times N$ matrix, and returns an $N(N+1)/2$ vector containing the matrix elements of the upper diagonal (including the diagonal itself) stacked from left to right. Then, we define the dependence between $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\alpha}_t$ as follows

$$\begin{aligned} \boldsymbol{\Sigma}_t &= e^{\boldsymbol{\Theta}_t}, \\ \boldsymbol{\Theta}_t &\equiv \text{vec}^{-1}(\boldsymbol{\alpha}_t), \end{aligned}$$

where we define $\text{vec}^{-1}(\boldsymbol{\alpha}_t)$ to be the inverse function of $\text{vec}(\boldsymbol{\Theta}_t)$. For symmetric $\boldsymbol{\Theta}_t$, it can be shown that $e^{\boldsymbol{\Theta}_t}$ is positive definite for all symmetric matrices $\boldsymbol{\Theta}_t$. For a proof, we refer to Kawakatsu (2006). This proof also gives a way of efficiently computing the symmetric matrix exponential by means of the spectral theorem.

In this paper, we mainly study the case where $N = 2$. As this implies a 3-dimensional driving process, EIS likelihood maximization becomes computationally infeasible without parameter restrictions. Because of this, we only consider diagonal specifications for the parameter-driven implementations. Since the observation-driven implementation is substantially less computationally intensive to estimate, this restriction does not apply to this model class. For the case where $N = 2$, both the score as well as the information are available in closed form. We refer to the appendix for a full derivation, as it is quite extensive.

5.3 Fitting GAS to Parameter-Driven Specifications

As mentioned in the introduction, one of the questions that we answer in this paper is how well parameter-driven models can be approximated by observation-driven models. This is done by fitting the observations to a generalized autoregressive score (or GAS model) introduced in Section 4.2. Note that a parameter-driven DGP implies that the driving process $\boldsymbol{\alpha}_t$ is unobserved. Following Koopman, Lucas and Scharth (2016), we can instead estimate the model as if $\boldsymbol{\alpha}_t$ follows the GAS specification given in (8).

Here some freedom exists the in the choice for the scaling matrix $\boldsymbol{S}_t$, as well as the parameterization of $\mathbf{A}, \mathbf{B}$. Common choices in the literature are based on the information matrix, as is described in Section 4.2. In earlier work by Koopman, Lucas and Scharth (2016), the choice of scaling matrix was fixed during pa-

rameter estimation, and set to the Cholesky decomposition of the inverse information matrix. An advantage of this approach is that it gives the score unit variance. Other choices for the scaling matrix are the inverse information matrix itself, as well as the identity matrix.

As for the parameterization, the most obvious candidates are a full specification and a diagonal specification. Here it is interesting to see if the intuitive idea holds that diagonal and full parameter-driven specifications are best approximated by diagonal and full observation-driven specifications, respectively.

6 Simulation Study

In this section, we apply the methodology discussed earlier, and describe a simulation study. This study contains multiple parts. First of all, we test consistent parameter estimation for both the parameter-driven as well as the observation-driven models. After that, we fit multiple variants of observation-driven specifications to sample data generated by a known parameter-driven model, and consider how well the hidden states $\boldsymbol{\alpha}_t$ are approximated. This can be regarded as a multidimensional analogue to the analysis performed in Koopman, Lucas and Scharth (2016).

6.1 Consistency Study

A desirable property for models to have is consistent parameter estimation. In a nutshell, this means that for a large enough sample size, the estimated model parameters $\widehat{\boldsymbol{\Psi}}$ will be arbitrarily close to the true model parameters $\boldsymbol{\Psi}_0$. In order to show empirical evidence for this property, we perform Monte Carlo simulation, where we repeatedly generate series according to some known model $m(\boldsymbol{\Psi}_0)$. Then, the generated data series is used to estimate the model parameters for m to obtain $m(\widehat{\boldsymbol{\Psi}})$. This is done multiple times for different sample sizes. As a indicator for consistency, we require that the mean squared error (MSE) of the parameter estimates decreases as the sample size increases. This estimation is performed for both the parameter-driven variant, as well as the observation-driven variant for all three models.

6.1.1 Choice Of Parameters

As the models considered are larger relative to those considered in Koopman, Lucas and Scharth (2016), we are unable to perform the same number of replications. This problem is mainly caused by the parameter-driven models, since T regressions with $(k+1)(k+2)/2$ variables need to be performed for every EIS iteration, which becomes computationally intensive for k larger than 1. Because of this, we perform 500 replications ($R = 500$). To keep our simulations computationally feasible, we choose our sample size $T \in \{250, 500, 1000\}$.

As reasoned in Section 3.2.2, if the variance of the hidden process becomes too large, the approximate EIS method becomes unfeasible for some model classes. In order to construct a simulation study which is computationally feasible, we are required to choose parameter combinations for which the OLS approximation in the EIS regressions performs well. Furthermore, for the larger models we choose diagonal specifications for $\boldsymbol{\Phi}$ and $\mathbf{Q}$, to further reduce the computational complexity. For the SMV model, we study a specification with a relatively large correlation of 0.8 in the innovations. The choices for the parameter-driven true specification

are given below.

SMV:

$$\boldsymbol{\omega} = \begin{bmatrix} 0.00 \\ 0.10 \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} 0.80 & 0.00 \\ 0.00 & 0.90 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0.010 & 0.008 \\ 0.008 & 0.010 \end{bmatrix},$$

SCC & SEV:

$$\boldsymbol{\omega} = \begin{bmatrix} -0.10 \\ 0.10 \\ -0.10 \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} 0.90 & 0.00 & 0.00 \\ 0.00 & 0.90 & 0.00 \\ 0.00 & 0.00 & 0.90 \end{bmatrix}, \quad \mathbf{Q} = \begin{bmatrix} 0.01 & 0.00 & 0.00 \\ 0.00 & 0.01 & 0.00 \\ 0.00 & 0.00 & 0.01 \end{bmatrix}.$$

Since the likelihood for observation-driven models is available in closed form, we do not have to worry about the earlier mentioned limitations that apply to the parameter-driven models. In concrete terms, we choose the following parameterization for our observation-driven models:

SMV:

$$\boldsymbol{\omega} = \begin{bmatrix} -0.10 \\ 0.10 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0.10 & 0.00 \\ 0.00 & 0.10 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.90 & 0.00 \\ 0.00 & 0.90 \end{bmatrix},$$

SCC & SEV:

$$\boldsymbol{\omega} = \begin{bmatrix} -0.10 \\ 0.10 \\ -0.10 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 0.10 & 0.00 & 0.00 \\ 0.00 & 0.10 & 0.00 \\ 0.00 & 0.00 & 0.10 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.90 & 0.00 & 0.00 \\ 0.00 & 0.90 & 0.00 \\ 0.00 & 0.00 & 0.90 \end{bmatrix}.$$

6.1.2 Results

Using the parameterizations described in Section 6.1.1, we find 500 parameter estimates for the 6 different models. The bias and RMSE of these results are formatted in Tables 1 to 3.

Table 1: Monte Carlo parameter statistics for the SMV models.

Parameter-Driven SMV	Bias			RMSE		
Parameter	$T = 250$	$T = 500$	$T = 1000$	$T = 250$	$T = 500$	$T = 1000$
μ_1	-0.0032	-0.0006	-0.0009	0.0505	0.0337	0.0235
μ_2	-0.0892	-0.0556	-0.0316	0.1602	0.1170	0.0817
$\Phi_{1,1}$	-0.2583	-0.1810	-0.0986	0.4624	0.3443	0.2347
$\Phi_{2,2}$	-0.2871	-0.2089	-0.1253	0.5465	0.4350	0.3268
$Q_{1,1}$	0.0179	0.0142	0.0089	0.0253	0.0223	0.0176
$Q_{1,2}$	-0.0096	-0.0065	-0.0046	0.0214	0.0161	0.0109
$Q_{2,2}$	0.0180	0.0135	0.0098	0.0247	0.0214	0.0188

Observation-Driven SMV	Bias			RMSE		
Parameter	$T = 250$	$T = 500$	$T = 1000$	$T = 250$	$T = 500$	$T = 1000$
μ_1	-0.0080	-0.0046	-0.0020	0.0989	0.0696	0.0524
μ_2	-0.0247	-0.0150	-0.0041	0.1555	0.1132	0.0782
$A_{1,1}$	0.0005	-0.0003	0.0001	0.0335	0.0234	0.0155
$A_{2,2}$	0.0025	0.0029	0.0026	0.0771	0.0449	0.0285
$B_{1,1}$	-0.0477	-0.0219	-0.0099	0.1321	0.0706	0.0327
$B_{2,2}$	-0.1820	-0.0794	-0.0298	0.3900	0.2232	0.1181

Table 2: Monte Carlo parameter statistics for the SCC models.

Parameter-Driven SCC	Bias			RMSE		
Parameter	$T = 250$	$T = 500$	$T = 1000$	$T = 250$	$T = 500$	$T = 1000$
μ_1	-0.0098	-0.0061	-0.0031	0.1061	0.0764	0.0533
μ_2	0.0063	0.0106	0.0057	0.0940	0.0661	0.0477
μ_3	-0.0140	-0.0040	-0.0049	0.1146	0.0781	0.0562
$\Phi_{1,1}$	-0.1223	-0.0889	-0.0595	0.1632	0.1424	0.1138
$\Phi_{2,2}$	-0.1288	-0.1003	-0.0515	0.1788	0.1597	0.1121
$\Phi_{3,3}$	-0.1253	-0.0854	-0.0658	0.1685	0.1384	0.1204
$\mathrm{Q}_{1,1}$	0.0056	0.0061	0.0055	0.0132	0.0135	0.0121
$\mathrm{Q}_{2,2}$	0.0097	0.0110	0.0059	0.0209	0.0228	0.0153
$\mathrm{Q}_{3,3}$	0.0078	0.0057	0.0059	0.0150	0.0130	0.0128

Observation-Driven SCC	Bias			RMSE		
Parameter	$T = 250$	$T = 500$	$T = 1000$	$T = 250$	$T = 500$	$T = 1000$
μ_1	-0.0115	-0.0002	-0.0004	0.1611	0.1072	0.0734
μ_2	0.0131	0.0069	0.0035	0.1250	0.0807	0.0593
μ_3	-0.0098	-0.0028	-0.0010	0.1487	0.1004	0.0691
$\mathrm{A}_{1,1}$	-0.0007	0.0004	0.0006	0.0456	0.0309	0.0230
$\mathrm{A}_{2,2}$	-0.0021	0.0003	0.0012	0.0358	0.0250	0.0178
$\mathrm{A}_{3,3}$	-0.0058	-0.0006	0.0005	0.0476	0.0316	0.0243
$\mathrm{B}_{1,1}$	-0.1187	-0.0437	-0.0177	0.3116	0.1540	0.0753
$\mathrm{B}_{2,2}$	-0.0699	-0.0290	-0.0125	0.1919	0.0909	0.0412
$\mathrm{B}_{3,3}$	-0.1047	-0.0316	-0.0135	0.2893	0.1030	0.0510

Table 3: Monte Carlo parameter statistics for the SEV models.

Parameter-Driven SEV	Bias			RMSE		
Parameter	$T = 250$	$T = 500$	$T = 1000$	$T = 250$	$T = 500$	$T = 1000$
μ_1	-0.0197	-0.0194	-0.0116	0.1117	0.0804	0.0537
μ_2	-0.0009	-0.0013	-0.0008	0.0937	0.0631	0.0454
μ_3	-0.0224	-0.0129	-0.0128	0.1118	0.0733	0.0548
$\Phi_{1,1}$	-0.1269	-0.0971	-0.0725	0.1624	0.1452	0.1239
$\Phi_{2,2}$	-0.1499	-0.1143	-0.0669	0.1968	0.1712	0.1226
$\Phi_{3,3}$	-0.1401	-0.1146	-0.0771	0.1750	0.1562	0.1274
$\mathrm{Q}_{1,1}$	0.0068	0.0072	0.0074	0.0140	0.0141	0.0144
$\mathrm{Q}_{2,2}$	0.0125	0.0121	0.0090	0.0273	0.0246	0.0188
$\mathrm{Q}_{3,3}$	0.0087	0.0075	0.0079	0.0161	0.0150	0.0153

Observation-Driven SEV	Bias			RMSE		
Parameter	$T = 250$	$T = 500$	$T = 1000$	$T = 250$	$T = 500$	$T = 1000$
μ_1	-0.0237	-0.0056	-0.0032	0.1515	0.1032	0.0705
μ_2	0.0072	0.0045	0.0023	0.1292	0.0854	0.0622
μ_3	-0.0246	-0.0099	-0.0053	0.1531	0.1022	0.0708
$\mathrm{A}_{1,1}$	-0.0082	0.0007	0.0013	0.0615	0.0360	0.0273
$\mathrm{A}_{2,2}$	-0.0042	-0.0014	0.0002	0.0422	0.0298	0.0220
$\mathrm{A}_{3,3}$	-0.0086	-0.0010	0.0008	0.0579	0.0365	0.0268
$\mathrm{B}_{1,1}$	-0.1828	-0.0727	-0.0252	0.4118	0.2169	0.0872
$\mathrm{B}_{2,2}$	-0.0871	-0.0304	-0.0143	0.2415	0.1012	0.0492
$\mathrm{B}_{3,3}$	-0.1651	-0.0503	-0.0227	0.4013	0.1612	0.0928

We see that for all models and all model parameters the bias as well as the root mean square error, decreases as we increase the sample size T. This is informal evidence that estimation for all model classes is consistent. We also notice that both the bias and RMSE for the parameter-driven models are often larger than their observation-driven counterpart. This is mainly because the observation-driven models have a known driving process $\boldsymbol{\alpha}_t$, while the driving process of parameter-driven models is hidden. This extra uncertainty leads to a larger error in the estimated parameters. Another effect is that the likelihood for parameter-driven processes does not have a closed form, which is why it has to be estimated. This estimation also increases the total error made during parameter estimation.

For the parameter-driven SMV model, we see that especially the autoregressive component $\boldsymbol{\Phi}$ has a relatively large bias and RMSE compared to the other parameter-driven models. This can be caused by multiple factors. First of all, in order for the simulation to be viable, the autoregressive parameters were not allowed to be too close to 1 in absolute value. Because of this, the hidden process is more noisy compared to the other models, which makes it harder to identify the model parameters. Another reason might be more closely related to the nature of the model itself. A large upwards shock can be caused by either a upwards movement of μ_t or σ_t^2. Since the signal itself is hidden, it is difficult to identify the cause of the shock. This uncertainty in the hidden process would therefore once again lead to larger errors in the parameter estimates.

Overall, we find that both models show signs of being asymptotically consistent, although this convergence is much slower for the the parameter-driven specifications compared to the observation-driven specifications. Furthermore, we found that the observation-driven models were much quicker to estimate, mainly because of the closed form likelihood.

6.2 Approximation Study

Following Koopman, Lucas and Scharth (2016), we perform a simulation study to find how well parameter-driven models can be approximated by similar observation-driven model. Furthermore, we also study how well parameter estimates which are found by using the Bellman filter perform. Here a sample is generated multiple times according to a known specification. First, we set this known data generating process equal to a parameter-driven model. Using EIS to approximate the likelihood, an estimate is made of the model parameters $\widehat{\boldsymbol{\Psi}}$, and the filter described in Section 3.2.3 is used to estimate the hidden driving process $\hat{\boldsymbol{\alpha}}_t^{(PD)}|\mathbf{y}_{1:t-1}$. This is also done for the Bellman filter. The same generated data is then used to estimate the parameters of a GAS specification. Using the GAS filter, $\hat{\boldsymbol{\alpha}}_t^{(OD)}|\mathbf{y}_{1:t-1}$ can be found under this specification.

For the parameters of the true specification, we choose the the same generating processes as described in Section 6.1.1 for our parameter-driven model. Furthermore, we also use sample sizes of $T \in \{500, 1000, 2000\}$. Here, half of the sample is used to estimate the model parameters, such that the other half can be used to analyze out-of-sample performance. This is done for $R = 500$ repetitions.

One of the desirable properties of this type of study is that the normally unobserved driving process $\boldsymbol{\alpha}_t$ is known. Because of this, comparisons in quality between two estimates can be made. We quantify this as the total mean squared predication error. Assuming we have R repetitions each containing T observations, we estimate the MSPE for model m as follows

$$\widehat{\text{MSPE}}^{(m)} = \frac{1}{kRT}\sum_{r=1}^{R} \text{Tr}\left(\boldsymbol{\Xi}_r \boldsymbol{\Xi}_r'\right),$$
$$\boldsymbol{\Xi}_r^{(m)} \equiv \left[\hat{\boldsymbol{\alpha}}_{1,r}^{(m)}|\mathbf{y}_{0,r} - \boldsymbol{\alpha}_{1,r} \quad \dots \quad \hat{\boldsymbol{\alpha}}_{T,r}^{(m)}|\mathbf{y}_{1:T-1,r} - \boldsymbol{\alpha}_{T,r}\right],$$

where $\mathbf{y}_{t,r}$ and $\boldsymbol{\alpha}_{t,r}$ correspond respectively to the observed and hidden process at time t for replication r. This experiment is then performed for all three models described in Section 5. For all three different models, the three different choices for scaling matrix $\boldsymbol{S}_t$ are tested. As a comparison, the MSPE of the corresponding parameter-driven model is also noted. Here we analyze two parameter-driven filtrations, one obtained by using the true parameters, and one using the estimated parameters.
Another interesting question one could ask is what choice of specification for $\mathbf{A}$ and $\mathbf{B}$ leads to the best approximation. For example, a full specification for $\mathbf{A}$ and $\mathbf{B}$ is the most straightforward approach, and would also be able to capture the most general effects compared to more sparse choices. However, it could also lead to overfitting in small samples. Because of this, we also consider a diagonal specification for $\mathbf{A}$ and $\mathbf{B}$. For every choice of scaling matrix, both the full and diagonal specification for the model parameters is implemented, leading to 6 different GAS approximations in total. While a scalar model specification ($\mathbf{A} = a\mathbf{I}_k, \mathbf{B} = b\mathbf{I}_k$) would also be an option, we argue that it unlikely to perform well. This is because

for the selected models the asymmetric effects that each hidden variable has on the observed process. A scalar implementation would be more interesting if all hidden variables influenced the same aspect of the data, for example only the mean or only the variance, instead of both the mean and variance or variance and correlation.

6.2.1 Results

As described at the start of Section 6.2, we analyze the performance of different models when the data generating process is a parameter-driven specification. While we find relatively normal results, similar to the earlier work by Koopman, Lucas and Scharth (2016), we found unexpected behavior in the out-of-sample filtrations. While the in-sample filtrations obtained by a observation-driven specification were always stable, we found that this did not hold for the out-of-sample performance. In many cases, the filtered driving process became unstable. Because of this, the out-of-sample performance MSPE measure becomes impractical. Therefore, we report the in-sample MSPE, as well as the percentage of times the driving process became unstable. We measure instability of an estimate by considering how much the out-of-sample RMSE of the run deviates from the mean in-sample RMSE. Here we consider a run to become unstable if the out-of-sample RMSE is 10 standard deviations of the mean in-sample RMSE. As the RMSE of a run can be arbitrarily large with positive probability, this implies that we might classify stable runs as unstable. However, we found that in practice this did not seem to happen, as the probability of such an event is very small. Using this classification, the results of the study are reported in Table 4.

Table 4: Results of the approximation study, where we fit multiple observation-driven specification to a known parameter-driven specification. This is done for the SMV, SCC and SEV model classes. The tables contain the relative RMSE, as well as the earlier defined probability of instability of for all relevant model classes.

SMV Model	Parameter-Driven			Observation-Driven					
Relative RMSE				Diagonal			Full		
Sample Size	True Param.	EIS Est.	Bellman Est.	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$
$T = 250$	1.0000	1.4499	1.3237	1.6864	1.6445	1.6809	2.8121	2.6774	2.7905
$T = 500$	1.0000	1.2457	1.2427	1.3759	1.3969	1.3867	1.6334	1.6500	1.6810
$T = 1000$	1.0000	1.1474	1.0850	1.2715	1.2899	1.2851	1.3204	1.3143	1.3096

SMV Model	Parameter-Driven			Observation-Driven					
Prob. Instability.				Diagonal			Full		
Sample Size	True Param.	EIS Est.	Bellman Est.	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$
$T = 250$	0.0%	0.0%	0.0%	23.8%	20.0%	22.6%	64.8%	57.0%	57.8%
$T = 500$	0.0%	0.0%	0.0%	7.0%	5.4%	7.2%	32.2%	27.8%	27.4%
$T = 1000$	0.0%	0.0%	0.0%	1.4%	0.8%	0.6%	10.4%	8.8%	8.4%

SCC Model	Parameter-Driven			Observation-Driven					
Relative RMSE				Diagonal			Full		
Sample Size	True Param.	EIS Est.	Bellman Est.	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$
$T = 250$	1.0000	1.1341	1.1757	1.171	1.1512	1.1398	1.7011	1.6762	1.7186
$T = 500$	1.0000	1.0931	1.1135	1.1085	1.0960	1.0837	1.4634	1.4813	1.4760
$T = 1000$	1.0000	1.0578	1.0638	1.0638	1.0437	1.0321	1.2267	1.2416	1.2208

SCC Model	Parameter-Driven			Observation-Driven					
Prob. Instability.				Diagonal			Full		
Sample Size	True Param.	EIS Est.	Bellman Est.	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$
$T = 250$	0.0%	0.0%	0.0%	5.6%	5.6%	7.8%	83.0%	83.0%	83.6%
$T = 500$	0.0%	0.0%	0.0%	1.2%	0.6%	0.8%	60.8%	66.4%	62.2%
$T = 1000$	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	23.4%	24.6%	24.2%

SEV Model	Parameter-Driven			Observation-Driven					
Relative RMSE				Diagonal			Full		
Sample Size	True Param.	EIS Est.	Bellman Est.	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$
$T = 250$	1.0000	1.1294	1.1435	1.1528	1.1341	1.1244	1.8217	1.7787	1.7032
$T = 500$	1.0000	1.0765	1.0904	1.1081	1.1008	1.0964	1.5732	1.5626	1.5204
$T = 1000$	1.0000	1.0485	1.0487	1.0583	1.0545	1.0530	1.3009	1.2990	1.2911

SEV Model	Parameter-Driven			Observation-Driven					
Prob. Instability.				Diagonal			Full		
Sample Size	True Param.	EIS Est.	Bellman Est.	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{I}}_{t\|t-1}^{-1}$	$\boldsymbol{S}_t = \boldsymbol{\mathcal{J}}_{t\|t-1}$	$\boldsymbol{S}_t = \mathbf{I}_k$
$T = 250$	0.0%	0.0%	0.0%	12.8%	4.8%	2.4%	80.8%	82.8%	82.4%
$T = 500$	0.0%	0.0%	0.0%	3.2%	0.4%	0.4%	69.6%	68.8%	67.6%
$T = 1000$	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	30.4%	29.2%	29.6%

Let us first discuss the convergence problem. Over all model classes, we see two general patterns. First of all, the full specifications have a noticeably lower rate of convergence compared to the diagonal specifications.

We found that this is likely due to overfitting for the full models. Furthermore, it seems that due to the increased number of parameters, the full model is more likely to get stuck in a local optimum compared to the diagonal model. Secondly, as the sample size increases, the rate of convergence always decreases. This can be explained by the following reasoning: suppose that we look at finitely many possible specifications to fit to a time series, and for every specification, consider the first time that the filtration behaves unstable (which one can define as breaching some absolute limit). If this time point is included in the optimization, that specification will be linked to a very low likelihood compared to specifications which remain stable over the entire time period. As we increase the sample size T, more and more of these unstable specifications will be filtered out, until only stable specifications remain.

An interesting question to ask is why this problem of convergence did not occur for the models studied by Koopman, Lucas and Scharth (2016). We note that several factors might cause this effect. First of all, we note that all the models studied by Koopman, Lucas and Scharth (2016) had a unidimensional driving process. In general, this might lead to a simpler relation between the driving process and the observed variables compared to multidimensional models, reducing the chance of convergence problems. Furthermore, we note that because of the smaller model size, they were able to use larger sample sizes in their simulation study, further increasing the probability of convergence.

Because of the earlier mentioned convergence issues, out-of-sample MSPE is no longer an accurate measure of performance. For example, if we choose to throw all replications where convergence failed, then an unfair advantage might be given to models with a low convergence rate, which would be undesirable. Furthermore, if we choose to include these samples, then a small number of outliers is likely to have a large influence on the measured MSPE, which also would not be preferable. Because of this, we choose to report the in-sample MSPE.

In general, multiple interesting patterns can be observed in these tables. First of all, we note that the filtration obtained using the true parameterization of the parameter-driven models yields the lowest RMSE. This can be explained by the fact that the filtration $\widehat{\boldsymbol{\alpha}}_{t+1} \equiv \mathbb{E}(\boldsymbol{\alpha}_{t+1}|\mathbf{y}_{1:t})$ minimizes the expected value $\mathbb{E}\left((\widehat{\boldsymbol{\alpha}}_{t+1} - \boldsymbol{\alpha}_{t+1}|\mathbf{y}_{1:T})^2\right)$. If we exclude the true parameter specification, we see that the parameter-driven models with EIS and BF estimated model parameters often perform the best. We see that the Bellman filter is able to compete with the EIS estimated models, while only requiring a fraction of the estimation time. Furthermore, we note that the observation-driven models have a similar performance for the SCC and SEV models, and in some even outperform the parameter-driven specifications. Especially the diagonal specifications perform well for all models, while the full models perform noticeably worse. This can again be (partly) explained by the fact that the full specifications often had trouble finding the global optimum. Furthermore, the fact that the data generating process contains only diagonal matrices for the SCC and SEV models also seem to favor diagonal over full observation driven models, as this disparity is much smaller for the SMV model.

As for the estimation time, we found similar results to those found by Koopman, Lucas and Scharth (2016). Most observation-driven models were estimated between 10-100 times faster than their parameter-driven counterpart. The Bellman estimation was also significantly faster, with the parameter estimation being approximately 10-30 times faster. As multiple information matrix evaluations are needed per time point for

the Bellman filter, it is a bit slower than the observation-driven approximation. However, both are still much faster than the original EIS likelihood maximization process. Furthermore, we found that the larger models with 3 driving variables had the largest difference in estimation time. As mentioned in Section 6.1.1, the number of auxiliary EIS regressors increases quadratically. When k becomes relatively large, these auxiliary regressions become computationally expensive. Because of this, we find that that for larger k, the observation-driven models are preferable when computational intensity is taken into account.

Concerning the choice of scaling matrix, we find that there does not exist one choice which is the best for all models. For example, it seems that for the SMV model, the best choice of scaling matrix is the inverse information matrix, while for the SCC and SEV models the identity yields the lowest MSPE. Furthermore, we also find that the importance of the scaling matrix varies noticeably per model. Here we see that for the SMV and SCC model, the scaling choice has a relatively large influence on the performance, while this effect is smaller for the SEV model.

Let us consider the SMV approximation study more deeply. One aspect that stands out is that the observation-driven models perform worse compared to the other observation density choices. This is an indicator that observation-driven models are not able to approximate the off-diagonal effects in a parameter-driven specification very well. We do note that the difference in performance between the diagonal and full specification is much smaller compared to the SCC and SEV approximations. Furthermore, the RMSE of the full models also decreases quite rapidly as we increase the sample size T. We also hypothesize that for larger sample sizes, the positive effect of more model flexibility beats the negative effect of additional parameter uncertainty for the full models. This likely leads to full models outperforming the diagonal models when the data generating process contains cross effects.

It is also interesting to consider the distribution of the RMSE, and if it they are similar in shape over all the model classes. For example, it could be the case that a small percentage of outliers strongly affects the RMSE. In order to further investigate this, we consider the empirical cumulative distribution function of the RMSE for the SMV observation density and $T = 1000$. Here we choose the best performing diagonal and full observation-driven, as well as the EIS and Bellman parameter-driven specifications. These empirical cumulative distributions can be found in Figure 1.

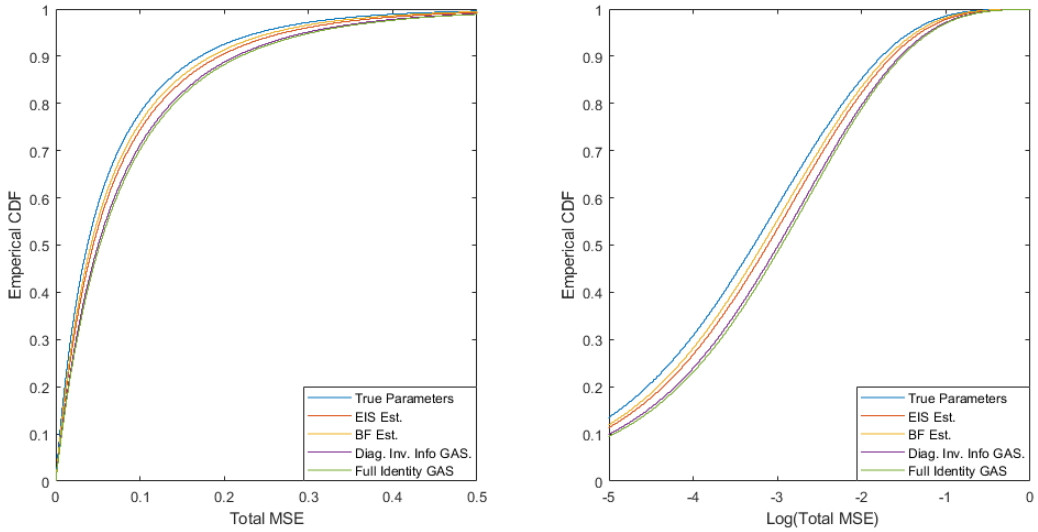

Figure 1: Cumulative distribution function of the RMSE for the SMV observation density and $T = 1000$. Here all EIS variants, as well as the best performing diagonal and full observation driven models, are shown.

As can be seen in the figure, the CDFs are all shaped similar to each other. The main difference in performance does not seem to lie around a single point, but instead seems to happen gradually. Furthermore, we note that for this model, the EIS and BP filters perform very similarly, as do the best performing GAS models.

7 Empirical Study

In order to test the real world applications of the models given in the paper, we perform a small empirical study on data from stock indices. The first element of the study consists of fitting the SMV model to return series, and compare the performance to other alternative specifications by means of Value at Risk analysis. The second element is similar to the minimum variance portfolio simulation study. Here we fit both the SCC and SEV specifications to the demeaned return series, and estimate the minimum variance portfolio. This will be compared to other well-known specifications, such as a moving window covariance estimate and the $1/N$ portfolio.

For these empirical applications, we implement multiple parameter- and observation-driven specifications. For the parameter-driven specifications, we choose to mainly consider the EIS maximum likelihood estimation and filtration instead of the Bellman filter. This is mostly done because the Bellman filter has a similar performance to the EIS method. Furthermore, filtering functions of the unobserved states is relatively straightforward under an EIS framework (see Section 3.2.3), while this technique for the Bellman filter still has to be developed. While the Bellman filter finds a good estimate for $\mathbb{E}(\boldsymbol{\alpha_{t+1}}|\mathbf{y}_t)$, this does not easily translate into a good estimate for $\mathbb{E}(f(\boldsymbol{\alpha_{t+1}})|\mathbf{y}_t)$ for some arbitrary non-affine function f.

7.1 Data

The data set that is chosen consists of the daily returns of both the Dutch AEX, as well as the American S&P500 index. The data set contains the closing prices of the respective indices from 1-1-2010 to 31-12-2019. After removing days on which one or both of the indices were closed, as well as transforming the prices into log returns, we are left with 2490 observations. Plots and descriptive statistics for both stock indices can be found in Figure 2 and Table 5 respectively.

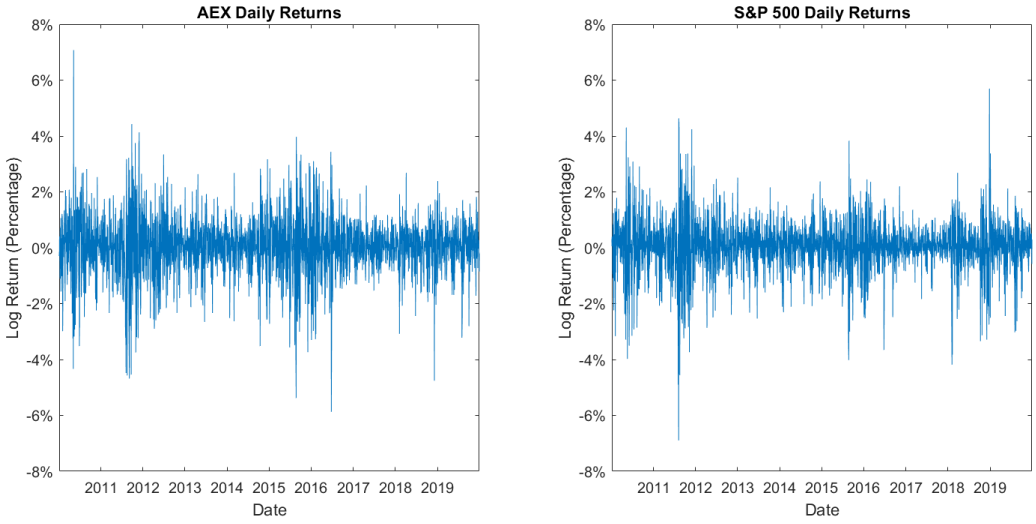

Figure 2: Log returns of the AEX and S&P500 indices, from 1-1-2010 to 31-12-2019, in percentages.

Table 5: Descriptive statistics of the log returns of the AEX and S&P500 indices, from 1-1-2010 to 31-12-2019, in percentages.

Test Statistics	AEX	S&P500
Mean	0.023	0.042
Variance	1.120	0.880
Correlation	0.628	0.628
Skewness	-0.273	-0.468
Kurtosis	6.159	7.759

These figures exhibit some interesting and well-known properties of stock returns. Note that both the AEX as well as the S&P500 index exhibit periods of increased volatility, which can be linked to periods of uncertainty in the markets. For example, the heightened volatility between 2010 and 2012 can be explained by the European debt crisis. Furthermore, these periods of higher volatility seem to correlate between indices.

Now let us consider the values in Table 5. We see that both indices have a slight positive mean, although only that of the S&P500 is significantly different from zero, where we reject if $p < 0.05$. We also observe that the correlation between the stock index returns is 0.628, which implies that approximately 39% of the variance in one of the stock index returns can be explained by the other. Furthermore, we see that both stock indices

have negative skewness and excess kurtosis, which are well-known properties of stock returns. We also reject the hypothesis of normality for both indices using the Jarque–Bera test, with p-values smaller than 0.001.

Lastly, we note that the AEX index has a higher return variance, as well as a lower mean compared to the S&P500 index, which would contradict the idea of a volatility premium. This can be partly explained by the fact that while the US experienced the worst parts of the mortgage crisis around 2008-2009, Europe was hit a few years later during the debt crisis in the 2010's.

7.2 SMV Performance

To analyze the performance of our stochastic mean variance, or SMV model, we fit the model on the returns of the AEX. This is done using both the parameter-driven as the GAS specification. For our parameter and observation-driven specification, we implement a diagonal and full specification. For our scaling matrix, we choose the type which performed the best in the simulation study. we acknowledge that this comparison is not perfect, since our data is unlikely to follow a parameter-driven specification. However, the alternative of including 6 different parameter-driven models would give an unfair advantage to these models. It is interesting to see how our full models include the earlier mentioned negative correlation between volatility and returns. This effect would be observable in the off-diagonal elements of the $\mathbf{\Phi}$ and $\mathbf{A}, \mathbf{B}$ matrices for the parameter and observation-driven specifications respectively.

Apart from parameter and observation-driven models, we also include some well-known specifications. This is done in order to compare the performance of our models, and see if they are able to beat simpler implementations. The first choice we introduce is the GARCH$(1,1)$ model with mean, and normal innovations (Bollerslev, 1986). While it is able to capture a varying volatility, the mean μ is held constant. If we define $\sigma^2_{t|t-1}$ as the filtered estimate for the volatility at time t for the GARCH model, the $p \times 100\%$ Value at Risk can be expressed in the following way

$$\text{VaR}_{p,t} = \mu + \sigma_{t|t-1} z_p,$$

where z_p is the pth quantile of the standard normal distribution.

For the second model, a moving window mean and variance is implemented, with normal innovations. For the size of the moving window, 30 trading days are chosen. This way the persistence of the variance is captured, while also allowing a quick adaptation. We use the unbiased method of moments estimators for the mean $\mu_{t-29:t}$ and variance $\sigma^2_{t-29:t}$. If these are estimated with observations between $t-29$ and t, the $p \times 100\%$ Value at Risk estimate for $t+1$ is given by the following formula

$$\text{VaR}_{p,t} = \mu_{t-29:t} + \sigma_{t-29:t} z_p.$$

The data is again partitioned into equally sized parts. The first half of the observations are used to estimate the model parameters, while the second half is used to test the performance of the model.Due to the nature of the (empirical) data, the true values for the driving process $\boldsymbol{\alpha}_t$ are unknown. Because of this, a different measure of model performance needs to be used. In this case, we evaluate the quality of a model by means of Value at Risk analysis. Let $\widehat{\alpha}^{(m)}_{p,t}$ define the $p \times 100\%$ VaR estimate for model m conditional on observations $\text{y}_1, \ldots, \text{y}_{t-1}$. If we assume that model m is a good approximation of the true DGP, then we should

approximately have that the following holds

$$H_0 : \mathbb{P}(\mathrm{y}_t < \widehat{\alpha}_{p,t}^{(m)}) = p,$$
$$H_a : \mathbb{P}(\mathrm{y}_t < \widehat{\alpha}_{p,t}^{(m)}) \neq p.$$

If we set this as our null hypothesis, then the following asymptotic test statistic given in McNeil (2015) can be constructed using the sample data $\mathrm{y}_{T_0:T}$

$$\widehat{p} \equiv \frac{1}{T_0 - T} \sum_{t=T_0}^{T} \mathbb{I}(\mathrm{y}_t < \widehat{\alpha}_{p,t}^{(m)}),$$
$$Z \equiv \sqrt{T - T_0} \frac{\widehat{p} - p}{\sqrt{p(1-p)}} \sim N(0,1) \text{ asymptotically under } H_0.$$

Here we define T_0 to be the start of the testing sample. For our applications, $T_0 = \left\lfloor \frac{T}{2} \right\rfloor$. The Value at Risk series and the corresponding test statistic are then calculated for both the parameter-driven as the observation-driven specification. The tests are performed for $p \in \{0.01, 0.05, 0.10\}$. Note that for a given p, we can also compare the performance of different models. Here the best performing model is the one with the test statistic closest to zero.

7.2.1 Results

As is described in the section above, diagonal and full specifications for both the parameter as the observation-driven model classes were estimated on approximately five years of AEX return data. The resulting parameter estimates can be found below:

Table 6: Parameter estimates for (full and diagonal) parameter and observation-driven models. These were estimated on daily return data of the AEX index.

Model Type	μ_1	μ_2	$\Phi_{1,1}$	$\Phi_{2,1}$	$\Phi_{1,2}$	$\Phi_{2,2}$	$\mathrm{Q}_{1,1}$	$\mathrm{Q}_{2,2}$
Diagonal Parameter-Driven	0.0640	-0.1936	0.0955	0.0000	0.0000	0.9611	0.0215	0.0555
Full Parameter-Driven	0.0254	-0.3764	0.2497	-0.6500	-0.0135	0.9440	0.0992	0.0025

Model Type	μ_1	μ_2	$\mathrm{A}_{1,1}$	$\mathrm{A}_{2,1}$	$\mathrm{A}_{1,2}$	$\mathrm{A}_{2,2}$	$\mathrm{B}_{1,1}$	$\mathrm{B}_{2,1}$	$\mathrm{B}_{1,2}$	$\mathrm{B}_{2,2}$
Diagonal Observation-Driven	0.0398	0.0123	-0.0193	0.0000	0.0000	0.0741	0.9700	0.0000	0.0000	0.9723
Full Observation-Driven	0.0168	-0.0790	0.0197	-0.2465	-0.0002	0.0305	0.8817	0.4301	0.0155	0.9123

Let us first discuss the parameter-driven models. For both models, we see that the volatility persistence factor $\Phi_{2,2}$ lies around 0.95. This is in line with earlier research on stochastic volatility models by Liesenfeld and Richard (2003), which found similar large coefficients. For the diagonal model, we see that the persistence factor $\Phi_{1,1}$ for the mean is very close to zero. This indicates that any deviation from the mean expected returns quickly disappears. When a full specification is chosen, we see that the cross effect that the mean has on the volatility $\Phi_{2,1}$ is relatively large and negative. In comparison, the cross effect that the volatility has on the mean, given by $\Phi_{1,2}$, is very close to zero. This implies that the correlation between negative returns and volatility is one directional: negative returns increase the volatility, but a high volatility does not affect the unconditional mean. Furthermore, we see that the unconditional mean log volatility μ_2 is smaller for the full specification in comparison to the diagonal variant. Furthermore, we also see a large decrease in the variance of the volatility process. This can be explained by studying the process that the mean follows under a full specification. Notice that the variance and persistence (given by $\mathrm{Q}_{1,1}$ and $\Phi_{1,1}$ respectively) are larger for the full model than for the diagonal model. Since this increases the unconditional variance of the mean, we hypothesize that for the full model, the volatility is largely driven by the hidden mean process through the cross factor $\Phi_{2,1}$.

Concerning the observation-driven models, we see some similarities and some differences. For example, we again see very high persistency factors ($\mathrm{B}_{2,2}$ for this model class) for the volatility process. However, we also see high persistency factors for a much higher persistency factor for the mean, compared to the parameter-driven models. This is mainly due to the different way both models handle outliers. For the parameter-driven models, outliers can be caused by the uncertainty in the observed process itself, which does not necessarily impact the filtered values for the driving process much. However, for the observation-driven models, outliers especially affect the the driving process. Because of this difference, we expect that the persistency factors are also different. Lastly, we notice that the cross persistency factor $\mathrm{B}_{2,1}$ is positive, while the cross scaling factor $\mathrm{A}_{2,1}$ is negative. This has an interesting implication. On average, a negative filtered mean for the previous day has a decreasing effect on the volatility the next day. However, when the observed returns is more negative than expected, the volatility increases. After some algebra, one can derive that the difference between the observed and expected value of y_t needs to be at least $\frac{\mathrm{B}_{2,1}}{\mathrm{A}_{2,1}} \approx 1.74$ times lower than μ_t in order to increase the future volatility σ_{t+1} through this cross effect.

As mentioned at the start of Section 7.2, we evaluate the performance of the different models by means of Value at Risk analysis. This was also done for a GARCH$(1,1)$ and historical mean variance specifications with normal innovations. In order to test if the difference between the correct and found level of coverage is significant, a Z-test is used, where we reject if the p-value is smaller that 0.05. The results of this analysis can be found in Table 7, given below:

Table 7: Unconditional Value at Risk coverage for both parameter as observation-driven models, as well as a GARCH(1,1) and historical mean variance specification.

Model Type	10% VaR			5% VaR			1% VaR		
	$\widehat{p}$	Z-Score	p-value	$\widehat{p}$	Z-Score	p-value	$\widehat{p}$	Z-Score	p-value
Diagonal Parameter-Driven (EIS)	0.1092	1.0864	0.2773	0.0522	0.3576	0.7206	0.0129	1.0112	0.3119
Full Parameter-Driven (EIS)	0.1398	4.6763	0.0000	0.0859	5.8192	0.0000	0.0313	7.5624	0.0000
Diagonal Observation-Driven	0.0956	-0.5196	0.6034	0.0490	-0.1625	0.8709	0.0169	2.4354	0.0149
Full Observation-Driven	0.0843	-1.8422	0.0655	0.0386	-1.8530	0.0639	0.0129	1.0112	0.3119
Garch$(1,1)$ with Mean	0.0932	-0.8030	0.4220	0.0490	-0.1625	0.8709	0.0169	2.4354	0.0149
Historical Mean Variance	0.1100	1.1809	0.2377	0.0675	2.8283	0.0047	0.0289	6.7079	0.0000

Let us discuss the parameter-driven models first. We see that the diagonal model performs quite well, with no hypothesis rejections at any level, while we reject the hypothesis of correct coverage at all levels for the full parameter-driven specification. This difference in quality of estimates has likely to do with the fact that parameter-driven models require a relatively large sample size in order to produce credible estimates. When more parameters are added, this seems to quickly limit the quality of the estimated parameters. In order to test this hypothesis, larger sample sizes may be analyzed in order to find out if the same problem occurs.

For the observation-driven model, we see for the diagonal model performs similar to the GARCH$(1,1)$ model. For both model classes, the hypothesis of correct coverage is not rejected at the 10% and 5% levels, but we do reject at the 1% level. This implies that both the GARCH$(1,1)$ and diagonal observation-driven specification are unable to give correct Values at Risk for the far left quantiles of the return distribution. We do note that the full observation-driven model performs better at this coverage level. This might be related that these extremes are often observed in times of crisis, which are characterized by a high volatility and large negative returns. By incorporating cross effects in our model, the mean and volatility are likely able to adapt more quickly to these economic downturns.

In general, only two models the hypothesis of correct coverage at all levels is not rejected, namely the diagonal parameter-driven model as well as he full observation-driven model. However, we note that the hypothesis of correct coverage at the 10% and 5% levels are close to being rejected for the full observation-driven model, while this is not the case for the diagonal parameter-driven model. Therefore, we note that the diagonal parameter-driven model is preferable based on this data. One reason for this difference in performance is that the parameter-driven specification is able to correctly assess the uncertainty in the hidden process. For example, the uncertainty in the mean may be much larger as the uncertainty in the volatility. If one would ignore this uncertainty, then it is likely that the Values at Risk obtained are incorrect, and not low enough. For situations where this uncertainty is large enough, we expect to find that model which incorporate this uncertainty outperform models which do not.

7.3 SCC & SEV Performance

Because the SCC and SEV specifications model multivariate volatility, an interesting and widely applied performance test is the construction of minimum variance porfolios. If we assume that the returns $\mathbf{y}_t$ have a conditional (time-varying) covariance matrix $\mathbf{\Sigma}_t$, then a commonly asked question is how to best choose the portfolio weights in order to minimize the portfolio variance. In mathematical terms, this can be seen as the

following optimization problem

$$\begin{aligned}&\arg\min_{w_t}\left\{\mathbb{E}\left(\boldsymbol{w}_t'\boldsymbol{\Sigma}_t\boldsymbol{w}_t\right)\right\}=\arg\min_{w_t}\left\{\boldsymbol{w}_t'\mathbb{E}\left(\boldsymbol{\Sigma}_t\right)\boldsymbol{w}_t\right\},\\&\text{s.t }\boldsymbol{w}_t'\boldsymbol{\iota}=1,\end{aligned}$$

where $\boldsymbol{w}_t$ are the weights at time t. This can be solved by means of Lagrange multipliers, with solution

$$\boldsymbol{w}_t=\frac{\boldsymbol{\iota}'\mathbb{E}\left(\boldsymbol{\Sigma}_t\right)^{-1}\boldsymbol{\iota}}{\mathbb{E}\left(\boldsymbol{\Sigma}_t\right)^{-1}\boldsymbol{\iota}}.$$

The challenge in constructing global minimum variance portfolios is finding a good estimate for $\mathbb{E}\left(\boldsymbol{\Sigma}_t\right)$. Therefore, if a volatility model is able to give good predictions of next periods' covariance matrix, then it should also be able to construct good minimum variance portfolios.

After demeaning the return series, we use the new bivariate daily return series, denoted by

$$\mathbf{y}_t\equiv\begin{bmatrix}\mathrm{y}_{t,AEX}\\\mathrm{y}_{t,S\&P}\end{bmatrix}$$

to estimate a full SCC and SEV specification. This is done for both parameter-driven as well as a observation-driven specifications. One again, we include a historical moving window model, which estimates the covariance matrix for observation t based on the previous 30 observations. Another well-known reference portfolio is the $1/N$ allocation. This corresponds to setting all portfolio weights equal to each other. For our study, this implies that

$$\boldsymbol{w}_t=\begin{bmatrix}0.5\\0.5\end{bmatrix},$$

for all points in time t.

How the covariance matrix is estimated depends on the underlying process. If the underlying process is observation-driven, the covariance matrix for time t is known at time $t-1$ by construction. Therefore, it can be found using the GAS filter. For parameter-driven models, we use the fact that covariance matrix is function of the hidden process. This implies that the expected value $\mathbb{E}\left(\boldsymbol{\Sigma}_t\right)$ can be found by means of the filtering approach given in Section 3.2.3. For this study, we perform the same sample split as was done in section 7.2. The performance measures are therefore out-of-sample.

7.3.1 Results

As was described at the start of Section 7.3, we construct global minimum variance portfolios under multiple different specifications. In order to test the quality of our portfolios, we also consider allocations formed under a moving window covariance matrix, as well as the $1/N$ portfolio. The resulting variances of the portfolio returns during the test period is given in Table 8.

Table 8: Estimated out-of-sample variance of the constructed minimum variance portfolios.

Model Type	σ_p^2
Diagonal Parameter-Driven SCC (EIS)	0.6552
Diagonal Parameter-Driven SEV (EIS)	0.6765
Diagonal Observation-Driven SCC	0.6387
Diagonal Observation-Driven SEV	0.6481
Full Observation-Driven SCC	0.6181
Full Observation-Driven SEV	0.6507
30 Day Moving Window	0.6290
$1/N$ Allocation	0.6713

Let us first discuss the results in general. We note that for all types considered, the SEV model is always outperformed by the SCC specification. This, in combination with easier to interpret parameters, leads us to prefer the SCC specification over the SEV specification. Secondly, we also see that the parameter-driven models are outperformed by the observation-driven models. Since maximum likelihood estimation is also much faster for observation-driven models, we conclude that this model type is better suited for modeling covariance compared to the corresponding parameter-driven models.

Surprisingly, we see that the relatively simple moving window covariance estimate generates weights which yield the second lowest portfolio variance. One of the reasons for this performance may be the that a moving window is able to adapt well to a changing data generating process, where the other models have to follow the parameter estimates of the training period.

The model that performed the best is the observation-driven SCC model with full parameter matrices. However, when studying the cross effects in the parameter matrices $\mathbf{A}$ and $\mathbf{B}$, we find that the cross effects between σ_1 and σ_2 found in $\mathbf{A}$ are especially large, signaling that a positive shock in the variance of one index also positively influences the variance of the other index.

8 Conclusion

In summary, we investigate the performance of observation-driven models in comparison with similar parameter-driven models. This study is done by performing two simulation and two empirical studies.

In the first simulation study, we analyze the convergence properties of the parameters, and find evidence that all estimators considered are consistent. The second simulation study is an approximation study similar to the one performed for one dimensional driving processes by Koopman, Lucas and Scharth (2016), where we find similar results. In most cases, observation-driven models are able to predict the hidden driving process with similar accuracy as their corresponding parameter-driven variant. However, when there exists strong cross effects between the hidden states, the approximation becomes less accurate. It is noted that for small sample sizes and models with relatively many parameters, convergence issues are likely to arise for the observation-driven models. However, these problems quickly decrease in severity when the sample size is increased. Lastly, the Bellman filter seems to be a good compromise between the two methods, both able to

handle strong cross effects while remaining computationally fast.

In our empirical studies, Value at Risk analysis and minimum variance portfolio construction was studied. The data used for these studies is approximately ten years of daily log returns of the AEX and S&P500 indices, from 2010 to 2020.

Value at Risk analysis on AEX return series was performed. Here we see that the diagonal parameter-driven models perform slightly better than the observation-driven models. However, when a full specification is chosen for the parameter-driven model, the model performs very poorly. This is likely because the increased number of parameters that need to be estimated also increases the uncertainty in the estimates.

For the construction of minimum variance portfolios, diagonal parameter-driven SCC and SEV specifications were studied, as well as diagonal and full observation-driven SCC and SEV models. For the construction of the minimum variance portfolios, we found that the SCC models always outperform the SEV models. Furthermore, we found that the observation-driven models always outperform their parameter-driven counterpart.

In conclusion, we note that in almost all cases studied, observation-driven models had a similar or better performance as their parameter-driven counterpart. This holds for cases where the data generating process is a known parameter-driven specification, as well as on empirical data. Because of the large difference in computational complexity when estimating the model parameters, observation-driven specifications are preferable in most situations. This holds especially for models with three or more driving parameters, as parameter-driven specifications become very computationally difficult to estimate at this point. We also conclude that for observation-driven models, there does not exists one scaling matrix that works best for all specifications. In many cases, we recommend that one tries multiple different scaling choices to see which works the best.

There are multiple interesting directions for future research. One of these is how to best handle the uncertainty of which scaling matrix is the best. For example, it is also possible to make this choice part of the optimization process. This can be done by introducing an auxiliary model parameter θ, which impacts the choice of scaling matrix in the following matter

$$\boldsymbol{S}_t \equiv \boldsymbol{\mathcal{I}}_{t|t-1}^{-\delta}, \quad \mathbf{X}^{\delta} \equiv \mathbf{Q}\boldsymbol{\Lambda}^{\delta}\mathbf{Q}',$$

with $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$ being the spectral decomposition of $\mathbf{X}$, and defining the power of a diagonal matrix to be the diagonal matrix containing the diagonal elements raised to the same power. We can then maximize the likelihood over δ, similar to the other model parameters.

While the parameter-driven models are relatively slow to estimate compared to similar observation-driven models, they are still interesting to consider. While the NAIS method of Koopman, Lucas and Scharth (2015) is already a lot faster than the EIS method, it relies on a very low-dimensional signal variable, which is not applicable for our models. It would be interesting to study if this method can also be extended to two, three or even higher-dimensional problems, and if there is a similar gain in computational efficiency as was found for the unidimensional case.

Lastly, it would also be interesting to see if parameter-driven models are able to support a discrete parameter structure. In Richard and Zhang (2007) most integrals in the EIS algorithm can be replaced by sums over a finite number of values. For example, consider a dynamic non-linear non-Gaussian model with Markov switching, such that the observed variable $\mathbf{y}_t$ has the following specification

$$\begin{aligned}
&\mathbf{y}_t|\boldsymbol{\alpha}_t \sim p(\mathbf{y}_t|\boldsymbol{\theta}(\boldsymbol{\alpha}_t)),\\
&\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, s_{t+1} \sim g_{s_{t+1}}(\boldsymbol{\alpha}_{t+1}|\boldsymbol{\alpha}_t, \boldsymbol{\Psi}),\\
&s_t \in \{1, \ldots, S\} \text{ is first order Markovian with transition matrix } \mathbf{P},
\end{aligned}$$

where the model coefficients of g_{s_t} depend on the state s_t at that moment. Note that for this example, both $\boldsymbol{\alpha}_t$ and s_t are unobserved. If one can model the hidden state $[\boldsymbol{\alpha}_t \; s_t]$' in a similar matter as is normally done in EIS, then the likelihood of jointly observing $\mathbf{y}_{1:T}$ can be found.

References

Bauwens, L., Laurent, S., Rombouts, J. V. (2006). Multivariate GARCH models: A survey. *Journal of Applied Econometrics, 21*(1), 79-109.

Chesnay, F., Jondeau, E. (2001). Does correlation between stock returns really increase during turbulent periods? *Economic Notes, 30*(1), 53-80.

Creal, D., Koopman, S. J., Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics, 28*(5), 777-795.

De Zela, F. (2014). Closed-form expressions for the matrix exponential. *Symmetry, 6*(2), 329-344.

Giot, P. (2005). Relationships between implied volatility indexes and stock index returns. *The Journal of Portfolio Management, 31*(3), 92-100.

Glynn, P. W., Iglehart, D. L. (1989). Importance sampling for stochastic simulations. *Management Science, 35*(11), 1367-1392.

Harvey, A. C. (2013). Dynamic models for volatility and heavy tails: With applications to financial and economic time series (Vol. 52). *Cambridge University Press.*

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35-45.

Kawakatsu, H. (2006). Matrix exponential GARCH. *Journal of Econometrics, 134*(1), 95-128.

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

Koopman, S. J., Lucas, A., Scharth, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *Review of Economics and Statistics, 98*(1), 97-110.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian non-linear state-space models. *Journal of Computational and Graphical Statistics, 5*(1), 1-25.

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

Liesenfeld, R., Richard, J. F. (2003). Univariate and multivariate stochastic volatility models: Estimation and diagnostics. *Journal of Empirical Finance, 10*(4), 505-531.

Malagò, L., Pistone, G. (2015). Information geometry of the Gaussian distribution in view of stochastic optimization. *Proceedings of the 2015 ACM Conference on Foundations of Genetic Algorithms, 8*(1), 150-162.

McNeil, A. J., Frey, R., Embrechts, P. (2015). Quantitative risk management: Concepts, techniques and tools, revised edition. *Princeton University Press.*

Poole, D. (2014). Linear algebra: A modern introduction. Cengage Learning.

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

Richard, J.F., Zhang, W. (1996). Econometric modeling of UK house prices using accelerated importance sampling. *Oxford Bulletin of Economics and Statistics 58*, 601–613.

Richard, J.F., Zhang, W. (1997). Accelerated Monte Carlo integration: An application to dynamic latent variable models. In: Mariano, R., Weeks, M., Schuermann, T. (Eds.), Simulation-Based Inference in *Econometrics: Methods and Application. Cambridge University Press, Cambridge* 47–70.

Richard, J.F., Zhang, W. (1998). Efficient high-dimensional Monte Carlo importance sampling. Unpublished manuscript, *Department of Economics, University of Pittsburgh.*

Appendix

Derivation Integration Constant

In this subsection, the integration constant $\chi_{t+1}\left(\boldsymbol{\alpha}_t|\boldsymbol{\beta}_{t+1},\boldsymbol{\Psi}\right)$ is derived analytically for the following kernel:

$$
\begin{aligned}
&k_t(\boldsymbol{\alpha}_t,\boldsymbol{\alpha}_{t-1}) \propto g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1},\boldsymbol{\Phi})\zeta(\boldsymbol{\alpha}_t)\\
&\zeta(\boldsymbol{\alpha}_t) \equiv \exp\left(\boldsymbol{\lambda}_t'\boldsymbol{\alpha}_t + \boldsymbol{\alpha}_t'\boldsymbol{\Gamma}_t\boldsymbol{\alpha}_t\right)\\
&\boldsymbol{\beta}_t \equiv \{\boldsymbol{\lambda}_t,\boldsymbol{\Gamma}_t\}
\end{aligned}
$$

First we note that the kernel of g can be written in the following manner:

$$
\begin{aligned}
&g(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1},\boldsymbol{\Phi}) \propto\\
&\exp\left\{-\frac{1}{2}\left(\boldsymbol{\alpha}_t-\boldsymbol{\omega}-\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right)'\mathbf{Q}^{-1}\left(\boldsymbol{\alpha}_t-\boldsymbol{\omega}-\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right)\right\} =\\
&\exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right)'\mathbf{Q}^{-1}\left(\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right) - 2\boldsymbol{\alpha}_t'\mathbf{Q}^{-1}\left(\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right) + \boldsymbol{\alpha}_t'\mathbf{Q}^{-1}\boldsymbol{\alpha}_t\right]\right\}
\end{aligned}
$$

Here it can be seen that the product of two k-variate Gaussian kernels is itself k-variate Gaussian. Multiplying ζ and g yields the following kernel

$$
\begin{aligned}
&k_t(\boldsymbol{\alpha}_t,\boldsymbol{\alpha}_{t-1}) \propto\\
&\exp\left\{-\frac{1}{2}K_1\right\}\cdot\exp\left\{-\frac{1}{2}\left\{-2\boldsymbol{\alpha}_t'\left[\boldsymbol{\lambda}_t+\mathbf{Q}^{-1}\left(\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right)\right]+\boldsymbol{\alpha}_t'\left(\mathbf{Q}^{-1}-2\boldsymbol{\Gamma}_t\right)\boldsymbol{\alpha}_t\right\}\right\} =\\
&\exp\left\{-\frac{1}{2}K_1\right\}\cdot\exp\left\{-\frac{1}{2}\left\{-2\boldsymbol{\alpha}_t'\widetilde{\boldsymbol{\Sigma}}_t^{-1}\widetilde{\boldsymbol{\mu}}_t+\boldsymbol{\alpha}_t'\widetilde{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\alpha}_t\right\}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&K_1 \equiv \left(\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right)'\mathbf{Q}^{-1}\left(\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right),\\
&\widetilde{\boldsymbol{\Sigma}}_t \equiv \left(\mathbf{Q}^{-1}-2\boldsymbol{\Gamma}_t\right)^{-1},\\
&\widetilde{\boldsymbol{\mu}}_t \equiv \widetilde{\boldsymbol{\Sigma}}_t\left(\boldsymbol{\lambda}_t+\mathbf{Q}^{-1}\left(\boldsymbol{\omega}+\boldsymbol{\Phi}\boldsymbol{\alpha}_{t-1}\right)\right),
\end{aligned}
$$

which shows that k_t is a k-variate Gaussian kernel with mean $\widetilde{\boldsymbol{\mu}}_t$ and covariance $\widetilde{\boldsymbol{\Sigma}}_t$. This kernel can then be rewritten in the following manner:

$$
\begin{aligned}
&\exp\left\{-\frac{1}{2}K_1\right\}\cdot\exp\left\{-\frac{1}{2}\left\{-2\boldsymbol{\alpha}_t'\widetilde{\boldsymbol{\Sigma}}_t^{-1}\widetilde{\boldsymbol{\mu}}_t+\boldsymbol{\alpha}_t'\widetilde{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\alpha}_t\right\}\right\} =\\
&\exp\left\{-\frac{1}{2}K_1+\frac{1}{2}\widetilde{\boldsymbol{\mu}}_t'\widetilde{\boldsymbol{\Sigma}}_t^{-1}\widetilde{\boldsymbol{\mu}}_t\right\}\left(2\pi\det\left(\widetilde{\boldsymbol{\Sigma}}_t\right)\right)^{k/2}\cdot\left(2\pi\det\left(\widetilde{\boldsymbol{\Sigma}}_t\right)\right)^{-k/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{\alpha}_t-\widetilde{\boldsymbol{\mu}}_t\right)'\widetilde{\boldsymbol{\Sigma}}_t^{-1}\left(\boldsymbol{\alpha}_t-\widetilde{\boldsymbol{\mu}}_t\right)\right\} \equiv\\
&\chi_t\left(\boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t,\boldsymbol{\Psi}\right)\cdot m_t\left(\boldsymbol{\alpha}_t,\boldsymbol{\alpha}_{t-1}|\boldsymbol{\beta}_t,\boldsymbol{\Psi}\right).
\end{aligned}
$$

Derivation Score and Information Matrix for SMV

Let us first derive the score for the SMV model, which is given as the derivative of the log pdf with respect to the driving parameters

$$\log\left(p\left(\mathrm{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right) = -\frac{1}{2}\left\{\log(2\pi) + \alpha_{t,2} + \left(\mathrm{y}_t - \alpha_{t,1}\right)^2 e^{-\alpha_{t,2}}\right\} \implies$$
$$\boldsymbol{\nabla}_t \equiv \frac{\partial \log\left(p\left(\mathrm{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)}{\partial \boldsymbol{\alpha}_t} = \begin{bmatrix} \left(\mathrm{y}_t - \alpha_{t,1}\right) e^{-\alpha_{t,2}} \\ -\frac{1}{2} + \frac{1}{2}(\mathrm{y}_t - \alpha_{t,1})^2 e^{-\alpha_{t,2}} \end{bmatrix}.$$

As for the information matrix, we make use of the following identity:

$$\boldsymbol{\mathcal{I}}_{t|t-1} \equiv \mathbb{E}_{t-1}\left[\boldsymbol{\nabla}_t \boldsymbol{\nabla}_t'\right] = -\mathbb{E}_{t-1}\left[\frac{\partial \log\left(p\left(\mathrm{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)}{\partial \boldsymbol{\alpha}_t \boldsymbol{\alpha}_t'}\right].$$

Computing the Hessian of the log probability density function, and taking the expectation yields the following formula for the information matrix:

$$\frac{\partial \log\left(p\left(\mathrm{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)}{\partial \boldsymbol{\alpha}_t \boldsymbol{\alpha}_t'} = -\begin{bmatrix} e^{-\alpha_{t,2}} & \left(\mathrm{y}_t - \alpha_{t,1}\right) e^{-\alpha_{t,2}} \\ \left(\mathrm{y}_t - \alpha_{t,1}\right) e^{-\alpha_{t,2}} & \frac{1}{2}(\mathrm{y}_t - \alpha_{t,1})^2 e^{-\alpha_{t,2}} \end{bmatrix},$$
$$\boldsymbol{\mathcal{I}}_{t|t-1} = \begin{bmatrix} e^{-\alpha_{t,2}} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}.$$

Derivation Score and Information Matrix for SCC

Let us first derive the score for the SCC model, which is given as the derivative of the log pdf with respect to the driving parameters

$$\log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)=-\frac{1}{2}\left\{N\log(2\pi)+\log\left(\det\left(\boldsymbol{\Sigma}_t\right)\right)+\mathbf{y}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{y}_t\right\},$$

where we have that for the SCC model,

$$\boldsymbol{\Sigma}_t=\begin{bmatrix} e^{\alpha_{t,1}} & e^{\frac{1}{2}(\alpha_{t,1}+\alpha_{t,2})}\tanh(\alpha_{t,2}) \\ e^{\frac{1}{2}(\alpha_{t,1}+\alpha_{t,2})}\tanh(\alpha_{t,2}) & e^{\alpha_{t,3}} \end{bmatrix}.$$

By plugging in this formula for $\boldsymbol{\Sigma}_t$ in the formula for the score, we obtain a closed form expression for the score. For ease of notation, define the $(i,j)^{th}$ element of $\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,i}}(x,y,z)$ as the derivative of the $(i,j)^{th}$ element of $\boldsymbol{\Sigma}_t^{-1}$ with respect to $\alpha_{t,i}$, evaluated in evaluated in $\boldsymbol{\alpha}_t=[x\ y\ z]'$. Then, by direct computation, we find that the following formulas for $\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,i}}(x,y,z)$ hold

$$\begin{aligned}
\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,1}}(x,y,z)&=\frac{1}{4}\begin{bmatrix} -2e^{-x}\left(\cosh\left(2y\right)+1\right) & \sinh\left(2y\right)e^{-\frac{1}{2}(x+z)} \\ \sinh\left(2y\right)e^{-\frac{1}{2}(x+z)} & 0 \end{bmatrix}, \\
\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,2}}(x,y,z)&=\begin{bmatrix} \sinh(2y)e^{-x} & -\cosh\left(2y\right)e^{-\frac{1}{2}(x+z)} \\ -\cosh\left(2y\right)e^{-\frac{1}{2}(x+z)} & \sinh(2y)e^{-z} \end{bmatrix}, \\
\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,3}}(x,y,z)&=\frac{1}{4}\begin{bmatrix} 0 & \sinh\left(2y\right)e^{-\frac{1}{2}(x+z)} \\ \sinh\left(2y\right)e^{-\frac{1}{2}(x+z)} & -2e^{-z}\left(\cosh\left(2y\right)+1\right) \end{bmatrix}.
\end{aligned}$$

Then, using this convention, the score of the SCC model with respect to the driving process can be expressed as follows

$$\begin{aligned}
\frac{\partial\log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)}{\partial\alpha_{t,1}}&=-\frac{1}{2}-\frac{1}{2}\mathbf{y}_t'\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,1}}(\alpha_{t,1},\alpha_{t,2},\alpha_{t,3})\mathbf{y}_t, \\
\frac{\partial\log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)}{\partial\alpha_{t,2}}&=\tanh(\alpha_{t,2})-\frac{1}{2}\mathbf{y}_t'\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,2}}(\alpha_{t,1},\alpha_{t,2},\alpha_{t,3})\mathbf{y}_t, \\
\frac{\partial\log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)}{\partial\alpha_{t,3}}&=-\frac{1}{2}-\frac{1}{2}\mathbf{y}_t'\frac{\partial\boldsymbol{\Sigma}_t^{-1}}{\partial\alpha_{t,3}}(\alpha_{t,1},\alpha_{t,2},\alpha_{t,3})\mathbf{y}_t.
\end{aligned}$$

As for the information matrix, we refer to Harvey (2015), where they note that the information matrix for the SCC model can be given as follows

$$\boldsymbol{\mathcal{I}}_{t|t-1}=\frac{1}{4}\begin{pmatrix} \frac{\rho_t^2-2}{(\rho_t^2-1)} & -2\rho_t & \frac{\rho_t^2}{(\rho_t^2-1)} \\ -2\rho_t & 4\rho_t^2+4 & -2\rho_t \\ \frac{\rho_t^2}{(\rho_t^2-1)} & -2\rho_t & \frac{\rho_t^2-2}{(\rho_t^2-1)} \end{pmatrix}, \quad \rho_t\equiv\tanh(\alpha_{t,2}).$$

Derivation Score and Information Matrix for SEV

As was given in section 4.2, the score is defined as follows

$$\boldsymbol{\nabla}_t = \frac{\partial \log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_t\right)\right)}{\partial \boldsymbol{\alpha}_t}.$$

First we find a closed form expression for the $\log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_t\right)\right)$. Note that p is assumed to be the bivariate Gaussian distribution ($N = 2$) with parameters $\boldsymbol{\mu} = \mathbf{0}_2$ and $\boldsymbol{\Sigma}$ defined as follows

$$\begin{aligned}
\boldsymbol{\Sigma}_t &= e^{\boldsymbol{\Theta}_t}, \\
\boldsymbol{\Theta}_t &\equiv \begin{bmatrix} \alpha_{t,1} & \alpha_{t,2} \\ \alpha_{t,2} & \alpha_{t,3} \end{bmatrix}.
\end{aligned}$$

Using this formulation, we have that the log probability density can be expressed as follows

$$\log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t, \mathcal{F}_t\right)\right) = -\frac{1}{2}\left\{N\log(2\pi) + \log\left(\det\left(\boldsymbol{\Sigma}_t\right)\right) + \mathbf{y}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{y}_t\right\}.$$

Since the first term disappears as we differentiate with respect to $\boldsymbol{\alpha}_t$, let us focus on the second term first. Note that $\det(e^{\boldsymbol{\Theta}_t}) = e^{\mathrm{Tr}(\boldsymbol{\Theta}_t)}$, which can easily be proven by applying the spectral theorem and the fact that that the determinant is invariant under permutation. Therefore, we have that the derivative of the second term with respect to $\boldsymbol{\alpha}_t$ is given as follows

$$\frac{\partial \log\left(\det\left(\boldsymbol{\Sigma}_t\right)\right)}{\partial \boldsymbol{\alpha}_t} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}.$$

Now, let us consider the last term term. Since it can be seen as a 1x1 matrix, we have that

$$\mathbf{y}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{y}_t = \mathrm{tr}(\mathbf{y}_t'\boldsymbol{\Sigma}_t^{-1}\mathbf{y}_t) = \mathrm{tr}(\mathbf{y}_t\mathbf{y}_t'\boldsymbol{\Sigma}_t^{-1}) = \mathrm{tr}\left(\mathbf{B}e^{-\boldsymbol{\Theta}_t}\right), \quad \mathbf{B} = \mathbf{y}_t\mathbf{y}_t'.$$

Because both the trace and matrix multiplication are linear operators, it is sufficient to find the derivative of the symmetric matrix exponential with respect to the elements of $\boldsymbol{\alpha}_t$. Furthermore, if we have a formula of the derivative of $e^{\boldsymbol{\Theta}_t}$, the derivative of $e^{-\boldsymbol{\Theta}_t}$ is also known. Because of this, we first derive the derivative of $e^{\boldsymbol{\Theta}_t}$. As a shortcut, we note that

$$\begin{aligned}
&\boldsymbol{G}(\alpha_{t,1}, \alpha_{t,2}, \alpha_{t,3}) \equiv e^{\boldsymbol{\Theta}_t} = \mathbf{P}\boldsymbol{G}(\alpha_{t,3}, \alpha_{t,2}, \alpha_{t,1})\mathbf{P}, \quad \mathbf{P} \equiv \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \implies \\
&\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,3}}(x, y, z) = \mathbf{P}\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,1}}(z, y, x)\mathbf{P},
\end{aligned}$$

where we define $\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,i}}(x, y, z)$ as the derivative of the matrix function $\boldsymbol{G}$ with respect to $\alpha_{t,i}$, evaluated in x, y, z. This relation can be proven by noting that $\mathbf{P}$ is its own inverse, and that applying $\mathbf{P}$ to both sides permutes the diagonal elements of a 2×2 matrix. We are now only required to find the derivative with respect to $\alpha_{t,1}$ and $\alpha_{t,2}$, as the derivative with respect to $\alpha_{t,3}$ follows from the relation.

In our derivation, we consider two separate cases. First consider the case where $\boldsymbol{\Theta}_t$ is a multiple of the

2×2 identity matrix ($\boldsymbol{\Theta}_t = \delta \mathbf{I}_2$ for some real δ). For diagonal matrices with on the diagonal δ_1 and δ_2, it can be shown that the following holds

$$\exp\left\{\begin{bmatrix}\delta_1 & 0\\ 0 & \delta_2\end{bmatrix}\right\} = \begin{bmatrix}e^{\delta_1} & 0\\ 0 & e^{\delta_2}\end{bmatrix}.$$

Using this fact, it can be shown using the definition of a derivative that the derivative of $e^{\boldsymbol{\Theta}_t}$ with respect to $\alpha_{t,1}$ is given by

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,1}}(\delta, 0, \delta) = \begin{bmatrix}e^{\delta} & 0\\ 0 & 0\end{bmatrix}, \quad \frac{\partial \boldsymbol{G}}{\partial \alpha_{t,3}}(\delta, 0, \delta) = \begin{bmatrix}0 & 0\\ 0 & e^{\delta}\end{bmatrix}.$$

For the derivative with respect to the second state variable more work is required. Note that the following equalities hold:

$$\begin{bmatrix}\delta & 0\\ 0 & \delta\end{bmatrix} = \frac{1}{2}\mathbf{R}\begin{bmatrix}\delta & 0\\ 0 & \delta\end{bmatrix}\mathbf{R}, \quad \begin{bmatrix}\delta & dx\\ dx & \delta\end{bmatrix} = \frac{1}{2}\mathbf{R}\begin{bmatrix}\delta + dx & 0\\ 0 & \delta - dx\end{bmatrix}\mathbf{R}, \quad \mathbf{R} \equiv \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}.$$

This can be derived by the spectral theorem. Furthermore, note that $e^{\mathbf{PAP}^{-1}} = \mathbf{P}e^{\mathbf{A}}\mathbf{P}^{-1}$ for all square matrices $\mathbf{P}, \mathbf{A}$ (assuming $\mathbf{P}$ is invertible). Noting that $\frac{1}{\sqrt{2}}\mathbf{R}$ is its own inverse, and writing out the definition, we find the following

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,2}}(\delta, 0, \delta) = \lim_{dx \to 0} \frac{1}{dx}\left[\exp\left\{\begin{bmatrix}\delta & dx\\ dx & \delta\end{bmatrix}\right\} - \exp\left\{\begin{bmatrix}\delta & 0\\ 0 & \delta\end{bmatrix}\right\}\right] =$$
$$\frac{1}{2}\mathbf{R}\lim_{dx \to 0}\frac{1}{dx}\left[\begin{bmatrix}e^{\delta + dx} & 0\\ 0 & e^{\delta - dx}\end{bmatrix} - \begin{bmatrix}e^{\delta} & 0\\ 0 & e^{\delta}\end{bmatrix}\right] = \frac{1}{2}\mathbf{R}\begin{bmatrix}e^{\delta} & 0\\ 0 & -e^{\delta}\end{bmatrix}\mathbf{R} = \begin{bmatrix}0 & e^{\delta}\\ e^{\delta} & 0\end{bmatrix}.$$

Now we consider the case where $\boldsymbol{\Theta}_t$ is not a scalar multiple of the identity matrix. In this case, the characteristic equation of $\boldsymbol{\Theta}_t$ has discriminant $(\alpha_{t,1} - \alpha_{t,3})^2 + \alpha_{t,2}^2$, and is always nonnegative. This ensures that there exists two distinct eigenvalues, denoted by λ_1, λ_2. In this case, the following formula for the matrix exponential holds (De Zela, 2014)

$$e^{\boldsymbol{\Theta}_t} = e^{s}\left(\left(\cosh(q) - s\frac{\sinh(q)}{q}\right)\mathbf{I}_2 + \frac{\sinh(q)}{q}\boldsymbol{\Theta}_t\right),$$
$$s = \frac{\alpha_{t,1} + \alpha_{t,3}}{2}, \quad q = \frac{1}{2}\sqrt{(\alpha_{t,1} - \alpha_{t,3})^2 + 4\alpha_{t,2}^2}$$

Now we differentiate with respect to α_1 and α_2, to obtain the following derivatives:

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,1}}(x,y,z)_{1,1} =$$
$$\frac{e^s}{4q^3}\left(2q^2\left(q\cosh\left(q\right)+\left(x-s\right)\sinh\left(q\right)\right)-q\left(s-x\right)\left(x-z\right)\cosh\left(q\right)+\left(q^2\left(x-z+2\right)+\left(s-x\right)\left(x-z\right)\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,1}}(x,y,z)_{1,2} =$$
$$\frac{ye^s}{4q^3}\left(q\left(x-z\right)\cosh\left(q\right)+\left(2q^2-x+z\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,1}}(x,y,z)_{2,2} =$$
$$-\frac{e^s}{4q^3}\left(-2q^2\left(q\cosh\left(q\right)+\left(z-s\right)\sinh\left(q\right)\right)+q\left(s-z\right)\left(x-z\right)\cosh\left(q\right)+\left(q^2\left(-x+z+2\right)-\left(s-z\right)\left(x-z\right)\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,2}}(x,y,z)_{1,1} =$$
$$\frac{ye^s}{q^3}\left(-q\left(s-x\right)\cosh\left(q\right)+\left(q^2+s-x\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,2}}(x,y,z)_{1,2} =$$
$$\frac{e^s}{q^3}\left(qy^2\cosh\left(q\right)+\left(q^2-y^2\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,2}}(x,y,z)_{2,2} =$$
$$\frac{ye^s}{q^3}\left(-q\left(s-z\right)\cosh\left(q\right)+\left(q^2+s-z\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,3}}(x,y,z)_{1,1} =$$
$$-\frac{e^s}{4q^3}\left(-2q^2\left(q\cosh\left(q\right)+\left(x-s\right)\sinh\left(q\right)\right)-q\left(s-x\right)\left(x-z\right)\cosh\left(q\right)+\left(q^2\left(x-z+2\right)+\left(s-x\right)\left(x-z\right)\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,3}}(x,y,z)_{1,2} =$$
$$\frac{ye^s}{4q^3}\left(-q\left(x-z\right)\cosh\left(q\right)+\left(2q^2+x-z\right)\sinh\left(q\right)\right),$$

$$\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,3}}(x,y,z)_{2,2} =$$
$$\frac{e^s}{4q^3}\left(2q^2\left(q\cosh\left(q\right)+\left(z-s\right)\sinh\left(q\right)\right)+q\left(s-z\right)\left(x-z\right)\cosh\left(q\right)+\left(q^2\left(-x+z+2\right)-\left(s-z\right)\left(x-z\right)\right)\sinh\left(q\right)\right),$$

with $\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,k}}(x,y,z)$ defined as the i,jth element of the derivative of $e^{\boldsymbol{\Theta}_t}$ with respect to $\alpha_{t,k}$, evaluated in x,y,z. Using these derivatives, we are also able to find the derivative of $e^{-\boldsymbol{\Theta}_t}$. For example, suppose that we want to find the derivative of the i,jth element of $e^{-\boldsymbol{\Theta}_t}$ with respect to $\alpha_{t,1}$. Then, by definition, we have:

$$\frac{\partial e^{-\boldsymbol{\Theta}_t}}{\partial \alpha_{t,1}}(x,y,z)_{i,j} \equiv \lim_{dx\to 0}\frac{1}{dx}\left(e^{-\boldsymbol{\Theta}_t - dx\boldsymbol{e}_1\boldsymbol{e}_1'} - e^{-\boldsymbol{\Theta}_t}\right)_{i,j} = \tag{12}$$

$$-\lim_{dx\to 0}\frac{1}{dx}\left(e^{-\boldsymbol{\Theta}_t + dx\boldsymbol{e}_1\boldsymbol{e}_1'} - e^{-\boldsymbol{\Theta}_t}\right)_{i,j} = -\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,1}}(-x,-y,-z)_{i,j}, \tag{13}$$

with similar derivations for the other state variables. Then by direct computation, the derivative of the third term in the log likelihood with respect to $\alpha_{t,i}$ variable is given by

$$\frac{\partial \log\left(p\left(\mathbf{y}_t|\boldsymbol{\alpha}_t,\mathcal{F}_t\right)\right)}{\partial \alpha_{t,i}} = -\frac{1}{2}\mathbb{I}(i\neq 2) + \frac{1}{2}\mathbf{y}_t'\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,i}}(-\alpha_{t,1},-\alpha_{t,2},-\alpha_{t,3})\mathbf{y}_t. \tag{14}$$

As for the information matrix, we use the following formula given in Malagò et al. (2015) for the $(i,j)^{th}$ element of $\boldsymbol{\mathcal{I}}_{t|t-1}$

$$\mathcal{I}_{i,j} = \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,i}}\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{G}}{\partial \alpha_{t,j}}\right),$$

which can be computed using the earlier derived functions.