Bayesian Risk Evaluation using Importance Sampling

Agnieszka Borowska

Tinbergen Institute

Department of Econometrics, VU University Amsterdam

Supervisors: *Prof. S. J. Koopman, Dr. L. F. Hoogerheide*

August 2015

Abstract

We consider the evaluation of two financial risk measures, Value at Risk and Expected Shortfall. Our analysis is performed in a Bayesian fashion where we adopt a model-based approach. We employ the *Quick Evaluation of Risk using Mixture of t approximation* algorithm (QERMit) of Hoogerheide and van Dijk (2010) due to its accuracy and efficiency, and we upgrade its basic framework in two ways. First, we replace the originally used posterior approximation algorithm with a superior, flexible technique. We report a substantial gain in the accuracy and the precision of estimates in our empirical application based on the daily S&P 500 returns. Second, we extend the basic QERMit framework to allow for latent variables in the underlying model. In this way, the developed technique can be applied to the class of the parameter driven models. We illustrate the procedure using a series of daily IBM returns. Noticeably, all the employed methods are based on importance sampling, which allows for fast computations and is not subject to convergence problem inherent to the alternative Markov Chain Monte Carlo methods.

Keywords: Bayesian inference; Value at Risk; Expected Shortfall; Efficient importance sampling; mixture of Student's t distributions; Nonlinear state space models.

Contents

1 Introduction

We consider the evaluation of two popular financial risk measures, Value at Risk and Expected Shortfall. The focus is laid on the market risk related to changes in the portfolio value due to the fluctuations of market factors. Our analysis is performed in a Bayesian fashion, where the object of interest is a posterior predictive density. We adopt a model-based approach, which means that the estimation of the dynamic features of the volatility and of the implied risk measures is performed given a chosen model. Such an approach is considered as a standard in the literature as it provides a more systematic analysis compared to other ad hoc procedures (cf. Jungbacker and Koopman, 2009 and Hoogerheide and van Dijk, 2010). The key feature of the model-based approach is an explicit specification of the time-varying properties of the volatility, including its conditional distribution, using a parametric model. We employ the *Quick Evaluation of Risk using Mixture of t approximation* algorithm (QERMit) of Hoogerheide and van Dijk (2010) due to the accuracy of the estimates it provides and the efficiency of the computations it is based on.

The original QERMit method heavily relies on two elements: the algorithm used to approximate the candidate posterior density and the class of models it admits. As the former the authors take the *Adaptive Mixture of t* (AdMit) method of Hoogerheide et al. (2007). AdMit uses a mixture of the multivariate Student's t distributions to approximate (a kernel of) a target distribution which can be highly non-elliptical (e.g. multimodal or skewed). The final mixture is constructed iteratively, by adding consecutive components to the previous mixture starting with a single multivariate Student's t density. Despite its relatively good performance, the AdMit method suffers from two serious shortcomings. First, adding of a new component is carried out without updating of the parameters of the components in the current mixture, which may lead to a faulty location of the new component. Second, the number of degrees of freedom is fixed to 1 for all the components[1]. Together, these limitations make AdMit a rather difficult tool to work with, often yielding not fully reliable results.

The latter ingredient of the QERMit approach, i.e. the class of models which it can be applied to, is restricted to the observation driven models (cf. Cox, 1981). These models gained a substantial popularity due to their simple estimation. In these models parameters are stochastic processes, which are perfectly predictable given the current observation set, so that likelihood function is available in a closed-form. Typical representatives of this class are variants of the GARCH model stemming from the works of Engle (1982) and Bollerslev (1986). A straightforward estimation may come at a price of an inferior out-of-sample performance as compared to more complex models[2]. Given out main focus on the predictive densities, it is crucial to allow for models which potentially can perform better in terms of yielding more accurate forecasts.

In this thesis we upgrade the basic QERMit algorithm by removing the both above-mentioned limitations. The exposition is carried out in two steps. First, we implement the QERMit method where we replace AdMit with a superior approximation technique, the *Mixture of t by Importance Sampling weighted Expectation Maximization* (MitISEM) algorithm of Hoogerheide et al. (2012). Not only is MitISEM a more flexible tool for constructing an approximation to a given distribution of interest, but also it is computationally more efficient. MitISEM allows for a simultaneous adjusting of the parameters of all mixture components, including the number of the degrees of freedom. This proves out to be crucial for the approximation accuracy. We compare the performance of the modified QERMit approach based on MitISEM with the one of the original QERMit method based on AdMit. Moreover, we contrast the

[1] In our application we modified this restriction allowing for the number of the degrees of freedom to be specified prior to the computations. However, we did not alter the main assumption on the number of degrees of freedom, i.e. that they remain fixed during the program run.

[2] For instance, Hol and Koopman (2002) find that without the information on intraday volatility, there is room for improvement of the GARCH-based forecasts. This, however, requires using of different classes of models.

results generated with both versions of QERMit with these from an unsophisticated evaluation, called the *direct approach*.

Second, we extend the basic QERMit framework to allow for the latent variables in the underlying model, so that the technique can be applied to the class of the parameter driven models. A fundamental example of a model from this class is the Stochastic Volatility (SV) model (cf. Taylor, 1986, Harvey et al., 1994). In the parameter driven models the unobserved volatility process is subject to idiosyncratic innovations, resulting in two error processes driving the observations dynamics. In consequence, both the in-sample and the out-of-sample model performance is likely to be ameliorated as compared to the one of the observation driven models. Incorporating of an unobserved state in the model poses, however, substantial computational difficulties. The reason for the practical problems is that, in general, the likelihood function is not given in a closed form in these models. To overcome this obstacle, numerous approximative methods for inference have been proposed[3]. We adopt the *Numerically Accelerated Importance Sampling* (NAIS) method of Koopman et al. (2015), which has been proven to be numerically more efficient than the competing approximation techniques.

We stress the significance of the importance sampling (IS) as the main sampling tool used in our study. These methods have a long history in statistics, being considered already in Kahn and Marshal (1953), Marshall (1956) and Hammersley and Handscomb (1964). Their introduction to Bayesian econometrics is due to Kloek and van Dijk (1978). Not only underlie the IS techniques the methods used in the computations, MitISEM and NAIS, but also serve to obtain the ultimate VaR and ES estimates. An alternative class of algorithms is provided by the MCMC methods, which have been intensively used for computations in various fields, including time series econometrics. We opt for the IS-based techniques for several reasons. First, they are not subject to the chain convergence problem, inherent to the MCMC algorithms (cf. Gelman, 1995). Second, they easily allow for parallelisation, which substantially boosts the speed of computations. Third, a substantial interest has recently been devoted to the on-line inference problems, which can be efficiently approached using via the Sequential Monte Carlo methods based on IS (cf. Doucet et al., 2001, for a comprehensive study).

The outline of the paper is as follows. Section 2 defines the two risk measures of interest and presents the Bayesian approach to the financial risk estimation based on IS. In Section 3 we discuss the QERMit technique, already adopting a generalised perspective. The key methods used in our analysis, MitISEM and NAIS, are described in Section 4. Section 5 presents our first contribution, i.e. the replacement of AdMit as the posterior approximation algorithm in QERMit by MitISEM, and illustrates the achieved gain in the risk estimation accuracy using a series of daily S&P 500 returns. The extension of the allowed model class to the parameter driven models is discussed in Section 6, where we also indicate some numerical problems which we encountered in the application of our novel method to a series of IBM stock daily returns. Section 7 concludes.

2 Bayesian Risk Estimation

A correct and precise evaluation of the financial risk is without doubts desirable, not to say necessary, for all types of agents active in the global economy. In particular, the development of the derivatives markets, relying on complex index or stock based instruments, calls for an accurate estimation of the *market* risk[4]. The recent experiences from the Great Recession clearly show what the consequences of

[3] In the context of the SV model, Harvey et al. (1994) suggested the so called Quasi-Maximum Likelihood approach, where the estimation procedure is based on the approximative linear, Gaussian state space model. Alternative approximative methods are based on the numerical integration of the likelihood function, as in Kitagawa (1987) and Fridman and Harris (1998). A vast strand in the literature has been devoted to Markov chain Monte Carlo (MCMC) estimation, cf. Jacquier et al. (1994) and Kim et al. (1998). Finally, the methods based on Monte Carlo likelihood evaluation in the context of the SV model were developed in Danielsson (1994) and Sandmann and Koopman (1998).

[4] We define various types of risk below

the financial risk underestimation might be, with banks bankruptcies, liquidity spirals and the overall financial distress being just the first few to mention. On the other hand, an excessively prudent approach to risk evaluation is likely to prohibit investors from exploring profitable asset allocation opportunities. This explains why the financial risk estimation stands in the centre of attention of a vast strand of the econometric literature.

In addition, in the constantly changing economic environment, often driven by market sentiments and *beliefs* about the current and the future state of the system, it is necessary to be able to promptly update the forecasts and to use the "expert knowledge" (external to the model) in predictions. Bayesian analysis satisfies these two requirements, because it allows for incorporating of the prior beliefs, as well as it for a sequential updating of the estimates when new data becomes available. The key aspect of the Bayesian estimation is that the observations are treated as fixed, while all the unobserved elements (such as model parameters, latent states or the future observations) are treated as random variables. Hence, the Bayesian approach, with its focus on the joint posterior (predictive) densities, is particularly suited for the financial risk evaluation.

2.1 Risk Measurement

There are different types of financial risk, and for each of them different tools for the quantitative measurement have been developed. As discussed in Jorion (2007), the two main risk categories are the market risk and the credit risk. *Market risk* concerns changes in the investment value in response to the moves of the market risk factors such as stock prices, interest rates and foreign exchange rates. *Credit risk*, also known as default risk, is related to the losses occurring due to a default of the transaction's counterparty, which is mainly considered in case of bond evaluation. Other types of risk include *operational risk* or *liquidity risk*. In this thesis, we focus on the first category, i.e. the market risk. Two standard measures used in this context are *Value at Risk* (VaR) and *Expected Shortfall* (ES), which we discuss below.

2.1.1 Value at Risk and Expected Shortfall

Value at Risk can be intuitively understood as an answer to a question about the worst loss within a certain time horizon from a probabilist's perspective, i.e. given a specified confidence level. More precisely, VaR is the quantile of the predictive distribution of losses over the horizon of interest. An important example of VaR is the one accepted by the Basel Committee (Basel Committee on Banking Supervision, 1995) and defined as the 99% quantile (1% lower tail) of a loss distribution for a two-week (10-day-ahead) horizon. Following Jorion (2007), we formally define the $100\alpha\%$ VaR as the $100(1-\alpha)\%$ quantile of the percentage return's distribution, i.e.

$$100\alpha\% VaR \equiv \inf \left\{ x \in \mathbb{R} : \mathbb{P}[X \leq x] \geq \alpha \right\},$$

where X is a random variable of interest.

There are several advantages of using the VaR as a risk measure, such as its conceptual simplicity, a possibility of its estimation via different methods and the fact that nowadays it is considered as a standard tool in risk management. From a theoretical point of view, however, it has some undesirable properties. First of all, in general it is not a *coherent* risk measure[5], since it fails to be subadditive. This might negatively affect an investor's propensity to diversification. Another issue with the VaR is that because it is barely a quantile, it gives no insight into the properties of the tail it determines. Hence, an alternative

[5] Artzner et al. (1999) axiomatise the concept of coherent risk measure using four properties, desirable for a measure of the financial risk. These are: monotonicity, invariance to translation, homogeneity and subadditivity. In particular, the interpretation of the latter is that merging of portfolios shall not increase the overall risk.

risk measure has been developed, which summarises the losses exceeding the VaR, i.e. the Expected Shortfall (cf. McNeil and Frey, 2000). The ES answers the question, what loss can be expected, given that it overreaches the VaR. Formally, the $100\alpha\%$ ES is the conditional expected loss given that the loss exceeds $100\alpha\%$ VaR, i.e.

$$100\alpha\% ES \equiv \mathbb{E}\big[X|X < 100\alpha\% VaR\big].$$

In contrast to the VaR, the ES subadditive, which allows for a conservative risk assessment of a portfolio. Taking into account the popularity of the VaR and the advantages of the ES over the VaR, we consider both risk measures in the thesis.

2.1.2 Direct Estimation

Below, we describe the *direct approach* to the Bayesian estimation of the risk measures defined in the previous Subsection. Suppose we have a sample of n historical logreturns, $y = \{y_t\}_{t=1}^{n}$, and let θ denote the vector of all the model parameters[6]. We are interested in the h-day-ahead forecast of the $100\alpha\%$ VaR or ES, which are determined by the profit-loss function PL, mapping the h-vector of the future logreturns $y^* = \{y_{n+1}, \dots, y_{n+h}\}$ into a scalar. Because we consider 100× the logreturns, the PL function can be defined as

$$PL(y^*) \equiv PL(y_{n+1}, \dots, y_{n+h}) = 100 \left[\exp\left(\sum_{s=1}^{h} y_{n+s}/100 \right) - 1 \right]. \tag{2.1}$$

Obviously, it is positive for profits and negative for losses. The forecasting density of y^* is given by $p(y^*|\theta, y)$. A straightforward way to estimate the h-day-ahead in a Bayesian fashion is to proceed as follows.

1. Simulate $\theta^{(i)}$, $i = 1, \dots, N$, a set of parameter draws, from the posterior distribution of the parameters, $p(\theta|y)$.

2. Generate $y^{*(i)}$, $i = 1, \dots, N$, the corresponding paths of the future logreturns, given the parameter draws $\theta^{(i)}$, $i = 1, \dots, N$, and observations y, i.e.

$$y^{*(i)} \sim p(y^*|\theta^{(i)}, y), \qquad i = 1, \dots, N.$$

3. Compute $PL(y^{*(i)})$, $i = 1, \dots, N$, the corresponding values of the profit-loss function.

4. Sort the values of the profit-loss function ascendingly, denoting the resulting permutation $PL^{(j)} := PL(y^{*(j)})$, $j = 1, \dots, N$.

5. Compute the $100\alpha\%$ VaR and ES estimates as

$$\widehat{VaR}_{DA} = PL^{((1-\alpha)N)}, \tag{2.2}$$

$$\widehat{ES}_{DA} = \frac{1}{(1-\alpha)N} \sum_{j=1}^{(1-\alpha)N} PL^{(j)}. \tag{2.3}$$

The meaning of formulae (2.2) and (2.3) is straightforward. The $100\alpha\%$ VaR is estimated as the $(1-\alpha)N$ lowest profit-loss value. Then, the corresponding ES is obtained by averaging the profit-loss values which do not exceed the computed VaR.

[6] Which may also include the latent state variables.

Although conceptually lucid, the direct approach is clearly inefficient. Because one is ultimately interested in the estimation of tail events, it is suboptimal to mainly focus on the "complement of the tail" (i.e. the non-extreme events) and sample from the tail only occasionally. Given the simulated sample of length N, only a small subsample of $(1-\alpha)N$ draws is used in the estimation, which negatively affect the precision of the estimates. Consequently, if one wishes to obtain a certain, high level of precision, a substantial increase in the number of draws is required. To overcome this problem, Hoogerheide and van Dijk (2010) developed the VaR and ES estimation method based on importance sampling. Their approach focuses specifically on the so-called *high-loss* subspace, which is in line with the theoretical result on the optimal sampling design of Geweke (1989). The next two Sections provide the discussion of the fundamentals of this methodology.

2.2 Importance Sampling Estimation

As discussed in Subsection 2.1.2, to overcome the inefficiency of the direct approach to VaR and ES estimation, Hoogerheide and van Dijk (2010) suggested a superior method based on importance sampling. This approach provides a focus on an *important* part of the posterior distribution, which is obtained first, by an appropriate weighting of draws, and second, by generating them from an optimal, tail-focused density. Before explaining the details of the method in Subsection 2.2.2, we briefly recall the principles of the IS estimation in Subsection 2.2.1. The discussion of the second question, i.e. the construction of the optimal importance density, is provided in Section 2.3.

2.2.1 IS Principles

Let X and Y be random variables, where X takes values in $\mathcal{X} \subset \mathbb{R}^d$. For simplicity assume that the joint distribution of X and Y allows for the joint density $p(x, y)$. We will denote by $p(\cdot|\cdot)$ and $p(\cdot)$ the conditional and marginal densities, respectively. Suppose we are interested in estimation of the (conditional) mean of an arbitrary measurable function of X, $f : \mathcal{X} \to \mathbb{R}$, given by

$$\begin{aligned}\bar{f} &= \mathbb{E}\left[f(X)|Y\right] \\ &= \int f(x)p(x|y)dx. \end{aligned} \tag{2.4}$$

If one could sample directly from $p(x|y)$, the *Monte Carlo* (MC) estimate (2.4) would be given by straightforward expression

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}), \tag{2.5}$$

where $x^{(1)}, \dots, x^{(N)}$ are independent draws from $p(x|y)$. By the Strong Law of Large Numbers, (2.5) is strongly consistent, while by the Central Limit Theorem it is asymptotically normal, provided that the variance $\mathrm{Var}_p f(X)$ exists.

However, it is typically difficult to sample directly from $p(x|y)$, therefore in practice one usually resorts to drawing from the so called *importance density* $q(x_t|y_t)$ with the support including the one of the density of interest $p(x|y)$. It is assumed the sampling from $q(x_t|y_t)$ is relatively easy and inexpensive. This method of simulation based estimation is called *importance sampling* (IS). Below we present the basic idea behind this MC method.

To start with, notice that (2.4) can be expressed using $q(x|y)$ in the following way

$$\begin{aligned}\bar{f} &= \int f(x)\frac{p(x|y)}{q(x|y)}q(x|y)dx \\ &= \mathbb{E}_q\left[f(X)\frac{p(X|Y)}{q(X|Y)}\right] \\ &= \mathbb{E}_q\left[f(X)W(X,Y)\right], \end{aligned} \tag{2.6}$$

where $\mathbb{E}_q$ stands for expectation with respect to density q and

$$W(x,y) = \frac{p(x|y)}{q(x|y)} \tag{2.7}$$

is known as the *importance weight* function. Notice, that since the importance weight function is defined as the likelihood ratio, it is the Radon–Nikodým derivative of the true distribution $p(\cdot)$ with respect to the importance distribution $q(\cdot)$. Generally, it depends on x and y, however, in the remaining part of the work we skip the arguments for notational convenience. Hence, with some abuse of notation, we will use the same symbols to denote the weight functions as function of random variables and of real numbers.

Since the joint density factorises as $p(x,y) = p(x|y)p(y)$, one can express (2.7) as

$$\begin{aligned} W &= \frac{1}{p(y)}\frac{p(x,y)}{q(x|y)} \\ &= \frac{1}{p(y)}w, \end{aligned} \tag{2.8}$$

with $w = p(x,y)/q(x|y)$, so that W is w corrected for the (unconditional) observation density. Then, (2.6) becomes

$$\bar{f} = \frac{1}{p(y)}\mathbb{E}_q\left[f(X)w\right], \tag{2.9}$$

Notice, that by taking $f \equiv \mathbf{1}$ one can obtain from (2.9) that

$$\mathbb{E}_q\left[w\right] = p(y),$$

which implies that (2.9) can be rewritten as follows

$$\bar{f} = \frac{\mathbb{E}_q\left[f(X)w\right]}{\mathbb{E}_q\left[w\right]}. \tag{2.10}$$

Next, we can estimate (2.10) using a random sample $x^{(1)},\dots,x^{(N)}$ drawn from the importance distribution $q(x|y)$. The required estimate has the form

$$\begin{aligned}\hat{f} &= \frac{N^{-1}\sum_{i=1}^N f(x^{(i)})w^{(i)}}{N^{-1}\sum_{i=1}^N w^{(i)}} \\ &= \frac{\sum_{i=1}^N f(x^{(i)})w^{(i)}}{\sum_{i=1}^N w^{(i)}} \end{aligned} \tag{2.11}$$

with

$$w^{(i)} = \frac{p(x^{(i)},y)}{q(x^{(i)}|y)}, \tag{2.12}$$

the importance weight of the draw $x^{(i)}$.

2.2.2 IS Risk Estimation

To estimate $100\alpha\%$ VaR using importance sampling one first needs to specify the function f in (2.11). By definition, the $100\alpha\%$ VaR is a $100(1-\alpha)\%$ quantile of the profit-loss function which means it is implicitly defined as

$$\mathbb{P}\big[PL(X) \leq VaR\big] = 1 - \alpha.$$

For the indicator function of a set $C = (-\infty, c]$, $c \in \mathbb{R}$, it holds that

$$\mathbb{P}[X \in C] = \mathbb{E}\left[\mathbb{I}_C(X)\right],$$

so one can consider $f(x;c) = \mathbb{I}_{\{PL(x)\leq c\}}(x)$ with $c = \widehat{VaR}$, since

$$\begin{aligned}\mathbb{E}[f(X;c)] &= \mathbb{E}\left[\mathbb{I}_{\{PL(X)\leq c\}}\right] \\ &= \mathbb{P}\left[PL(X) \leq \widehat{VaR}\right].\end{aligned}$$

Then, the IS estimator $\widehat{VaR}_{IS}$ of the $100\alpha\%$ VaR is obtained by solving

$$\widehat{\mathbb{E}[f(X)]}_{IS} = 1 - \alpha. \tag{2.13}$$

In practice, solving of (2.13) amounts to the following simple procedure.

1. Simulate $\theta^{(i)}$, $i = 1, \ldots, N$, a set of parameter draws[7], from $q(\theta|y)$, the importance distribution for the posterior of the parameters.
2. Compute $w^{(i)}$, $i = 1, \ldots, N$, the importance weights of the draws $\theta^{(i)}$, using (2.12).
3. Generate $y^{*(i)}$, $i = 1, \ldots, N$, the corresponding paths of the future logreturns, given the parameter draws $\theta^{(i)}$, $i = 1, \ldots, N$, and observations y, i.e.

$$y^{*(i)} \sim p(y^*|\theta^{(i)}, y), \qquad i = 1, \ldots, N.$$

4. Compute $PL(y^{*(i)})$, $i = 1, \ldots, N$, the corresponding values of the profit-loss function.
5. Sort the values of the profit/loss function ascendingly, denoting the resulting permutation $PL^{(j)} := PL(y^{*(j)})$, $j = 1, \ldots, N$.
6. Find the value $PL(y^{*(k)})$ for which it holds that

$$\sum_{j=1}^{k} w(\theta^{(j)}) \leq 1 - \alpha, \qquad \text{and} \qquad \sum_{j=1}^{k+1} w(\theta^{(j)}) > 1 - \alpha,$$

and take this value as the $100\alpha\%$ VaR estimate $\widehat{VaR}_{IS}$.

Given the IS estimate $\widehat{VaR}_{IS}$, the k-th value of the sorted profit/loss function as discussed above, the IS estimator $\widehat{ES}_{IS}$ of corresponding ES is simply computed as

$$\widehat{ES}_{IS} = \sum_{j=1}^{k} w(x^{(j)}) PL(x^{(j)}) / \sum_{j=1}^{k} w(x^{(j)}),$$

i.e. it is the weighted average of the profit-loss values up to the VaR estimate.

[7] Which contains the future error terms, and may also include the latent state variables.

2.2.3 Numerical Standard Errors

To assess the accuracy of the IS estimator (2.11) Geweke (1989) considers its numerical standard error (NSE), which is a square root of the estimate of the asymptotic variance of the IS estimator[8]. However, in the case of the IS estimation of VaR or ES, the required NSEs do not follow from Geweke (1989). Regarding the NSE of the former, Hoogerheide and van Dijk (2010) show how to derive it by applying the delta rule. We present their approach in Appendix A.

For the NSE of $\widehat{ES}_{IS}$, however, no analytical (approximative) formula is known, and one needs to resort to simulations. The reason for this is that although ES itself can be expressed as an integral (i.e. the expectation) over a certain set, the boundary of this set, i.e. VaR, is a random variable. Hence, in the case of the ES estimation, we not only estimate the ultimate integral, but also the region over which the integration takes place. The MC procedure to estimate the NSE of $\widehat{ES}_{IS}$, developed by Hoogerheide and van Dijk (2010), is discussed in Appendix A.

2.3 Optimal Importance Density

The choice of the candidate density is of a crucial importance for the performance of the IS estimation. The optimal importance distribution ought to minimise, given the specified number of draws, the numerical standard error of the IS estimator $\bar{f} \equiv \mathbb{E}\left[f(X)\right]$, where f is the function of interest of the random variable X, which has the density $\tilde{p}(x)$ with the kernel p. According to Geweke (1989), this optimal importance distribution has the kernel[9]

$$q_{opt} \propto |f(x) - \bar{f}|p(x), \tag{2.14}$$

provided that $\mathbb{E}[|f(X) - \bar{f}|] < \infty$. For the case of $f(x) = \mathbb{I}_S(x)$, i.e. the indicator function of the set S, we have

$$\mathbb{E}[f(X)] = \mathbb{P}[X \in S] =: \bar{p}$$

and the optimal importance density is given by

$$q_{opt}(x) \propto \begin{cases} (1-\bar{p})p(x), & \text{for } x \in S \\ \bar{p}p(x), & \text{for } x \notin S \end{cases}, \qquad \text{or} \qquad q_{opt}(x) = \begin{cases} c(1-\bar{p})\tilde{p}(x), & \text{for } x \in S \\ c\bar{p}\tilde{p}(x), & \text{for } x \notin S \end{cases},$$

where c is a constant, which results in[10]

$$\int_{x \in S} q_{opt}(x)dx = \int_{x \notin S} q_{opt}(x)dx = \frac{1}{2}. \tag{2.15}$$

Condition (2.15) implies that the half of the total mass of the candidate shall be located in the region of interest S, while the remaining half – outside that region. Such a split is the consequence of using only

[8] Geweke (1989) shows that under the standard regularity conditions

$$\sqrt{n}\left(\widehat{\mathbb{E}[f(X)]}_{IS} - \mathbb{E}[f(X)]\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{IS}\right).$$

[9] Geweke (1989) also points out three considerable difficulties related to this form of the optimal importance density. First, it depends on the particular form of the function f in question. Second, it involves the estimate of interest $\bar{f}$. Third, it preassumes that sampling from it is feasible.

[10] This is obtained by noting that

$$\int_{x \in S} q_{opt}dx = c(1-\bar{p})\int_{x \in S} \tilde{p}(x)dx = c(1-\bar{p}) = c\bar{p}(1-\bar{p}) = c\bar{p}\int_{x \notin S} \tilde{p}(x)dx = \bar{p}\int_{x \notin S} q_{opt}dx.$$

while $\int_{x \in S} q_{opt}(x)dx + \int_{x \notin S} q_{opt}(x)dx = 1$.

the kernel of the target distribution and not its proper density, which makes it necessary to adequately normalise the weights via sampling from the whole domain.

The above result was derived by Hoogerheide and van Dijk (2010), who apply it in the context of the VaR estimation. Then, S is interpreted as the "high loss region", i.e. the subspace of the profit-returns space with the $100(1-\alpha\%)$ lowest values, while the optimal importance density prescribes that 50% of draws shall represent high losses while the other 50% the remaining profit-loss realisations. Figure 2.1 illustrates the construction of the optimal candidate for the VaR estimation.

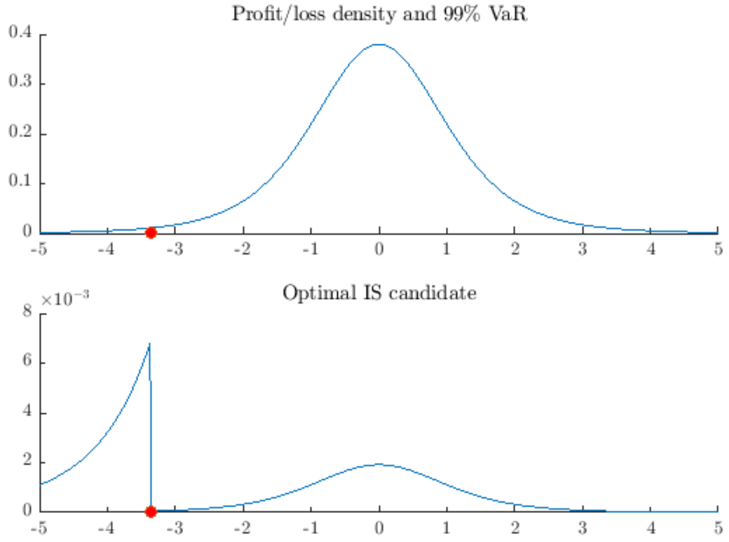


Figure 2.1: Construction of the optimal IS density. Exemplary profit/loss function (Student's t with 5 degrees of freedom) and the implied 99% VaR (top). The optimal importance candidate density for the VaR estimation (bottom).

As pointed out in Hoogerheide and van Dijk (2010), the formula (2.14) cannot be applied to deriving the optimal importance density for the ES estimation as the latter is based on the unknown value of VaR, making the problem nonstandard. One can notice, however, that such an optimal importance density would be characterised by fatter tails than the optimal candidate for the VaR estimation. In practice, however, if the latter has sufficiently thick tails, i.e. reflects the conservative approach to risk estimation, the same candidate may be used for both estimation purposes.

The above result gives rise to an efficient estimation procedure called the Quick Risk Evaluation by Mixture of t Approximation (QERMit) developed by Hoogerheide and van Dijk (2010). Since its thorough discussion is postponed until Section 3, here we only present the main idea behind this approach. Basically, QERMit consists of two steps. First, one approximates the optimal candidate density q_{opt} by $\hat{q}_{opt}$. This approximation requires obtaining of a preliminary estimate of $100\alpha\%$ VaR, which serves as a rough "separator" of the high-loss subspace and its compliment (cf. the red dot in the top panel in Figure 2.1). The construction of $\hat{q}_{opt}$ is additionally complicated by the fact that q_{opt} is bimodal and asymmetric. In Section 4.1 we present the MitISEM algorithm due to Hoogerheide et al. (2012), which allows for approximating such nonstandard density shapes. Second, one performs the IS estimation as discussed in Section 2.2.2 using $\hat{q}_{opt}$ as the importance density.

3 Quick Evaluation of Risk using Mixture of t Approximation

In the Introduction we stressed the two main ingredients of the original QERMit algorithm of Hoogerheide and van Dijk (2010). However, the underlying idea of the inference on the tail of the posterior profit/loss function, i.e. IS estimation using the candidate focused on the high-loss region, specifies neither the algorithm to construct the approximation in question, nor the class of models to which the technique can be applied.

The aim of this section is to present the two steps of the QERMit method: the approximation step and the IS estimation step. However, taking the above remark into consideration, we do not simply restate the procedure described by Hoogerheide and van Dijk (2010). Instead, we discuss the QERMit method from a more general perspective, already incorporating its intended modification and extension, which are the subject of this thesis.

3.1 QERMit Step 1: Approximation

As concluded in Subsection 2.3, construction of the approximation $\hat{q}_{opt}$ to the optimal importance density q_{opt} is not a straightforward task. The reason is that typically it is multidimensional and bimodalHoogerheide and van Dijk (2010) point out that in some cases the optimal importance density can be even trimodal. Such a shape arises e.g. in short selling of a straddle of options., and therefore far from being elliptical. In consequence, the optimal candidate needs to be tailored to the problem at hand, so that the speed and the reliability of the chosen method becomes even more crucial. Below we adopt the method of Hoogerheide and van Dijk (2010) who suggested the following procedure to obtain the approximation to the optimal importance density.

1. Approximate the kernel of the posterior density of parameters by $q_1(\theta)$ using a chosen, reliable method.
2. Sample $\theta^{(i)}$, $i = 1, \dots, N$, a set of parameter draws from $q_1(\theta)$ using the independence chain MH.
3. Obtain a preliminary $100\alpha\%$ VaR estimate

   2.1. Generate $y^{*(i)}$, $i = 1, \dots, N$, the corresponding paths of the future logreturns, given the parameter draws $\theta^{(i)}$, $i = 1, \dots, N$, and observations y, i.e.

   $$y^{*(i)} \sim p(y^*|\theta^{(i)}, y), \qquad i = 1, \dots, N.$$

   2.2. Compute $PL(y^{*(i)})$, $i = 1, \dots, N$, the corresponding values of the profit-loss function.

   2.3. Sort the values of the profit-loss function ascending, denoting the resulting permutation $PL^{(j)} := PL(y^{*(j)})$, $j = 1, \dots, N$.

   2.4. Compute $\widehat{VaR}_{prelim}$, a preliminary $100\alpha\%$ VaR estimate as $\widehat{VaR}_{prelim} = PL^{((1-\alpha)N)}$.

4. Approximate $p_2(\theta, y^*)$, the kernel of the joint high loss density of parameter vector θ and future returns y^*, by $q_2(\theta, y^*)$, using the same method as in Step 1. This is done by imposing the restriction $PL(y^*) \leq \widehat{VaR}_{prelim}$ on the joint density $q(\theta, y^*)$.

In the original QERMit paper, the method employed in Step 1 was the *Adaptive Mixture of t* (AdMit) algorithm of Hoogerheide et al. (2007). As already indicated in the Introduction, we replace AdMit with a superior procedure due to Hoogerheide et al. (2012), the *Mixture of t by Importance Sampling weighted Expectation Maximisation* (MitISEM) algorithm. The motivation behind this modification is given in Section 4.1, where we investigate methods for posterior approximation.

3.2 QERMit Step 2: IS Estimation

The approximation $\hat{q}_{opt}$ to the optimal importance density q_{opt} is given by a mixture of the form

$$\hat{q}_{opt}(\theta, y^*) = \frac{1}{2} q_1(\theta) p(y^*|\theta, y) + \frac{1}{2} q_2(\theta, y^*), \tag{3.1}$$

where the mixing weights correspond to condition (2.15) for the optimal mass allocation in the optimal candidate density. The second term on the right-hand side in (3.1) is self-evident, as it corresponds to the high-loss region, i.e. the realisation of y^* which lead to the profit-loss values equal or lower than $\widehat{VaR}_{prelim}$. The first term deserves somewhat more comment. First, it describes the joint posterior of θ and y^*, which is not the compliment of the high-loss subspace, as required in (2.15). This means that ultimately one oversamples from the high-loss region. This is not harmful, however, as it allows for a more "conservative" risk evaluation.

Second, to improve sampling efficiency, the joint distribution of θ and y^* is factorised into the posterior for θ and the conditional forecasting density of y^*, given θ. The latter does not need to be approximated since it is implied by the chosen model. In this way only the density of θ is approximated, which obviously has a lower dimensionality than the joint distribution. This in turn reduces the computational time.

As argued in Hoogerheide and van Dijk (2010), the reason why one considers the joint distribution for θ and y^*, even though the profit-loss mapping is a function of y^* only, is that the kernel of the predictive density for y^* is usually unavailable explicitly. On contrary, the kernel of the joint posterior can be derived as

$$\begin{aligned} p(\theta, y^*|y) &= p(\theta|y) p(y^*|\theta, y) \\ &\propto p(\theta) p(y|\theta) p(y^*|\theta, y), \end{aligned}$$

where $p(\theta)$ is the prior density for the model parameters.

The final remark concerns the actual variable used to approximate the high-loss density. The above discussion referred to the densities (joint, marginal) of the future *returns*, i.e. we analysed the realisation of y^*. However, an equivalent (in terms of the implied estimates) procedure can be based on the future *disturbances* $\varepsilon^* = \{\varepsilon_{T+1}, \ldots, \varepsilon_{T+h}\}$. The latter approach has the advantage of being much easier to perform in practice, as the relationships between the financial returns are typically highly complex, e.g. characterised by volatility clustering, while the error terms are usually assumed to be serially independent. Hence, we will implicitly understand $p(\theta, \varepsilon^*)$ when referring to $p(\theta, y^*)$.

4 Methods for Modification and Extension of QERMit

In this Section we explain two methods which play a key role in the modification and extension of the basic QERMit algorithm. We begin with a general discussion of the density approximation by Mixture of Student's t distributions in Section 4.1. Then, we present the details of MitISEM, which is used to replace AdMit as the approximation algorithm in QERMit. Since we aim at incorporating into the analysis the latent signal process, inherent to the parameter driven models, the original MitISEM procedure needs to be modified. In this respect we follow Barra et al. (2014), who notice that, basically, the only required change consists in a different computation of the importance weights. Now, the latter become also a function of the unobserved signal, not only of the model parameters. This, in turn, calls for the adoption of the techniques used in the analysis of nonlinear non-Gaussian state space models, which we discuss in Section 4.2.

Prior to proceeding to the core of the analysis, however, we briefly recall the framework of the general state space models (SSM), which are considered throughout the analysis. The SSM is a class which incorporates numerous models of interest and provides a flexible tool to approach diverse real-life problems. The main reason for this versatility in the context of time-series modelling is the focus of SSM on the *state* or signal vector driving the dynamics of the investigated system. Furthermore, *nonlinear* dynamics frequently arises in science, engineering, economics and a number of other fields, while *non-Gaussian* disturbances are either natural or practical in applications. As in Koopman et al. (2015) we restrict our attention to the class of models with a linear Gaussian state transition equation.

The consequence of such a generality is a nontrivial inference for these models, especially from the Bayesian perspective, which has been approached by numerous stands of literature. In particular, particle filtering methods, dating back to Kitagawa (1996) and Gordon et al. (1993), very popular in e.g. signal processing (cf. e.g. Arulampalam et al., 2002), have gained a substantial attention also in econometrics (cf. e.g. Pitt et al., 2012). However, due to our simplifying assumption on the linear Gaussian nature of the state process, we do not need to employ these computationally intensive methods. Instead, we focus on the numerically more efficient importance sampling techniques (cf. Barra et al., 2014), in which we follow Shephard and Pitt (1997), Durbin and Koopman (1997), Richard and Zhang (2007) and Koopman et al. (2015).

For a time series of observations $\{y_t\}_{t=1}^n$, consider the nonlinear non-Gaussian SSM consisting of the observation density and the linear Gaussian state transition equation

$$y_t|\theta_t \sim p(y_t|x_t;\theta), \qquad x_t = Z_t\alpha_t, \qquad t = 1,\dots,n, \tag{4.1}$$

$$\alpha_{t+1} = d_t + T_t\alpha_t + \eta_t, \qquad \alpha_1 \sim N(a_1, P_1), \qquad \eta_t \sim N(0, Q_t), \tag{4.2}$$

where x_t is the latent $q \times 1$ signal vector, α_t is the $m \times 1$ state vector, η_t is the vector of uncorrelated Gaussian innovations. The dynamic properties of the stochastic vectors y_t, x_t and α_t are characterised by the potentially time-varying yet deterministic system matrices: the $m \times 1$ constant vector d_t, the $m \times m$ transition matrix T_t, the $m \times m$ variance matrix Q_t. The similar assumptions also apply to the $q \times m$ selection matrix Z_t and the parameters of the initial distribution of the state α_t, i.e. the mean vector a_1 and the variance matrix P_1. Finally, θ denotes the unknown vector of the model parameters, consisting of the coefficient of the observation density $p(y_t|x_t;\theta)$ and the system variables. Below, for notational convenience we often consider the stacked vectors $y = (y_1^T, \dots, y_n^T)$, $x = (x_1^T, \dots, x_n^T)$ and $\alpha = (\alpha_1^T, \dots, \alpha_n^T)$.

Importantly, the signal x_t is a linear function of the state α_t, and is assumed to be low-dimensional, which is in contrast to the potentially high dimensionality of the latter. Notice that although the state transition is linear Gaussian, the framework can still accommodate for a wide range of nonlinear non-Gaussian models due to a nonlinear non-Gaussian observations equation.

4.1 Posterior Approximation

The choice of the candidate density is crucial for the performance of the IS estimation. Clearly, as pointed out in Geweke (1989), the importance density should resemble the target density at the same time remaining easy to sample from. Moreover, the tails of the importance density need to be thicker than those of the target density, in order to minimise the risk of omitting subsets of the target's support. Finding of an appropriate candidate becomes particularly cumbersome when the shape of the target density is non-elliptical. Such shapes, however, are commonly encountered in the Bayesian analysis, where posterior densities often exhibit multimodality or skewness.

4.1.1 Mixtures of Student's t Distributions

A standard approach to overcome this problem is to approximate the target density with a mixture of basis densities[11]. Below, following Hoogerheide et al. (2007) and Hoogerheide et al. (2012), we use mixtures of Student's t densities. The primary reason for this choice of the basis functions is that Student's t distributions have thicker tails than the normal distributions, which, as discussed above, is an important property of a candidate distribution. This property makes them also robust to outliers, i.e. the importance weights assigned to atypical observations are reduced (cf. Peel and McLachlan, 2000), which makes the IS estimation more efficient and stable. Last but not least, mixtures of Student's t distributions are easy to sample from, and the subsequent draws evaluation is quick, making the whole estimation procedure effective.

Several methods to construct the approximating mixture of Student's t has been developed, cf. Peel and McLachlan (2000), Svensén and Bishop (2005), Hoogerheide et al. (2007) and Hoogerheide et al. (2012). Below, we will mainly focus on the last approach, called Mixture of t by Importance Sampling weighted Expectation Maximization (MitISEM). Because we will compare its performance with the one of an earlier algorithm, called the Adaptive Mixture of t (AdMit) of Hoogerheide et al. (2007), here, we provide a brief discussion of the differences between both approaches. First, the objective function in AdMit is the coefficient of variation of the importance weights (i.e., the standard deviation divided by the mean), which is directly minimised via numerical optimisation. On contrary, MitISEM aims at minimising the Kullback-Leibler divergence, which is an indirect way to minimise the variance of the IS estimator. This makes the latter method quicker and more reliable, as it avoids the computationally intensive numerical optimisation step. Second, MitISEM is a "fully adaptive" algorithm, as each time a new candidate component is added to the old mixture, the parameters of all the components in the new mixture are jointly optimised, whereas in AdMit only the parameters of the new component are optimised, with those of the old mixture not being adjusted any more. Third, the only inputs to MitISEM are draws from the candidate density and their importance weights, while in AdMit one needs to use the kernel of the joint target density. Thus, the latter method cannot be applied to conditional or marginal densities, which makes it useless in our Bayesian analysis based on the factorisation of the joint posterior density.

4.1.2 MitISEM Algorithm

Below, we discuss the basic MitISEM algorithm, which is one of two key tools in our methodology. The exposition is given in a slightly modified way compared to the original one in Hoogerheide et al. (2012), in order to account for potential treatment of the state space models. In this respect, we follow the approach of Barra et al. (2014), who refer to their procedure as to Extended Mixture of t by Importance Sampling weighted Expectation Maximization (EMitISEM).

Because we adopt the Bayesian perspective, we treat all the unknown quantities (i.e. the model parameters, the potential latent states, and the future observation realisations) as random variables and refer to them as to *parameters* (cf. Durbin and Koopman, 2012, chapter 13, who distinguish state parameters and additional parameters, and Hoogerheide and van Dijk, 2010, where model parameters and future returns are estimated jointly). We denote such an augmented parameter vector by $\tilde{\theta}$ to distinguish it from the model parameter vector θ used in the remaining part of this thesis. Consequently, q_{ζ} refers to the joint distribution of the augmented parameter vector. We explain these issues in more detail in Subsection 4.1.3.

The idea behind the MitISEM algorithm is to start with a single Student's t component and iteratively augment it with new components until a specified convergence criterion is met. The mixture parameters

[11] Zeevi and Meir (1997) show that such mixtures can provide an arbitrarily close approximation to any strictly positive density over a compact domain.

at a given iteration step are derived using the importance sampling weighted version of the Expectation-Maximisation (EM) algorithm of Dempster et al. (1977), which we will refer to as ISEM (the details of the ISEM step are given in Appendix B.2).

1. **Initialisation**

   1.1. Define the naive candidate density $q_{\zeta(\bullet)}$ (as discussed below) .

   1.2. Simulate N augmented draws $\tilde{\theta}^{(\bullet,1)}, \dots, \tilde{\theta}^{(\bullet,N)}$ from $q_{\zeta(\bullet)}$. This may require first drawing of model parameters $\theta^{(\bullet,1)}, \dots, \theta^{(\bullet,N)}$ and then conditional simulation of the corresponding signal paths.

   1.3. Evaluate the corresponding importance weights $w^{(\bullet,1)}, \dots, w^{(\bullet,N)}$.

2. **Adaptation**

   2.1. Adapt the candidate $q_{\zeta(\bullet)}$ to $q_{\zeta(0)}$ by setting its mean and variance equal to the IS estimates of the mean μ_0 and the variance Σ_0, based on the draws $\tilde{\theta}^{(\bullet,1)}, \dots, \tilde{\theta}^{(\bullet,N)}$ from $q_{\zeta(\bullet)}$, and setting the number of degrees of freedom for the adapted candidate ν_1 to a specified fixed value (e.g. 5).

   2.2. Simulate N augmented draws $\tilde{\theta}^{(0,1)}, \dots, \tilde{\theta}^{(0,N)}$ from $q_{\zeta(0)}$. This may require first drawing of model parameters $\theta^{(0,1)}, \dots, \theta^{(0,N)}$ and then conditional simulation of the corresponding signal paths.

   2.3. Evaluate the corresponding importance weights $w^{(0,1)}, \dots, w^{(0,N)}$.

3. **ISEM**

   3.1. Adapt the candidate for the model parameters $q^{\theta}_{\zeta(0)}$ to $q^{\theta}_{\zeta(1)}$ by performing the ISEM step (cf. Appendix B.2) based on draws $\tilde{\theta}^{(0,1)}, \dots, \tilde{\theta}^{(0,N)}$ from $q_{\zeta(0)}$. If there are no latent states in the model, $q^{\theta}_{\zeta(0)} = q_{\zeta(0)}$, otherwise $q^{\theta}_{\zeta(0)}$ is the marginal density over the state vector. Call the obtained mode, scale matrix and number of degrees freedom μ_1, Σ_1 and ν_1, respectively.

   3.2. Simulate N augmented draws $\tilde{\theta}^{(1,1)}, \dots, \tilde{\theta}^{(1,N)}$ from $q_{\zeta(1)}$. This may require first drawing of model parameters $\theta^{(1,1)}, \dots, \theta^{(1,N)}$ and then conditional simulation of the corresponding signal paths.

   3.3. Evaluate the corresponding importance weights $w^{(1,1)}, \dots, w^{(1,N)}$.

   3.4. Calculate $CoV^{(1)}$, the coefficient of variation (CoV) of the weights $w^{(1,1)}, \dots, w^{(1,N)}$, where

$$CoV^{(h)} = \frac{\sqrt{\mathbf{E}[(w^{(h)})^2] - \mathbf{E}[w^{(h)}]^2}}{\mathbf{E}[w^{(h)}]}, \tag{4.3}$$

$$\mathbf{E}[(w^{(h)})^m] = \frac{1}{N}\sum_{i=1}^{N}(w^{(h,i)})^m, \qquad m = 1, 2,$$

   is the CoV of the weights obtained for the mixture of h components.

   3.5. Set $H = 2$ and $CoV^{(2)} = \infty$.

4. **Iteration on the number of components**
   While the relative change between $CoV^{(H)}$ and $CoV^{(H-1)}$ is greater than the chosen threshold (e.g. 0.01) keep adding new components to the mixture in the following way.

   4.1. Use a chosen fraction (e.g. $[0.1N]$) of the draws $\tilde{\theta}^{(H-1,1)}, \dots, \tilde{\theta}^{(H-1,N)}$ from the previous mixture $q_{\zeta(H-1)}$ corresponding to the highest IS weights to compute the IS mean and variance. Use these parameters as the starting mode and scale parameters for the new mixture component, μ_H and Σ_H.

4.2. Update the mixture probabilities: assign the starting value for the new component probability η_H (e.q. 0.1) and multiply the old mixture probabilities $\eta_1, \ldots, \eta_{H-1}$by η_H. Set the number of degrees of freedom for the new component ν_H to a specified fixed value (e.g. 5).

4.3. Given the staring parameters of the new mixture $\zeta(H) = \{\mu_h, \Sigma_h, \nu_h, \eta_h\}_{h=1}^{H}$, adapt the candidate for the model parameters $q^{\theta}_{\zeta(H-1)}$ to $q^{\theta}_{\zeta(H)}$ by performing the ISEM step based on the draws from the previous mixture $\tilde{\theta}^{(H-1,1)}, \ldots, \tilde{\theta}^{(H-1,N)}$ and the corresponding weights $w^{(H-1,1)}, \ldots, w^{(H-1,N)}$. If there are no latent states in the model, $q^{\theta}_{\zeta(H-1)} = q_{\zeta(H-1)}$, otherwise $q^{\theta}_{\zeta(H-1)}$ is the marginal density over the state vector.

4.4. Simulate N augmented draws $\tilde{\theta}^{(H,1)}, \ldots, \tilde{\theta}^{(H,N)}$ from $q_{\zeta(H)}$. This may require first drawing of model parameters $\theta^{(H,1)}, \ldots, \theta^{(H,N)}$ and then conditional simulation of the corresponding signal paths.

4.5. Evaluate the corresponding importance weights $w^{(H,1)}, \ldots, w^{(H,N)}$.

4.6. Calculate $CoV^{(H)}$, the coefficient of variation (CoV) of the weights $w^{(H,1)}, \ldots, w^{(H,N)}$.

4.1.3 Details of the Algorithm

The proposal density, and hence the weight formula (2.7), essentially depend on the type of a model under consideration. For *observation driven* models, the likelihood function is available in closed-form and the time varying parameters are perfectly predictable one-step-ahead (conditional on the current information set). Hence, there are no latent variables, which means that the only parameters from the Bayesian perspective are the model parameters θ. Then, the posterior density is simply $p(\theta|y)$ and it can be approximated be $q_\zeta(\theta|y)$, a mixture of Student's t distributions, yielding the importance weights

$$w(\theta^{(i)}) = \frac{p(\theta^{(i)}|y)}{q_\zeta(\theta^{(i)}|y)}, \tag{4.4}$$

with $\theta^{(i)} \overset{i.i.d.}{\sim} q_{\zeta_{old}}(\theta|y)$, where $q_{\zeta_{old}}$ is an old candidate density (we refer to Appendix B.1 for the details of the derivation).

In *parameter driven* models the time-varying parameters are subject to an idiosyncratic noise, leading to the likelihood function being unavailable in closed form, hence neither is the kernel of the posterior density $p(\theta|y)$. This class of models includes the nonlinear non-Gaussian state space models, which we are particularly interested in. As already noticed, Bayesian treatment of these models requires to consider the whole latent process $x = \{x_t\}_{t=1}^{n}$ as an additional parameter, implying that it enters the posterior density $p(\theta, x|y)$. Following Barra et al. (2014), we use as the importance density $q_\zeta(\theta, x|y)$, which we decompose as

$$q_\zeta(\theta, x|y) = q(x|\theta, y) q_\zeta(\theta|y) \tag{4.5}$$

The first term on the right-hand side in (4.5) targets the smoothed state density and the draws from it are obtained via a simulation smoother. For this purpose, we employ the NAIS method of Koopman et al. (2015), which is discussed in Section 4.2. The second term on the right-hand side in (4.5) is approximated by a mixture of Student's t distributions, with the parameters ζ derived with MitISEM. However, since the kernel of the (marginal) posterior $p(\theta|y)$ is unknown, we cannot use the standard weight formula (4.4) to obtain the necessary inputs to the algorithm. As shown in Barra et al. (2014), the required modified weights are given by

$$\begin{aligned} w(\theta^{(i)}, x^{(i)}) &= \frac{p(\theta^{(i)}, x^{(j)}|y)}{q_{\varsigma}(\theta^{(i)}, x^{(j)}|y)} \\ &\propto \frac{p(y|\theta^{(i)}, x^{(i)})p(x^{(i)}|\theta^{(i)})p(\theta^{(i)})}{q(x^{(i)}|\theta^{(i)}, y)q_{\varsigma}(\theta^{(i)}|y)} \\ &= q(y|\theta^{(i)})\frac{p(y|\theta^{(i)}, x^{(i)})}{q(y|\theta^{(i)}, x^{(i)})}\frac{p(\theta^{(i)})}{q_{\varsigma}(\theta^{(i)}|y)}, \end{aligned} \tag{4.6}$$

as $p(x|\theta) = q(x|\theta)$ due to our restriction to the class of state space models with linear Gaussian state equation. In (4.6), $p(y|\theta, x)$ is the conditional observation density implied by the model; $q(y|\theta, x)$ is the importance observation density, given the augmented parameter vector $\tilde{\theta} = (\theta, x)$; $p(\theta)$ is the prior on the model parameter vector θ; $q_{\varsigma}(\theta|y)$ is the marginal proposal density for the model parameters based on the previous mixture; $q(y|\theta)$ is the likelihood of the auxiliary model marginalised over x. The latter can be efficiently obtained via the Kalman filter when the importance density for the states is based on an auxiliary linear Gaussian model, as it will be the case in the remaining part of the thesis.

Regarding the initial (naive) candidate used in Step 1.1. of the MitISEM algorithm, in case of an observation driven model one can simply take a Student's t density with the mode equal to the mode of the target density, and with the scale set to the inverse Hessian of the logkernel density of the target, computed at the mode. For a parameter driven model, one can proceed in a similar way, however the mode is obtained by the simulated maximum likelihood method (cf. Subsection 4.2.3), while the scale is equal to minus the inverse Hessian of the simulated loglikelihood, computed at the mode parameter estimates.

The final remark relates to the chosen convergence criterion, i.e. the relative change in the *coefficient of variation* (CoV) of the IS weights. The CoV of a given sample of IS weights $w^{(\cdot,1)}, \dots, w^{(\cdot,N)}$, defined by formula (4.3), is its standard deviation divided by its mean. We use the CoV to evaluate the distribution of the IS weighs, in which we follow Hoogerheide et al. (2012). The main reason for this choice is its intuitiveness: loosely speaking the lower CoV, the better the approximation to target the candidate provides. To see this, notice that if the candidate and the target coincide, then the CoV becomes 0; if the candidate deviates a lot from the target, thus poorly approximating the latter, the distribution of weights is uneven, with some being extremely high and some being close to 0, which results in large values of CoV. For a more in-depth discussion of the desirable properties of the CoV we refer to Ardia et al. (2009).

4.2 Estimation of Parameter Driven Models

In this Section we consider the problem of likelihood evaluation for nonlinear non-Gaussian state space models. We focus on the IS-based methods originating from Shephard and Pitt (1997) and Durbin and Koopman (1997), which we will refer to as SPDK. In these papers the importance density is constructed based on an auxiliary linear Gaussian model, yielding a local approximation to the original model. This approach was further developed in Richard and Zhang (2007), who proposed the Efficient Importance Sampling (EIS). Their technique is also based on a linear Gaussian auxiliary model, yet provides a global approximation to the true model. Recently, Koopman et al. (2015) constructed a novel IS approach, incorporating the advantages of the numerical integration into the EIS framework, called the Numerically Accelerated Importance Sampling (NAIS). As this method is numerically and computationally efficient, we adopt it in the remaining part of the thesis as the state sampler.

4.2.1 IS Likelihood Evaluation

The likelihood function corresponding to (4.1) and (4.2) is given by

$$L(y;\theta) = \int p(\alpha, y;\theta) d\alpha = \int \prod_{t=1}^{T} p(y_t|x_t;\theta) p(\alpha_t|\alpha_{t-1};\theta) d\alpha_1 \dots d\alpha_T, \tag{4.7}$$

where $p(\alpha, y;\theta)$ is the joint density of y and α following from (4.1) and (4.2). In all the cases but when $p(y_t|x_t;\theta)$ is a Gaussian density with mean $x_t = Z_t\alpha_t$ and variance H_t, the integral (4.7) is analytically intractable. In the mentioned special case it can be evaluated using the Kalman filtering and smoothing (KFS) methods, but in general it needs to be computed using numerical techniques. A common approach to the likelihood evaluation in the context of state space models is importance sampling. Although it is a special case of the general framework discussed in Subsection 2.2.1, we provide details of the derivations for future reference.

The importance density is taken to be Gaussian and we let it factorise as follows

$$q(\alpha, y;\theta) = q(y|\alpha;\theta)q(\alpha;\theta), \tag{4.8}$$

where the two densities on the right hand side are assumed to be Gaussian as well. From the specification of the transition equation in (4.2) it follows that $q(\alpha;\theta) \equiv p(\alpha;\theta)$, which, together with (4.8), allows us to rewrite (4.7) as

$$\begin{aligned} L(y;\theta) &= \int \frac{p(\alpha, y;\theta)}{q(\alpha, y;\theta)} q(\alpha, y;\theta) d\alpha \\ &= g(y;\theta) \int \frac{p(y|x;\theta)p(\alpha;\theta)}{q(y|x;\theta)q(\alpha;\theta)} q(\alpha|y;\theta) d\alpha \\ &= q(y;\theta) \int \omega(x, y;\theta) q(\alpha|y;\theta) d\alpha. \end{aligned} \tag{4.9}$$

In (4.9) $q(y;\theta)$ is the observation likelihood function as implied by the importance Gaussian model which does not depend on x and can be treated as a constant; $\omega(x, y;\theta)$ is the importance weight function defined as

$$\omega(x, y;\theta) = \frac{p(y|x;\theta)}{q(y|x;\theta)}. \tag{4.10}$$

The IS evaluation of (4.9) is performed by generating S independent signal trajectories $x^{(1)}, \dots, x^{(S)}$ from the Gaussian importance density $q(x|y;\theta)$. Under mild regularity conditions (cf. e.g. Geweke, 1989), the weak law of large numbers guarantees that the likelihood estimate

$$\hat{L}(y;\theta) = q(y;\theta) \frac{1}{S} \sum_{s=1}^{S} w(x^{(s)}, y;\theta), \tag{4.11}$$

where importance weight $w(x^{(s)}, y;\theta)$ is the importance function evaluated at the draw $x^{(s)}$, converges in probability to the true likelihood value $L(y;\theta)$ when $S \to \infty$.

4.2.2 Gaussian Importance Density

Up to now the only characterisation of the importance density was via (4.8), which is of limited practical use. However, as always the case of IS estimation, the exact specification of the candidate density is crucial for the quality of the estimate. The common approach is to base $q(\alpha, y;\theta)$ on an auxiliary linear

Gaussian model as discussed below, because this allows for applying KFS methods to compute various quantities of interest.

Notice that the Gaussian importance density (4.8) can be decomposed as

$$q(\alpha, y; \theta) = \prod_{t=1}^{T} q(y_t|x_t; \theta) q(\alpha_t|\alpha_{t-1}; \theta), \tag{4.12}$$

where $q(\alpha_t|\alpha_{t-1}; \theta)$ corresponds to the state transition equation from (4.2). Richard and Zhang (2007) suggested that the importance observation density $q(y_t|\alpha_t; \theta)$ can be expressed as

$$q(y_t|x_t; \theta) \equiv \exp\left\{a_t + b_t^T x_t - \frac{1}{2} x_t^T C_t x_t\right\}, \tag{4.13}$$

with the coefficients a_t, b_t and C_t, depending on the observations y and the model parameters ψ, need to be specified. First, the scalars a_t, $t = 1, \dots, T$, are chosen to guarantee that (4.12) is a proper density function, i.e. that it integrates to one, and hence are given by

$$a_t = \frac{1}{2} \log |C_t| - \frac{1}{2} \log 2\pi - \frac{1}{2} b_t^T C_t^{-1} b_t.$$

Second, the vectors b_t and the matrices C_t, $t = 1, \dots, T$, which characterise the unique set of IS parameters

$$\chi = \{b_1, \dots, b_T, C_1, \dots, C_T\},$$

need to be determined. SPDK considered representing (4.13) as the smoothed density of the linear Gaussian state space model for the artificial observation $y_t^* \equiv C_t^{-1} b_t$, with the observation equation

$$y_t^* = x_t + \varepsilon_t, \qquad x_t = Z_t \alpha_t, \qquad \varepsilon_t \sim N(0, C_t^{-1}), \qquad t = 1, \dots, T$$

and the transition equation equivalent to (4.2). The logdensity of the artificial Gaussian model is equivalent to the log of the importance density (4.13), since the former has the form

$$\begin{aligned} \log q(y_t^*|\alpha_t; \theta) &= -\frac{1}{2} \left(\log 2\pi + \log |C_t|^{-1} + (y_t^* - \theta_t)^T C_t^{-1} (y_t^* - \theta_t)\right) \\ &= -\frac{1}{2} \left(\log 2\pi + \log |C_t|^{-1} + (C_t^{-1} b_t - \theta_t)^T C_t^{-1} (C_t^{-1} b_t - \theta_t)\right), \\ &= \underbrace{-\frac{1}{2} \left(\log 2\pi - \log |C_t| + b_t^T C_t^{-1} b_t\right)}_{a_t} + b_t x_t - \frac{1}{2} x_t^T C_t x_t, \end{aligned}$$

which indeed is the log of (4.13). This in turn implies that

$$\begin{aligned} q(x, y|\theta) &\equiv q(x, y^*; \theta), \\ q(\alpha, y|\psi) &\equiv q(\alpha, y^*; \theta), \end{aligned}$$

where $y^* = (y_1^{*T}, \dots, y_T^{*T})$. The equivalence between the artificial model and the importance density allows us to employ efficient KFS techniques together with the related simulation smoothing method to generate the draws from $q(x|y; \theta)$, as required to compute the IS estimate $\hat{L}(y; \theta)$. Regarding the IS parameters χ, in the SPDK approach they are set to ensure that the mode (mean) estimate of x with respect to the artificial $q(x|y^*; \theta)$ equals the mode estimate of x with respect to the true $p(x|y; \theta)$. For that reason, this method provides barely the *local approximation* of the integral in question.

The alternative approach of Richard and Zhang (2007), which will be referred to as EIS (abbreviation of efficient importance sampling), has an advantage of delivering the *global approximation* to $p(y|x; \theta)$.

Their technique sets χ to minimize the variance of the log importance weights $\log \omega(x, y; \theta) =: \lambda(x, y; \theta)$. Thus, the task boils down to solving of the following variance minimization problem

$$\min_{\chi} \int \lambda^2(x, y; \theta)\omega(x, y; \theta)q(x|y; \theta)d\theta. \tag{4.14}$$

Due to multidimensionality and hence infeasibility of (4.14), is needs to be approximated, which in the EIS method is done by its reduction to a series of minimization problems, at each time point separately, as follows

$$\min_{\chi_t} \int \lambda^2(x_t, y_t; \theta)\omega(x_t, y_t; \theta)q(x_t|y; \theta)dx_t, \qquad t = 1, \dots, T, \tag{4.15}$$

with $\chi_t = \{b_t, C_t\}$. Although much simpler than (4.14), the integral in (4.15) still needs to be approximated, which in the EIS approach is again done via IS. The importance draws are generated using the simulation smoothing from $q(\alpha|y; \theta)$ and are used to iteratively update the IS parameters via weighted least squares computations. Hence, this technique relies on simulations in two ways: to approximate the likelihood (4.7) and to obtain the optimal IS parameters for the importance density (4.13), which introduces substantial simulation noise to the analysis.

A superior method compared to EIS was developed by Koopman et al. (2015). Their Numerically Accelerated Importance Sampling (NAIS) technique replaces the second simulation step in EIS by numerical integration. The key insight is that the smoothing density $q(x_t|y; \theta) \equiv q(x_t|y^*; \theta)$ for the artificial model has a closed-form representation which allows us to apply numerical integration when minimizing (4.15). Indeed, for a given χ and for a scalar signal it holds

$$q(x_t|y^*; \theta) = N(\tilde{x}_t, V_t) = \frac{1}{\sqrt{2\pi V_t}} \exp\left\{ -\frac{1}{2} V_t^{-1}(x_t - \tilde{x}_t)^2 \right\}, \qquad t = 1, \dots, T, \tag{4.16}$$

where $\tilde{x}_t$ and V_t stand for the smoothed (conditional) mean and variance, respectively, and can be obtained via KFS techniques. The details of the numerical integration are given in Appendix C.

4.2.3 Simulated Maximum Likelihood

Once we are able to estimate the likelihood function (4.7) for the model (4.1) and (4.2) via (4.11), the numerical maximisation of this simulated likelihood can be employed to deliver the Simulated Maximum Likelihood (SML) estimate of θ. Typically, maximisation is performed using a quasi-Newton method, e.g. the BFGS algorithms and is applied to the loglikelihood for numerical stability. Starting values for the optimisation can be determined through running the whole NAIS procedure with $S = 0$, i.e. via performing only the numerical integration. These values are the SML estimates of the auxiliary model as they maximise its likelihood $q(y; \theta)$. Then, the full NAIS algorithm is carried out, with number of simulations set e.g. to $S = 200$ (cf. Koopman et al., 2015). Finally, notice that to guarantee a smooth likelihood function in θ, necessary for convergence of the quasi-Newton algorithm, the common random number need to be used each time one applies the NAIS algorithm with $S > 0$.

5 Observation Driven Models

As pointed out in the Introduction, the two key elements of the original QERMit methods were the posterior approximation via AdMit and the class of models limited to the observation driven models. The starting point of our analysis is modification of the first of these ingredients while remaining within the original class of models. The main purpose of this Section is thus to compare the results obtained with QERMit based on different posterior approximation techniques. We aim to show that the chosen MitISEM algorithm leads to more accurate estimates. In addition to MitISEM and AdMit, we consider the quantitative results obtained with the direct approach, which despite its simplism is still used in the literature, constituting therefore an important benchmark. All the algorithms were implemented in MATLAB (the codes are available upon request) and the computations were performed using version 2014b.

5.1 ARCH(1)

As the first application, we estimate the 1-day-ahead 99% VaR and ES using the basic, simple ARCH(1) (Engle, 1982) model. The simplicity of this example suits it particularly well to illustrate the key idea behind the QERMit approach. The model under consideration is given by

$$\begin{aligned} y_t &= \sqrt{h_t}\varepsilon_t, \\ \varepsilon_t &\overset{i.i.d.}{\sim} \mathcal{N}(0,1), \\ h_t &= \omega + \alpha y_{t-1}^2, \end{aligned}$$

where for numerical stability we impose the variance targeting constraint, i.e. we assume that

$$\omega = S^2(1-\alpha),$$

with S^2 being the sample variance of y_t. This results in a reduction of model parameters to one, namely α. As in the original QERMit paper (Hoogerheide and van Dijk, 2010), we use the data on daily logreturns of the S&P 500, from January 2, 1998 to April 14, 2000, which constitutes a sample of $T = 576$ observations, with $y_T = -6.0045$ and $S^2 = 1.6256$, cf. Figure 5.1. Finally, we set a flat prior on $[0, 1)$.

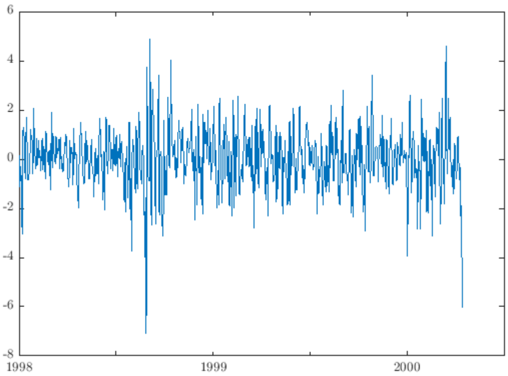


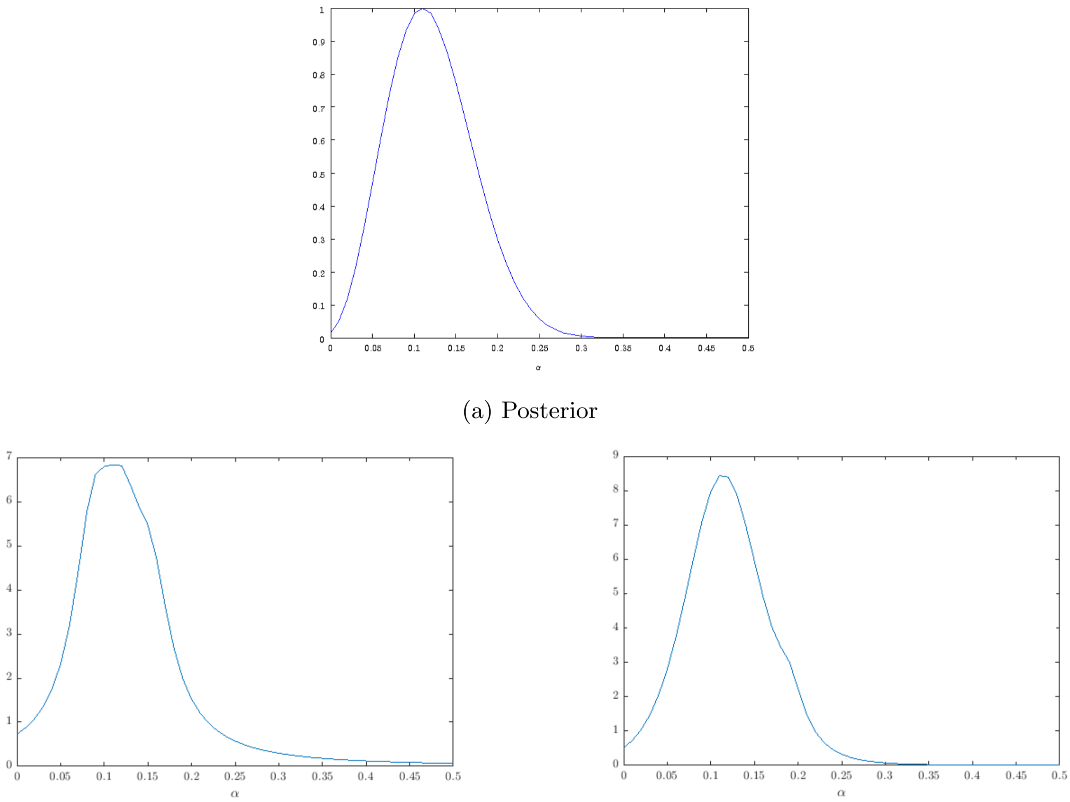
Figure 5.1: S&P 500 log-returns, from January 2, 1998 to April 14, 2000.

Within this simple setting, the QERMit algorithm proceeds as follows. First, we approximate the posterior function for α

$$p(\alpha|y) = \left(\frac{1}{\sqrt{2\pi}}\right)^T \prod_{t=1}^{T} \frac{1}{\sqrt{\omega+\alpha y_{t-1}^2}} \exp\left(-\frac{y_t^2}{2(\omega+\alpha y_{t-1}^2)}\right),$$

with $q_{1,Mit}(\alpha)$, a mixture of Student's t distributions using both, AdMit and MitISEM. Figure 5.2a presents the posterior density $p(\alpha)$, while Figures 5.2b and 5.2c illustrate the approximations to $p(\alpha)$ obtained using AdMit and MitISEM, respectively. Even though the two latter plots appear to be rather similar, it can be inferred from Table 5.1 that they differ quite substantially. Recall from Subsection

4.1.3 that the quality of a candidate can be measured by the CoV it delivers (roughly speaking, the lower the CoV, the better the approximation). Moreover, if the approximation is based on an iteratively augmented mixture, like in our case, the computational efficiency calls for using as little components as possible to obtain a required level of the approximation accuracy.



(a) Posterior

(b) AdMit

(c) MitISEM

Figure 5.2: The posterior parameter density $p(\alpha|y)$ and the approximations to it.

The candidate obtained with AdMit is a mixture of 5 components, which yields the coefficient of variation equal to 0.28. On the other hand, MitISEM needs only 4 components and delivers a much lower CoV of 0.15. These results confirm that MitISEM targets the posterior parameter density more efficiently and more accurately than AdMit does.

	Admit	MitISEM
‘Whole’ space component q_1		
No. of components	5	4
CoV	0.2791	0.1462
High-loss component q_2		
No. of components	6	4
CoV	0.8227	0.4052

Table 5.1: Quality of the mixture-based approximations to the components of the optimal importance density as reflected by the final CoV and the required number of components.

In the second step, we generate $N = 10,000$ draws $\alpha^{(i)}$, $i = 1, \ldots, N$, from the posterior $p(\alpha|y)$ using the independence chain Metropolis-Hastings algorithm with $q_{1,Mit}(\alpha)$ taken as the proposal distribution.

The corresponding 1-day-ahead forecast of the logreturns are obtained as $y^{*(i)} \sim \mathcal{N}\left(0, \omega + \alpha^{(i)} y_T^2\right)$, i.e. $y^{(i)} \sim \mathcal{N}\left(0, 1.6256(1-\alpha^{(i)}) + 36.0542\alpha^{(i)}\right) = \mathcal{N}\left(0, 1.6256 + 34.4286\alpha^{(i)}\right)$. To derive the preliminary VaR estimate, we sort the profit-loss values $PL(y^{*(i)})$ (cf. formula (2.1)) in ascending order and take the $(1-\alpha)N = 100$-th one. As it can be inferred from Table 5.2, both algorithm delivered comparable initial VaR estimates: $\widehat{VaR}_{prelim}$ obtained with AdMit was equal to -5.62, while with MitISEM -5.64. However, the acceptance rate obtained using the approximation based on MitISEM is higher that the one recorded with AdMit (equal to 0.93 and 0.87, respectively). This confirms the results reported in Table 5.1 that the former algorithm provides a better approximation to the posterior density.

The last element of the first step of QERMit concerns approximation of the high-loss density. As already indicated in Section 3.2, we consider $p(\alpha, \varepsilon_{T+1})$ instead of $p(\alpha, y_{T+1})$. Hence, we need to characterise those $\varepsilon_{T+1}^{(i)}$, which lead to the profit/loss $PL(y^{(i)})$ values not exceeding the preliminary VaR estimate $\widehat{VaR}_{prelim}$, given the draws of $\alpha^{(i)}$. For a given future return $y_{T+1}^{(i)}$, the profit-loss value falls into the high-loss region if it satisfies

$$y_{T+1}^{(i)} = PL(\widehat{VaR}_{prelim}) = 100 \log\left(\widehat{VaR}_{prelim}/100 + 1\right).$$

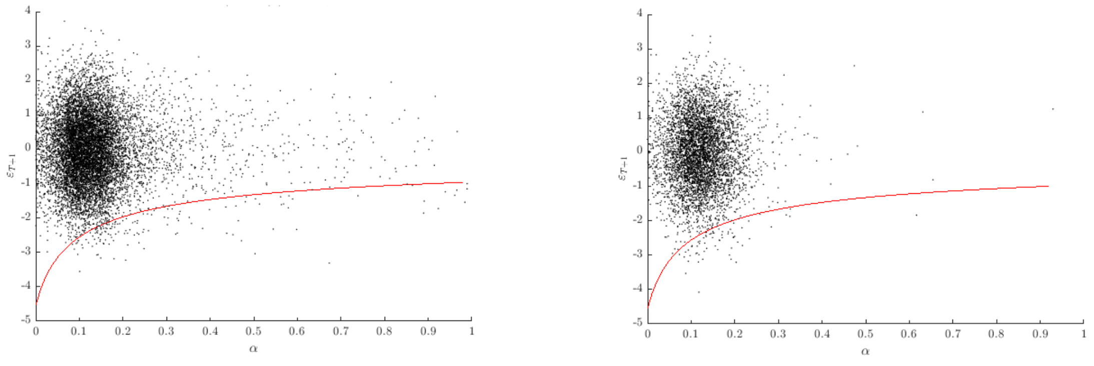
Since

$$\begin{aligned} y_{T+1}^{(i)} &= \sqrt{h_{T+1}^{(i)}} \varepsilon_{T+1}^{(i)}, \\ h_{T+1}^{(i)} &= \omega + \alpha^{(i)} y_T \\ &= 1.6256 + 34.4286\alpha^{(i)}, \end{aligned} \tag{5.1}$$

we arrive at the following condition describing the border of the high-loss subspace

$$\varepsilon_{T+1}^{(i)} = \frac{100 \log\left(\widehat{VaR}_{prelim}/100 + 1\right)}{\sqrt{1.6256 + 34.4286\alpha^{(i)}}}.$$

Figure 5.3 present the draws from the approximation to joint distribution $p(\alpha, \varepsilon_{T+1})$, together with the border of the high-loss region, obtained with both algorithms. One can see that indeed roughly 100 draws fall "below" the red line, as one would expect for the 99% VaR. It can also be noticed that the approximation generated with AdMit yields draws more spread along the α dimension. This is due to the fixed, fat-tails of Student's t components in the AdMit approach, where the number of degrees of freedom is set to 1 and never updated. MitISEM, on contrary, optimises the components' number of degrees of freedom, which allows for more flexibility in fitting the targets of interest. This superiority of MitISEM over AdMit is reflected in the properties of the approximation to the high-loss density, as reported in Table 5.1. Again, the former approach is able to deliver a lower CoV of 0.40 with using only 4 components, while AdMit needs 6 components to converge, and yielded a higher CoV of 0.82.

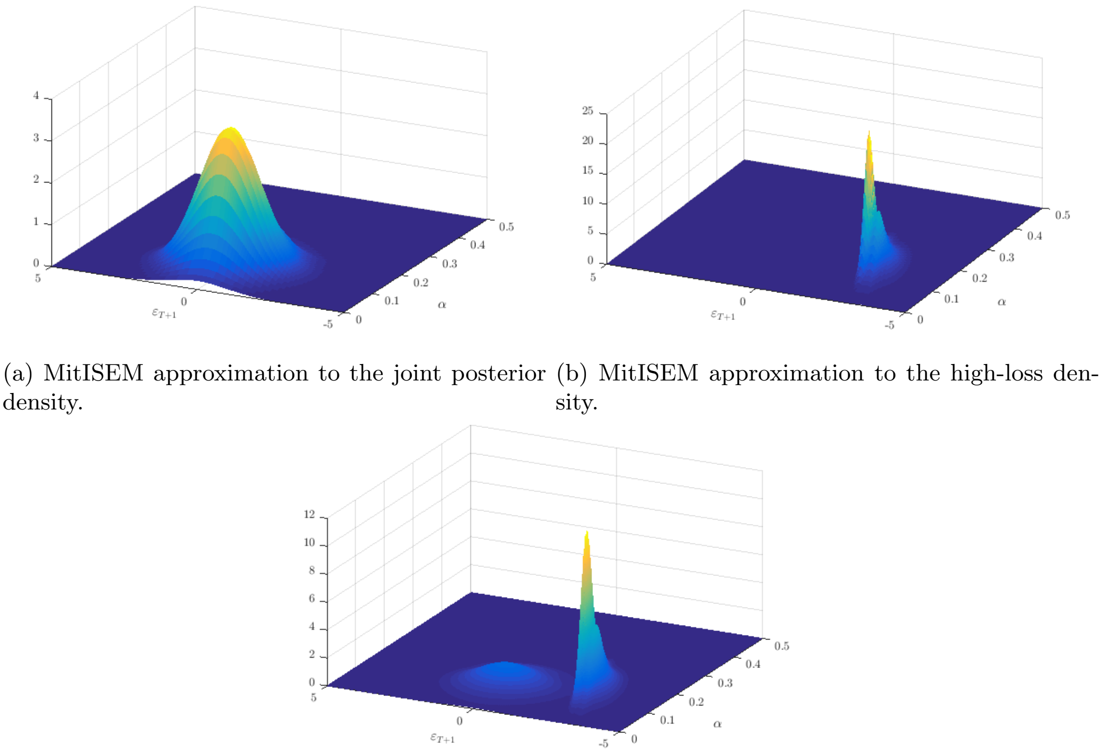


(a) AdMit

(b) MitISEM

Figure 5.3: Draws from the approximation to the joint density $p(\alpha, \varepsilon_{T+1})$ and the high-loss region.

Figure 5.4 presents the construction of the approximation to the optimal importance density, as discussed in Section 3.2, using the MitISEM algorithm[12]. Figure 5.4a displays the approximation to the joint density, i.e. $q_{1,Mit}(\alpha)p(\varepsilon_{T+1}|\alpha)$, while Figure 5.4b – to the high-loss density $q_{2,Mit}(\alpha,\varepsilon_{T+1})$. Then, the approximation to the optimal candidate density, obtained as a 50–50 mixture of these two approximations, is depicted in Figure 5.4c.



(a) MitISEM approximation to the joint posterior density.

(b) MitISEM approximation to the high-loss density.

(c) MitISEM approximation to the optimal importance density.

Figure 5.4: Construction of the approximation to the optimal candidate density.

Having constructed the optimal candidate densities with both QERMit method, we compare the 99% VaR and ES estimates computed using AdMit and MitISEM, as well as with the direct approach. The direct approach is based on a Student's t candidate for the draws of parameter α and direct sampling for future errors ε_{T+1}, given the parameter draws. Regarding the parameters of the direct candidate, its mode is set to the mode of the target density equal to 0.1099, the scale – to the inverse Hessian of the logkernel density of the target computed at the mode equal 0.0029 and we choose 1 degrees of freedom.

Table 5.2 shows that the estimates of both risk measures delivered by all three approaches are in line with each other . The 1-day-ahead 99% VaR estimate delivered by MitISEM is equal to -5.68 and by AdMit to -5.66, which is almost identical to the direct estimate. However, both IS-based methods with the optimal importance candidate density outperform the direct approach in terms of delivering the NSE one order of magnitude lower. The NSE occurring in the VaR estimation with the direct approach is equal to 0.0798, while when the IS methods are adopted it becomes 0.0073 and 0.0048 for AdMit and MitISEM, respectively. Also in the case of the ES, the estimates from all three methods are comparable and equal -6.52, -6.56 and -5.68 for the direct, AdMit and MitISEM algorithms, respectively. The NSE of MitISEM again is lower than the one from AdMit (0.0126 compared to 0.0145), and both are roughly 10 times lower than the NSE occurring in the direct estimation. We conclude that MitISEM is

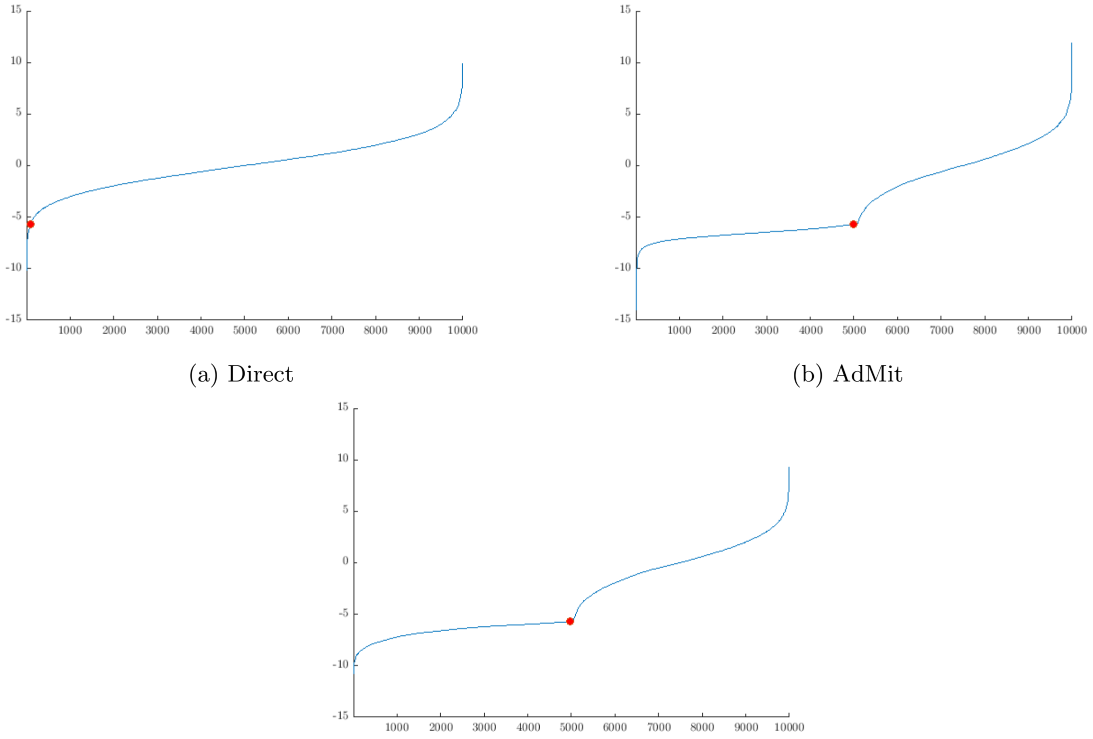
[12] Visually, the three dimensional plots generated with AdMit and with MitISEM look very much alike, therefore we have abandoned the initial idea of presenting both versions. Nevertheless, the quantitative results obtained with both algorithms differ considerably.

the most precise tool for risk evaluation, outperforming not only the direct approach but also the AdMit method.

	Direct	Admit	MitISEM
VaR prelim	–	-5.6260	-5.6405
Acceptance rate	0.8069	0.8722	0.9300
VaR estimate	-5.6558	-5.6622	-5.6843
VaR NSE	0.0798	0.0073	0.0048
ES estimate	-6.5198	-6.5572	-6.5761
ES NSE	0.1227	0.0145	0.0126

Table 5.2: Estimates of 1-day-ahead 99% VaR and ES for S&P 500 in the ARCH(1) model.

Figure 5.5 illustrates the grounds for the superiority of the IS based methods, in particular of the MitISEM approach. The horizontal axis shows the indices i of draws ($i = 1, \ldots, 10000$), while the vertical axis shows the i-th sorted profit/loss value. One can see that the focus on the high-loss density allows to obtain a more precise insight into the shape of the lower tail of the profit/loss distribution as opposed to the direct approach which performs inference based on barely few samples. Moreover, with MitISEM this shape is characterised by a higher curvature as compared to the one obtained with AdMit, as expected for the fat-tailed financial returns series.



(a) Direct

(b) AdMit

(c) MitISEM

Figure 5.5: Sorted future profit/losses values $PL(y_{T+1}^{(i)})$ for the ARCH(1,1) model and the corresponding 99% VaR estimate (red dot).

5.2 GARCH(1,1)-t

As the second application of our modified QERMit approach, we consider the estimation of 1-day-ahead[13] 99% VaR and ES for the S&P500 daily logreturns from January 2, 1998 to December 31, 2007 using the GARCH(1,1) model with Student's t error. The data is presented in Figure 5.6. The model specification is similar to the one used in Hoogerheide and van Dijk (2010) and is given by

$$\begin{aligned} y_t &= \sqrt{\rho h_t} \varepsilon_t, \\ \varepsilon_t &\sim t(\nu), \\ \rho &:= \frac{\nu - 2}{\nu}, \\ h_t &= \omega + \alpha y_t^2 + \beta h_{t-1}, \end{aligned}$$

where $t(d)$ denotes the Student's-t distribution with d degrees of freedom. For numerical stability, we used variance targeting and set

$$\omega = S^2(1 - \alpha - \beta),$$

where S^2 is the sample variance, in this case equal to 1.29. We set flat priors on ω, α and β

$$\omega > 0, \qquad \alpha \in (0,1), \qquad \beta \in (0,1)$$

and we impose the restriction that $\alpha + \beta < 1$. For the number of degrees of freedom, we specify the proper uninformative exponential prior for $\nu - 2$ (the restriction $\nu > 2$ ensures that the variance is finite). We collect the model parameters in the vector $\theta = (\alpha, \beta, \mu, \nu)^T$.

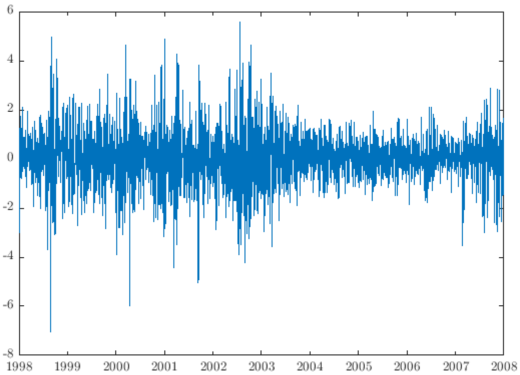


Figure 5.6: S&P 500 log-returns, from January 2, 1998 to December 31, 2007.

Differently than in the case of the ARCH(1,1) model, to compute the implied 1-day-ahead volatility h_{T+1} in the GARCH(1,1)-t model, we need to obtain, given the draws $\theta^{(i)}$, $i = 1, \ldots, N$, from the posterior $p(\theta|y)$, the paths of the volatility $\{h_1^{(i)}, \ldots, h_T^{(i)}\}$ corresponding to the draw $\theta^{(i)}$. Obviously, also the prediction of y_{T+1} becomes more complex and ceases to have a simple formula (5.1). Now, it is recursively computed, as implied by the model, given y_T, $h_{T+1}^{(i)}$ and $\theta^{(i)}$, $i = 1, \ldots, N$. Despite these complications and the fact that now we need to find an approximation to 4 and 5 dimensional densities (corresponding to the posterior of the parameters and the joint posterior of parameters and disturbances), the core idea of QERMit remains unchanged so we refrain from its detailed description similar to the one for the ARCH(1) model.

[13] Ultimately, we would be interested in the estimation of 10-day-ahead risk measures, as required by the Basel standards and as done in Hoogerheide and van Dijk (2010). However, for comparative purposes and due to the computational limitations (the high-loss density in that case would be 14-dimensional), we decided to focus our attention on 1-day-ahead measures.

The properties of the mixtures used in two QERMit methods are presented in Table 5.3. Regarding the joint component corresponding to the whole profit/loss space, both AdMit and MitISEM use three multivariate Student's t components. However, the mixture provided by MitISEM is characterised by a more than twice lower CoV (equal to 0.36) than the one obtained with AdMit (where the CoV is equal 0.80). For the high-loss component we set the maximum number of components to 2 in order to speed up the computation. Hence, it is likely that neither algorithm achieved convergence, yet the CoV generated by MitISEM is already substantially lower than the one from AdMit (0.66 and 0.78 respectively).

	AdMit	MitISEM
'Whole space' component q_1		
No. of components	3	3
CoV	0.8061	0.3626
High loss component q_2 *		
No. of components	2	2
CoV	0.7854	0.6592

*The maximum number of components for the high loss density approximation was upfront limited to 2 to speed up the computations.

Table 5.3: Quality of the mixture-based approximations to the components of the optimal importance density as reflected by the final CoV and the required number of components.

Table 5.4 compares the results delivered under three considered method, the two QERMit approaches discussed above and the direct approach. The latter is based on the one Student's t component to sample parameters followed by the direct sampling of the future observation disturbances given the draws of the parameter vector. The Student's t component has one degree of freedom, the mode equal to the mode of the target density, and the scale set to the inverse Hessian of the logkernel density of the target computed at the mode. The preliminary VaR estimate obtained with MitISEM is equal to -2.75, while the one generated using AdMit amounts to -2.69. The result delivered by MitISEM seems to be more reliable, as it originates from a mixture which more closely approximates the target of interest: the acceptance rate of the independence MH performed with the MitISEM candidate is almost 80%, while roughly speaking the AdMit candidate allows for accepting only every second draw (acceptance rate of 53%).

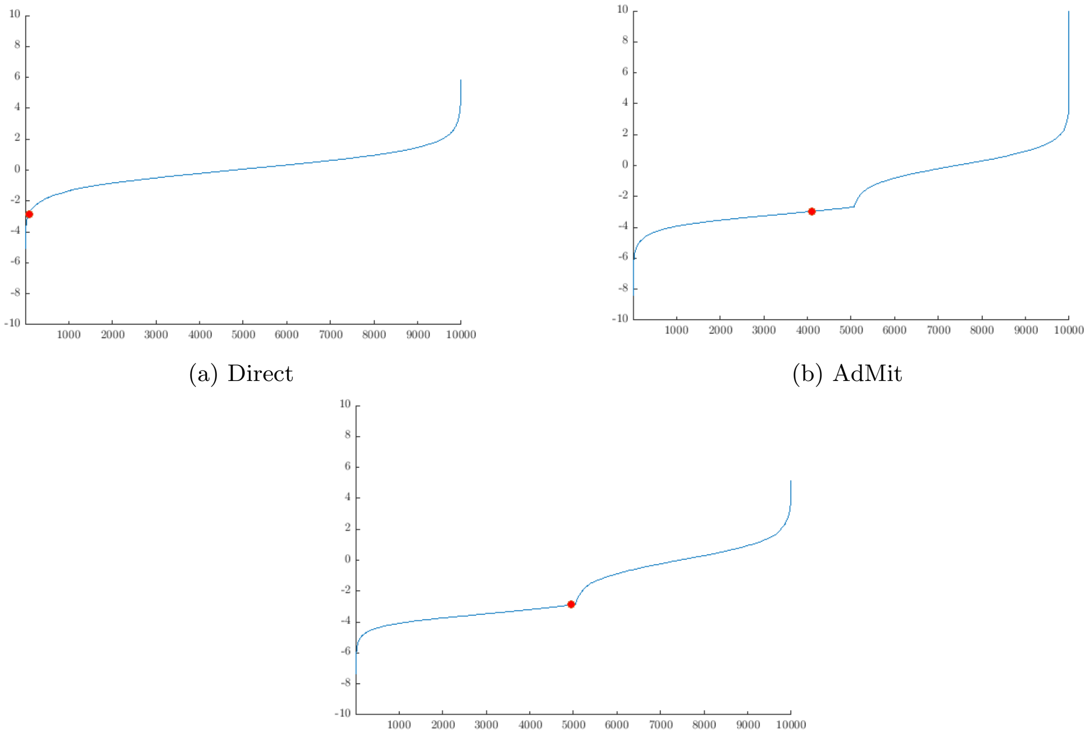
Differently than in the ARCH(1) example, the VaR estimates in the current case noticeably differ among themselves. The direct approach gives the value of -2.79, while the AdMit yields a much lower number of almost -3.00. The estimate from MitISEM lies in-between the two former results and is equal to -2.88. These discrepancies are even more suspicious when one considers the corresponding NSE values, which for the IS-based methods are again much lower than for the direct approach. However, the ranges determined by NSE around the VaR estimates obtained for AdMit and for MitISEM are disjoint, each being very precise. A potential explanation for this outcome is a considerable difference in the preliminary VaR estimates from both methods. Even though the one generated using AdMit is higher than the one obtained with MitISEM, the fact that AdMit does update neither the number of degrees of freedom, nor the remaining parameters of the previous components, may pose difficulties for this algorithm to correctly approximate the target, especially in higher-dimensional spaces. The differences between the ES estimates are also noticeable, yet less worrying than these in the VaR estimates. This is because the numerical error ranges determined by the NSE around the estimates are overlapping, so the conclusions based on all there methods are rather similar. As in the case of ARCH(1) model, the MitISEM based QERMit outperforms the competing approaches. The NSE generated when MitISEM is employed is equal to 0.0284, which is the lowest value among these obtained with the considered methods (with 0.0145 and

0.1227 for AdMit and the direct approach, respectively).

	Direct	AdMit	MitISEM
VaR prelim	–	-2.6873	-2.7461
Acceptance rate	0.5188	0.5268	0.7989
VaR estimate	-2.7947	-2.9626	-2.8766
VaR NSE	0.0635	0.0106	0.0075
ES estimate	-3.4388	-3.6353	-3.5345
ES NSE	0.1159	0.0426	0.0284

Table 5.4: Estimates of 1-day-ahead 99% VaR and ES for S&P 500 in the GARCH(1,1)-t model.

As for the ARCH(1) model, we conclude with an illustration of the key conceptual difference between the direct approach and the IS-based methods, in particular the one using the MitISEM algorithm. In Figure 5.7 we consider the sorted series of the obtained profit/loss values. The horizontal axis shows the indices i of draws ($i = 1, \dots, 10000$), while the vertical axis shows the i-th sorted profit/loss value. Clearly, no precise inference on the nature of the high-loss region can be carried out using the direct approach, which bases the estimates on a handful of draws. In contrast, the focus on the high-loss subspace of both QERMit methods is likely to result in an accurate insight into this subspace. Nevertheless, the approximation to the posterior profit/loss density delivered by AdMit is rather strange-looking, with very wide right tail. The MitISEM-based curve exhibits a much more desirable behaviour.



(a) Direct

(b) AdMit

(c) MitISEM

Figure 5.7: Sorted future profit/losses values $PL(y_{T+1}^{(i)})$ for the GARCH(1,1)-t model and the corresponding 99% VaR estimate (red dot).

6 Parameter Driven Models

The contribution of the previous section was upgrading the original QERMit method through altering the posterior approximation algorithm which it is based on. In the current section we aim to make QERMit applicable to a broader class of models, namely to allow for the risk evaluation in the parameter driven models. Due to the demonstrated superiority of the MitISEM-based QERMit over the two other methods, in this section we adopt MitISEM as our tool to approximate the model parameters posterior density and the high loss density.

In Section 3 we discussed the general QERMit method treating the parameter vector very broadly. In principle, it can also include the unobserved states of a state space model, as it is the case in the Bayesian analysis. Hence, *conceptually*, the QERMit algorithm applied to the parameter driven models remains unchanged compared to the case of the observation driven models. However, the necessity of dealing with multidimensional conditional smoothed state densities in the former class of models poses considerable *practical* difficulties, so that the whole QERMit procedure becomes much more complex. For this reason, below we present step-by-step the QERMit extension to the models including latent variables.

Furthermore, we discuss the technical problems which we encountered in the implementation of the algorithm. These issues prevented us from generating point estimates of VaR and ES, i.e. the counterparts of Tables 5.1–5.4 obtained for the observation driven models. We leave the task of obtaining such results for the parameter driven models for further research. Finally, we illustrate the steps of a simplified version of the general QERMit algorithm using two standard models: the basic Stochastic Volatility (SV) model and the Stochastic Volatility model with the Student's t observation errors (SVt).

6.1 Extended QERMit algorithm

The starting point for the procedure is constructing of $q_{1,\zeta_1}(\theta, x|y)$, i.e. the approximation to the posterior density of the augmented parameter vector (θ, x). To this end, we can use the decomposition (4.5) given by $q_{1,\zeta}(\theta, x|y) = q_1(x|\theta, y)q_{1,\zeta_1}(\theta|y)$. As in the Extended MitISEM algorithm of Barra et al. (2014), discussed in Subsection 4.1.2, we adopt a mixture of Student's t densities, parametrised by vector ζ_1, to approximate the posterior distribution of states and the conditional Gaussian density to target the smoothed state density.

Since the profit/loss function depends on the future (observed) realisations of the logreturns, and these are determined by the future realisation of the unobserved logvolatility process, to obtain the preliminary h-days-ahead $100(1 - \alpha\%)$ VaR estimate, in addition to the model parameters, one needs to analyse the disturbances both of the state and the observation processes, denoted by $\eta^* = \{\eta_{T+1} \ldots, \eta_{T+h}\}$ and $\varepsilon^* = \{\varepsilon_{T+1} \ldots, \varepsilon_{T+h}\}$, respectively. Notice that this is one of the main differences as compared to the observation driven models, where only the observation disturbances need to be modelled. Given the model parameter draws $\theta^{(i)}$, $i = 1, \ldots, N$, the corresponding NAIS importance sampling parameters and the conditional simulated signal paths, as well as the draws of both disturbances, one can calculate the profit/loss values $PL(y^{*(i)})$ and derive the desired preliminary estimate of the $100(1 - \alpha)\%$ VaR.

The next step consists in approximating the high loss region, where the losses exceed the preliminary VaR estimate. In the parameter driven models this becomes a more challenging task than in the observation driven models, again due to the latent nature of the logvolatility process. To determine whether a given draw of the model parameter vector $\theta^{(i)}$ can lead to extreme losses one needs to compute the NAIS importance sampling parameters $\chi^{(i)}$ and employ the simulation smoother to sample one corresponding signal path $x^{(i)}$. Similarly as in the standard QERMit version, one is interested in the joint Bayesian estimation of (augmented) parameter vector and future returns. Therefore, now the approximation to the joint posterior high loss density takes the form $q_{2,\zeta_2}(\theta, x, \eta^*, \varepsilon^*|y)$, which factorises as

$q_2(x|\theta, \eta^*, \varepsilon^*|y)q_{2,\zeta_2}(\theta, \eta^*, \varepsilon^*|y)$. This means that the mixture of Student's t densities can serve as an approximation to the joint posterior density of model parameter vector and future disturbances, while the latent states can be smoothed conditionally on the parameter and disturbances draws, and the observations. As a result one obtains a sample $\{\theta, x, \eta^*, \varepsilon^*\}^{(i)}$, $i = 1, \ldots, N$, where each draw generates extreme losses.

Having both approximations, one applies the 50-50 formula (3.1) to evaluate both subsamples, from $q_1(x|\theta, y)q_{1,\zeta_1}(\theta|y)p(\eta^*)p(\varepsilon^*)$ and from $q_2(x|\theta, \eta^*, \varepsilon^*|y)q_{2,\zeta_2}(\theta, \eta^*, \varepsilon^*|y)$, on the optimal candidate density

$$q_{opt}(\theta, x, \eta^*, \varepsilon^*) = \frac{1}{2}q_1(x|\theta, y)q_{1,\zeta_1}(\theta|y)p(\eta^*)p(\varepsilon^*) + \frac{1}{2}q_2(x|\theta, \eta^*, \varepsilon^*, y)q_{2,\zeta_2}(\theta, \eta^*, \varepsilon^*|y).$$

Notice, that given the data $y = \{y_1, \ldots, y_T\}$ one can only determine the first T NAIS parameters $\chi^{(i)} = \{\chi_1^{(i)}, \ldots, \chi_T^{(i)}\}$ and hence smooth the signal up to time T. Thus, η^* and ε^* cannot help in predicting the smoothed signal and so the simulation smoothing is performed independently from the future disturbances. This allows us to reformulate the optimal candidate as follows

$$\begin{aligned} q_{opt}(\theta, x, \eta^*, \varepsilon^*) &= \frac{1}{2}q_1(x|\theta, y)q_{1,\zeta_1}(\theta|y)p(\eta^*)p(\varepsilon^*) + \frac{1}{2}q_2(x|\theta, y)q_{2,\zeta_2}(\theta, \eta^*, \varepsilon^*|y) \\ &= \frac{1}{2}q(x|\theta, y)\Big[q_{1,\zeta_1}(\theta|y)p(\eta^*)p(\varepsilon^*) + q_{2,\zeta_2}(\theta, \eta^*, \varepsilon^*|y)\Big], \end{aligned}$$

because we take the same conditional Gaussian density $q(x|\theta, y)$ to target the signal process.

To obtain the IS estimates of the $100(1-\alpha)\%$ VaR and ES, in addition to both subsamples, one needs the corresponding importance weights. The kernel evaluation is obtained by computing the joint posterior density of a given draw implied by the model, i.e.

$$\begin{aligned} p(\theta, x, \eta^*, \varepsilon^*|y) &\propto p(y|\theta, x, \eta^*, \varepsilon^*)p(x|\theta, \eta^*, \varepsilon^*)p(\theta)p(\eta^*)p(\varepsilon^*) \\ &= p(y|\theta, x)p(x|\theta)p(\theta)p(\eta^*)p(\varepsilon^*), \end{aligned}$$

due to the independence of the future disturbances assumed in the model. Then, the standard formula (2.7) determines the importance weight function yielding

$$\begin{aligned} w(\theta, x, \eta^*, \varepsilon^*|y) &= \frac{p(\theta, x, \eta^*, \varepsilon^*|y)}{q_{opt}(\theta, x, \eta^*, \varepsilon^*|y)} \\ &\propto \frac{p(y|\theta, x)p(x|\theta)}{q(x|\theta, y)} \frac{p(\theta)p(\eta^*)p(\varepsilon^*)}{\Big[q_{1,\zeta_1}(\theta|y)p(\eta^*)p(\varepsilon^*) + q_{2,\zeta_2}(\theta, \eta^*, \varepsilon^*|y)\Big]} \\ &= q(y|\theta)\frac{p(y|\theta, x)}{q(y|\theta, x)} \frac{p(\theta)p(\eta^*)p(\varepsilon^*)}{\Big[q_{1,\zeta_1}(\theta|y)p(\eta^*)p(\varepsilon^*) + q_{2,\zeta_2}(\theta, \eta^*, \varepsilon^*|y)\Big]} \qquad (6.1) \end{aligned}$$

where in (6.1) we used the fact that

$$q(x|\theta, y) = \frac{q(y|\theta, x)q(x|\theta)}{q(y|\theta)}$$

and $p(x|\theta) = q(x|\theta)$ (cf. Barra et al., 2014). Notice that the formula (6.1) corresponds to the EMitISEM weight (4.6), while the second factor in (6.1) is the importance weight (4.10) of the signal from the likelihood evaluation based on NAIS. Once the IS weights are determined, one can proceed as discussed in Subsection 2.2.2 to obtain the IS estimates of two risk measures of interest.

In contract with the IS risk evaluation, the direct approach for the parameter driven models can be characterised in a much simpler way. One first draws the models parameters from some candidate density, e.g. a multivariate Student's t distribution, then samples the corresponding signal paths using a simulation smoother. The disturbances vectors are drawn independently. The implied future profit/losses values lead to the direct VaR and ES estimates as discussed in Subsection 2.1.2.

6.2 Applications

The evidence of a poor performance of the direct approach applied to the observation driven models calls for employing of the IS VaR and ES estimation also in the parameter driven models. In practice, however, some numerical problems may occur, as it was in our case. A very common problem arising in the IS analysis is that the importance weights are likely to either become extremely small or large, which may lead to only few draws being assigned non-negligible weights. Depending on the context, different approaches are developed to deal with this problem[14].

We also experienced the ill-behaviour of the importance weights, which we can attribute to the marginalised likelihood $q(y|\theta^{(i)})$ and the signal importance weights $p(y|\theta, x^{(i)})/q(y|\theta, x^{(i)})$ varying heavily among the model parameter draws $\theta^{(i)}$, $i = 1, \ldots, N$. Given the limited scope of the project, so far we were not able to develop a way to overcome this difficulty, which we leave for further research. Even though the inability of computing the importance weights prevents us from obtaining the final point IS estimates of the h-day-ahead $100(1-\alpha)\%$ VaR and ES, below we present the idea behind the Extended QERMit algorithm based on the simplified approach without the importance weights. We illustrate the consecutive steps of the method to finally obtain a posterior profit/loss distribution for the SV and SVt models resembling the ones obtained in the previous section for the observation driven models. This suggests that our innovative approach is likely to deliver much more accurate results than the direct approach.

6.2.1 SV

The second class of models used to analyse volatility of financial return series is the stochastic volatility (SV) model (cf. Taylor, 1986, Harvey et al., 1994). Similarly to the GARCH model, the SV model is able to capture volatility clustering present in time series of financial logreturns. The key difference between both types of models is that in the SV model the variance of the logreturns is subject to an unobserved innovation. Hence, the current information set does provide an explicit characterisation of the underlying latent stochastic process for the log-volatility. As pointed out in Jungbacker and Koopman (2009), the SV model is known to outperform the GARCH models in terms of volatility forecasts and to be linked to option pricing theory.

For the time series of financial log-returns y_t the simplest version of the SV model is given by

$$\begin{aligned} y_t &= \sigma_t \varepsilon_t & \varepsilon_t &\sim \mathcal{N}(0,1) \\ \sigma_t &= \exp\left(\frac{1}{2}x_t\right), & x_t &= c + \alpha_t, \\ \alpha_{t+1} &= \phi\alpha_t + \sigma_\eta \eta_t, & \eta_t &\sim \mathcal{N}(0,1). \end{aligned} \tag{6.2}$$

The univariate signal x_t is interpreted as the unobserved log-volatility. The parameter vector θ is given by $(c, \sigma^2, \phi)^T$, where c is the unconditional mean of the log-volatility; $\phi \in (0,1)$ is a persistence parameter, which typically exceeds 0.8; $\sigma_\eta^2 > 0$ is the variance of the log-volatility process. The unconditional variance of the log-volatility is equal to $\frac{\sigma_\eta^2}{1-\phi^2}$ and it characterises the "volatility of volatility". Regarding

[14] In the Sequential Monte Carlo literature these problems are known as the weight degeneracy and sample impoverishment. There, the basic technique to overcome these obstacles is the Sequential Importance Resampling, cf. Gordon et al., 1993.

the prior specification, we follow Omori et al. (2007) and set the prior distribution as below

$$
\begin{aligned}
c &\sim \mathcal{N}(0,1),\\
\frac{\phi+1}{2} &\sim \text{Beta}(20,1.5),\\
\frac{1}{\sigma_\eta^2} &\sim \text{Gamma}\left(\frac{5}{2},\frac{0.05}{2}\right)
\end{aligned}
$$

We consider the estimation of 1-day-ahead 99% VaR and ES for the IBM stock daily logreturns from January 3, 2007 to December 30, 2011, which are presented in Figure 6.1. The sample consists of 1259 observation and is characterised by the excess kurtosis (equal to 7.0063) and a slight positive skewness (equal to 0.1051).

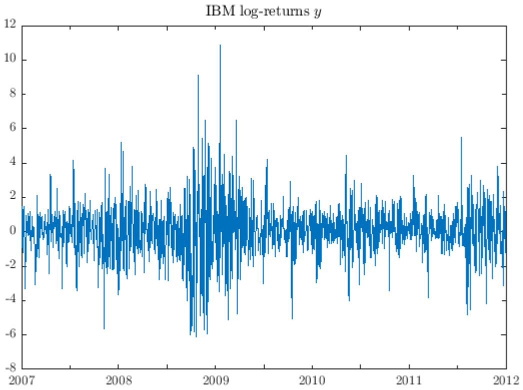


Figure 6.1: IBM log-returns, from January 3, 2007 to December 30, 2011.

Approximation to the parameter posterior density

As discussed in Subsection 6.1, the first step of the Extended QERMit procedure consists in constructing of the candidate importance sampling density for the posterior density corresponding to the whole profit/loss space. We begin with building of a mixture of the Student's t densities approximating the model parameter density using the MitISEM algorithm as discussed in Subsection 4.1.2. To initialise the procedure, we perform the SML estimation of the model parameter vector $\theta = (c, \phi, \sigma_\eta^2)$ (cf. Subsection 4.1.2). The numerical optimisation of the simulated likelihood is carried out with respect to the transformed parameter vector $\tilde{\theta} = T(\theta)$, where the inverse of the mapping T, $T(c, \phi, \sigma_\eta^2) = \left(c, \log\left(\frac{\phi}{1-\phi}\right), \log \sigma_\eta^2\right)$ has the form

$$
T^{-1}(\tilde{c}, \tilde{\phi}, \tilde{\sigma}_\eta^2) = \left(\tilde{c}, \frac{1}{1+\exp(-\tilde{\phi})}, \exp \tilde{\sigma}_\eta^2\right).
$$

The inverse of the Jacobian of T^{-1} is given by

$$
J^{-1}(T) = \text{diag}\left(1, \left(1+\exp(-\tilde{\phi})\right)\left(1+\exp(\tilde{\phi})\right), \exp(-\tilde{\sigma}_\eta^2)\right)
$$

and it is necessary to obtain the Hessian of the loglikelihood as the function of the original parameters, which is used to construct the scale of the initial naive component. The SML estimates are reported in Table 6.1 are in line with the results in the literature (cf. Barra et al., 2014). We take them as the mode of the initial component.

The inverse of the Hessian of the loglikelihood at the SML estimates, corrected by the inverse of Jacobian

Parameter	Estimate	St. deviation
c	0.5116	7.1254
ϕ	0.9677	0.3677
σ^2_η	0.0524	0.5308

Table 6.1: SML estimation results for the parameters of the SV model.

of the reverse transformation, is given by

$$\Sigma = \begin{bmatrix} 50.7707 & 0.0890 & -0.1982 \\ 0.0890 & 0.1352 & -0.1338 \\ -0.1982 & -0.1338 & 0.2817 \end{bmatrix}.$$

To enhance the numerical performance of the algorithm, we replace element $\Sigma_{1,1}$ by 1, so that the draws of parameter c are more concentrated around its SML estimate. We take this modified matrix as the scale matrix of the initial Student's t density, where we set the number of degrees of freedom to 5.

We simulate $N = 10,000$ model parameter draws θ^i, $i = 1, \dots, N$ from the initial proposal density. For each draw we perform the NAIS algorithm to determine $\chi^{(i)}$, the optimal importance sampling parameters for the Gaussian density, which we use in the simulation smoothing of the corresponding logvolatility paths $x^{(i)}$. Finally, we compute the importance weights $w^{(i)}$ of the augmented draws $(\theta^{(i)}, x^{(i)})$ using the formula (4.6). The numerical problems with the weights discussed above prevent us from performing further steps of the MitISEM algorithm, hence we take the initial candidate as our final approximation to the model parameters posterior density. The corresponding NAIS parameters define the conditional Gaussian density for the signal.

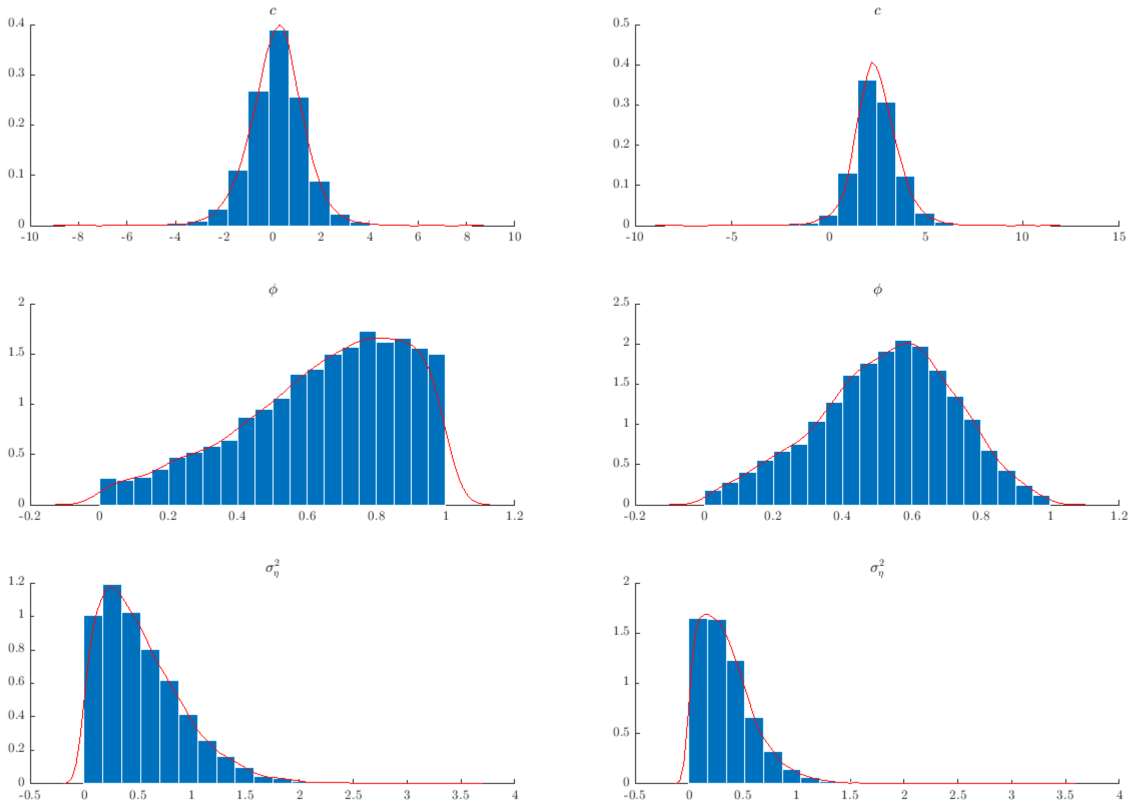
Figure 6.2a depicts the simulated values of the SV model parameters based on the draws from the approximation to the posterior parameter density $q_{1,\zeta}(\theta|y)$. One can see that the shapes of the histograms reflect the prior assumptions on the model parameters. However, the draws of ϕ and σ^2_η are spread substantially more widely than indicated by the prior assumptions. This can be contributed to the draws being generated from the naive candidate (i.e. the one which was not updated) with the large elements of the scale matrix.

The weighted averages of the NAIS parameters $b = \{b_t\}_{t=1}^T$ and $C = \{C_t\}_{t=1}^T$ are shown in Figure 6.3 (the blue lines). As the weights we use only the weights for the model parameter $p(\theta)/q_{1,\zeta}(\theta|y)$. The reason is that the NAIS parameters do not depend on the draws of the signal paths simulated from the Gaussian model determined by them. In contrast, the signal paths should be weighted using the EMitISEM weights to reflect the probability of obtaining both, the underlying model parameter vector and the realised signal trajectory. Due to the problems in computing the extended weights, Figure presents 6.4 the ordinary average of the signal paths simulated using the model parameter draws (the blue line). One can see that it roughly corresponds to the smoothed signal obtained from the KFS based on the model parameters set to their SML estimates (the black line), although the averaged signal is slightly higher. This might be the consequence of using the ordinary average and not the one based on the extended weights.

Preliminary VaR estimation

In the second step, we simulate N independent draws of the future disturbances of states η_{T+1} and observations ε_{T+1} from the standard normal distribution as implied by model (6.2). These, together with the augmented draws $(\theta^{(i)}, x^{(i)})$, determine the future profit/loss values. The preliminary 1-day-ahead 99% VaR estimate is then obtained in a standard way, i.e. as the $0.01N$-th of the ascending sorted profit/loss values and is equal to

$$\widehat{VaR}_{\text{prelim}} = -4.6510.$$



(a) Draws from the approximation to the posterior parameter density $q_{1,\varsigma}(\theta|y)$.

(b) Draws from the approximation to the high-loss density $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$.

Figure 6.2: Simulated values of SV model parameters.

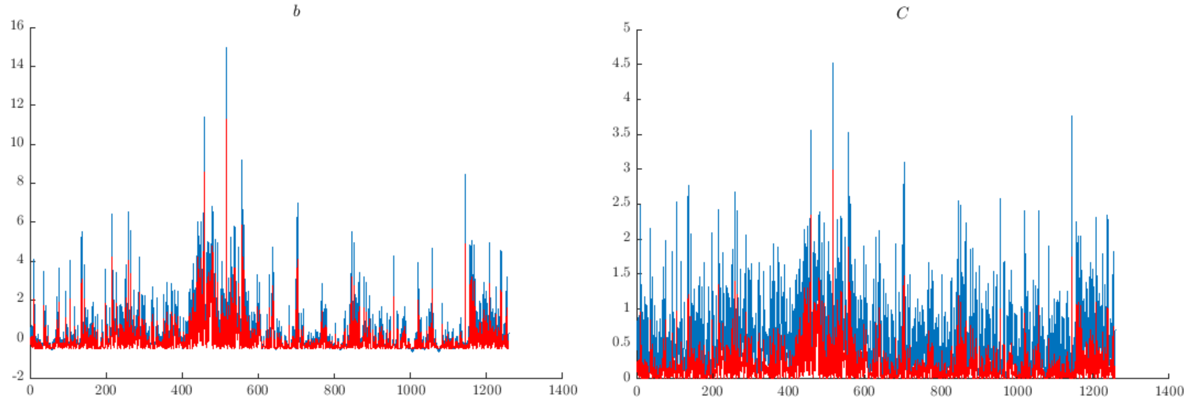


Figure 6.3: Weighted average of the NAIS parameters given the draws from $q_{1,\varsigma}(\theta|y)$ and $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$ for the SV model corresponding to the whole profit/loss space (blue) and to the the high-loss region (red).

High-loss density approximation

The approximation to the 5-dimensional high-loss density of parameter and disturbances was obtained iteratively and jointly with the conditional Gaussian density for the states. Figure (6.2b) presents the simulated values of the SV model parameters corresponding to the high-loss density, i.e. based on the

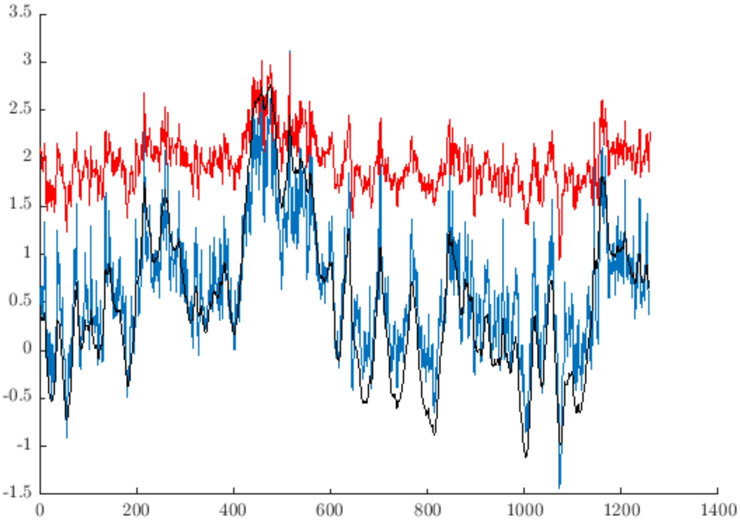


Figure 6.4: Average signal path given the draws from $q_{1,\varsigma}(\theta|y)$ and $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$ for the SV model corresponding to the whole profit/loss space (blue) and to the the high-loss region (red) together with the smoothed signal for the SML model parameter estimates (black).

draws from the approximation to the posterior parameter density $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$. One can see a considerable change in the histogram shapes for all model parameters. The mode of the posterior high-loss draws of c moved substantially to the right, now being equal to around 2.22, while for the whole profit/loss region the draws of c were centred around 0. The mass of the high-loss posterior distribution of ϕ visibly shifted from around 0.8 towards 0, concentrating around 0.6. The draws of σ_η^2 from the high-loss density are spread much less widely, and now they predominantly lie below 0.5.

The high-loss-region model parameters shall imply different sample properties of the NAIS importance parameters. This is indeed the case, as it can be seen in Figure 6.3. The average NAIS parameters corresponding to the high-loss subspace are lower and less volatile. Since the NAIS parameter $C = \{C_t\}_{t=1}^T$ defines the inverse of the variance of the observation disturbance in the linear Gaussian auxiliary model, a lower C in the high-loss region means that the artificial observation from the approximative Gaussian model are more volatile.

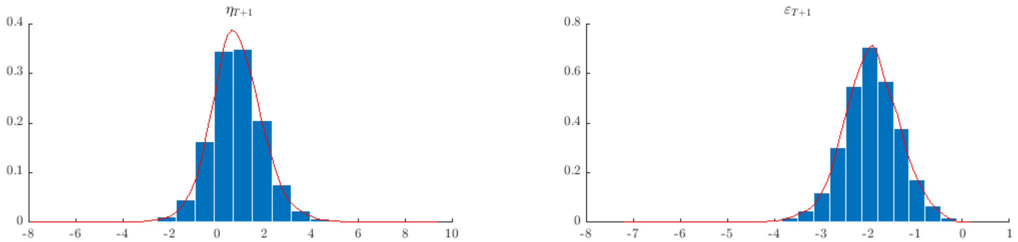


Figure 6.5: Simulated values of the future state (left) and observation (right) disturbances drawn from $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$.

Regarding the simulated values of the two error terms, Figure 6.5 shows that in the high-loss region the state disturbance has a positive mean of around 0.9, with the standard deviation similar to the one implied by the model. On contrary, the observation disturbances which lead to high losses are highly negative, with the average value of around -2. They are also more concentrated around the mean compared to the observation disturbances for the whole profit/loss space.

One can conclude that the high-loss region is characterised by much higher average logvolatility, lower logvolatility persistence and lower volatility of the logvolatility. The unobserved logvolatility process is, however, subject to a larger noise. These properties of the high-loss logvolatility process are captured in Figure 6.4 (the red line). Such a signal, together with considerable negative observation disturbances, leads to extreme losses.

Optimal candidate density construction

As in the standard case, the optimal candidate density in the extended QERMit method is obtained using 50-50 formula (2.15). For a given set of draws from the whole space and from the high-loss region, q_{opt} plays a role in determining their importance weights. The latter are necessary to correctly determine the location in the sorted profit/loss values for the combined sample of the profit/loss value corresponding to the quantile of interest of the posterior profit/loss distribution for the whole space. Loosely speaking, this amounts to locating the red dot in the plot of the ordered profit/losses, as in Figures 5.5 and 5.7 for the ARCH and GARCH-t cases.

The problems with the weight determination, which we have discussed in the introduction to this section, prevented us from obtaining this very last step of the QERMit procedure. Hence, we need to limit our analysis to a more qualitative discussion, leaving the task of obtaining of the quantitative results for further research. Figure 6.6 displays the sorted profit/loss values for the combined sample from the whole profit/loss space and from the high-loss region. The horizontal axis shows the indices i of draws ($i = 1, \dots, 20000$), while the vertical axis displays the i-th sorted profit/loss value. One can see that the shape of the obtained curve follows a similar pattern to the ones observed in the previous section. However, compared to the models analysed there, there is remarkable difference in the shape of the high-loss part, which now is steeper and characterised by a higher curvature. This can be attributed to the main difference between the SV models and the GARCH-type models, which is the number of error processes driving the observation dynamics (i.e. treating the logvolatility of a latent process). This result confirms that the SV model, even with the Gaussian observation errors, is able to replicate the fat tails of the profit/loss distribution observed in the data.

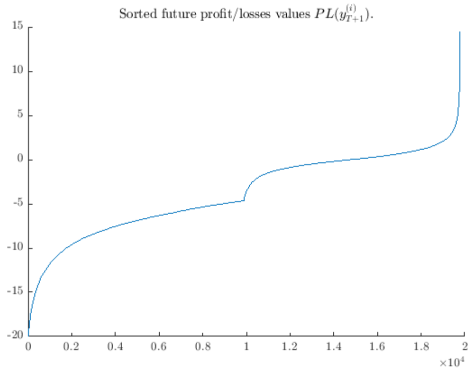


Figure 6.6: Sorted future profit/losses values $PL(y_{T+1}^{(i)})$ for the SV model.

6.2.2 SVt

As the second application of the Extended QERMit method to a parameter driven model we consider the extension of the basic Gaussian SV model, the SV model with t observation disturbances (SVt). Its specification is given by

$$\begin{aligned} y_t &= \sqrt{\rho}\sigma_t\varepsilon_t \qquad\qquad \varepsilon_t \sim t(\nu) \\ \rho &:= \frac{\nu-2}{\nu}, \\ \sigma_t &= \exp\left(\frac{1}{2}x_t\right), \quad x_t = c + \alpha_t, \\ \alpha_{t+1} &= \phi\alpha_t + \sigma_\eta\eta_t, \quad \eta_t \sim \mathcal{N}(0,1). \end{aligned} \tag{6.3}$$

and allows it to capture the heavy tails of the financial returns since the number of degrees of freedom ν of the distribution of the observation disturbances parametrises the fatness of the tails. This modification leads to the model logdensity no longer being Gaussian but Student's t and hence taking the following form

$$\begin{aligned} \log p(y|x) =& T\cdot\left(\log\Gamma\left(\frac{\nu+1}{2}\right) - \log\Gamma\left(\frac{\nu}{2}\right) - \frac{1}{2}\log(\nu-2)\right) \\ & - \frac{1}{2}\sum_{t=1}^{T}x_t - \frac{1}{2}\sum_{t=1}^{T}(\nu+1)\log\left(1+\frac{y_t^2}{(\nu-2)\sigma_t^2}\right). \end{aligned}$$

We set the same priors for parameters c, ϕ and σ_η^2 as for the SV model. For parameter ν we proceed as in the case of the GARCH-t model specifying the proper uninformative exponential prior for $\nu - 2$, so that the variance of ν exists. We consider the estimation of 1-day-ahead 99% VaR and ES and use the same IBM dataset, which we used in the previous Subsection. Because the general outline of the QERMit procedure is the same as for the SV model, we refrain here from a detailed discussion of each step, focusing instead on the modifications in the implementation of the basic method and the differences in the results.

Approximation to the parameter posterior density

We base the candidate importance sampling density for the posterior density corresponding to the whole profit/loss space on the Student's t density with the mode equal to the SML estimates of the model parameter vector $\theta = (c, \phi, \sigma_\eta^2, \nu)$, the scale set to the inverse of the Hessian computed at the mode and 5 degrees of freedom. Again, the numerical optimisation of the simulated likelihood is carried out with respect to the transformed parameter vector $\tilde{\theta} = T(\theta)$, where now the inverse of the mapping T, $T(c, \phi, \sigma_\eta^2, \nu) = \left(c, \log\left(\frac{\phi}{1-\phi}\right), \log\sigma_\eta^2, \log(\nu-2)\right)$ has the formula

$$T^{-1}(\tilde{c}, \tilde{\phi}, \tilde{\sigma}_\eta^2, \tilde{\nu}) = \left(\tilde{c}, \frac{1}{1+\exp(-\tilde{\phi})}, \exp\tilde{\sigma}_\eta^2 \exp(\tilde{\nu}) + 2,\right).$$

The inverse of the Jacobian of T^{-1} is given by

$$J^{-1}(T) = \operatorname{diag}\left(1, \left(1+\exp(-\tilde{\phi})\right)\left(1+\exp(\tilde{\phi})\right), \exp(-\tilde{\sigma}_\eta^2), \exp(-\tilde{\nu})\right).$$

Table 6.2 present the SML estimates together with the corresponding standard errors. Notice that the estimates for the three parameters from the basic SV model change only slightly, with the estimates for c and ϕ increasing, while for σ_η^2 decreasing. The estimated value of the degrees of freedom is high (equal to 11.17), yet still similar to the values reported in literature (cf. e.g. Chib et al., 2002,Jacquier et al., 1994). The huge standard error of the estimate of ν reflects the fact that this parameter is hard to precisely estimate in practice.

The inverse of the Hessian of the loglikelihood at the SML estimates, corrected by the inverse of Jacobian

Parameter	Estimate	St. deviation
c	0.5565	8.2203
ϕ	0.9783	0.3073
σ_η^2	0.0320	0.4134
ν	11.1708	136.7771

Table 6.2: SML estimation results for the SVt model parameters.

of the reverse transformation, is given by

$$\Sigma = \begin{bmatrix} 67.5734 & 0.1085 & -0.2109 & -118.3921 \\ 0.1085 & 0.0944 & -0.0906 & -11.8130 \\ -0.2109 & -0.0906 & 0.1708 & 20.9784 \\ -118.3921 & -11.8130 & 20.9784 & 18707.9746 \end{bmatrix}.$$

Similarly as in the SV case, we tune matrix Σ so that it generates more reliable results: we set $\Sigma_{1,1}$ to 1 and $\Sigma_{4,4}$ to around 46. The latter choice reflects the high uncertainty inherent in the estimation of the number of degrees of freedom.

As in the case of the SV model, we simulate $N = 10,000$ model parameter draws θ^i, $i = 1, \ldots, N$ from the initial proposal density and for each draw we compute the NAIS importance parameters χ^2. Then, we use them to simulate the corresponding logvolatility paths $x^{(i)}$. Again, we take the initial candidate as our final approximation to the model parameters posterior density, due to the numerical problems with the computation of the importance weights defined by (4.6). The corresponding NAIS parameters define the conditional Gaussian density for the signal.

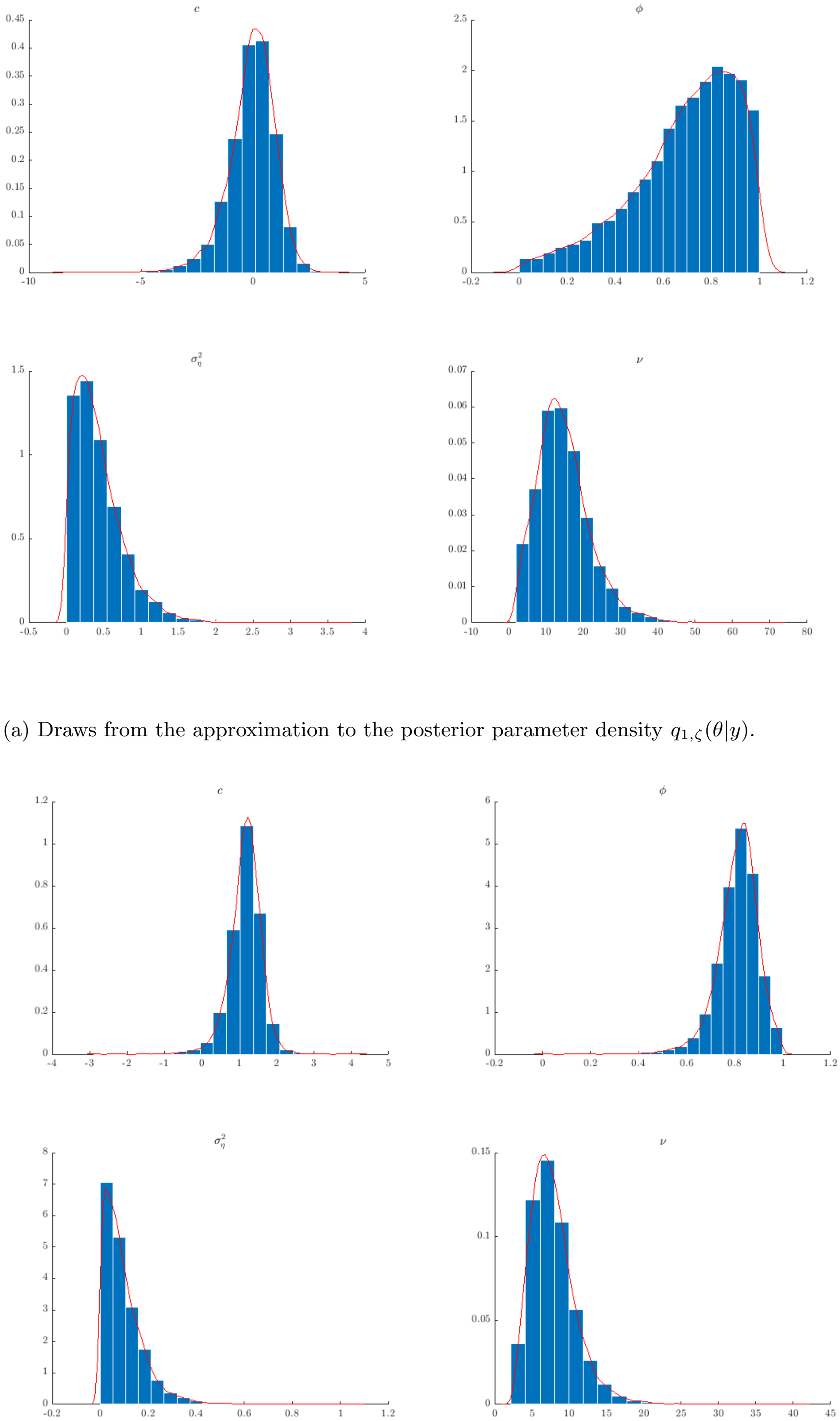
The simulated values of the SVt model parameters based on the draws from the approximation to the posterior parameter density $q_{1,\varsigma}(\theta|y)$ are shown in Figure 6.7a. Compared to the SV model case, the draws of the parameter c are spread more widely, with a slight negative skew. A noticeable change can also be observed for the distribution of σ_η^2, which shrinks towards 0. The simulated values of ϕ exhibit fairly any change compared to the ones obtained in the previous subsection. As far as the draws of ν are concerned, the shape of the histogram closely reflects both, the prior assumption, and the chosen modification in the scale matrix of the initial Student's t component.

Figure 6.8 shows the weighted averages of the NAIS parameters $b = \{b_t\}_{t=1}^T$ and $C = \{C_t\}_{t=1}^T$ (the blue lines). Remarkably, both b and C for the SVt model are on average much lower than for the SV model (cf. the ranges of the vertical axes) and less volatile. The average signal paths simulated using the model parameter draws is shown in Figure 6.9 (the blue line). As in the SV case, it behaves similarly to the smoothed signal obtained from the KFS based on the model parameters set to their SML estimates (the black line). Differently than it was previously, however, now the averaged signal is slightly lower than the smoothed signal based on the SML estimates.

Preliminary VaR estimation

As previously, the second step consists in the simulation of N independent draws of the future disturbances of states η_{T+1} and observations ε_{T+1}. This time, however, the latter are drawn from the Student's t distribution, as implied by model (6.3). The number of degrees of freedom used for each draw corresponds to the parameter draws $\theta^{(i)}$, $i = 1, \ldots, N$. The state disturbances are simulated from the standard normal distribution. The preliminary 1-day-ahead 99% VaR estimate based on the profit/loss values implied by the parameter and the disturbance draws is given by

$$\widehat{VaR}_{\text{prelim}} = -3.8420.$$



(a) Draws from the approximation to the posterior parameter density $q_{1,\zeta}(\theta|y)$.

(b) Draws from the approximation to the high-loss density $q_{2,\zeta}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$.

Figure 6.7: Simulated values of SV model parameters.

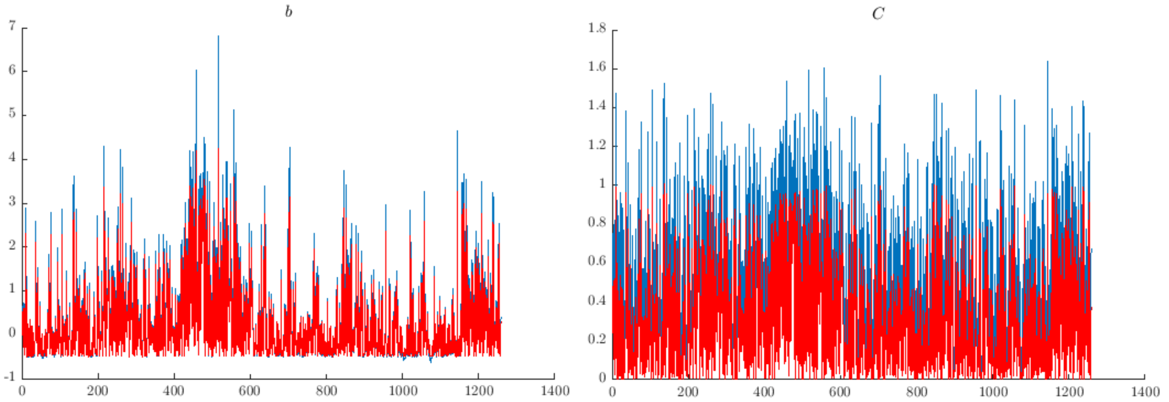


Figure 6.8: Average NAIS parameters given the draws from $q_{1,\varsigma}(\theta|y)$ and $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$ for the SVt model corresponding to the whole profit/loss space (blue) and to the high-loss region (red).

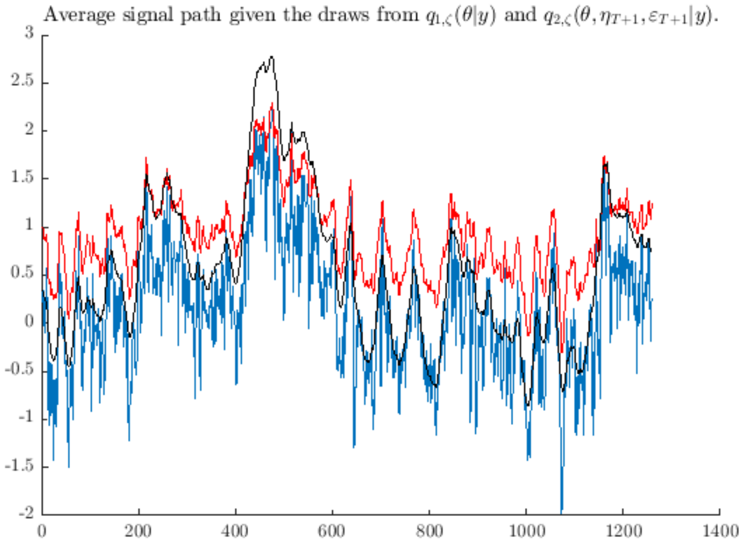


Figure 6.9: Average signal path given the draws from $q_{1,\varsigma}(\theta|y)$ and $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$ for the SV model corresponding to the whole profit/loss space (blue) and to the high-loss region (red) together with the smoothed signal for the SML model parameter estimates (black).

High-loss density approximation

Figure (6.7b) presents the simulated values of the SV model parameters corresponding to the high-loss density, which are constructed using the draws from the approximation to the posterior parameter density $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$. The most striking observation is that all the high-loss draws are much more concentrated around the modes than their 'whole' space counterparts. The extend of this shrinkage is noticeably higher than in the SV model. As far as properties of the individual parameters are concerned, the mode of the parameter c moves from 0 to 1, which is less than in the SV model. The behaviour of the parameter ϕ is yet far more different than in the previous model, as now the logvolatility persistence increases for the high-loss scenarios, with its mass predominantly concentrated around 0.85. The shape of the histogram of σ_η^2 draws largely corresponds to the shape of the prior for σ_η^2, as now the vast majority of draws of this parameter lies below 0.2. Finally, the mode of the simulated values of the number of degrees of freedom parameter moved from around 11 for the 'whole' space to around 7 for the high-loss region. This indicates that the observation disturbances leading to extreme losses are characterised by fatter tails than in these of the observation errors for the 'standard' profit/loss values.

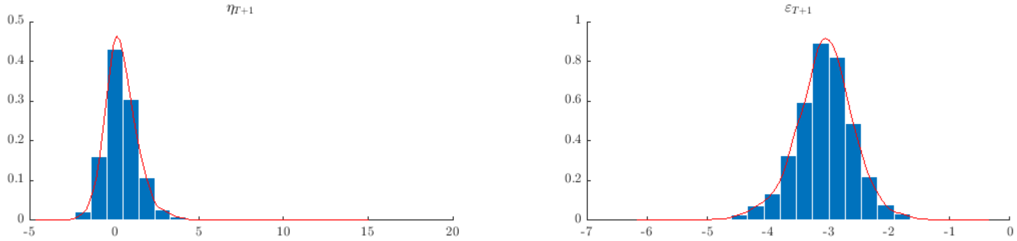


Figure 6.10: Simulated values of the future state (left) and observation (right) disturbances drawn from $q_{2,\varsigma}(\theta, \eta_{T+1}, \varepsilon_{T+1}|y)$.

Figure 6.10 displays the posterior distribution of the two error terms for the high-loss scenarios, which noticeably differ from their SV counterparts. First, one can see that now the state disturbance in the high-loss region has the mode which is not much different from the one implied by the model (6.3), i.e. 0. Interestingly, however, now the state disturbances are slightly skewed to the right. Second, the observation disturbances which lead to high losses are again highly negative, even more than in the case of the SV model. Moreover, one can indeed spot the tails which are fatter than when these disturbances were Gaussian.

Summing up, the high-loss region arising in the SVt setting features the average logvolatility higher than the whole profit/loss space and considerably lower volatility of volatility. These properties were also reported for the corresponding parameters of the SV model. Differently than in the latter, now the persistence of logvolatility in the high-loss region in higher and less dispersed. The observation errors leading to the high-loss scenarios have fatter tails than usually and are even more negative than the ones from the SV model. Finally, the state noise in the high-loss region has only slightly positive mean, which states in contrast to the SV model case, where it was highly positive.

One can conclude that the high-loss region is characterised by much higher average logvolatility, lower logvolatility persistence and lower volatility of the logvolatility. The unobserved logvolatility process is, however, subject to a larger noise. These properties of the high-loss logvolatility process are captured in Figure 6.4 (the red line). Such a signal, together with considerable negative observation disturbances, leads to extreme losses.

Optimal candidate density construction

For the same reasons as in the previous subsection, we report only the qualitative outcomes of the SVt study, leaving establishing of the quantitative ones for further research. Figure 6.11 shows that in principle, the shape of the sorted future profit/loss curve is similar to the one obtained with the normal SV model. Again, the horizontal axis shows the indices i of draws ($i = 1, \ldots, 20000$), while the vertical axis displays the i-th sorted profit/loss value. What is different, is that the high-loss part is noticeably flatter as compared to the previous case. This is rather suspicious, as one would expect that modelling of the observation errors with the Student's t distribution shall lead to even fatter tails than in the normal SV model.

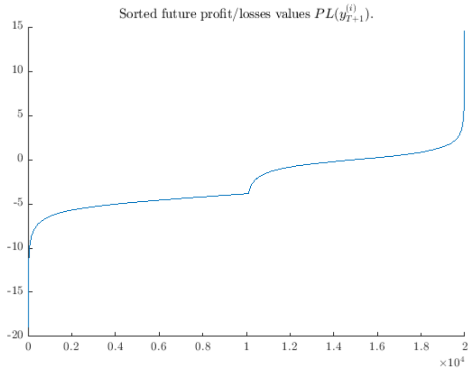


Figure 6.11: Sorted future profit/losses values $PL(y_{T+1}^{(i)})$ for the SVt model.

7 Conclusions

We have considered Bayesian risk estimation using the QERMit algorithm of Hoogerheide and van Dijk (2010), which we have modified and extended in two steps. First, we have replaced the unreliable and numerically unstable AdMit algorithm by the robust and accurate MitISEM method for constructing of the approximation to the posterior density. The results of our empirical study based on a series of daily S&P500 returns confirm the importance of our modification, showing a substantial gain in the accuracy and the precision of the VaR and ES estimates when the posterior predictive densities are targeted using MitISEM rather than with AdMit.

Second, we have developed a fundamental extension of the basic QERMit approach which allows for the application of QERMit to a broader class of models, including the parameter driven models. The incorporation of the latent volatility process into the analysis heavily relies on the importance sampling methods for nonlinear, non-Gaussian state space models. In the application of the extended QERMit method, we have encountered numerical difficulties in computing the required importance weights, overcoming of which we leave for further research. Nonetheless, we have illustrated our new method using a series of daily IBM stock returns. This application suggests promising results for the risk evaluation based the nonlinear non-Gaussian models, given the obstacles in the importance weights are removed.

Obtaining of the quantitative results for the parameter driven models constitutes the main focus for further research. Subsequently, we want to extend the time horizon for which we calculate VaR and ES, in particular to obtain the 10-days-ahead estimates, as required by the Basel standards. Finally, the developed framework can serve to answer numerous interesting research questions. For instance, we aim to compare the QERMit results generated with the parameter driven models and with the observation driven models. This would contribute to the long-standing debate about the superiority of one class of models over another.

Bibliography

Ardia, D., L. F. Hoogerheide, and H. K. van Dijk (2009), "Adaptive Mixture of Student-t Distributions as a Flexible Candidate Distribution for Efficient Simulation: The R Package AdMit." *Journal of*

Statistical Software, 29, 1–32.

Artzner, P., F. Delbaen, J. M. Eber, and D. Heath (1999), “Coherent Measures of Risk.” *Mathematical Finance*, 9, 203–228.

Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002), “A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking.” *IEEE Transactions on Signal Processing*, 50, 174–188.

Barra, I., L. F. Hoogerheide, S. J. Koopman, and A. Lucas (2014), “Joint Bayesian Analysis of Parameters and States in Nonlinear, Non-Gaussian State Space Models.” *TI Discussion Paper, Amsterdam, Tinbergen Institute*, 14-118/III.

Basel Committee on Banking Supervision (1995), “An Internal Model-based Approach to Market Risk Capital Requirements.” *The Bank for International Settlements, Basel, Switzerland.*

Bollerslev, T. (1986), “Generalised Autoregressive Conditional Heteroskedasticity.” *Journal of Econometrics*, 51, 307—-327.

Chib, S., F. Nardari, and N. Shephard (2002), “Markov chain Monte Carlo methods for stochastic volatility models.” *Journal of Econometrics*, 108, 281–316.

Cox, D. R. (1981), “Statistical Analysis of Time Series: Some Recent Developments.” *Scandinavian Journal of Statistics*, 8, 93–115.

Danielsson, J. (1994), “Stochastic Volatility in Asset Prices, Estimation with Simulated Maximum Likelihood.” *Journal of Econometrics*, 64, 375–400.

Dempster, A.P., N. M. Laird, and D. B. Rubin (1977), “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Doucet, A., N. de Freitas, and N. Gordon (2001), *Sequential Monte Carlo Methods in Practice.* Information Science and Statistics, Springer.

Durbin, J. and S. J. Koopman (1997), “Monte Carlo Maximum Likelihood Estimation for non-Gaussian State Space Models.” *Biometrika*, 84, 669–684.

Durbin, J. and S. J. Koopman (2012), *Time Series Analysis by State Space Methods: Second Edition.* Oxford Statistical Science Series, OUP Oxford.

Engle, R. F. (1982), “Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the United Kingdom Inflation.” *Econometrica*, 50, 987—-1007.

Fridman, M. and L. Harris (1998), “A Maximum Likelihood Approach for Non-Gaussian Stochastic Volatility Models.” *Journal of Business and Economic Statistics*, 16, 284—291.

Gelman, A. (1995), “Inference and Monitoring Convergence.” In *Markov Chain Monte Carlo in Practice* (K. Gilks, S. Richardson, and D. J. Spiegelhalter, eds.), 131–143, Chapman & Hall.

Geweke, J. (1989), “Bayesian Inference in Econometric Models using Monte Carlo Integration.” *Econometrica*, 57, 1317–1739.

Gordon, N. J., D. J. Salmond, and A. F. M. Smith (1993), “Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation.” *IEE Proceedings F on Radar and Signal Processing*, 140, 107—113.

Hammersley, J. M. and D. C. Handscomb (1964), *Monte Carlo Methods.* Methuen.

Harvey, A. C., E. Ruiz, and N. Shephard (1994), “Multivariate Stochastic Variance Models.” *Review of Economic Studies*, 61, 247—6.

Hol, E. and S. J. Koopman (2002), “Stock Index Volatility Forecasting with High Frequency Data.” Tinbergen Institute Discussion Papers 02-068/4, Tinbergen Institute.

Hoogerheide, L. and H. K. van Dijk (2010), “Bayesian Forecasting of Value at Risk and Expected Shortfall using Adaptive Importance Sampling.” *International Journal of Forecasting*, 26.2, 231–247.

Hoogerheide, L. F., J. F. Kaashoek, and H. K. van Dijk (2007), “On the Shape of Posterior Densities and Credible Sets in Instrumental Variable Regression Models with Reduced Rank: an Application of Flexible Sampling Methods using Neural Networks.” *Journal of Econometrics*, 139, 154–180.

Hoogerheide, L. F., A. Opschoor, and H. K. van Dijk (2012), “A Class of Adaptive Importance Sampling Weighted EM Algorithms for Efficient and Robust Posterior and Predictive Simulation.” *Journal of Econometrics*, 171.2, 101–120.

Jacquier, E., N. Polson, and P. Rossi (1994), “Bayesian Analysis of Stochastic Volatility Models (with Discussion).” *Journal of Business and Economic Statistics*, 12, 371—-417.

Jorion, P. (2007), *Value at Risk: the New Benchmark for Managing Financial Risk*, volume 3. McGraw-Hill New York.

Jungbacker, B. and S. J. Koopman (2009), “Parameter estimation and practical aspects of modeling stochastic volatility.” In *Handbook of Financial Time Series* (T. G. Andersen, R. A. Davis, J. P. Kreiss, and Th. V. Mikosch, eds.), 61–79, Springer-Verlag.

Kahn, H. and A. W. Marshal (1953), “Methods of Reducing Sample Size in Monte Carlo Computations.” *Journal of the Operational Research Society of America*, 46, 263–271.

Kim, S., N. Shephard, and S. Chib (1998), “Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models.” *Review of Economic Studies*, 65, 361—-393.

Kitagawa, G. (1987), “Non-Gaussian State-Space Modeling of Nonstationary Time Series.” *Journal of the American Statistical Association*, 82, 284—291.

Kitagawa, G. (1996), “Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models.” *Journal of Computational and Graphical Statistics*, 5, 1—25.

Kloek, T. and H. K. van Dijk (1978), “Bayesian Estimates of Equation System Parameters: an Application of Integration by Monte Carlo.” *Econometrica*, 46, 1–20.

Koopman, S. J., A. Lucas, and M. Scharth (2015), “Numerically Accelerated Importance Sampling for Nonlinear Non-Gaussian State Space Models.” *Journal of Business and Economic Statistics*, 33, 114–127.

Kullback, S. and R. A. Leibler (1951), “On Information and Sufficiency.” *The Annals of Mathematical Statistics*, 1, 79–86.

Marshall, A. W. (1956), “The Use of Multi-Stage Sampling Schemes in Monte Carlo Computations.” In *Symposium on Monte Carlo Methods* (M. Meyer, ed.), 123–140, Wiley.

McNeil, A. J. and R. Frey (2000), “Estimation of Tail-Related Risk Measures for Heteroscedastic Financial Time Series: an Extreme Value Approach.” *Journal of Empirical Finance*, 7, 271–300.

Omori, Y., S. Chib, N. Shephard, and J. Nakajima (2007), “Stochastic Volatility with Leverage: Fast and Efficient Likelihood Inference.” *Journal of Econometrics*, 140, 425–449.

Peel, D. and G. McLachlan (2000), “Robust Mixture Modeling using the t-Distribution.” *Statistics and Computing*, 10, 339–348.

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

Richard, J. and W. Zhang (2007), “Efficient High-Dimensional Importance Sampling.” *Journal of Econometrics*, 141, 1385–1411.

Sandmann, G. and S. J. Koopman (1998), “Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood.” *Journal of Econometrics*, 87, 271—-301.

Shephard, N. and M. Pitt (1997), “Likelihood Analysis of Non-Gaussian Measurement Time Series.” *Biometrika*, 84, 653–667.

Svensén, M. and C. M. Bishop (2005), “Robust Bayesian Mixture Modeling.” *Neurocomputing*, 64, 339–348.

Taylor, S. J. (1986), *Modelling Financial Time Series.* Wiley, Chicheste.

Zeevi, A. J. and R. Meir (1997), “Density Estimation through Convex Combinations of Densities; Approximation and Estimation Bounds.” *Neural Networks*, 10, 99—-106.

A NSE

A.1 VaR

The NSE for the IS estimator for $100\alpha\%$ VaR needs to be obtained using the delta method. This is because the true value of $100\alpha\%$ VaR is not known with certainty, so that the expectation of $\mathbb{I}\{PL(X) \leq VaR\}$ is not an unconditional expectation of a function of a random variable X with the known density kernel. The required formula is given by

$$\hat{\sigma}_{IS,VaR} = \frac{\hat{\sigma}_{IS,\mathbb{P}[PL \leq \widehat{VaR}_{IS}]}}{\hat{p}_{PL}(\widehat{VaR}_{IS})}. \tag{A.1}$$

The derivation of (A.1) can be found in Hoogerheide and van Dijk (2010). The nominator of (A.1) is the numerical standard error for the IS estimator of the probability $\mathbb{P}[PL(x) \leq c]$, for $c = \widehat{VaR}_{IS}$. It follows from Geweke (1989), as the value c is this time fixed to the preliminary estimate of $100\alpha\%$ VaR.

The denominator is the estimator of the density of the PL function at $\widehat{VaR}_{IS}$ and in general it lacks an explicit formula. It can be estimated by

$$\frac{\mathbb{P}[PL(X) \leq c + \varepsilon] - \mathbb{P}[PL(X) \leq c - \varepsilon]}{2\varepsilon}, \qquad \varepsilon > 0.$$

It is advisable to compute the above expression for several values of ε and use the ε which yields the *lowest* estimate of $\hat{p}_{PL}(\widehat{VaR}_{IS})$, so the one ends up with the highest, more conservative estimate of $\hat{\sigma}_{IS,VaR}$.

A.2 ES

If the VaR were known with certainty, the IS estimation of ES would be a standard estimation of a function of a random variable distributed according to a truncated distribution, i.e. of $PL(X)$, where

$X \sim p(x)\mathbb{I}\{PL(X) \leq VaR\}$. In such a standard case the NSE would follow from Geweke (1989). However, only the IS estimate of $100\alpha\%$ is available, which means that explicit formula for the NSE does not hold. To compute the required NSE one can proceed as follows. First, construct a grid of VaR values, e.g. ranging from $\widehat{VaR}_{IS} - 4\hat{\sigma}_{IS,\widehat{VaR}}$ to $\widehat{VaR}_{IS} + 4\hat{\sigma}_{IS,\widehat{VaR}}$. Second, for each value on the grid compute the NSE of the IS estimator ES if the true VaR were the value from the grid (i.e. conditional NSE), as well as estimate the conditional probability $\hat{p}(\widehat{ES}_{IS}|VaR)$ of the corresponding ES estimator. Third, estimate the density of the ES estimator as the weighted average of the conditional NSEs from the previous step, with the weights equal to the conditional densities of the VaR values from the grid. Finally, compute the NSE of $\widehat{ES}_{IS}$ as the standard deviation of the estimated density $p(\widehat{ES})_{IS}$.

B MitISEM

B.1 Approximation with Mixtures of Student's t Distributions

We want to approximate the target density $\tilde{p}(\theta)$ of which only the kernel $p(\theta)$ is required with the candidate density $q_\zeta(\theta)$ such that the *Kullback-Leibler divergence* (Kullback and Leibler, 1951)

$$\int p(\theta)\log p(\theta)d\theta - \int p(\theta)\log q_\zeta(\theta)d\theta \tag{B.1}$$

is minimised. The target density p will usually be the posterior density given the data y, but we omit the conditioning on y for the notational convenience. Moreover, we will take as the candidate q_ζ the mixture of Student's t distributions, so that the minimisation will be carried out with respect to the mixture parameters ζ and the number of mixture components H. Since the first term in (B.1) does not depend on ζ, the minimisation of (B.1) amounts to the maximisation of

$$\begin{aligned}\int \log q_\zeta(\theta)p(\theta)d\theta &= \int \log q_\zeta(\theta)\frac{p(\theta)}{q_\zeta(\theta)}q_\zeta(\theta)d\theta\\ &= \mathbb{E}_{q_\zeta}\left[\log g(\theta)\frac{p(\theta)}{q_\zeta(\theta)}\right],\\ &\approx \frac{1}{N}\sum_{i=1}^{N}\log q_\zeta(\theta^{(i)})\frac{p(\theta^{(i)})}{q_\zeta(\theta^{(i)})}\\ &= \frac{1}{N}\sum_{i=1}^{N}\log q_\zeta(\theta^{(i)})w(\theta^{(i)}),\end{aligned}$$

where $\theta^{(i)} \overset{i.i.d.}{\sim} q_{\zeta_{old}}(\theta)$ were drawn from the previous candidate, and

$$w(\theta^{(i)}) = \frac{p(\theta^{(i)})}{q_\zeta(\theta^{(i)})}. \tag{B.2}$$

Importantly, the draws $\theta^{(i)}$, $i = 1, \dots, N$, and their weights $w(\theta)^{(i)}$ are fixed during the optimization and they do not depend on ζ.

B.2 EM Step in MitISEM

Consider a mixture of H Student-t densities

$$q_\zeta(\theta) = \sum_{h=1}^{H}\eta_h t(\theta|\mu_h, \Sigma_h, \nu_h), \tag{B.3}$$

where $t(\theta|\mu,\Sigma,\nu)$ denotes the d-dimensional Student-t density

$$t_d(\theta|\mu,\Sigma,\nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{nu}{2}\right)(\pi\nu)^{d/2}}|\Sigma|^{-1/2}\left(1+\frac{(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)}{\nu}\right)^{-(d+\nu)/2}$$

and $\zeta = \{\mu_h, \Sigma_h, \nu_h, \eta_h\}_{h=1}^H$ is the set of the mixture parameters: modes, scale matrices, degrees of freedom and mixing probabilities. The aim is to maximise the weighted log-density

$$\frac{1}{N}\sum_{i=1}^{N} w^{(i)} \log q_\zeta(\theta^{(i)}), \tag{B.4}$$

with respect to ζ, where $w^{(i)} = w(\theta^{(i)}) = \frac{p(\theta^{(i)})}{q_\zeta(\theta^{(i)})}$ is the importance weight of the draw $\theta^{(i)}$. Using the fact the a Student's t distribution can be represented as a mixture of normal distributions with the covariance matrices scaled by the random variables following an Inverse-Gamma distribution, one can equivalently represent the draws $\theta^{(i)}$ from the mixture (B.3) in (B.4) as

$$\theta^{(i)} \sim \mathcal{N}(\mu_h, \kappa_h^{(i)}\Sigma_h), \qquad \text{if } z_h^{(i)} = 1,$$

where $z^{(i)} \in \mathbb{R}^H$ is a latent vector from the standard base with one on the place corresponding to the component h which the draw $\theta^{(i)}$ has been drawn from. The probability $\mathbb{P}[z^{(i)} = e_h]$ of belonging to the component h is given by η_h. The scaling factor $\kappa_h^{(i)}$ follows the Inverse-Gamma distribution

$$\kappa_h^{(i)} \sim \mathcal{IG}(\nu_h/2, \nu_h/2).$$

Such a representation introduces the latent data $\tilde{\theta} = \{z_h, \kappa_h\}_{h=1}^H$ into the logdensity $\log p(\theta)$, so that the standard numerical maximisation of the data-augmented $\log p(\theta, \tilde{\theta}|\zeta)$ density is infeasible. To find the optimal mixture parameters ζ one can resort to the Expectation-Maximisation (EM) algorithm of Dempster et al. (1977), which allows for the maximum likelihood estimation for the incomplete data problems. The core of the procedure is to iterate between two steps, the Expectation step and the Maximisation step. In the former, one calculates the conditional expectation of the loglikelihood function with respect to the latent variables $\tilde{\theta}$, given the parameter values from the previous iteration, ζ. In the latter, the expected loglikelihood is maximised with respect to the parameters.

The conditional expectations in the **Expectation** step are given by

$$\begin{aligned}
\tilde{z}_h^{(i)} &\equiv \mathbb{E}\left[z_h^{(i)}\middle|\theta^{(i)},\zeta\right] = \frac{\eta_h t(\theta^{(i)}|\mu_h,\Sigma_h,\nu_h)}{\sum_{l=1}^H \eta_l t(\theta^{(i)}|\mu_l,\Sigma_l,\nu_l)}, \\
\widetilde{z/\kappa}_h^{(i)} &\equiv \mathbb{E}\left[\frac{z_h^{(i)}}{\kappa_h^{(i)}}\middle|\theta^{(i)},\zeta\right] = \tilde{z}_h^{(i)}\frac{d+\nu_h}{\rho_h^{(i)}+\nu_h}, \\
\tilde{\xi}_h^{(i)} &\equiv \mathbb{E}\left[\log\kappa_h^{(i)}\middle|\theta^{(i)},\zeta\right] \\
&= \left[\log\left(\frac{\rho_h^{(i)}+\nu_h}{2}\right) - \psi\left(\frac{d+\nu_h}{2}\right)\right]\tilde{z}_h^{(i)} + \left[\log\left(\frac{\nu_h}{2}\right) - \psi\left(\frac{\nu_h}{2}\right)\right](1-\tilde{z}_h^{(i)}), \\
\tilde{\delta}_h^{(i)} &\equiv \mathbb{E}\left[\frac{1}{\kappa_h^{(i)}}\middle|\theta^{(i)},\zeta\right] = \widetilde{z/\kappa}_h^{(i)} + (1-\tilde{z}_h^{(i)}),
\end{aligned}$$

where $\rho_h^{(i)} = (\theta^{(i)}-\mu_h)^T\Sigma_h^{-1}(\theta^{(i)}-\mu_h)$ and ψ denotes the digamma function.

The updates at the iteration L of the **Maximisation** step are as follows

$$\begin{aligned}
\mu_h^{(L)} &= \left[\sum_{i=1}^{N} w^{(i)} \widetilde{z/\kappa_h}^{(i)}\right]^{-1} \left[\sum_{i=1}^{N} w^{(i)} \widetilde{z/\kappa_h}^{(i)} \theta^{(i)}\right], \\
\Sigma_h^{(L)} &= \frac{\sum_{i=1}^{N} \kappa^{(i)} \widetilde{z/\kappa_h}^{(i)} (\theta^{(i)} - \mu_h^{(L)})(\theta^{(i)} - \mu_h^{(L)})^T}{\sum_{i=1}^{N} w^{(i)} \tilde{z}_h^{(i)}}, \\
\eta_h^{(L)} &= \frac{\sum_{i=1}^{N} w^{(i)} \tilde{z}_h^{(i)}}{\sum_{i=1}^{N} w^{(i)}},
\end{aligned}$$

while the updates for the degrees of freedom $\nu_h^{(L)}$ parameters come from solving of the first-order conditions with respect to ν_h

$$-\psi(\nu_h/2) + \log(\nu_h/2) + 1 - \frac{\sum_{i=1}^{N} w^{(i)} \xi_h^{(i)}}{\sum_{i=1}^{N} w^{(i)}} - \frac{\sum_{i=1}^{N} w^{(i)} \delta_h^{(i)}}{\sum_{i=1}^{N} w^{(i)}} = 0.$$

A more detailed discussion of the MitISEM algorithm can be found in Hoogerheide et al. (2012).

C NAIS

The NAIS approach relies on the *Gauss-Hermite quadrature* based on M nodes z_j and related weights $h(z_j)$, $j = 1, \dots, M$, so that the minimization problem (4.15) is deterministically approximated with

$$\min_{\chi_t} \sum_{j=1}^{M} h(z_j) \exp(z_j^2) \varphi(\tilde{x}_{tj}), \tag{C.1}$$

$$\begin{aligned}
\varphi(\tilde{x}_{tj}) &= \lambda^2(\tilde{x}_{tj}, y_t; \theta) \omega(\tilde{x}_{tj}, y_t; \theta) g(\tilde{x}_{tj} | y_t^*; \theta), \qquad (C.2) \\
\tilde{x}_{tj} &= \tilde{x}_t + \sqrt{V_t} z_j.
\end{aligned}$$

Since $\tilde{x}_t$ and V_t are the moments of the smoothed density $g(x_t|y;\theta)$ given by (4.16), the last term on the right hand side in (C.2) becomes

$$g(\tilde{x}_{tj} | y_t^*; \theta) = \frac{1}{\sqrt{2\pi V_t}} \exp\left\{-\frac{1}{2} z_j^2\right\}.$$

Finally, having the necessary approximation (obtained given IS parameters), the minimisation (C.1) reduces to the weighted least squares regression, with the regressand $\log p(y_t|\tilde{x}_{tj};\theta)$, the regressors $(1, \tilde{x}_{tj}, -\tilde{x}_{tj}^2/2)$ and the weight $h(z_j)\exp(z_j^2)\omega(\tilde{x}_{tj}, y_t|\theta)$[15]. The estimated coefficients corresponding to the second and the third regressors become the updated IS parameters. These are then used to obtain the new integral approximation, which is in turn employed in the next weighted least squares computation. The whole iterative procedure is performed until the convergence criterion in χ is reached. To initialize the iterations, one can set $b_t = 0$, $C_t = 1$, $t = 1, \dots, T$.

[15] For numerical efficiency, the weights can be simplified to $h(z_j)\exp(z_j^2)$, which is referred to as the “fast” optimisation version.