Bayesian Risk Forecasting for Long Horizons *

Agnieszka Borowska[(a,b)], Lennart Hoogerheide[(a,b)] and Siem Jan Koopman[(a,b,c)]

[(a)] Vrije Universiteit Amsterdam

[(b)] Tinbergen Institute

[(c)] CREATES, Aarhus University

January 2019

Abstract

We present an accurate and efficient method for Bayesian forecasting of two financial risk measures, Value-at-Risk and Expected Shortfall, for a given volatility model. We obtain precise forecasts of the tail of the distribution of returns not only for the 10-days-ahead horizon required by the Basel Committee but even for long horizons, like one-month or one-year-ahead. The latter has recently attracted considerable attention due to the different properties of short term risk and long run risk. The key insight behind our importance sampling based approach is the sequential construction of marginal and conditional importance densities for consecutive periods. We report substantial accuracy gains for all the considered horizons in empirical studies on two datasets of daily financial returns, including a highly volatile period of the recent financial crisis. To illustrate the flexibility of the proposed construction method, we present how it can be adjusted to the frequentist case, for which we provide counterparts of both Bayesian applications.

Keywords: Bayesian inference; forecasting; importance sampling; numerical accuracy; long run risk; Value-at-Risk; Expected Shortfall.

*We would like to thank the participants at the 11th NESG meeting in Leuven and at the 10th CFE in Seville for their insightful comments.

1 Introduction

The global financial crisis stressed the importance of appropriate risk management, which requires accurate forecasts of the market risk related to fluctuations of stock or index prices. It also emphasised the necessity of precise predictions of the long-term financial risk: as noted by The Volatility Laboratory (2012)[1], the turbulent events of 2008 moved the focus of risk forecasting from solely short term horizons to longer ones. This is because most portfolios consist of assets that are held longer than just a few days, so that e.g. excess leverage is likely to pose a much higher risk in the long run than in the short run (Engle, 2009). Hence, increased attention has been recently devoted to risk forecasting for one-month-ahead or even one-year-ahead horizons, and not only the standard, 1-day-ahead or 10-days-ahead measures required by Basel Committee on Banking Supervision (1995).

One of the potential reasons why the main focus was previously on short run forecasts is the difficulty of obtaining precise risk predictions for long horizons. As noted by McNeil et al. (2015) and Embrechts et al. (2005), a straightforward approach to risk forecasting based on the so-called scaling rule, suitable for short term risk, might be inappropriate for long-term forecasts[2]. Furthermore, Christoffersen et al. (1998) state that generally conventional parametric models are ill-suited for extreme events analysis because they focus on "average" scenarios in order to obtain a high goodness of fit. This misperformance may be even more severe when the horizon of analysis increases.

McNeil and Frey (2000) distinguish three main approaches to forecasting tail related measures: non-parametric historical simulations (HS); parametric methods based on an econometric model where the volatility dynamics are explicitly specified; methods based on extreme value theory (EVT). They argue that a parametric model of volatility is essential in order to capture the volatility dynamics exhibited by financial returns, which allows for prediction of risk based on the current volatility background. Moreover, parametric

[1] As it describes itself, The Volatility Laboratory (V-Lab) of The Volatility Institute provides real time measurement, modelling and forecasting of financial volatility, correlations and risk for a wide spectrum of assets and it produces volatility forecasts up to a year in advance. The Volatility Institute was created at New York University Stern School of Business in 2009 under the direction of Professor R. F. Engle.

[2] The performance of the scaling rule crucially depends on the data generating process, in particular its "closeness" to a normal random walk model, where indeed a quantile of H-days-ahead distribution is given by the quantile of the 1-day-ahead distribution multiplied by $\sqrt{H}$ (see Daníelsson and Zigrand, 2006; Diebold et al., 1997).

time series models provide a framework to extrapolate the analysis beyond the observed data – as opposed to the HS methods. For these reasons it is a natural starting point for our analysis to build upon parametric methods from the second group. As the main drawback of these models McNeil and Frey (2000) indicate their common conditional normality assumption, which seems to be invalid for most financial series. Hence, they apply EVT to estimate extreme quantiles of the distribution of the standardised residuals from a normal GARCH model. The EVT approach for capturing the properties of extreme tails was also suggested by Christoffersen et al. (1998).

In this paper we decide to proceed differently: in order to address the issue of precise long-run risk evaluation we build upon the approach of Hoogerheide and van Dijk (2010). These authors suggest to forecast the probability of extreme events, for a given volatility model, via importance sampling (IS) based on a specially designed importance density focusing on the left tail. To cope with heavy tails of conditional return distributions we consider volatility models with Student's t distributed error terms. We propose an accurate and efficient approach to forecasting two standard measures of market risk, Value at Risk (VaR) and Expected Shortfall (ES), in a situation when the prediction horizon is long, e.g. 40, 100 or 250 days ahead. The latter is a noticeable contribution compared to Hoogerheide and van Dijk (2010), who proposed a method suited for standard short-run analysis[3]. To this end we first redesign the original approach of Hoogerheide and van Dijk (2010) using a more flexible approximation algorithm. Second, we suggest a novel sequential construction of the importance density, feasible thanks to employing that new algorithm. The construction of importance densities allows for "guiding" of the subsequent simulated returns over time so that the cumulative return falls in the "high-loss" region, making the analysis of long horizons feasible. In our approach the properties of the subsequent conditional importance densities depend on the previous simulated returns in the sense that at each step we take into consideration the cumulative return up to that time point. This allows us to assess how much the situation still needs to deteriorate in order to qualify for being a "high-loss" scenario. We focus on the 99% quantile of the profit-loss distribution, as required by the Basel Committee on Banking Supervision (1995); such an extreme tail is also more challenging to precisely predict than

[3]Hoogerheide and van Dijk (2010) note that the relative performance of their method may decrease with the prediction horizon length, due to the so-called "curse of dimensionality of importance sampling", and is likely to vanish for very long horizons, such as 100-days-ahead.

e.g. the 95% quantile, which is also commonly analysed.

It is important to stress that our method is universal, i.e. it can be applied for any chosen parametric volatility model. Hence, we abstract from the issue of model selection, but aim at a precise and efficient evaluation of risk implied by the given model. Nevertheless, our method is still highly advantageous in the context of model selection because by reducing the uncertainty related to the simulation noise the comparison between models is more likely to be based on their "true" quality.

As a variance reduction technique, IS has been already applied in the context of market risk evaluation. Importantly, Glasserman et al. (1999), Glasserman et al. (2000) and Glasserman et al. (2002) combine IS with stratified sampling to obtain precise estimates of VaR. They, however, do not consider time series models and carry out barely a "numerical example", not an empirical study with real data. Furthermore, they restrict their attention to a 10-days-ahead horizon and analyse portfolio loss probabilities from the frequentist perspective. However, risk forecasting, and especially for long horizons, is subject to a considerable parameter uncertainty. That is why the Bayesian approach seems to be particularly suited for long run risk analysis. In addition, not only it naturally captures parameter uncertainty but also provides a convenient starting point for considering model uncertainty via Bayesian model averaging. Therefore we follow Hoogerheide and van Dijk (2010) and focus primarily on the analysis from the Bayesian perspective. However, to illustrate the merits and the flexibility of the proposed method, we demonstrate how the method can be adjusted to the frequentist case, for which we provide the counterparts of the Bayesian applications.

The outline of the paper is as follows. In Section 2 we first recall the approach of Hoogerheide and van Dijk (2010) to show how IS can be applied in the context of Bayesian risk forecasting; second, we present how our proposed method allows to mitigate the "curse of dimensionality", inherent to IS, to allow for more accurate and efficient long run VaR and ES forecasts. We illustrate the performance of our novel method in Section 3 with two workhorse models, commonly used by practitioners, i.e. the Generalized Autoregressive Conditional Heteroscedasticity model (GARCH, Engle, 1982; Bollerslev, 1986) and the Generalised Autoregressive Score model (GAS, Creal et al., 2013), both with Student's t innovations. In Section 4 we consider the alternative, frequentist method for long run prediction of VaR and ES: we discuss the necessary methodology modifications and pro-

vide the frequentist counterparts of the Bayesian applications from Section 3. Section 5 concludes and presents an outline for the further research.

2 Bayesian risk evaluation using importance sampling

Let $\{y_t\}_{t\in\mathbb{Z}}$ be a time series of daily logreturns $y_t = 100(\log p_t - \log p_{t-1})$ on a financial asset with price p_t at the end of day t, with $y_{1:T} := \{y_1, \ldots, y_T\}$ denoting the observed data. We assume that $\{y_t\}_{t\in\mathbb{Z}}$ is subject to a dynamic stationary process parametrised by θ, on which we put a prior $p(\theta)$. Let $y^*_{1:H} = \{y_{T+1}, \ldots, y_{T+H}\}$ denote the vector of H future returns and consider the posterior predictive distribution of profit/loss $PL(y^*_{1:H}) = 100\left[\exp\left(\sum_{t=T+1}^{T+H} y_t/100\right) - 1\right]$ (converting the sum of the logreturns to the percentage return) defined as defined as

$$p\left(PL(y^*_{1:H})|y_{1:T}\right) = \int p\left(PL(y^*_{1:H})|y_{1:T}, \theta\right) p(\theta|y_{1:T}) d\theta, \tag{2.1}$$

obtained by marginalisation over the parameter with respect to the posterior distribution $p(\theta|y_{1:T})$. We are interested in Bayesian forecasting of the $100\alpha\%$ VaR, i.e. the $100(1-\alpha)\%$ quantile of the posterior predictive distribution of profit/loss within a horizon of the next H trading days, i.e.

$$100\alpha\%\,\mathrm{VaR} = \inf\left\{x \in \mathbb{R} : \mathbb{P}(PL(y^*_{1:H}) \geq x|y_{1:T}) \geq \alpha\right\}.$$

We also consider ES as an alternative risk measure, due to its advantageous properties compared to VaR, mainly sub-additivity (which makes ES a coherent risk measure in the sense of Artzner et al., 1999). Given $100\alpha\%$ VaR, the conditional ES is defined as

$$100\alpha\%\,\mathrm{ES} = \mathbb{E}\big[PL(y^*_{1:H})|PL(y^*_{1:H}) < 100\alpha\%\,\mathrm{VaR}\big].$$

Since (2.1) is usually analytically intractable, simulation based methods need to be applied in order to estimate VaR and ES. Following Hoogerheide and van Dijk (2010) we distinguish two approaches to that. The first one, which we will refer to as the *direct approach*, is straightforward:

1. draw a sample of model parameter $\theta^{(i)}$, $i = 1, \ldots, M$, from the posterior distribution

(using e.g. the Metropolis-Hastings algorithm);

2. generate the corresponding paths of H future log-returns $y^{*(i)} = \{y^{(i)}_{T+1}, \ldots, y^{(i)}_{T+H}\}$;

3. compute the resulting profits/losses $PL(y^{*(i)})$;

4. sort in ascending order the values of $PL(y^{*(i)})$ to obtain the permutation $PL^{(j)}$, $j = 1, \ldots, M$;

5. obtain the $100\alpha\%$ VaR and ES 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 Volatility Laboratory (2012) uses this direct approach for the non-Bayesian forecasting of long run VaR, where step 1 is replaced by frequentist estimation. The drawback of the direct approach is that it is subject to an inherent problem of rare events simulations, i.e. that most of the generated scenarios are not the ones of the ultimate interest, the extremely negative ones. This makes direct estimators very inefficient and the only way to increase their precision is to consider many more draws. Obviously, the latter is costly, in terms of both computing time and computing resources (e.g. the available memory).

To illustrate the problem, let us introduce a toy example of white noise returns[4]

$$y_t \sim \sqrt{\sigma^2}\varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0,1), \qquad \sigma^2 \sim p(\sigma^2),$$

where p is a conjugate prior distribution. Then, the future profits/losses follow $PL(y^*_{1:H}) \sim \mathcal{N}(0, H\sigma^2)$. If we treat σ^2 as known and equal to 1, i.e. under the assumption that the data were generated from a standard normal distribution, the value for the 10-days-ahead 99% VaR is given by $\Phi^{-1}(0.01)\sqrt{10} = -7.3566$, for 100-days-ahead it is equal to -23.2635, while for 250-days-ahead to -36.7828. Figure 2.1 presents the outcome of the direct approach for the shortest horizon of 10-days-ahead. One can see that – as discussed above – only a very small fraction of roughly 1/100 of the generated paths corresponds to the high losses that we are interested in, which indeed leads to a low efficiency.

[4]In this example we consider for simplicity the cumulative logreturn over $H = 10$ days (instead of the percentage return), so that the profit/loss is just the sum of the H logreturns, i.e. $PL(y^*_{1:H}) := \sum_{h=1}^{H} y^*_h$.

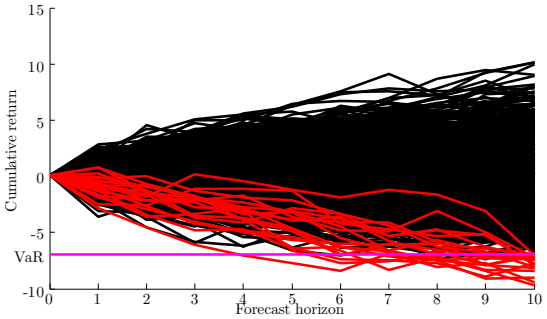


Figure 2.1: Direct simulation results in very few paths (the red ones) below the 99% VaR value (the violet horizontal line). White noise returns, 10-days-ahead horizon, simulated $10,000$ paths.

2.1 Tail focused importance density

To overcome the inefficiency of the direct approach, Hoogerheide and van Dijk (2010) suggest *importance sampling* (IS), a well known variance reduction technique. Its main merit is the potential focus on the *important* subspace by adopting an appropriate sampling density, which in the context of VaR and ES should be tail-focused. Hoogerheide and van Dijk (2010) propose the *Quick Evaluation of Risk using Mixture of t approximations* (QERMit) algorithm, where the key idea is to oversample the high-loss scenarios and to give them lower importance weights. The theoretical insight for their method comes from the properties of the optimal importance density for the Bayesian estimation of $\bar{f} \equiv \mathbb{E}[f(X)]$ for a variable X with density kernel $p(x)$, outlined by Geweke (1989)[5], which is given by $q_{opt}(x) \propto |f(x) - \bar{f}|p(x)$, 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

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

[5]Here, the optimality refers to minimisation, given the specified number of draws, of the numerical standard error of the IS estimator of $\bar{f} \equiv \mathbb{E}[f(X)]$, where f is the function of interest of the random variable X, which has the density $\tilde{p}(x)$ with the kernel $p(x)$.

where c is a constant, which results in[6]

$$\int_{x\in S} q_{opt}(x)dx = \int_{x\notin S} q_{opt}(x)dx = \frac{1}{2}. \tag{2.4}$$

Condition (2.4) implies that half of the total probability mass of the importance distribution shall be located in the region of interest S, and the remaining half outside that region. Such a split is the consequence of using only the kernel of the target distribution and not its proper density, which makes it necessary to adequately normalise the importance weights via sampling from the whole domain instead of merely sampling high loss scenarios, which is the optimal method in the frequentist approach that we consider in the sequel of this paper.

Hoogerheide and van Dijk (2010) apply the above result in the context of VaR and ES estimation. Then, S is interpreted as the "high loss region", i.e. the subspace of the profits/losses 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 2.2 illustrates the construction of the optimal importance density for the VaR estimation.

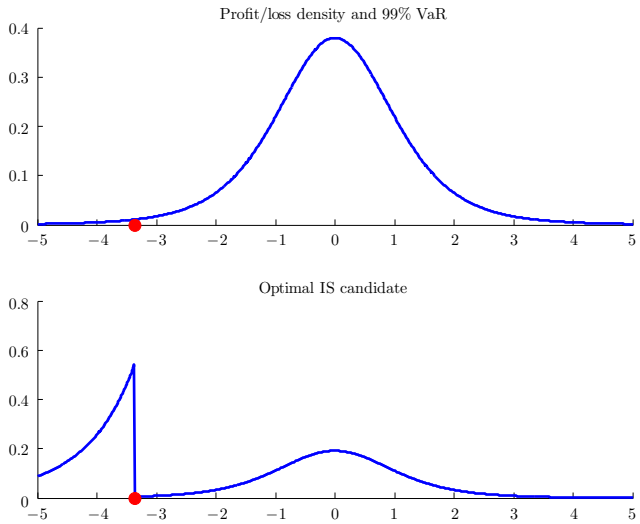


Figure 2.2: Construction of the optimal importance density. Exemplary density function (Student's t with 5 degrees of freedom) of profit/loss and the implied 99% VaR (top). The optimal importance density for the VaR estimation (bottom).

[6]This is obtained by noting that

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

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

Notice that in the case of Bayesian forecasting of VaR and ES we have a joint density $p(\theta, y^*_{1:H}|y_{1:T})$ of the parameters θ and future returns $y^*_{1:H}$ of which we have kernel

$$p(\theta, y^*_{1:H}|y_{1:T}) \propto p(\theta)p(y_{1:T}|\theta)p(y^*_{1:H}|\theta, y_{1:T}),$$

the product of the posterior density kernel and the future returns' density. The IS estimator $\widehat{VaR}_{IS}$ of the $100(1-\alpha)\%$ VaR is obtained by solving x in

$$\mathbb{P}[\widehat{PL(y^*_{1:H})} \leq x]_{IS} = 1 - \alpha,$$

which in practice can be done via the following procedure:

1. draw a sample of parameter vectors $\theta^{(i)}$ and corresponding future returns $y^{*(i)}_{1:H}$, $i = 1, \ldots, M$, from their joint importance density $q(\theta^{(i)}, y^{*(i)}_{1:H}|y_{1:T})$;
2. compute the corresponding importance weights $w^{(i)} = \frac{p(\theta^{(i)}, y^{*(i)}_{1:H}|y_{1:T})}{q(\theta^{(i)}, y^{*(i)}_{1:H}|y_{1:T})}$, $i = 1, \ldots, M$;
3. compute the resulting profits/losses $PL(y^{*(i)}_{1:H})$;
4. sort in ascending order the values of $PL(y^{*(i)}_{1:H})$ to obtain the permutation $PL^{(j)}$, $j = 1, \ldots, M$, with the corresponding weights $w^{(j)}$;
5. set $\widehat{VaR}_{IS}$ as $PL^{(k)}$ for which

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

and given $\widehat{VaR}_{IS}$

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

2.2 Approximations by mixtures of Student's t distributions

The choice of the importance density is crucial for the performance of the IS estimation. Clearly, as pointed out by Geweke (1989), the importance density should resemble the target density and at the same time remain 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 an appropriate importance density becomes particularly cumbersome when the shape of the target density is non-elliptical. As illustrated by Figure 2.2, the optimal importance density for Bayesian VaR estimation is generally bimodal.

A standard approach to overcome this problem is to approximate the target density with a mixture of basis densities[7], for which Student's t densities are often chosen. Several methods to construct the approximating mixture of Student's t have been developed, see Peel and McLachlan (2000), Svensén and Bishop (2005), Hoogerheide et al. (2007) and Hoogerheide et al. (2012). We employ the latter algorithm, Mixture of t by Importance Sampling weighted Expectation Maximization (MitISEM). This is a noticeable distinction compared to Hoogerheide and van Dijk (2010), whose original QERMit algorithm relies on another approximation algorithm, Adaptive Mixture of t (AdMit) of Hoogerheide et al. (2007). Our main motivation behind this change is that 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 target density of the parameters and future returns. We provide a more detailed comparison of both methods in the Online Appendix.

2.3 Sequential construction of importance densities

If the horizon of the future returns increases, then it becomes more difficult to obtain an appropriate importance density for the parameters and future returns. Hence, we want to construct an approximation "sequentially", in each future time period conditioning the properties of the current conditional importance density of the return on the simulated parameters and returns in the previous periods. Intuitively, the idea is to "guide" the draws to fall into the high-loss region: if so far certain losses have been recorded, we know by how much the situation must additionally deteriorate to end up in the tail. Such a sequential and conditional construction of the importance densities can be easily carried out using the Partial MitISEM (PMitISEM) method of Hoogerheide et al. (2012). This algorithm aims at approximating the joint target density indirectly, by approximating the product of marginal and conditional target densities of subsets of model parameters – and in our case future returns.

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

To explain how the "guiding" process is carried out, below we discuss the details of PMitISEM. We express the joint target density $p(\theta)$ as a product of a marginal density and conditional densities:

$$p(\theta) = p(\theta_S|\theta_{S-1}, \ldots, \theta_2, \theta_1) \ldots p(\theta_2|\theta_1)p(\theta_1),$$

where $(\theta_1, \ldots, \theta_S)$ is a partition of a k-dimensional vector θ into S subsets with respective dimensions k_s, $s = 1, \ldots, S$, where naturally $\sum_{s=1}^{S} k_s = k$. Then it may be desirable to iteratively approximate each of the marginal and conditional densities due to the implied dimensionality reduction for each of the sub-problems. In general, the basic MitISEM could be applied to each of them to optimise the *modes*, scale matrices, degrees of freedom and weights independently for each subset. However, this would naturally result in a very poor joint importance density (unless the subsets θ_s are independent) as the conditional structure would be neglected. In order to capture the interdependence between the subsets, in the PMitISEM algorithm the modes of the components in the subsequent conditional subsets are based on fitted values in the regression of the current subset parameters on (a function of) the parameters from the previous subsets (and potentially other "global" variables, e.g. functions of the data). PMitISEM optimises the *regression coefficients* for the conditional importance densities (corresponding to the subsets $\theta_2, \ldots, \theta_S$), instead of optimising their *modes*. Below we discuss the details of the regression.

The underlying idea comes from the basic result in multivariate regression theory. For the sake of simplicity of the exposition we restrict ourselves to the case $S = 2$; the extension to more subsets is straightforward. Consider the (asymptotically valid) approximating normal distribution $\mathcal{N}(\mu, \Sigma)$ for $\theta = (\theta_1^T, \theta_2^T)^T$, where $\mu = \arg\max_\theta f(\theta)$ and $\Sigma = -\mathcal{H}(\log f(\theta))^{-1}|_{\theta=\mu}$, where $f(\theta)$ is the posterior density kernel. Let

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then

$$\theta_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}),$$
$$\theta_2|\theta_1 \sim \mathcal{N}(\underbrace{\mu_2 + \Sigma_{22}^{-1}\Sigma_{21}(\theta_1 - \mu_1)}_{\beta X}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}).$$

The PMitISEM algorithm replaces both the marginal and conditional normal distributions with mixtures of Student's t distributions. The mixture for the marginal distribution for θ_1 is constructed with the basic MitISEM algorithm. The mixture for the conditional density for θ_2 given θ_1 is constructed with a modified version of the algorithm, based on a regression of the parameters θ_2 on a constant term and some functions of parameters from the subsequent subset θ_1 (and potentially the data), all kept in the matrix X. Then, the above mentioned modification pertains to the optimisation of the coefficients of regression β instead of the modes.

In the **basic MitISEM** algorithm the maximisation step for the modes and the covariance matrices of the c-th mixture component is given by

$$\mu_c^{(L)} = \left[\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i\right]^{-1} \left[\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i \theta^i\right],$$
$$\hat{\Sigma}_c^{(L)} = \frac{\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i (\theta - \mu_c^{(L)})(\theta^i - \mu_c^{(L)})^T}{\sum_{i=1}^{N} W^i \widetilde{z}_c^i},$$

where W^i are the importance weights, and where $\widetilde{z/w}_c^i$ and $\widetilde{z}_c^i$, $i = 1, \ldots, N$, are computed in the expectation step of the algorithm. The exact formulae for their computation, together with other details of the basic MitISEM algorithm are provided in the Online Appendix. In the **partial MitISEM** algorithm, the maximisation step for the regression coefficients β and the covariance matrices (for the conditional densities) of the c-th mixture component becomes as follows

$$(\beta_c^{(L)})^T = \left[\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i X_s^i (X_s^i)^T\right]^{-1} \left[\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i X_s^i (\theta^i)^T\right],$$
$$\hat{\Sigma}_c^{(L)} = \frac{\sum_{i=1}^{N} W^i \widetilde{z/w}_c^i (\theta^i - \beta_c^{(L)} X_s^i)(\theta^i - \beta_c^{(L)} X_s^i)^T}{\sum_{i=1}^{N} W^i \widetilde{z}_c^i}.$$

Notice that in the current partial setting each draw θ_s^i (of length k_s) from the subset s

$(s = 2, \ldots, S)$ has a different conditional mean $\mu_c^i = \beta_c X_s^i$, where X_s^i is an $r \times 1$ vector (with elements being a constant and some $r-1$ functions of y and $\theta_1, \ldots, \theta_{s-1}$) and β_c is a $k_s \times r$ matrix. Intuitively, for each subset (and each component h in the conditional importance density for this subset) β_c characterises the common dependence of the s-th subset of parameters θ_s (for component h) on the previous $s-1$ subsets of parameters (and on the data). The details of the procedure are presented in the Online Appendix.

Let us return to the introductory toy example of white noise returns, with only one model parameter σ^2 and H other "parameters" corresponding to the future disturbances $\varepsilon_1, \ldots, \varepsilon_H$. The sampling scheme is then as follows

$$\begin{aligned}
(\sigma^2, \varepsilon_1) &\sim q_1, \\
\varepsilon_2 | \sigma^2, \varepsilon_1 &\sim q_2, \\
\varepsilon_3 | \sigma^2, \varepsilon_1, \varepsilon_2 &\sim q_3, \\
&\vdots \\
\varepsilon_H | \sigma^2, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{H-1} &\sim q_H.
\end{aligned}$$

To construct the conditional mixture importance densities q_h with the PMitISEM algorithm we put for $h = 2, \ldots, H$

$$X_h = \left[\mathbf{1}, \sum_{t=1}^{h-1} y_t^* \right],$$

i.e. a column of ones and the cumulative returns in the previous periods. The latter choice is motivated by our aim to keep track of the evolution of the returns, i.e. how bad the situation has become up to now. In order to construct the marginal and conditional importance densities in the PMitISEM approach we need a preliminary set of parameter draws and corresponding high loss paths of future returns. For this purpose we use the high loss paths (and corresponding parameter draws) of a preliminary run of the direct approach (illustrated in Figure 2.1), which also yields a preliminary VaR forecast. Given the preliminary VaR, this can allow us to assess how much "down" we still need to go in order to get to the high loss region. In Figure 2.3 this aim can be seen as ending up below the violet line.

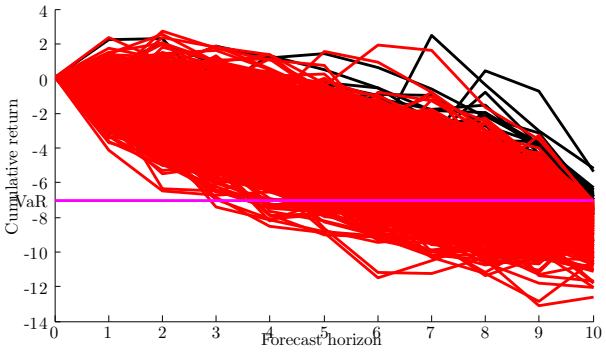


Figure 2.3: Simulations using PMitISEM result in almost all paths falling into the high-loss region (the red ones) below the 99% VaR value (the violet horizontal line). White noise returns, 10-days-ahead horizon, 10,000 simulated paths.

3 Empirical illustrations of Bayesian forecasting

In this section we discuss our key results for the 99% VaR and ES forecasting from the Bayesian perspective. We analyse two benchmark models of volatility, commonly employed by practitioners, the Generalized Autoregressive Conditional Heteroscedasticity model (GARCH, Engle, 1982; Bollerslev, 1986) and the Generalised Autoregressive Score model (GAS, Creal et al., 2013), both with Student's t innovations.

The main purpose of our applications is to illustrate the proposed IS-based forecasting method, i.e. how it is implemented and what efficiency gains it can yield. Keeping this in mind we apply each model to a different dataset, one used in the original paper of Hoogerheide and van Dijk (2010) and another one consisting of more recent data. Importantly, the former is a tranquil series, collected shortly before the financial crisis of 2008, while the latter contains the "wild" period of that financial distress, which makes the analysis much harder. Nevertheless, we record considerable efficiency gains for all the considered horizons also for that difficult dataset.

3.1 GARCH(1,1)-t

As our first illustration we consider the most advanced application from Hoogerheide and van Dijk (2010), where the authors apply the GARCH(1,1)-t model to the daily logreturns of the S&P 500, from January 2, 1998 to December 31, 2007 (2513 observations, Figure 3.1) to evaluate the 10-days-ahead 99% VaR and ES. This is a natural starting point for our analysis, as with the AdMit algorithm employed in the original paper it was already difficult to obtain 10-days-ahead forecasts, while with the MitISEM algorithm "shorter"

horizons, such as 10-day-ahead or 20-day-ahead, are easily reachable. Moreover, adopting the Partial MitISEM algorithm allows us to extend the original analysis much further, to record time–precision gains even for the one-year-ahead horizon.

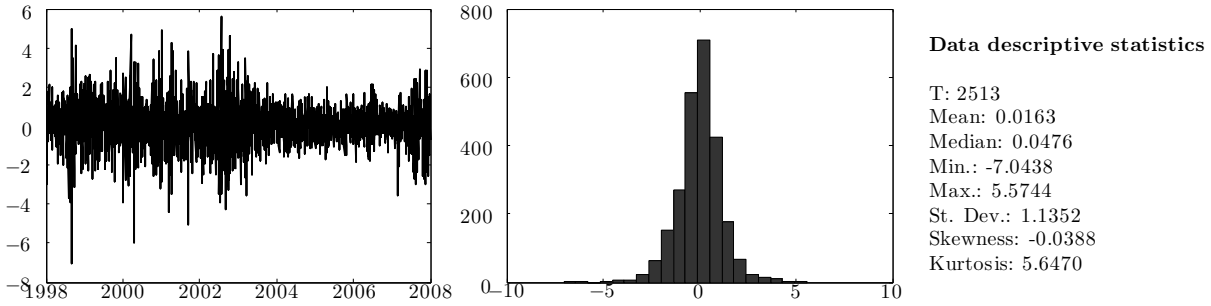


Figure 3.1: The data from the original Hoogerheide and van Dijk (2010) paper: the daily logreturns of the S&P 500, from January 2, 1998 to December 31, 2007.

The model is specified as follows:

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

and we stack the model parameters into the vector $\theta = (\omega, \alpha, \beta, \mu, \nu)$. We put flat priors on $\omega > 0$, $\alpha \in (0, 1)$, $\beta \in (0, 1)$ with $\alpha + \beta < 1$, to enforce that the conditional variance is positive and to ensure covariance stationarity, while for the degrees of freedom we set an uninformative yet proper prior: $\nu - 2 \sim \text{Exp}(0.01)$.

Table 3.1 presents the simulation results for the two direct approaches that we consider. In the naive-direct approach the candidate density is based on a single Student's t distribution with the mode equal to the MLE, the scale matrix equal to minus the inverse of the Hessian of the loglikelihood function evaluated at the mode, and the number of degrees of freedom set to 3 to allow for fat tails (as suggested by Geweke, 1989). To obtain the candidate with the adapted-direct approach we employ the MitISEM algorithm (Hoogerheide et al., 2012) to approximate the posterior of the model parameters with the resulting candidate being a two-component mixture of Student's t distributions. Here, and in the subsequent applications, computation times refer to computations performed on an Intel(R) Core(TM) i5–3470 processor with 3.20 GHz. The "adaptation" of the can-

	ML		MH (naive candidate)			MH (adapted candidate)		
Parameter	MLE	SD	Mean	SD	IF	Mean	SD	IF
ω	0.0082	0.0036	0.0091	0.0035	5.8174	0.0092	0.0034	5.4216
α	0.0726	0.0121	0.0702	0.0110	5.7170	0.0707	0.0109	4.7439
β	0.9238	0.0123	0.9241	0.0118	5.7776	0.9236	0.0117	4.8040
μ	0.0481	0.0169	0.0486	0.0171	5.5711	0.0489	0.0169	4.0058
ν	9.9964	1.9873	10.2582	1.9389	5.9288	10.2512	1.8897	4.2826
		AR		0.4376			0.6802	
		Time construction		0.93 s			60.89 s	
		Time sampling		10.86 s			13.72 s	
		No. of draws		10,000			10,000	

Table 3.1: Estimation results in the **GARCH(1,1)-t** model for Maximum Likelihood (ML) method (reported for comparison) and the Bayesian direct approach with naive (Student's t) and adapted (MitISEM mixture of Student's t) candidate distributions in the independence chain Metropolis-Hastings (MH) method: estimated posterior mean and standard deviation (SD), inefficiency factor (IF), acceptance rate (AR) in the MH method, and computing times for construction of the candidate distribution and for performing the direct approach.

didate takes around one minute but allows for much closer approximation to the posterior distribution. The acceptance rate (AR) in the independence Metropolis-Hastings (MH) with the adapted candidate is almost 70%, which is much higher than when the naive candidate is adopted, in which case the AR is roughly 44%. Similarly, the adapted candidate results in less autocorrelated draws as measured by the inefficiency factors (IF)[8].

For the 99% VaR and ES evaluation we consider, next to both direct methods, two QERMit (i.e. IS-based) approaches. In these methods we apply different methods to approximate the "high-loss" density. The first one uses the basic MitISEM algorithm and targets the posterior predictive density as a whole. For this reason, it usually becomes infeasible to use for prediction horizons longer than 20, because then the covariance matrices of the Student's t components are hard to work with. The second approximation algorithm is PMitISEM, based on the sequential construction of the marginal and conditional importance densities as discussed in Section 2.3, which allows to extend the analysis way further than the basic QERMit of Hoogerheide and van Dijk (2010). We refer to these two methods by subscripts *mit* and *pmit*, respectively. Table 3.2 presents the properties of the partial mixture generated by PMitISEM for the 10-days-ahead case.

[8]The inefficiency factor is defined as the variance of the parameter estimate divided by the variance in case the sampling scheme would generate independent posterior draws and it is the inverse of the relative numerical efficiency (see Pitt et al., 2012). For a sample of draws of a parameter ζ we compute IF as $\text{IF}(\zeta) = 1 + 2\sum_{\tau=1}^{\max\{L,1000\}} \rho_\tau(\zeta)$, where $\rho_\tau(\zeta)$ is the τ-th order autocorrelation in the sequence of draws of parameter ζ and L is the lowest order τ for which ρ_τ is not significant.

Subset	Parameters	No. of components	Weighted* μ or β^*
1	$\{(\theta, \varepsilon_1)\}$	4	[0.0089 0.0698 0.9250 0.0473 9.8126 -1.0284]
2	$\{\varepsilon_2\}$	5	[-1.0863 -0.0948]
3	$\{\varepsilon_3\}$	5	[-1.1816 -0.1083]
4	$\{\varepsilon_4\}$	5	[-1.2589 -0.1070]
5	$\{\varepsilon_5\}$	5	[-1.4608 -0.1390]
6	$\{\varepsilon_6\}$	5	[-1.6167 -0.1471]
7	$\{\varepsilon_7\}$	5	[-1.8912 -0.1697]
8	$\{\varepsilon_8\}$	5	[-2.4583 -0.2202]
9	$\{\varepsilon_9\}$	4	[-2.8833 -0.2568]
10	$\{\varepsilon_{10}\}$	5	[-5.0261 -0.4842]

*Weighted with the mixture weights.

****The mode μ (for subset 1) or the regression coefficients β (for the other subsets).

Table 3.2: Properties of the marginal and conditional importance densities from the PMitISEM method for $H = 10$ in the **GARCH(1,1)-t** model.

Similarly as in the "toy" example of white noise returns we regress the draws from the current conditional importance density s on

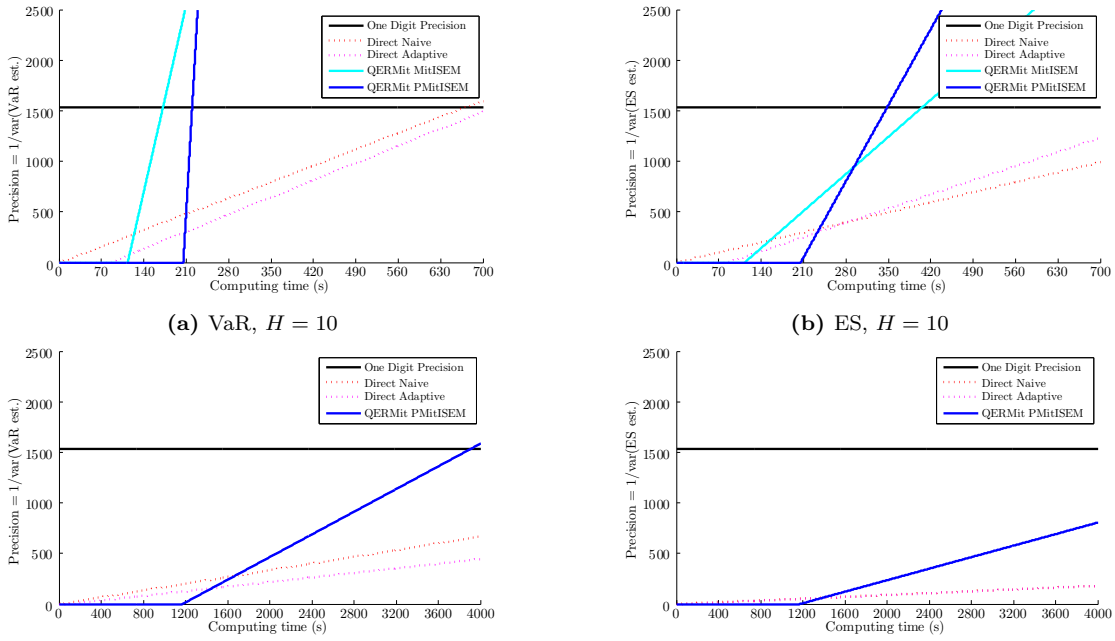
$$X_s = \left[\mathbf{1}, \sum_{t=1}^{s-1} y_t^* \right],$$

to update the mode of the current conditional density. The last column contains the weighted mode of the marginal importance density, i.e. for $s = 1$, and weighted coefficients of regression for the conditional importance densities, i.e. for $s = 2, \ldots, 10$. The latter show how PMitISEM "guides" the subsequent draws into the "high-loss region". As expected, the later the period, the more negative the regression coefficient (at the cumulative return up to period $s-1$), with a noticeable jump in the last period to guarantee that the whole scenario becomes a high-loss one.

Table 3.3 compares the results for the 99% VaR and ES forecasting for different horizons, for which a visualisation is provided in the Online Appendix. For each method, the results are based on $10,000$ draws, while to obtain the NSEs and interquantile ranges (IQR) we performed 20 Monte Carlo replications of the evaluation experiment. Here, and

in the next applications, we consider five horizon lengths, $H \in \{10, 20, 40, 100, 250\}$. This selection ranges from the standard, intermediate horizon of two weeks, required by Basel Committee on Banking Supervision (1995), through the one month horizon, up to the long run, one-year-ahead horizon. The QERMit based methods clearly outperform the direct approaches, with both RNEs and IQRs being roughly 6 times higher for $H = 10$ VaR and almost 3 times for $H = 10$ ES. For the longest horizon of $H = 250$ QERMit delivers two to three times more accurate results than its direct competitors, both for VaR and ES evaluations. As expected, for long horizons, with $H = 40$ or more, the basic MitISEM becomes infeasible, due to the too high dimensionality of the scale matrices of the mixture components it would need to tackle. Fortunately, owing to the partial candidate construction, PMitISEM is still able to deliver satisfactory results even for these long horizons. Notice that PMitISEM outperforms the basic MitISEM already for the shorter horizons ($H = 10$ and $H = 20$), where its VaR forecasts are over twice more accurate than those obtained with basic MitISEM; for the ES the relative advantage of PMitISEM over basic MitISEM is smaller, yet still existing (the results from the latter algorithm are almost 50% less accurate than these from the former one). Interestingly, a better approximation to the posterior does not need to lead to a better performance in the tail: in some cases the adapted direct approach yields worse results that its naive counterpart, in particular when one considers just the IQR and not the NSE (see the NSE and the IQR for the VaR at $H = 250$ or just the IQR for the VaR at $H = 10$ or the ES at $H = 20$). This confirms the remark of Christoffersen et al. (1998) that standard, goodness-of-fit-focused methods are not bound to succeed in the tail estimation problems.

Naturally, for any method it holds that the longer the horizon, the lower the prediction accuracy. Also the advantage of the QERMit method over the direct approach diminishes when the horizon gets extended. The crucial question is then whether there is still a gain, in terms of the time-precision trade-off, of adopting a more accurate but also a more complex and time consuming method. To quantify that trade-off we consider the gain in precision (defined as the inverse of the variance) for one unit of computing time. We refer to it as the *slope*, as it characterises the steepness of a function determining the dependence between precision and computing time. A method with a higher slope will eventually require less computing time to achieve a certain (high) precision, even after accounting for an inevitable fixed "investment cost" of time needed to construct a



(a) VaR, $H = 10$ **(b)** ES, $H = 10$

(c) VaR, $H = 250$ **(d)** ES, $H = 250$

Figure 3.2: Precision ($1/var$) of the predicted VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GARCH(1,1)-t** model, for the shortest and the longest horizon. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based importance density corresponds to a situation when it was not possible to construct such an importance density.

reliable importance density. The results of the investigation on this issue are presented in Table 3.4, and the plots corresponding to the shortest and longest horizon are presented in Figure 3.2 (the Online Appendix provides plots for all the horizons, with additional details on the plots construction). The QERMit based methods turn out to be not only more accurate but also more efficient than the direct approaches, in a sense that they require less computing time and fewer draws to achieve the same accuracy as the direct methods, or, stated differently, they yield higher precision in the same time and using the same number of draws. Importantly, the conditioning of partial MitISEM allows us to increase efficiency for all horizons, including the longest horizon of $H = 250$, for both, VaR and ES evaluations.

Finally, following Hoogerheide and van Dijk (2010) we also consider the benchmark of 1 digit precision with 95% confidence. It is defined as $1.96\text{NSE} \leq 0.05$, which corresponds to the required precision level of 1536. Then, the *time required* and *draws required* refer to the computing time and the number of draws necessary to achieve this precision level. Notice, that this benchmark is set somewhat arbitrarily and considering a higher confidence would mean a much higher required precision. For instance, changing of the confidence to 99% would raise it to 2654. Table 3.4 shows that even for the longest considered horizon of

H		VaR_{naive}	VaR_{adapt}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{adapt}	ES_{mit}	ES_{pmit}
10		-8.1484	-8.1257	-8.2091	-8.1808	-9.9134	-9.7853	-9.9209	-9.8759
	NSE	(0.1836)	(0.1748)	0.0531	(0.0267)	(0.2329)	(0.1922)	(0.1192)	(0.0838)
	IQR	[0.2066]	[0.2478]	[0.0840]	[0.0367]	[0.4104]	[0.2563]	[0.1543]	[0.1384]
	RNE	1.02	1.01	5.8198	12.37	1.59	1.60	11.28	44.66
20		-11.2028	-11.2846	-11.3024	-11.2265	-13.5991	-13.7225	-13.6589	-13.5866
	NSE	(0.2907)	(0.2151)	0.1454	(0.0626)	(0.3923)	(0.3436)	(0.1683)	(0.1141)
	IQR	[0.3382]	[0.3125]	[0.2157]	[0.0958]	[0.5424]	[0.6844]	[0.2118]	[0.1536]
	RNE	1.01	1.03	2.6315	8.75	1.63	1.65	2.23	12.05
40		-15.2151	-15.2188	–	-15.3329	-18.5758	-18.6593	–	-18.7022
	NSE	(0.3520)	(0.3094)	(–)	(0.1020)	(0.5806)	(0.5470)	(–)	(0.1991)
	IQR	[0.3605]	[0.3839]	[–]	[0.1213]	[0.9029]	[0.5279]	[–]	[0.2513]
	RNE	1.03	1.00	–	7.79	1.77	1.67	–	25.85
100		-22.6319	-22.6711	–	-22.6115	-28.4722	-28.3719	–	-28.6178
	NSE	(0.6497)	(0.4005)	(–)	(0.2433)	(0.8134)	(0.7701)	(–)	(0.3119)
	IQR	[0.8049]	[0.5865]	[–]	[0.4399]	[1.0842]	[1.2843]	[–]	[0.4846]
	RNE	1.03	1.05	–	5.43	1.64	1.65	–	9.08
250		-32.0179	-32.0471	–	-32.1617	-41.3818	-41.8261	–	-41.3818
	NSE	(0.6737)	(0.7966)	(–)	(0.3266)	(1.2958)	(1.2476)	(–)	(0.4583)
	IQR	[0.7134]	[0.9548]	[–]	[0.4905]	[2.2109]	[1.6169]	[–]	[0.5894]
	RNE	1.02	1.03	–	3.73	1.65	1.65	–	10.11

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 3.3: Results for the 99% VaR and ES, in the **GARCH(1,1)-t** model, based on $N = 10,000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

$H = 250$ the QERMit method is almost 2.5 times faster in estimating the 99% VaR with such a reasonable precision and requires over 4 times fewer draws to achieve that than the direct approach. For the ES the relative gain is even higher as QERMit turns out to be more than 5 times faster and nearly 8 times less draw-requiring than the naive direct approach. Notice that demanding a higher confidence on the precision would make QERMit even more advantageous relative to the direct approach.

	Direct		QERMit		Direct		QERMit	
H	Naive	Adapted	MitISEM	PMitISEM	Naive	Adapted	MitISEM	PMitISEM
	Total time							
10	13.89 s	98.65 s	127.00 s	218.56 s				
20	13.78 s	98.57 s	270.51 s	150.68 s				
40	13.96 s	98.61 s	–	328.77 s				
100	13.95 s	98.93 s	–	544.84 s				
250	14.09 s	99.20 s	–	1193.52 s				
	Construction time				Sampling time			
10	0.88 s	85.14 s	113.56 s	205.31 s	13.01 s	13.52 s	13.44 s	13.26 s
20	0.88 s	85.01 s	257.03 s	136.29 s	12.91 s	13.56 s	13.48 s	14.39 s
40	0.91 s	85.01 s	–	314.87 s	13.05 s	13.60 s	–	13.90 s
100	0.87 s	85.16 s	–	530.03 s	13.08 s	13.77 s	–	14.81 s
250	0.87 s	85.30 s	–	1176.81 s	13.22 s	13.90 s	–	16.72 s
	VaR slope*				**ES slope***			
10	**2.28**	**2.42**	**26.34**	**105.78**	**1.42**	**2.00**	**5.23**	**10.74**
20	**0.92**	**1.59**	**3.51**	**17.75**	**0.50**	**0.62**	**2.62**	**5.34**
40	**0.62**	**0.77**	–	**6.92**	**0.23**	**0.25**	–	**1.81**
100	**0.18**	**0.45**	–	**1.14**	**0.12**	**0.12**	–	**0.69**
250	**0.17**	**0.11**	–	**0.56**	**0.05**	**0.05**	–	**0.28**
	VaR time required**				ES time required**			
10	674.42 s	719.98 s	171.89 s	219.84 s	1,085.54 s	852.25 s	407.13 s	348.36 s
20	1,677.00 s	1,049.22 s	694.79 s	222.84 s	3,053.12 s	2,544.50 s	843.71 s	423.86 s
40	2,485.72 s	2,085.40 s	–	536.97 s	6,762.26 s	6,338.28 s	–	1,161.81 s
100	8,486.05 s	3,480.60 s	–	1,877.18 s	13,299.21 s	12,637.37 s	–	2,743.92 s
250	9,220.75 s	13,640.17 s	–	3,917.09 s	34,108.73 s	33,336.37 s	–	6,573.53 s
	VaR draws required**				ES draws required**			
10	517,761	469,580	43,392	10,959	833,801	567,412	218,386	107,921
20	1,298,426	711,093	324,786	60,162	2,364,446	1,813,836	435,278	199,891
40	1,903,737	1,470,764	–	159,768	5,180,194	4,597,643	–	609,227
100	6,486,508	2,465,138	–	909,605	10,165,937	9,113,098	–	1,494,828
250	6,975,069	9,750,193	–	1,639,192	25,803,439	23,917,916	–	3,228,229

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

*Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 3.4: Trade-off of precision versus computing time for the 99% VaR and ES in **GARCH(1,1)-t** model for different horizons.

3.2 GAS(1,1)-t

Having referred to the benchmark GARCH application of Hoogerheide and van Dijk (2010), in our second illustration we consider a more recently developed model for more recent data. Creal et al. (2013) propose an alternative approach to modelling volatility based on the updating of the time-varying parameter with the scaled score of the observation's contribution to the likelihood function. We employ their Generalised Autoregressive Score (GAS) model to the daily logreturns of the S&P 500, from January 3, 2005 to June 30, 2016 (2893 observations, Figure 3.3) to forecast the 99% VaR and ES at the same horizons as in the previous section[9]. The data span over the 2008 financial crisis resulting in very high sample kurtosis, so that one would expect potential difficulties in obtaining precise risk forecasts.

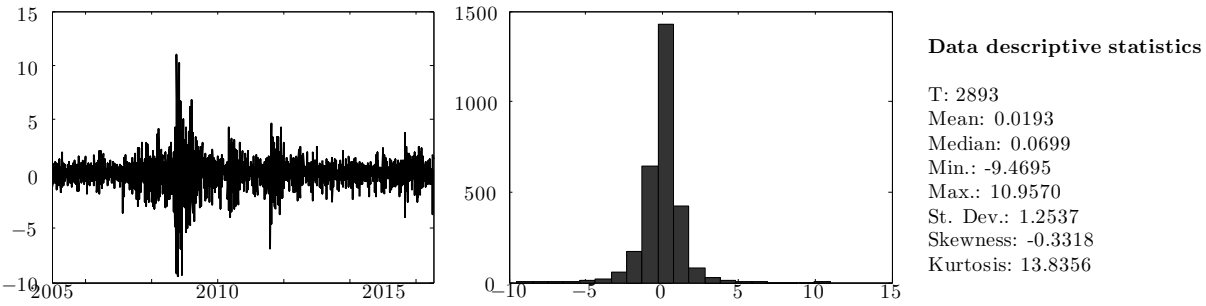


Figure 3.3: The series including the 2008 Financial Crisis: the daily logreturns of the S&P 500, from January 3, 2005 to June 30, 2016.

We adopt the following basic specification of the GAS model, referred to as GAS(1,1)-t,

$$
\begin{aligned}
y_t &= \mu + \sqrt{\rho h_t}\varepsilon_t, \\
\varepsilon_t &\sim t(\nu), \\
\rho &:= \frac{\nu-2}{\nu}, \\
h_t &= \omega + A\frac{\nu+3}{\nu}\Big(C_{t-1}(y_t-\mu)^2 - h_{t-1}\Big) + Bh_{t-1}, \\
C_t &= \frac{\nu+1}{\nu-2}\left(1+\frac{(y_{t-1}-\mu)^2}{(\nu-2)h_{t-1}}\right)^{-1},
\end{aligned}
$$

where we stack the model parameters into vector $\theta = (\mu, \omega, A, B, \nu)^T$. Finally, we put flat

[9]We also considered "complimentary" applications, i.e. employing the GAS model to the "old" dataset, as well as running the GARCH model on the "crisis" series. The former application performed better than the originally analysed model, yielding even more noticeable efficiency gains than those reported in Section 3.1. Regarding the latter, the GAS model as expected, provided a much better framework for modelling extreme returns present in the crisis data compared to the GARCH model, which is a result also reported by Jelsma and Lasak (2016).

priors on μ, ω, A and B, with $\omega > 0$ and $B \in (0, 1)$ to guarantee that the conditional variance is positive and to ensure covariance stationarity, and uninformative exponential prior on ν, $\nu - 2 \sim \text{Exp}(0.01)$.

Table 3.5: Estimation results in the **GAS(1,1)-t** model for Maximum Likelihood (ML) method (reported for comparison) and the Bayesian direct approach with naive (Student's t) and adapted (MitISEM mixture of Student's t) candidate distributions in the independence chain Metropolis-Hastings (MH) method: estimated posterior mean and standard deviation (SD), inefficiency factor (IF), acceptance rate (AR) in the MH method, and computing times for construction of the candidate distribution and for performing the direct approach.

	ML		MH (naive candidate)			MH (adapted candidate)		
Parameter	MLE	SD	Mean	SD	IF	Mean	SD	IF
μ	0.0702	0.0141	0.0738	0.0140	4.8736	0.0739	0.0140	3.6611
ω	0.0219	0.0050	0.0222	0.0048	4.9919	0.0221	0.0048	3.6370
A	0.0996	0.0111	0.1026	0.0111	4.7300	0.1022	0.0110	3.6902
B	0.9817	0.0061	0.9818	0.0059	4.6926	0.9819	0.0059	3.7480
ν	6.8979	1.0376	7.0853	1.0256	5.0386	7.0762	1.0163	3.7607
		AR		0.5547			0.7776	
	Time construction			0.98 s			108.83 s	
	Time sampling			17.24 s			17.80 s	
	No. of draws			10,000			10,000	

Table 3.5 presents the simulation results for the two direct approaches. This time, due to a bit longer series and a more complex volatility update formula, the adaptation of the direct candidate takes slightly more than 1.5 minutes. However, the resulting AR is much higher than in the previous application, reaching nearly 78%; it also exceeds the one obtained with the naive candidate, which somewhat exceeds 55%. The superiority of the adapted candidate is also reflected in lower IF values for all the parameters. Notice that the degrees of freedom for the observation disturbances ν are estimated at a lower level than in the previous application (around 7 compared to roughly 10 before), which corresponds to a much more volatile nature of the current dataset.

Table 3.6 presents the properties of the partial mixture generated by PMitISEM for the 10-day-ahead case. Given an uneasy character of the current time series it is interesting to notice that with the GAS model a lower number of mixture components was required by the PMitISEM algorithm to approximate the tails, compared to the previous application. Now two or three components are sufficient while with the GARCH model as many as four to five components were necessary – and this was the case for much more regular data. Again, the last column presents decreasing values of the regression coefficient (at

Subset	Parameters	No. of components	Weighted* μ or β*
1	$\{(\theta, \varepsilon_1)\}$	1	[0.0731 0.0225 0.1045 0.9823 7.0176 -1.0027]
2	$\{\varepsilon_2\}$	2	[-1.1023 -0.0975]
3	$\{\varepsilon_3\}$	2	[-1.1874 -0.0887]
4	$\{\varepsilon_4\}$	2	[-1.2748 -0.0966]
5	$\{\varepsilon_5\}$	2	[-1.4993 -0.1150]
6	$\{\varepsilon_6\}$	1	[-1.6575 -0.1231]
7	$\{\varepsilon_7\}$	2	[-1.9947 -0.1557]
8	$\{\varepsilon_8\}$	3	[-2.3280 -0.1755]
9	$\{\varepsilon_9\}$	3	[-2.9323 -0.2234]
10	$\{\varepsilon_{10}\}$	3	[-4.8901 -0.3954]

*Weighted with the mixture weights.

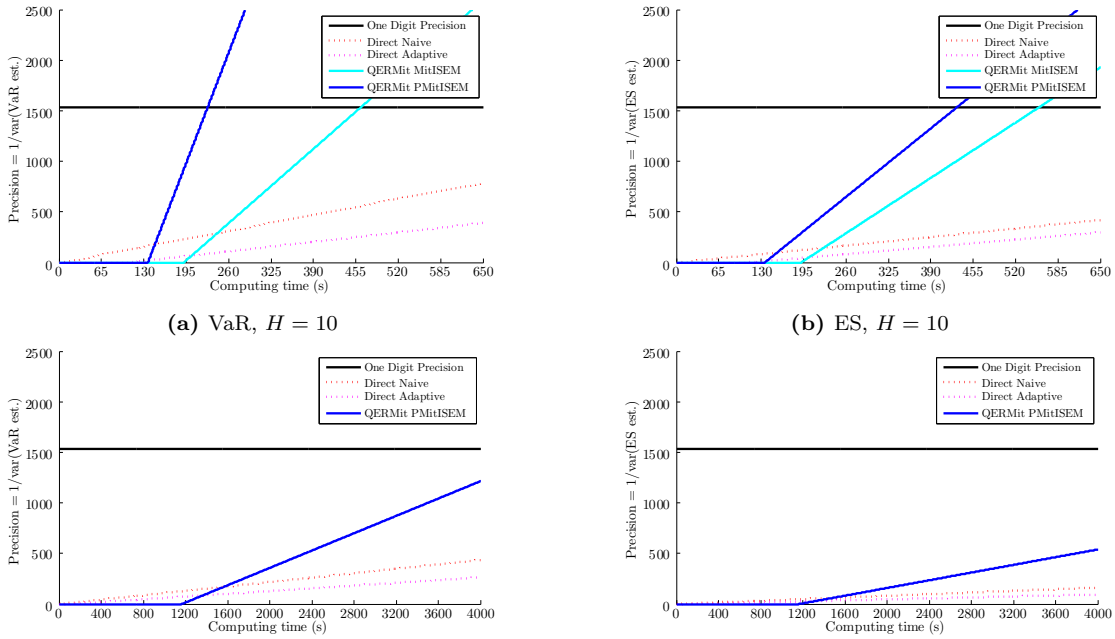
**The mode μ (for subset 1) or the regression coefficients β (for the other subsets).

Table 3.6: Properties of the marginal and conditional importance densities from the PMitISEM method for $H = 10$ in the **GAS(1,1)-t** model.

the cumulative return up to period $s-1$) used to determine the modes of subsequent conditional mixtures, exhibiting the process of "guiding" of the draws to the tail by PMitISEM.

Table 3.7 reveals that also this time we observe substantial accuracy gains for our proposed methods for all horizons, for both the 99% VaR and ES (for the corresponding visualisation we refer to the Online Appendix). For the VaR evaluations at $H = 10, 20, 40$ the NSE is around four times smaller, while for $H = 100$ and $H = 250$ it is roughly 2.5 times smaller. Again, the ES turns out to be somewhat harder to precisely estimate than the VaR, yet also in this case we report considerable gains. For $H \leq 40$ the computed NSEs are around three times lower with the PMitISEM based QERMit than with the direct approaches, while for the two longest horizons they diminish more than twice. Broadly speaking, a similar pattern pertains to the computed IQRs.

Finally, the most important results on time-precision trade-off are provided in Table 3.8 with the plots corresponding to the shortest and longest horizon presented in Figure 3.4 (the Online Appendix provides plots for all the horizons). For all horizons, for both the VaR and the ES, the slopes obtained with the PMitISEM algorithm are much higher



(a) VaR, $H = 10$

(b) ES, $H = 10$

(c) VaR, $H = 250$

(d) ES, $H = 250$

Figure 3.4: Precision ($1/var$) of the predicted VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GAS(1,1)-t** model, for the shortest and the longest horizon. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based importance density corresponds to a situation when it was not possible to construct such an importance density.

than in the case of the direct approach, often by more than one order of magnitude. Also basic MitISEM outperforms the direct approaches, but it is clearly inferior to PMitISEM. Eventually PMitISEM requires less time (and fewer draws) to achieve the same precision as the direct approaches. For instance, when the 1 digit precision with 95% confidence is considered, to accurately evaluate the 99% VaR and ES, the PMitISEM based QERMit needs, respectively, almost 3 and over 4 times less time than the naive direct approach (which outperforms the adaptive direct method).

H		VaR_{naive}	VaR_{adapt}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{adapt}	ES_{mit}	ES_{pmit}
10		-9.4284	-9.4076	-9.4290	-9.4358	-11.5862	-11.4901	-11.6038	-11.5870
	NSE	(0.2183)	(0.2793)	0.1040	(0.0601)	(0.2988)	(0.3205)	(0.1205)	(0.1078)
	IQR	[0.2697]	[0.3666]	[0.1865]	[0.0891]	[0.4290]	[0.5576]	[0.1183]	[0.1505]
	RNE	1.02	1.03	5.0467	9.92	1.67	1.59	8.82	26.39
20		-12.5332	-12.6962	-12.6807	-12.6483	-15.5819	-15.6293	-15.7741	-15.6556
	NSE	(0.3039)	(0.3145)	0.1569	(0.0686)	(0.4837)	(0.4070)	(0.3280)	(0.1310)
	IQR	[0.5002]	[0.3988]	[0.2253]	[0.1264]	[0.6433]	[0.5856]	[0.3339]	[0.1832]
	RNE	1.01	1.03	2.4818	8.43	1.65	1.60	4.62	24.17
40		-16.4218	-16.4804	–	-16.4626	-20.7435	-20.8218	–	-20.8775
	NSE	(0.3907)	(0.3582)	(–)	(0.0907)	(0.7497)	(0.5630)	(–)	(0.2182)
	IQR	[0.3910]	[0.5375]	[–]	[0.1363]	[0.7104]	[0.8472]	[–]	[0.2692]
	RNE	1.00	1.01	–	7.67	1.76	1.65	–	30.58
100		-21.7043	-21.7532	–	-21.6031	-28.3618	-28.6295	–	-28.4508
	NSE	(0.4918)	(0.5725)	(–)	(0.2135)	(1.1395)	(0.7797)	(–)	(0.4737)
	IQR	[0.6407]	[0.8066]	[–]	[0.3593]	[2.0389]	[1.1382]	[–]	[0.3251]
	RNE	1.04	1.02	–	5.73	1.71	1.69	–	14.57
250		-25.2962	-25.4476	–	-25.1630	-34.9541	-34.4421	–	-34.3317
	NSE	(0.7228)	(0.9014)	(–)	(0.3332)	(1.1825)	(1.5043)	(–)	(0.4997)
	IQR	[1.0707]	[1.2386]	[–]	[0.5608]	[1.8279]	[1.4069]	[–]	[0.5731]
	RNE	1.02	1.01	–	4.41	1.71	1.71	–	3.10

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 3.7: Results for the 99% VaR and ES, in the **GAS(1,1)-t** model, based on $N = 10000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

	Direct		QERMit		Direct		QERMit	
H	Naive	Adapted	MitISEM	PMitISEM	Naive	Adapted	MitISEM	PMitISEM
	Total time							
10	21.64 s	126.74 s	208.47 s	152.90 s				
20	21.52 s	126.74 s	189.78 s	164.66 s				
40	21.57 s	126.79 s	–	186.77 s				
100	21.65 s	126.86 s	–	297.01 s				
250	21.87 s	127.13 s	–	1191.88 s				
	Construction time				Sampling time			
10	4.41 s	108.94 s	192.11 s	136.38 s	17.23 s	17.81 s	16.36 s	16.52 s
20	4.28 s	108.92 s	171.95 s	147.82 s	17.24 s	17.82 s	17.83 s	16.83 s
40	4.29 s	108.94 s	–	169.85 s	17.28 s	17.85 s	–	16.92 s
100	4.29 s	108.91 s	–	279.05 s	17.36 s	17.94 s	–	17.96 s
250	4.30 s	108.96 s	–	1170.85 s	17.57 s	18.17 s	–	21.02 s
	VaR slope*				**ES slope***			
10	**1.22**	**0.72**	**5.66**	**16.76**	**0.65**	**0.55**	**4.21**	**5.21**
20	**0.63**	**0.57**	**2.28**	**12.61**	**0.25**	**0.34**	**0.52**	**3.46**
40	**0.38**	**0.44**	–	**7.18**	**0.10**	**0.18**	–	**1.24**
100	**0.24**	**0.17**	–	**1.22**	**0.04**	**0.09**	–	**0.25**
250	**0.11**	**0.07**	–	**0.43**	**0.04**	**0.02**	–	**0.19**
	VaR time required**				ES time required**			
10	1,266.27 s	2,243.21 s	463.78 s	228.07 s	2,368.24 s	2,919.50 s	557.28 s	431.26 s
20	2,450.42 s	2,817.30 s	846.93 s	269.72 s	6,201.54 s	4,643.55 s	3,120.15 s	591.50 s
40	4,057.15 s	3,628.94 s	–	383.75 s	14,926.56 s	8,803.87 s	–	1,408.28 s
100	6,455.40 s	9,146.02 s	–	1,537.72 s	34,635.95 s	16,871.56 s	–	6,473.64 s
250	14,112.09 s	22,790.17 s	–	4,756.85 s	37,756.56 s	63,277.38 s	–	9,238.90 s
	VaR draws required**				ES draws required**			
10	732,326	1,198,603	166,065	55,495	1,371,858	1,578,403	223,218	178,472
20	1,419,270	1,519,981	378,493	72,408	3,595,706	2,544,903	1,653,200	263,554
40	2,345,620	1,971,873	–	126,410	8,636,375	4,870,819	–	731,891
100	3,716,418	5,036,326	–	700,689	19,950,925	9,341,719	–	3,448,462
250	8,028,977	12,485,292	–	1,705,594	21,485,424	34,772,225	–	3,837,372

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

*Slope =Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 3.8: Trade-off of precision versus computing time for the 99% VaR and ES in **GARCH(1,1)-t** model for different horizons.

4 Empirical illustrations of frequentist QERMit

The theoretical reason for the 50%-50% formula (2.4) was that in Bayesian analysis usually only a kernel of the target density is available, i.e. the normalising constant for the posterior density is unknown, so that one needs to normalise the importance weights by their sum. If the target was known there would be no need to normalise the importance weights and the optimal sampling density would put all the probability mass into the region of interest. This is typically the case in frequentist inference, where we only need to simulate the future returns (or future innovations) of which we know the exact density (including the scaling constant) given the parameter vector θ. Let $p(\varepsilon^*)$ denote the target density of future disturbances, $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_H^*)$, and suppose that the vector of model parameters θ is fixed (this can be seen as either the "true" model parameters being known or the MLE being available). Then the optimal importance density is a function only of ε^*, given θ, and it is constructed solely over the tail.

In general, the optimal candidate density for estimation of $\mathbb{E}_p[g(X)]$ is given by

$$q_{opt}(x) = C|g(x)|p(x)$$

with the normalising constant $C = 1/\mathbb{E}_p[|g(X)|]$ (see Kahn and Marshall, 1953). In the case of estimating probability $\bar{p}$ of an event S we have $g(x) = \mathbb{1}_S(x)$, hence

$$q_{opt}(x) = \mathbb{1}_S(x)p(x)/\bar{p},$$

so it is a density proportional to the target over the set S. Then

$$\begin{aligned}\mathbb{E}_p[\mathbb{1}_S(X)] &= \int_S p(x)dx \\ &= \int_S \frac{p(x)}{q_{opt}(x)} q_{opt}(x)dx \\ &= \mathbb{E}_{q_{opt}}[\mathbb{1}_S(X)w(X)],\end{aligned}$$

where $w(x) = p(x)/q_{opt}(x)$, and its *unbiased* and consistent MC estimator is then given

by

$$\widehat{\mathbb{E}_p[\mathbb{1}_S(X)]} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_S(x^{(i)})w(x^{(i)}),$$

where $x^{(1)}, \ldots, x^{(N)}$ are i.i.d. draws from q_{opt}. Notice that using q_{opt} results in *zero-variance IS*, as $\frac{p(x)}{q_{opt}(x)}\mathbb{1}_S(x)$ is constant (equal to $\mathbb{E}_p[g(X)]$). Hence, in the case when the target density is known, there is *no limit* on the potential relative gain in precision from using IS rather than direct simulation[10], which means that the RNE can be unbounded. In practice, however, the problem is that to implement sampling based on q_{opt}, one needs to know $\bar{p}$, which obviously is infeasible as the evaluation of $\bar{p}$ is the goal of the undertaken analysis in the first place. In the context of risk forecasting, using the previously introduced notation, we would need to know the $100\alpha\%$ VaR. Hence, similarly as in the Bayesian case, we can approximate q_{opt} based on some preliminary value VaR_{prelim} obtained with the direct approach. To this end we again use a mixture of Student's t distributions delivered by MitISEM.

As already noted in the introduction, the advantages of IS as a variance reduction technique in the frequentist case have already been noticed in the literature. Glasserman et al. (1999) and Glasserman et al. (2000) combine IS with stratified sampling to obtain precise estimates of VaR, while Glasserman et al. (2002) extend their analysis to also include ES. They specify an importance density based on a quadratic "delta-gamma" approximation to the change in portfolio value. They, however, do not consider time series models and do not carry out an empirical study on the real data, which are of key interest to us. Hence, we do not consider their approach in our research, although some insights from those studies might be useful in further research.

4.1 GARCH(1,1)-t

Below we discuss the frequentist counterpart of the Bayesian analysis of the GARCH(1,1)-t model from Section 3.1. We fix the model parameters to their MLE values (reported in Table 3.1), compute the corresponding volatility for the last in-sample time period

[10] see Hoogerheide and van Dijk (2010), who derive the limit of the potential relative gain in precision from using IS rather than direct simulation for VaR evaluation in the Bayesian context, which is equal to $(4(1-\bar{p})\bar{p})^{-1}$. This is 25.25 for $\bar{p} = 0.01$, the case of the 99% VaR. Note that the relative precision gain may be higher for the ES.

and simulate only the i.i.d. future disturbances ε_h, $h = 1, \ldots, H$. Since there is no posterior density to approximate in this case, now we have only one direct approach, where we simulate ε_h directly from the target, which in this case is the standard Student's t density with roughly 10 degrees of freedom. The QERMit approaches are based on the approximations to the tail of the target, i.e. the tail of the predictive density.

Table 4.1 shows that also in the frequentist case we achieve noticeable improvements in the accuracy of the VaR and ES evaluations for all horizons (the corresponding plots are provided in the Online Appendix). This time, the NSEs for the VaR are four to eight times lower when computed using the PMitISEM based QERMit than in the case of direct sampling. For the ES PMitISEM outperforms the direct approach by more than three times. Notice that the RNEs for the QERMit based methods are astonishingly high, for the VaR ranging from over 3000 for $H = 10$, to 30 at $H = 250$, and for the ES from 250 to 10, respectively. This clearly demonstrates that in the frequentist case using IS rather than the direct simulation faces no limits on the relative precision gain and is "barely" constrained by our ability to construct an accurate candidate density.

Regarding the crucial time-precision trade-off, Table 4.2 shows that once again the slopes for the QERMit-based methods are higher than these of the direct approach (the Online Appendix provides the corresponding plots for all the horizons), usually two to three times. For some horizons, however, the increase in the slope due to adopting our IS-based method is much higher (for $H = 100$ it is over 7 for the VaR and over 13 for the ES). Interestingly, for $H = 10$ the superior algorithm turns out to be basic MitISEM and not PMitISEM, which delivers a slightly lower slope for the VaR than the direct approach.

An important remark must be made on the differences in sampling times between the Bayesian and the frequentist applications. *Any* frequentist method is extremely fast compared to any Bayesian method. Considering Table 3.4 for the Bayesian case and Table 4.2 for the frequentist case reveals that in the latter the sampling is usually faster by two orders of magnitude than in the former. The obvious reason for this speed of the frequentist sampling is that each logreturn draw $y^{*(i)}$, $i = 1, \ldots, M$, is based on the common value of the parameter θ, fixed at the MLE. Therefore, not only no time is spent on drawing parameters from the posterior, but also on calculating the implied time T volatilities $h_T^{(i)}$, necessary for prediction of the future volatilities. In the frequentist case the direct sampling time consists therefore barely of drawing i.i.d. variates from the

Student's t target (i.e. $\varepsilon_1^{(i)}, \ldots, \varepsilon_H^{(i)}$) and running the H-step-ahead recursion implied by the model to obtain the final $PL(y^{*(i)})$ value. When the QERMit methods are adopted, $\varepsilon_h^{(i)}$, $h = 1, \ldots, H$ are no longer independently drawn from a univariate target, but from more complex densities with an inner dependence structure, which makes the sampling more time consuming.

The fact that the direct approach is so fast in the frequentist case results in PMitISEM-based QERMit methods requiring relatively more time to reach the benchmark 1 digit precision (with 95% confidence), even though they are characterised by higher slopes. Fortunately, for QERMit based on basic MitISEM (when it is feasible) our method requires less time than the direct approach to achieve this benchmark precision level. Interestingly, however, both QERMit methods require far fewer draws than the direct approach to estimate 99% VaR and ES with the above specified precision, which again needs to be related to the differences in sampling time. Finally, recall once again that if more precise evaluations are required or a higher confidence for the precision is considered, the *time required* would of course change in favour of the QERMit-based methods, due to their higher slopes.

4.2 GAS(1,1)-t

Finally, we turn to the frequentist analysis of the GAS(1,1)-t model applied to the highly volatile "crisis" data from Section 3.2. As in the previous frequentist application we fix the model parameters at their MLE values (reported in Table 3.5). Hence, now the future observation disturbances are drawn from the Student's t distribution with roughly 7 degrees of freedom.

Table 4.3 presents the results for the VaR and ES evaluation (see the Online Appendix for the corresponding plots). One can see that also this time the QERMit-based methods generate much more accurate forecasts. For shorter horizons the NSE for the VaR is 5 to 6 times lower when evaluated with PMitISEM based QERMit than when computed directly, while for the ES the improvement ranges from 3 to 6 times. For both the VaR and the ES, the accuracy gain for long horizons is slightly lower, but still above 3 times. The IQR follows a similar pattern to the NSE, with the relative advantage of the QERMit-based methods being greater for the VaR than for the ES, and gradually slightly diminishing

H		VaR_{naive}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{mit}	ES_{pmit}
10		-7.9417	-7.8999	-7.8988	-9.5686	-9.5494	-9.5218
	NSE	(0.1496)	0.0256	(0.0179)	(0.2284)	(0.0900)	(0.0633)
	IQR	[0.1829]	[0.0235]	[0.0284]	[0.3452]	[0.1167]	[0.0812]
	RNE	1.00	1530.93	3129.10	1.00	123.57	249.28
20		-10.7175	-10.7775	-10.7904	-13.0425	-13.1050	-13.1076
	NSE	(0.2484)	(0.0786)	(0.0421)	(0.3270)	(0.0844)	(0.0969)
	IQR	[0.3229]	[0.0771]	[0.0404]	[0.4686]	[0.0834]	[0.0761]
	RNE	1.00	161.73	565.28	1.00	140.24	106.51
40		-14.5069	-14.4811	-14.5548	-17.7166	-17.8923	-17.8710
	NSE	(0.2999)	(0.1981)	(0.0630)	(0.5337)	(0.3440)	(0.0913)
	IQR	[0.3746]	[0.3215]	[0.0565]	[0.8246]	[0.2560]	[0.0760]
	RNE	1.00	25.49	251.59	1.00	8.45	120.01
100		-20.8270	–	-20.7822	-26.2797	–	-26.1151
	NSE	(0.7637)	(–)	(0.0882)	(1.1799)	(–)	(0.0956)
	IQR	[0.7992]	[–]	[0.0881]	[1.2810]	[–]	[0.1464]
	RNE	1.00	–	128.59	1.00	–	109.42
250		-27.6190	–	-27.3962	-35.6952	–	-35.4605
	NSE	(0.7819)	(–)	(0.1804)	(1.3967)	(–)	(0.3207)
	IQR	[1.0447]	[–]	[0.2453]	[2.1390]	[–]	[0.3624]
	RNE	1.00	–	30.73	1.00	–	9.73

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 4.1: Results for the 99% VaR and ES, in the **GARCH(1,1)-t** model, based on $N = 10000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

	Direct	QERMit		Direct	QERMit	
H	Naive	MitISEM	PMitISEM	Naive	MitISEM	PMitISEM
	Total time					
10	0.90 s	1.30 s	2.15 s			
20	0.92 s	2.08 s	6.85 s			
40	0.96 s	1.17 s	23.98 s			
100	1.04 s	–	146.35 s			
250	1.27 s	–	1015.70 s			
	Construction time			Sampling time		
10	0.88 s	1.26 s	2.01 s	0.02 s	0.04 s	0.15 s
20	0.89 s	2.04 s	6.53 s	0.03 s	0.05 s	0.33 s
40	0.90 s	1.09 s	23.35 s	0.06 s	0.08 s	0.62 s
100	0.89 s	–	144.71 s	0.15 s	–	1.63 s
250	0.90 s	–	1010.97 s	0.37 s	–	4.74 s
	VaR slope*			**ES slope***		
10	**2,464.45**	**42,196.32**	**21,071.42**	**1,056.97**	**3,405.84**	**1,678.66**
20	**490.24**	**3452.54**	**1,719.31**	**282.90**	**2,993.64**	**323.94**
40	**177.31**	**300.83**	**405.28**	**55.98**	**99.70**	**193.33**
100	**11.44**	–	**78.70**	**4.79**	–	**66.96**
250	**4.42**	–	**6.49**	**1.39**	–	**2.05**
	VaR time required**			ES time required**		
10	1.51 s	1.30 s	2.08 s	2.34 s	1.71 s	2.92 s
20	4.02 s	2.48 s	7.42 s	6.32 s	2.55 s	11.27 s
40	9.56 s	6.20 s	27.15 s	28.35 s	16.50 s	31.30 s
100	135.22 s	–	164.24 s	321.53 s	–	167.66 s
250	348.18 s	–	1,247.76 s	1,108.99 s	–	1,759.19 s
	VaR draws required**			ES draws required**		
10	343,912	10,037	4,911	801,869	124,356	61,643
20	948,367	95,010	27,184	1,643,393	109,575	144,275
40	1,381,752	602,821	61,078	4,376,720	1,818,906	128,039
100	8,962,809	–	119,498	21,394,002	–	140,435
250	9,395,216	–	500,025	29,977,657	–	1,580,009

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

*Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 4.2: Trade-off of precision versus computing time for the 99% VaR and ES in **GARCH(1,1)-t** model for different horizons.

with the length of the forecast horizon.

Regarding the time-precision trade-off, Table 4.4 shows that for all the measure-horizon combinations we obtain considerable efficiency gains from adopting tail-focused densities (the visualisation of the results can be found in the Online Appendix). This translates to fewer draws being required to achieve the 1 digit precision by the QERMit methods. Due to the specific nature of the frequentist sampling time, the PMitISEM-based QERMit in some cases requires more time for that purpose, yet with MitISEM we obtain gains also in this regards. A more demanding precision requirement would make both QERMit methods more competitive compared to the direct sampling both in terms of time and draws required.

5 Conclusions

We have proposed an efficient importance sampling based method for the Bayesian risk forecasting, given a chosen model of volatility. We focus on two standard risk measures, Value-at-Risk and Expected Shortfall. The proposed method enables an accurate forecasts even for long horizons, such as one-month or one-year-ahead. We have carried out two empirical studies for daily S&P 500 returns in different time periods, a tranquil period and a highly volatile crisis period. Both applications confirm that our method not only yields more accurate forecasts than the direct sampling approach, commonly used in practice (see The Volatility Laboratory, 2012), but also achieves this in a time efficient way, resulting in a considerable gain in terms of time-precision trade-off. This substantial extension of the applicability of importance sampling to the simulation of returns for long horizons is to be attributed to the sequential construction of the marginal and conditional importance densities, which are flexible mixtures of Student's t distributions.

The proposed method succeeds also for the frequentist applications, in terms of yielding a higher precision gain for a unit of computing time. However, due to generally very fast computations in the frequentist case, the advantage of the QERMit method relative to the direct approach depends on the required precision level or on the chosen confidence for the precision. We do stress that in the context of long run risk evaluation, Bayesian analysis provides a more natural framework due to accounting for parameter uncertainty.

H		VaR_{naive}	VaR_{mit}	VaR_{pmit}	ES_{naive}	ES_{mit}	ES_{pmit}
10		-9.3886	-9.3681	-9.3562	-11.4654	-11.4724	-11.4604
	NSE	(0.2346)	(0.0588)	(0.0346)	(0.2467)	(0.1043)	(0.0711)
	IQR	[0.2717]	[0.0677]	[0.0430]	[0.3244]	[0.1365]	[0.0797]
	RNE	1.00	288.88	835.34	1.00	91.88	198.07
20		-12.5140	-12.4591	-12.5526	-15.3418	-15.4223	-15.4870
	NSE	(0.3144)	(0.1222)	(0.0673)	(0.4080)	(0.2497)	(0.0658)
	IQR	[0.4695]	[0.1413]	[0.0562]	[0.5067]	[0.1851]	[0.0760]
	RNE	1.00	66.96	220.75	1.00	16.04	231.05
40		-16.2119	–	-16.3093	-20.4478	–	-20.4527
	NSE	(0.2933)	(–)	(0.0769)	(0.5824)	(–)	(0.0941)
	IQR	[0.4169]	[–]	[0.0991]	[0.9540]	[–]	[0.1494]
	RNE	1.00	–	169.30	1.00	–	112.83
100		-21.4720	–	-21.2891	-27.5570	–	-27.3327
	NSE	(0.6093)	(–)	(0.1695)	(0.8992)	(–)	(0.1591)
	IQR	[0.7680]	[–]	[0.1984]	[1.2750]	[–]	[0.2958]
	RNE	1.00	–	34.81	1.00	–	39.50
250		-24.3314	–	-24.2340	-32.3997	–	-32.2739
	NSE	(0.7701)	(–)	(0.2084)	(1.4013)	(–)	(0.2369)
	IQR	[0.9156]	[–]	[0.2901]	[1.5927]	[–]	[0.3614]
	RNE	1.00	–	23.02	1.00	–	17.81

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

Table 4.3: Results for the 99% VaR and ES, in the **GAS(1,1)-t** model, based on $N = 10000$ draws and 20 replications to obtain the numerical standard error (NSE) and the interquartile range (IQR). The RNE is the relative numerical efficiency, the inverse of the inefficiency factor. The results are obtained using the direct approach (with naive and adapted candidate distribution in the Metropolis-Hastings algorithm), and the QERMit method (with the basic MitISEM and PMitISEM methods), respectively.

	Direct	QERMit		Direct	QERMit	
H	Naive	MitISEM	PMitISEM	Naive	MitISEM	PMitISEM
	Total time					
10	4.35 s	0.67 s	3.12 s			
20	4.35 s	4.78 s	9.46 s			
40	4.38 s	–	23.71 s			
100	4.48 s	–	145.31 s			
250	4.85 s	–	988.06 s			
	Construction time			Sampling time		
10	4.33 s	0.64 s	2.98 s	0.02 s	0.03 s	0.14 s
20	4.31 s	4.73 s	9.16 s	0.03 s	0.05 s	0.29 s
40	4.31 s	–	23.10 s	0.06 s	–	0.61 s
100	4.32 s	–	143.79 s	0.15 s	–	1.52 s
250	4.47 s	–	983.40 s	0.38 s	–	4.66 s
	VaR slope*			**ES slope***		
10	**974.26**	**9,752.17**	**5789.43**	**880.80**	**3,101.69**	**1,372.74**
20	**293.21**	**1,389.32**	**749.19**	**174.13**	**332.71**	**784.13**
40	**182.63**	**–**	**276.93**	**46.30**	**–**	**184.56**
100	**17.45**	**–**	**22.90**	**8.01**	**–**	**25.98**
250	**4.48**	**–**	**4.94**	**1.35**	**–**	**3.82**
	VaR time required*			ES time required*		
10	5.91 s	0.80 s	3.24 s	6.08 s	1.14 s	4.09 s
20	9.55 s	5.83 s	11.22 s	13.14 s	9.35 s	11.12 s
40	12.73 s	–	28.65 s	37.50 s	–	31.43 s
100	92.40 s	–	210.89 s	196.11 s	–	202.93 s
250	347.47 s	–	1,294.69 s	1,140.28 s	–	1,385.65 s
	VaR draws required**			ES draws required**		
10	845,752	53,194	18,395	935,498	167248	77,581
20	1,519,296	229,489	69,609	2,558,257	958,281	66,507
40	1,321,478	–	90,766	5,212,265	–	136,195
100	5,705,631	–	441,394	12,423,807	–	388,979
250	9,112,650	–	667,648	30,176,110	–	862,735

Missing value (–): it was not possible to generate the particular result with the corresponding algorithm.

*Slope = increase in precision per unit of computing time.

**Required for % estimate with 1 digit of precision (with 95% confidence).

Table 4.4: Trade-off of precision versus computing time for the 99% VaR and ES in **GAS(1,1)-t** model for different horizons.

Bibliography

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

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.

Christoffersen, P. F., F. X. Diebold, and T. Schuermann (1998), "Horizon Problems and Extreme Events in Financial Risk Management." *Economic Policy Review*, 109–118.

Creal, D., S. J. Koopman, and A. Lucas. (2013), "Generalized Autoregressive Score Models with Applications." *Journal of Applied Econometrics*, 28, 777–795.

Daníelsson, J. and J. P. Zigrand (2006), "On Time-Scaling of Risk and the Square-Root-of-Time Rule." *Journal of Banking & Finance*, 30, 2701–2713.

Diebold, F. X., A. Hickman, A. Inoue, and T. Schuermann (1997), "Converting 1-Day Volatility to h-Day Volatitlity: Scaling by $\sqrt{h}$ is Worse Than You Think." Technical Report 97–34, Wharton Financial Institutions Center Working Papers.

Embrechts, P., R. Kaufmann, and P. Patie (2005), "Strategic Long-Term Financial Risks: Single Risk Factors." *Computational Optimization and Applications*, 32, 61–90.

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

Engle, R. F. (2009), "The Risk That Risk Will Change." *Journal of Investment Management*, 7, 24–28.

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

Glasserman, P., P. Heidelberger, and P. Shahabuddin (1999), *Importance Sampling and Stratification for Value-at-Risk*. IBM Thomas J. Watson Research Division.

Glasserman, P., P. Heidelberger, and P. Shahabuddin (2000), "Variance Reduction Techniques for Estimating Value-at-Risk." *Management Science*, 46, 1349–1364.

Glasserman, P., P. Heidelberger, and P. Shahabuddin (2002), "Portfolio Value-at-Risk with Heavy-Tailed Risk Factors." *Mathematical Finance*, 12, 239–269.

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, 101–120.

Hoogerheide, L. F. and H. K. van Dijk (2010), "Bayesian Forecasting of Value at Risk and Expected Shortfall using Adaptive Importance Sampling." *International Journal of Forecasting*, 26, 231–247.

Jelsma, H. and K. Lasak (2016), "Forecasting Volatility Using Long Memory Dynamics: How Effective Is the Use of a Realised Measure?" mimeo.

Kahn, H. and A. Marshall (1953), "Methods of Reducing Sample Size in Monte Carlo Computations." *Journal of the Operations Research Society of America*, 1, 263–278.

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.

McNeil, A. J., R. Frey, and P. Embrechts (2015), *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press.

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.

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

The Volatility Laboratory (2012), "V-Lab: Long Run Value at Risk Documentation." `https://vlab.stern.nyu.edu/doc/4?topic=apps`. Accessed: 30 November 2016.

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

Online Appendix for
Bayesian Risk Forecasting for Long Horizons

Agnieszka Borowska[(a,b)], Lennart Hoogerheide[(a,b)] and Siem Jan Koopman[(a,b,c)]

[(a)] Vrije Universiteit Amsterdam

[(b)] Tinbergen Institute

[(c)] CREATES, Aarhus University

January 2018

A MitISEM Algorithm

A.1 Comparison with the AdMit algorithm

To approximate the bimodal optimal target density for VaR estimation we employ the MitISEM algorithm of **?** which has a number of advantages over the AdMit algorithm of **?** used in the original QERMit method. To explain our motivation behind this change below we discuss the differences between both techniques.

First and most importantly from our perspective, the only inputs to MitISEM are draws from the importance density and their importance weights, whilst 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 target density of the parameters and future returns.

Second, 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 allows the optimization of the importance density to be performed with an EM algorithm, so that no Newton-Raphson based algorithm (such as the BFGS method) is needed.

Third, MitISEM is a "fully adaptive" algorithm, as each time a new 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.

A.2 Approximation by minimisation of Kullback-Leibler divergence

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

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

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, consisting of the mixture weights and the modes, scale matrices and degrees of freedom of each component as well as the number of mixture components H. Since the first term in (**??**) does not depend on ζ, the minimisation of (**??**) 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

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

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

A.3 EM step in MitISEM

Consider a mixture of C Student-t densities

$$q_\zeta(\theta) = \sum_{c=1}^{C} \eta_h t(\theta|\mu_c, \Sigma_c, \nu_c), \tag{A.3}$$

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 $\zeta = \{\mu_c, \Sigma_c, \nu_c, \eta_c\}_{h=1}^{H}$ 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

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

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 (**??**) in (**??**) as

$$\theta^{(i)} \sim \mathcal{N}(\mu_c, \kappa_c^{(i)}\Sigma_c), \qquad \text{if } z_c^{(i)} = 1,$$

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

$\mathbb{P}[z^{(i)} = e_c]$ of belonging to the component h is given by η_c. The scaling factor $\kappa_c^{(i)}$ follows the Inverse-Gamma distribution

$$\kappa_c^{(i)} \sim \mathcal{IG}(\nu_c/2, \nu_c/2).$$

Such a representation introduces the latent data $\tilde{\theta} = \{z_c, \kappa_c\}_{h=1}^{C}$ 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 **?**, 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

$$\begin{aligned}
\tilde{z}_c^{(i)} &\equiv \mathbb{E}\left[z_c^{(i)} \middle| \theta^{(i)}, \zeta \right] = \frac{\eta_c t(\theta^{(i)}|\mu_c, \Sigma_c, \nu_c)}{\sum_{l=1}^{H} \eta_l t(\theta^{(i)}|\mu_l, \Sigma_l, \nu_l)}, \\
\widetilde{z/\kappa_c}^{(i)} &\equiv \mathbb{E}\left[\frac{z_c^{(i)}}{\kappa_c^{(i)}} \middle| \theta^{(i)}, \zeta \right] = \tilde{z}_c^{(i)} \frac{d + \nu_c}{\rho_c^{(i)} + \nu_c}, \\
\tilde{\xi}_c^{(i)} &\equiv \mathbb{E}\left[\log \kappa_c^{(i)} \middle| \theta^{(i)}, \zeta \right] \\
&= \left[\log\left(\frac{\rho_c^{(i)} + \nu_c}{2} \right) - \psi\left(\frac{d + \nu_c}{2} \right) \right] \tilde{z}_c^{(i)} + \left[\log\left(\frac{\nu_c}{2} \right) - \psi\left(\frac{\nu_c}{2} \right) \right] (1 - \tilde{z}_c^{(i)}), \\
\tilde{\delta}_c^{(i)} &\equiv \mathbb{E}\left[\frac{1}{\kappa_c^{(i)}} \middle| \theta^{(i)}, \zeta \right] = \widetilde{z/\kappa_c}^{(i)} + (1 - \tilde{z}_c^{(i)}),
\end{aligned}$$

where $\rho_c^{(i)} = (\theta^{(i)} - \mu_c)^T \Sigma_c^{-1} (\theta^{(i)} - \mu_c)$ and ψ denotes the digamma function.

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

$$\mu_c^{(L)} = \left[\sum_{i=1}^N w^{(i)} \widetilde{z/\kappa_c}^{(i)}\right]^{-1} \left[\sum_{i=1}^N w^{(i)} \widetilde{z/\kappa_c}^{(i)} \theta^{(i)}\right],$$
$$\Sigma_c^{(L)} = \frac{\sum_{i=1}^N \kappa^{(i)} \widetilde{z/\kappa_c}^{(i)} (\theta^{(i)} - \mu_c^{(L)})(\theta^{(i)} - \mu_c^{(L)})^T}{\sum_{i=1}^N w^{(i)} \tilde{z}_c^{(i)}},$$
$$\eta_c^{(L)} = \frac{\sum_{i=1}^N w^{(i)} \tilde{z}_c^{(i)}}{\sum_{i=1}^N w^{(i)}},$$

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

$$-\psi(\nu_c/2) + \log(\nu_c/2) + 1 - \frac{\sum_{i=1}^N w^{(i)} \xi_c^{(i)}}{\sum_{i=1}^N w^{(i)}} - \frac{\sum_{i=1}^N w^{(i)} \delta_c^{(i)}}{\sum_{i=1}^N w^{(i)}} = 0.$$

A more detailed discussion of the MitISEM algorithm can be found in **?**.

B PMitISEM Algorithm

Below we present the details of the Partial MitISEM algorithm of **?**.

Step 0: Initialisation

$\theta^{(i)} \sim g_{\text{naive}}$, $i = 1, \ldots, N$

$(\mu_{\text{naive}} = \hat{\theta} \equiv \arg\max_\theta f(\theta)$, $\Sigma_{\text{naive}} = -\mathcal{H}^{-1} \left(\log f(\theta)\right)|_{\theta=\hat{\theta}}$ $)$

Step 1: Adaptation

Use $\{\theta^{(i)}\}_{i=1}^N$ to IS-estimate the mean and the covariance matrix of f by μ_{adapt} and Σ_{adapt}.

Use μ_{adapt} and Σ_{adapt} to construct g_{adapt}.

Set $g_0 = g_{\text{adapt}}$.

$\theta^{(i)} \sim g_0$, $i = 1, \ldots, N$.

$w_0^{(i)} = \frac{f(\theta^{(i)})}{g_0(\theta^{(i)})}$, $i = 1, \ldots, N$.

Step 2: Construction

for $s := 1$ **to** S **do**

Step 2a: ISEM

Run ISEM with $\{\theta^{(i)}\}_{i=1}^N$ and $\{w_0^{(i)}\}_{i=1}^N$ to optimise C_s components of g_s.
($g_s(\theta) = g(\theta_s|\theta_1, \ldots, \theta_{s-1})$ for $s = 2, \ldots, S$, and $g_s(\theta) = g(\theta_1)$ for $s = 1$)
Calculate the *current* weights[1] of the draws $\{\theta^{(i)}\}_{i=1}^N$ from g_0 using the optimised candidate with C_S components with formula:

$$w_{\text{curr}}^{(i)} = \frac{f(\theta^{(i)})}{\prod_{k=1}^S g_k(\theta^{(i)})}. \tag{B.1}$$

Compute CoV_s for g_s using $\{w_{\text{curr}}^{(i)}\}_{i=1}^N$.

Step 2b: Iterate

while CoV_s not converged **do**

Find $\tilde{\theta}^j$, $j \in \mathcal{M}$, where $\mathcal{M}$ is the set of indices of $x\%$ of the draws $\{\theta^{(i)}\}_{i=1}^N$ which correspond to the highest weights $\{w_{\text{curr}}^{(i)}\}_{i=1}^N$.
Use $\{\tilde{\theta}^{(j)}\}_{j\in\mathcal{M}}$ and $\{w_{\text{curr}}^{(j)}\}_{j\in\mathcal{M}}$ to IS-construct the mode/coefficients and the covariance matrix of the $C_s + 1$-th component of g_s as $\mu_{C_s+1}^s/\beta_{C_s+1}^s$ and $\Sigma_{C_s+1}^s$.
Update the current mixture g_s[2].

Run ISEM with $\{\theta^{(i)}\}_{i=1}^N$ and $\{w_0^{(i)}\}_{i=1}^N$ to optimise C_s+1 components of the updated g_s.

Calculate the new current weights $\{w_{\text{curr}}^{(i)}\}_{i=1}^N$ of the draws $\{\theta^{(i)}\}_{i=1}^N$ from g_0 using the latest candidate with formula (**??**).
(note that now g_s is an updated mixture of $C_s + 1$ components)

Compute CoV_s for the latest candidate.

end while

end for

Step 3: Resimulation and convergence check

$\theta^i \sim \prod_{s=1}^S g(s)(\theta)$, $i = 1, \ldots, N$.
$w^{(i)} = \frac{f(\theta^{(i)})}{\prod_{s=1}^S g(s)(\theta^{(i)})}$, $i = 1, \ldots, N$.
Update g to $\tilde{g}$:

for $s := 1$ **to** S **do**

Run ISEM with $\{\theta^{(i)}\}_{i=1}^N$ and $\{w^{(i)}\}_{i=1}^N$ to optimise components of g_s to obtain $\tilde{g}_s$.

end for

Calculate the new current weights $\{w_{\text{curr}}^{(i)}\}_{i=1}^N$ of the draws $\{\theta^{(i)}\}_{i=1}^N$ from g using the optimised candidate $\tilde{g}$ with formula (**??**).
If CoV has not converged set $g_0 := \tilde{g}$ and $w_0^{(i)} := w_{\text{curr}}^{(i)}$; else STOP.

---

[1] "Current" because these are not the "real" importance weights as the draws are fixed and coming from g_0, not from the updated candidate.

[2] Updating is done in the "standard" way: μ_h^s/β_h^s, Σ_h^s and ν_h^s for the old components $h = 1, \ldots, C_s$, remain unchanged; $\mu_{C_s+1}^s/\beta_{C_s+1}^s$ and $\Sigma_{C_s+1}^s$ for the new components is are set to the current estimates based on $\{\tilde{\theta}^{(j)}\}_{j\in\mathcal{M}}$; $\nu_{C_s+1}^s$ is set to some chosen initial value; $\eta_h^s := 0.9\eta_h^s$, $h = 1, \ldots, C_s$ and $\eta_{C_s+1} := 0.1$.

C Accuracy plots

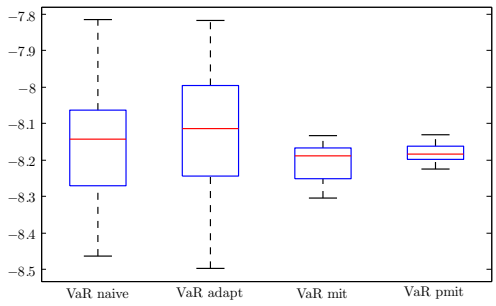
The plots in this appendix present the accuracy of 99% VaR and ES forecasts obtained with different algorithms. We consider two types of plots, standard box plots and error bar plots. The motivation behind this particular choice is that the former plot type is a popular and commonly used in dispersion visualisation, while the latter is more suited to illustrate the results based on QERMit. The underlying objective of QERMit is minimisation of the NSE, however this measure is not illustrated in a box plot, which focuses on the IQR instead. In a sense both plot types provide complimentary information regarding the accuracy of a certain evaluation method. Notice possible outliers as these might be of crucial importance in the context of risk forecasting.

C.1 Bayesian applications

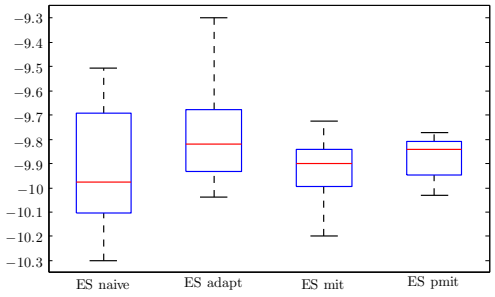
Figures **??** and **??** present the results for the GARCH(1,1)-t model, and Figures **??** and **??** for the GAS(1,1)-t (in both cases box plots and error bar plots, respectively). Clearly, the precision of the QERMit-based methods greatly exceeds the one from both direct approaches. Moreover, the latter often generate outliers, which may have serious practical consequences.

C.2 Frequentist applications

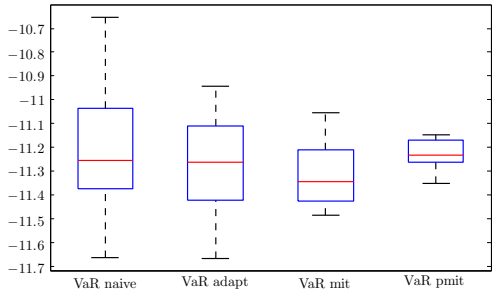
Figures **??** and **??** and Figures **??** and **??** are the frequentist counterparts of those presented in **??**. We can see that the outcomes are similar to the Bayesian ones, with a much higher accuracy achieved with QERMit than with the direct approach.



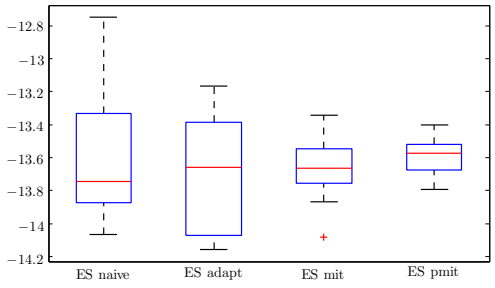
(a) VaR, $H = 10$



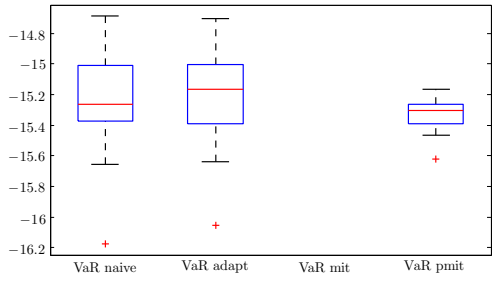
(b) ES, $H = 10$



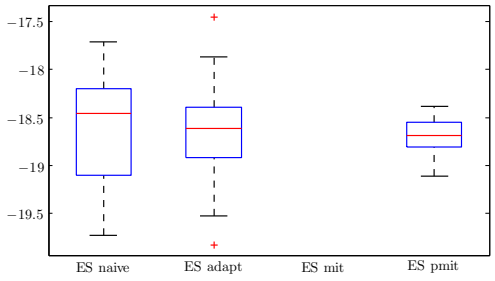
(c) VaR, $H = 20$



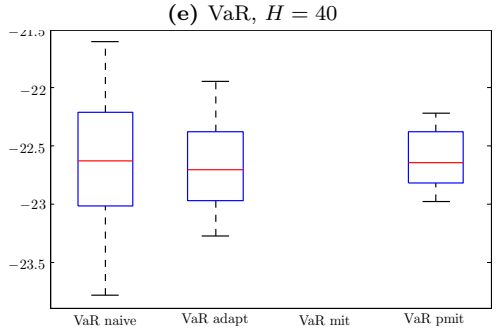
(d) ES, $H = 20$



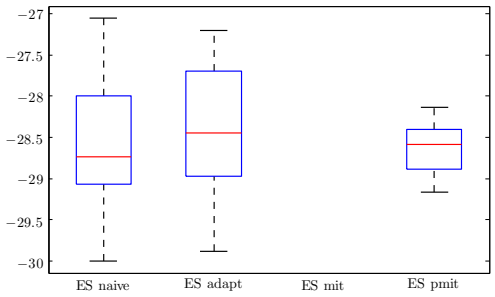
(e) VaR, $H = 40$



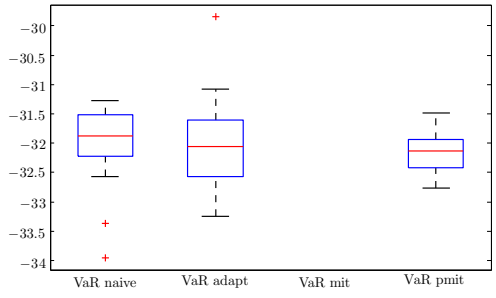
(f) ES, $H = 40$



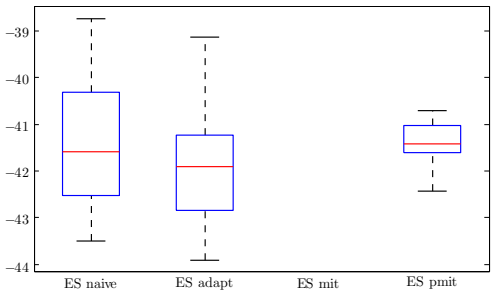
(g) VaR, $H = 100$



(h) ES, $H = 100$

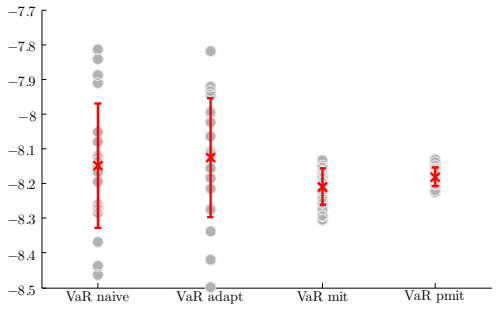


(i) VaR, $H = 250$

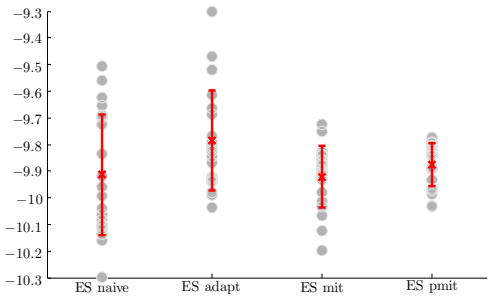


(j) ES, $H = 250$

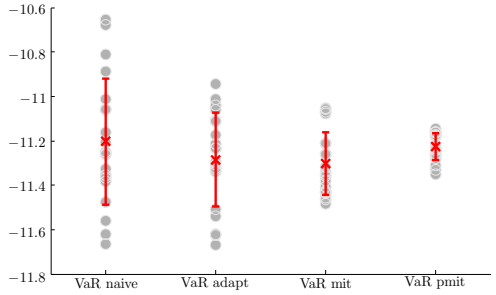
Figure C.1: Accuracy of 99% VaR (left) and ES (right) s for the **GARCH(1,1)-*t*** model for different horizons, based on 20 MC replications. Two left boxes correspond to the direct approach (based on the naive and adapted candidate, respectively), two right ones – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



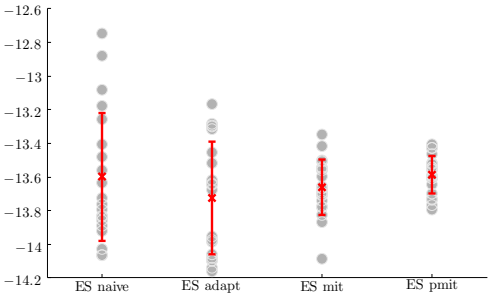
(a) VaR, $H = 10$



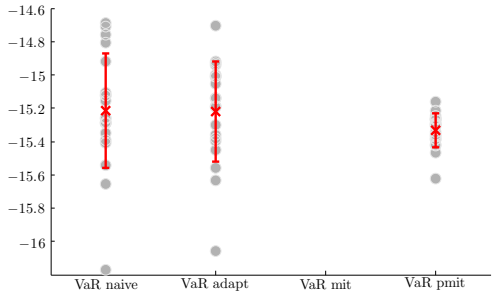
(b) ES, $H = 10$



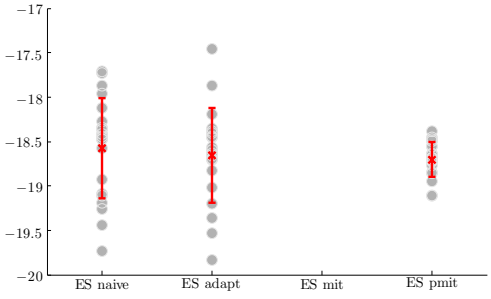
(c) VaR, $H = 20$



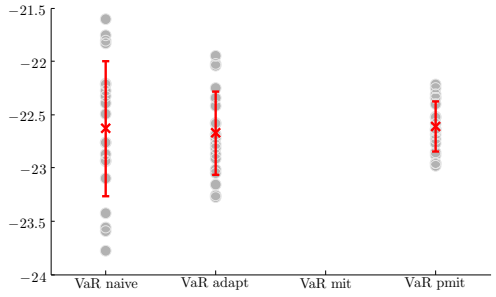
(d) ES, $H = 20$



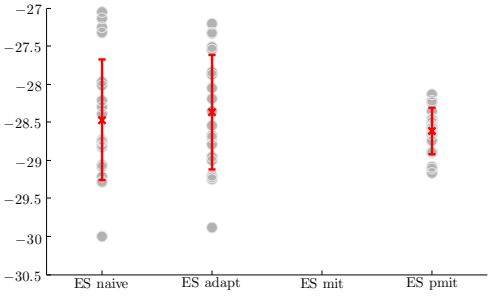
(e) VaR, $H = 40$



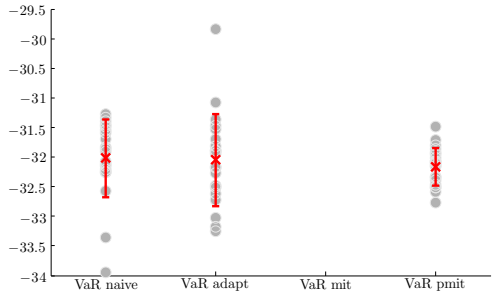
(f) ES, $H = 40$



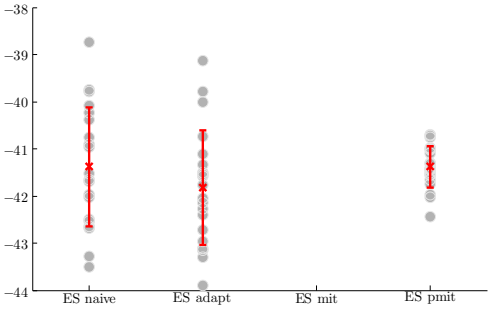
(g) VaR, $H = 100$



(h) ES, $H = 100$

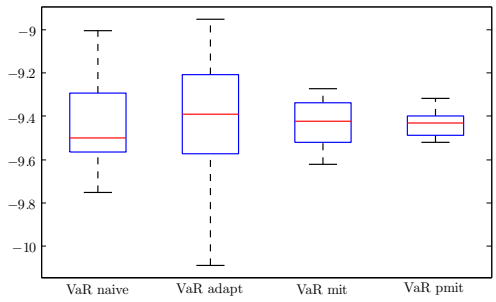


(i) VaR, $H = 250$

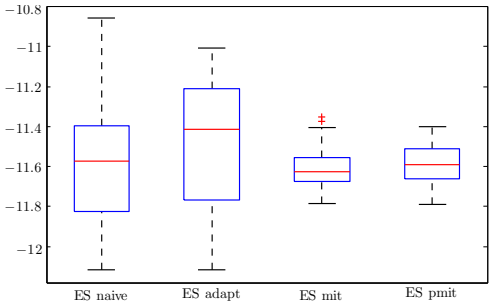


(j) ES, $H = 250$

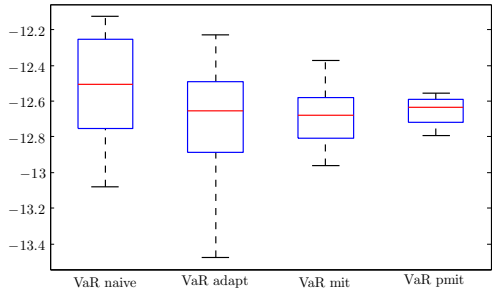
Figure C.2: Accuracy of 99% VaR (left) and ES (right) forecasts for the **GARCH(1,1)-t** model for different horizons, based on 20 MC replications. Two left error bars correspond to the direct approach (based on the naive and adapted candidate, respectively), two right ones – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing error bar for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



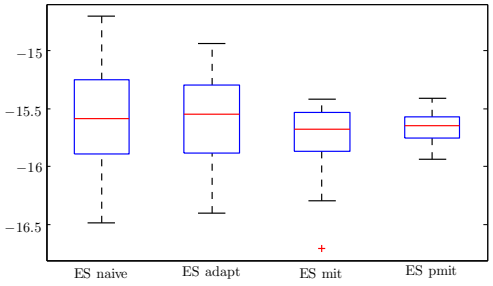
(a) VaR, $H = 10$



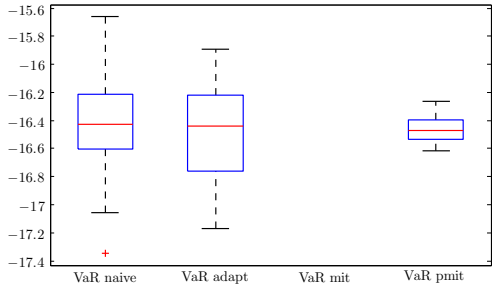
(b) ES, $H = 10$



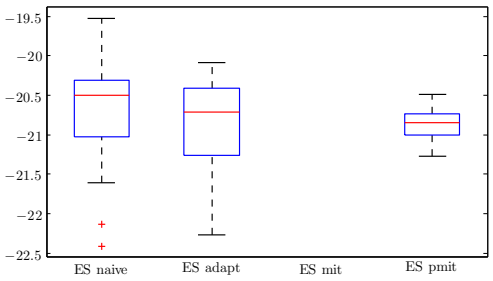
(c) VaR, $H = 20$



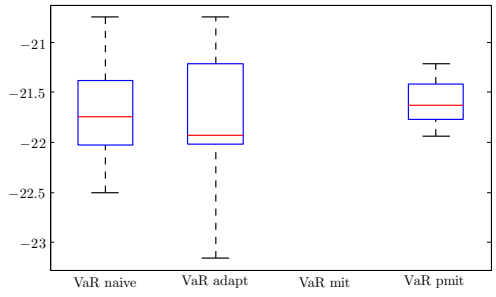
(d) ES, $H = 20$



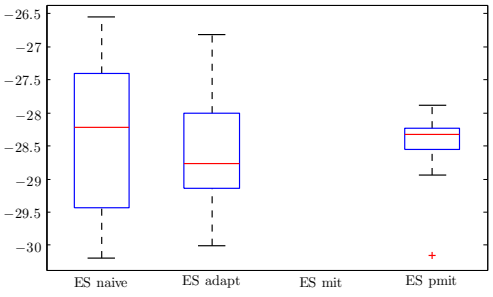
(e) VaR, $H = 40$



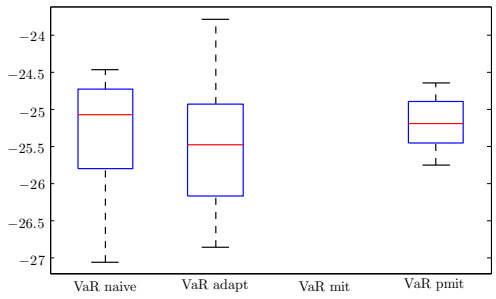
(f) ES, $H = 40$



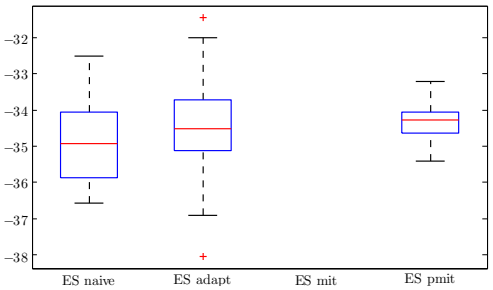
(g) VaR, $H = 100$



(h) ES, $H = 100$

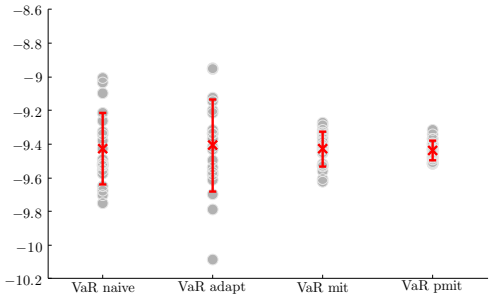


(i) VaR, $H = 250$

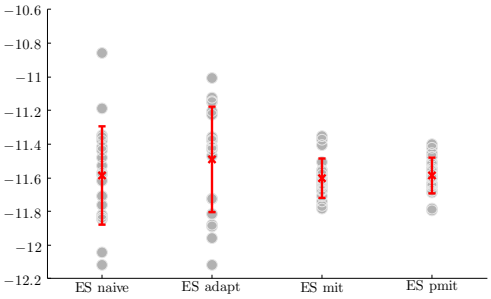


(j) ES, $H = 250$

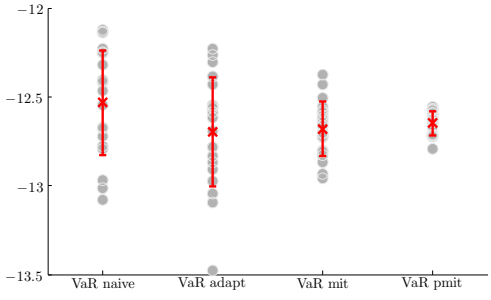
Figure C.3: Accuracy of 99% VaR (left) and ES (right) forecasts for the **GAS(1,1)-*t*** model for different horizons, based on 20 MC replications. Two left boxes correspond to the direct approach (based on the naive and adapted candidate, respectively), two right ones – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



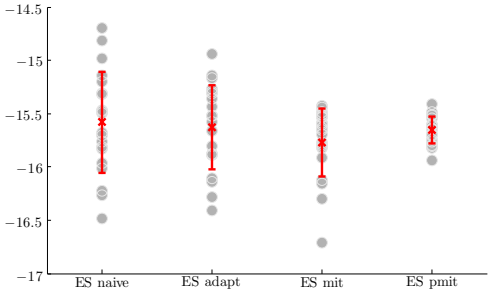
(a) VaR, $H = 10$



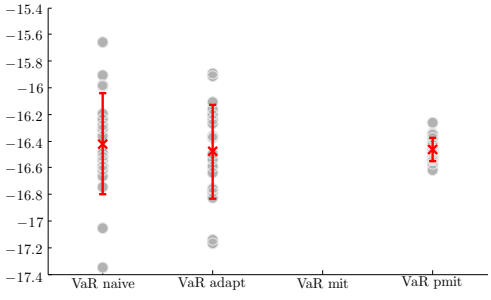
(b) ES, $H = 10$



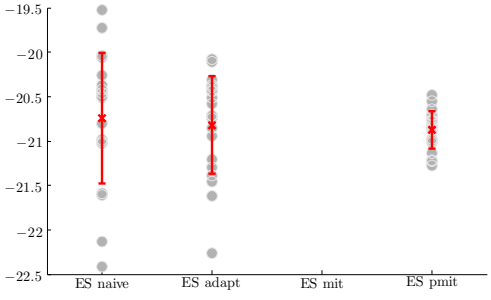
(c) VaR, $H = 20$



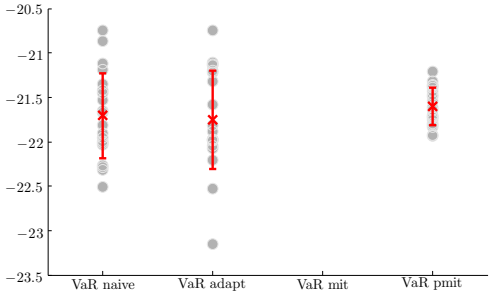
(d) ES, $H = 20$



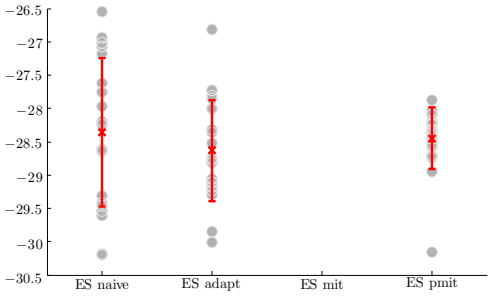
(e) VaR, $H = 40$



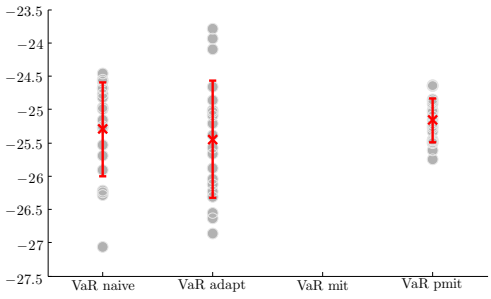
(f) ES, $H = 40$



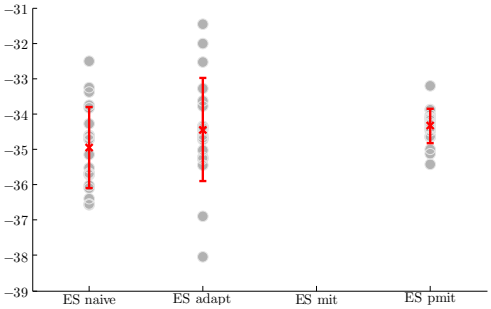
(g) VaR, $H = 100$



(h) ES,$H = 100$

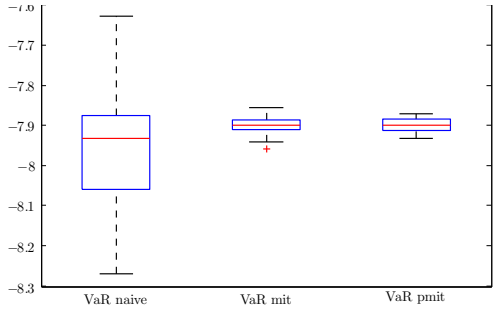


(i) VaR, $H = 250$

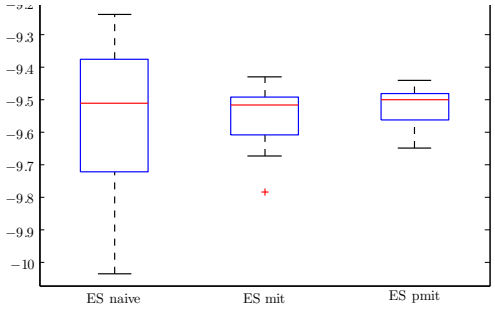


(j) ES, $H = 250$

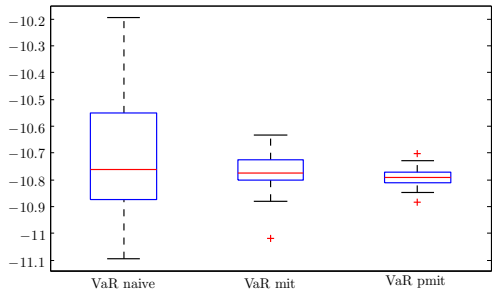
Figure C.4: Accuracy of 99% VaR (left) and ES (right) forecasts for the **GAS(1,1)-t** model for different horizons, based on 20 MC replications. Two left error bars correspond to the direct approach (based on the naive and adapted candidate, respectively), two right ones – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing error bar for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



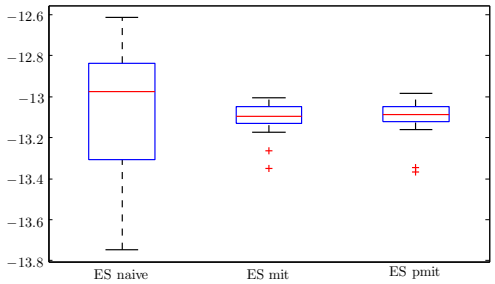
(a) VaR, $H = 10$



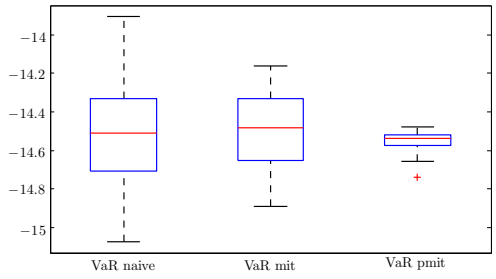
(b) ES, $H = 10$



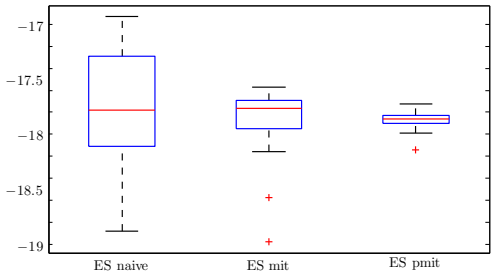
(c) VaR, $H = 20$



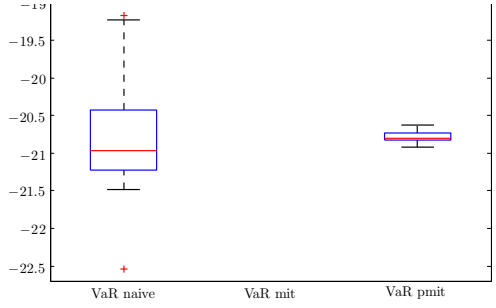
(d) ES, $H = 20$



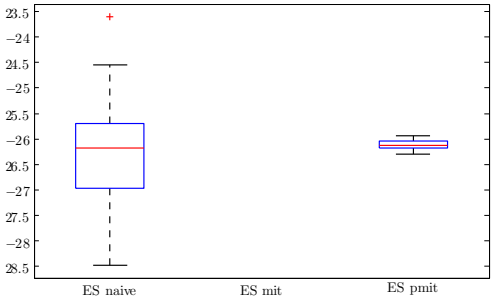
(e) VaR, $H = 40$



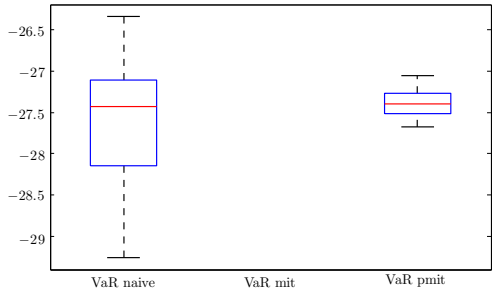
(f) ES, $H = 40$



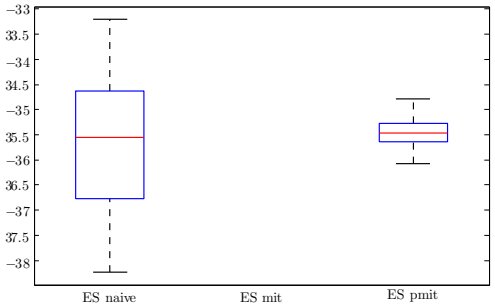
(g) VaR, $H = 100$



(h) ES, $H = 100$

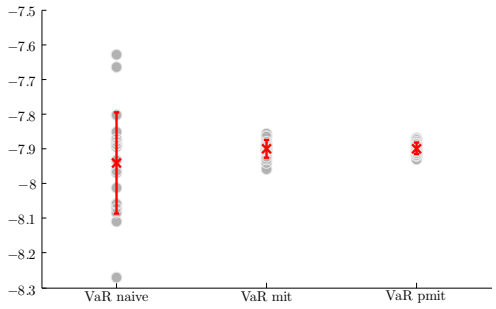


(i) VaR, $H = 250$

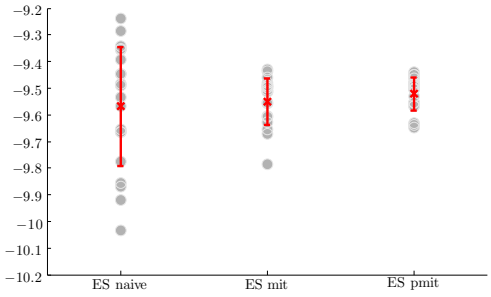


(j) ES, $H = 250$

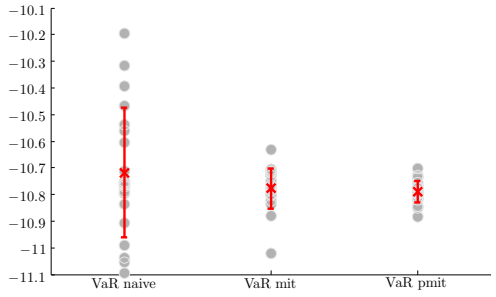
Figure C.5: Accuracy of 99% VaR (left) and ES (right) forecasts for the frequentist **GARCH(1,1)-t** model for different horizons, based on 20 MC replications. The left boxes correspond to the direct approach (based on the naive candidate), the middle and the right one – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



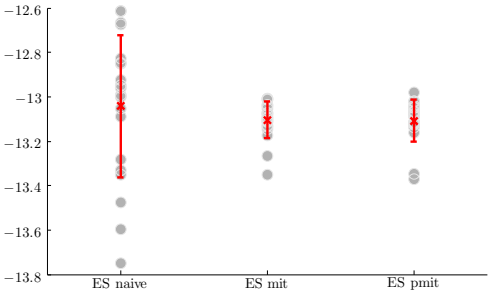
(a) VaR, $H = 10$



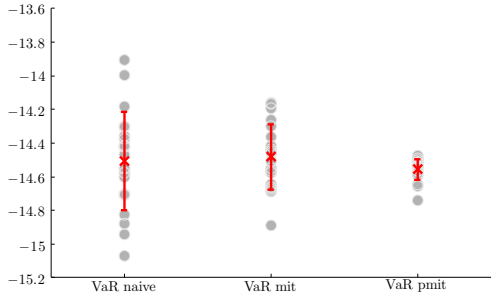
(b) ES, $H = 10$



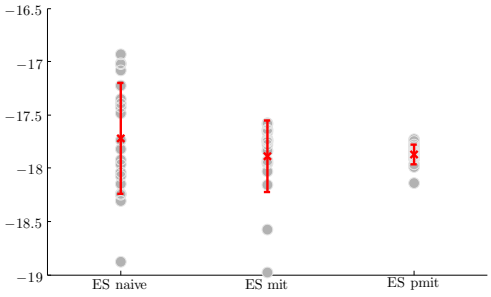
(c) VaR, $H = 20$



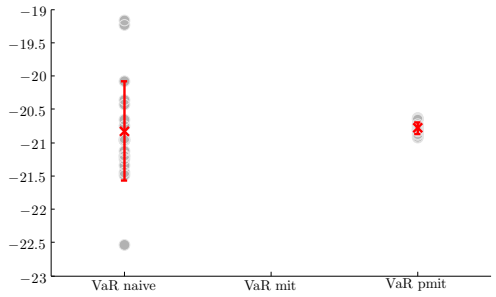
(d) ES, $H = 20$



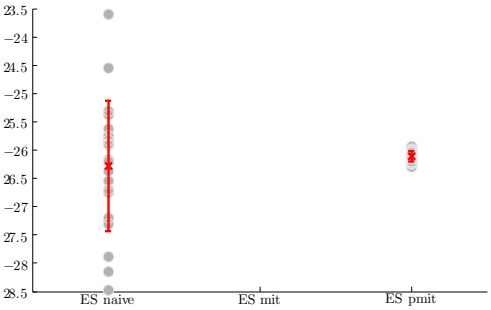
(e) VaR, $H = 40$



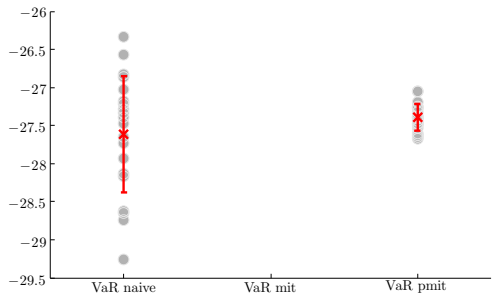
(f) ES, $H = 40$



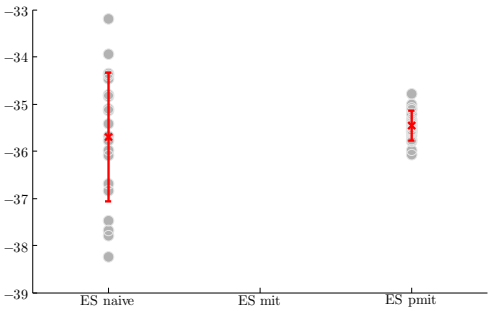
(g) VaR, $H = 100$



(h) eS, $H = 100$

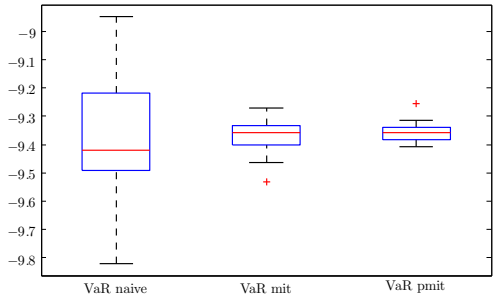


(i) VaR, $H = 250$

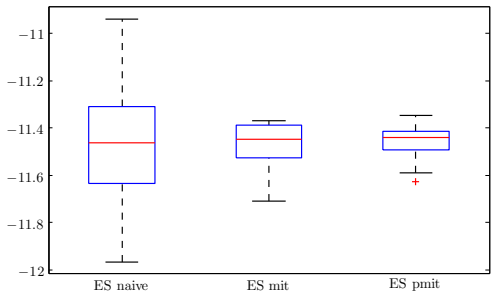


(j) ES, $H = 250$

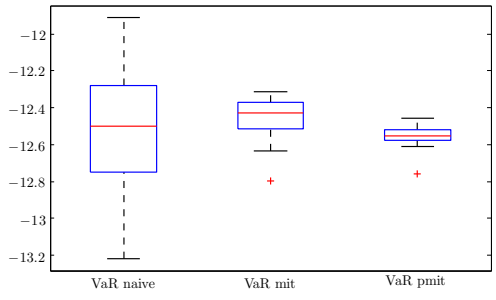
Figure C.6: Accuracy of 99% VaR (left) and ES (right) forecasts for the frequentist **GARCH(1,1)-t** model for different horizons, based on 20 MC replications. The left error bars correspond to the direct approach (based on the naive candidate), the middle and the right one – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing error bar for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



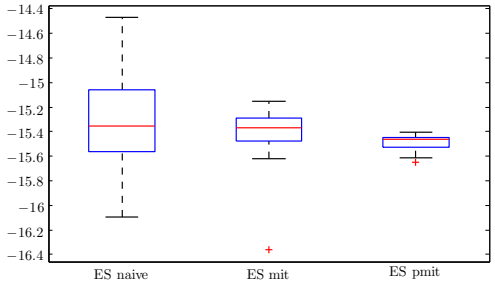
(a) VaR, $H = 10$



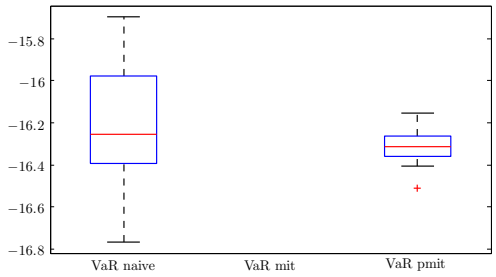
(b) ES, $H = 10$



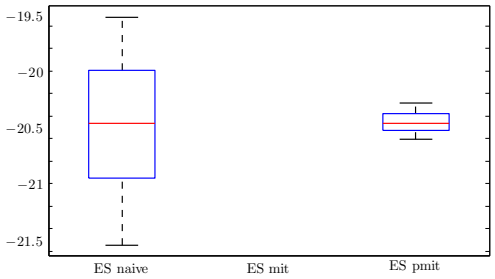
(c) VaR, $H = 20$



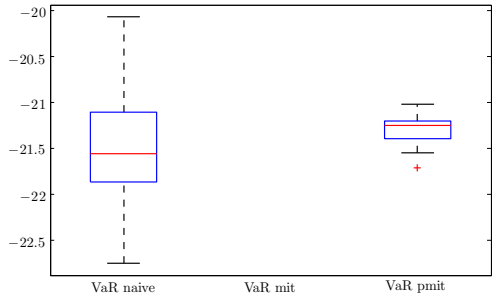
(d) ES, $H = 20$



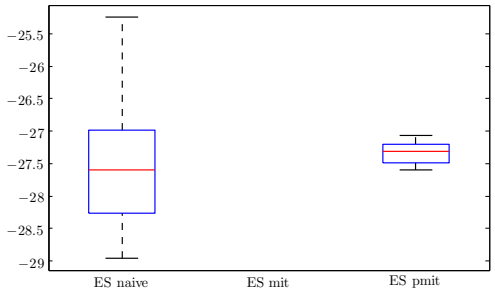
(e) VaR, $H = 40$



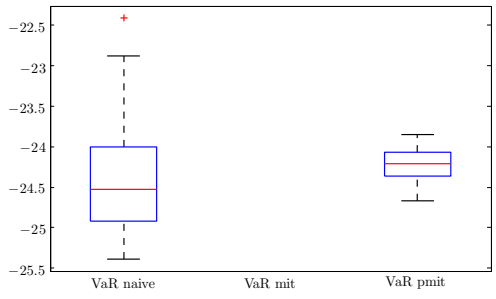
(f) ES, $H = 40$



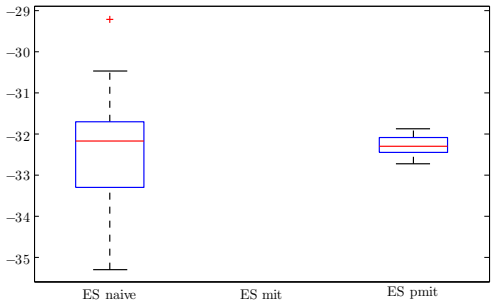
(g) VaR, $H = 100$



(h) ES, $H = 100$

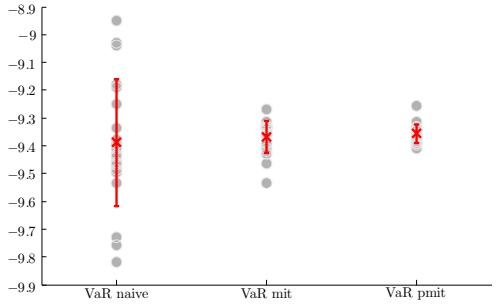


(i) VaR, $H = 250$

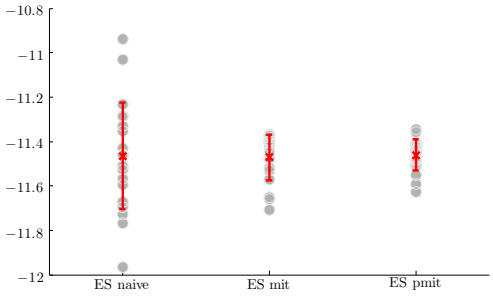


(j) ES, $H = 250$

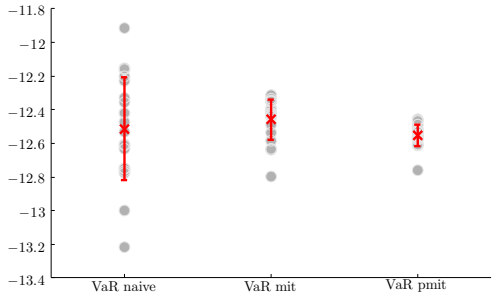
Figure C.7: Accuracy of 99% VaR (left) and ES (right) forecasts for the frequentist **GAS(1,1)-t** model for different horizons, based on 20 MC replications. The left boxes correspond to the direct approach (based on the naive candidate), the middle and the right one – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing box for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



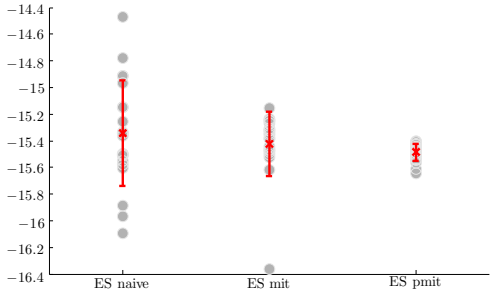
(a) VaR, $H = 10$



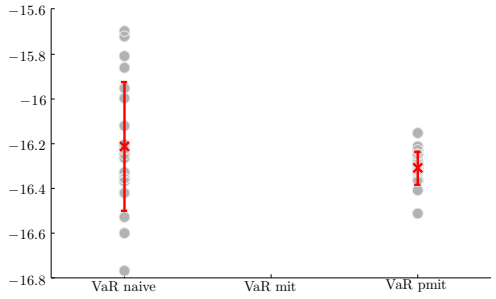
(b) ES, $H = 10$



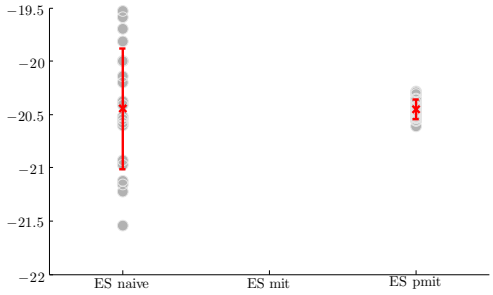
(c) VaR, $H = 20$



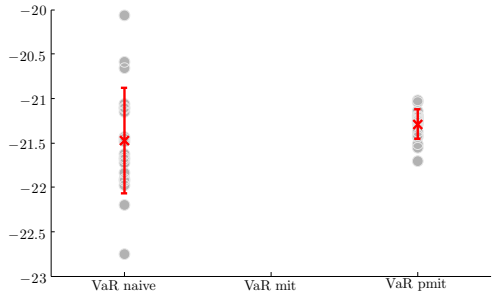
(d) ES,$H = 20$



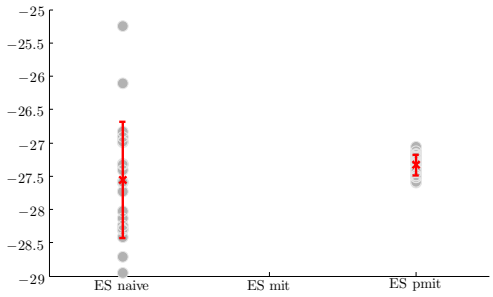
(e) VaR, $H = 40$



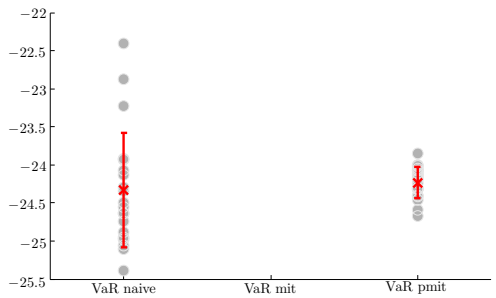
(f) ES, $H = 40$



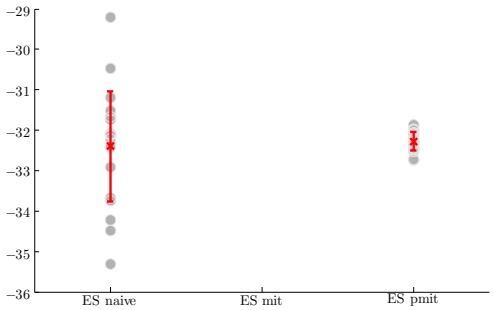
(g) VaR, $H = 100$



(h) ES, $H = 100$



(i) VaR, $H = 250$



(j) ES, $H = 250$

Figure C.8: Accuracy of 99% VaR (left) and ES (right) forecasts for the frequentist **GAS(1,1)-t** model for different horizons, based on 20 MC replications. The left error bars correspond to the direct approach (based on the naive candidate), the middle and the right one – to the QERMit approach (with the candidate constructed with MitISEM and PMitISEM, respectively). A missing error bar for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.

D Time-precision plots

The plots in this appendix illustrate the time-precision trade-off for 99% VaR and ES forecasts. Precision is defined as the inverse of the variance of the results obtained in a Monte Carlo study, where we carried out 20 computations of VaR and ES. Computing time includes the "fixed cost" for the candidate construction, for which the lines are flat (negligible yet non-zero for the direct Bayesian methods, and more noticeable for the QERMit methods). The "variable cost" of the computing time refers to the time needed to perform a single VaR and ES forecast, based on $N = 10,000$ parameter draws (for both, the direct and the QERMit approach). Note that the scales for the time axis differ among horizons. Then the slope of the non-flat part of each line is specified as the ratio of precision and sampling time. Following **?** we also consider the benchmark line of 1 digit precision with 95% confidence. It is defined as $1.96\text{NSE} \leq 0.05$, which corresponds to the required precision level of 1536, and is depicted in the plots as a black horizontal line.

D.1 Bayesian applications

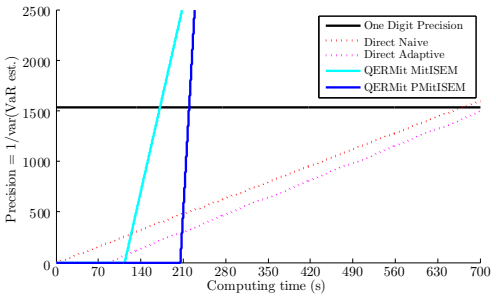
Figure **??** presents the results for the GARCH(1,1)-t model and Figure **??** for the GAS(1,1)-t. Importantly, for longer horizons ($H = 40$ and longer) there are no lines for QERMit based on the MitISEM algorithm, as it was not possible to apply it in such multidimensional cases. For both models the steepness of the QERMit methods is higher the for the direct approaches for both VaR and ES evaluations (see Tables 3.4 and 3.8 in the main paper for the quantitative results), which means that if a high precision is required, then the proposed QERMit based methods will need less computing time to achieve this.

D.2 Frequentist applications

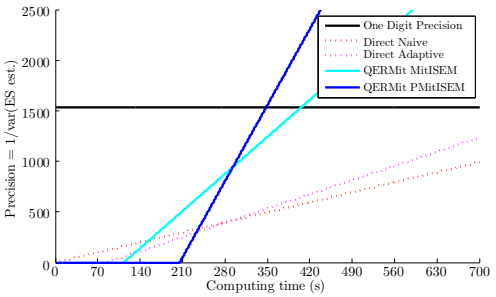
Figures **??** and **??** are the frequentist counterparts of the time-precision trade-off plots from Appendix **??**. The main difference between the current plots and the previous ones is that now there are at most three lines in each plot, as we do not consider the adaptive direct method for the frequentist applications. Moreover, the "fixed cost" for the direct approach is exactly zero because is based on sampling of i.i.d. variates from a univariate

standard Student's t distribution (with the number of degrees of freedom set equal to its MLE value) and not on a mixture of multivariate Student's t distributions which needs to be constructed. Again, for longer horizons ($H = 100$ and over for GARCH and for $H = 40$ and over for GAS) there are no lines for QERMit based on MitISEM due to its infeasibility in high dimensions.

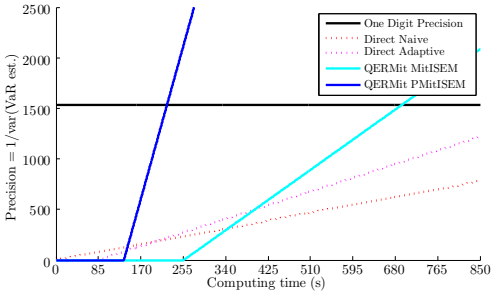
Even though for the longest horizons, $H \geq 100$ for GARCH and $H \geq 40$ for GAS, the direct approach is faster then QERMit in crossing the benchmark 1 digit precision line, higher slopes of the latter (see Tables 4.2 and 4.4 in the main paper) imply that eventually it is more efficient than the former. Notice, that the 1 digit precision line (with 95% confidence) was set somewhat arbitrarily and considering a higher confidence would mean a much higher line. For instance changing of the confidence to 99% would raise it from $1,536$ to $2,654$ so that in more cases less computing time would be needed to reach the required precision level with the QERMit approaches than with the direct one. This would be seen as more "crossings" of the lines for the direct and QERMit-based methods occurring below the required precision line.



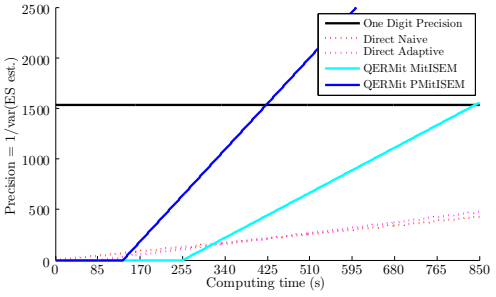
(a) VaR, $H = 10$



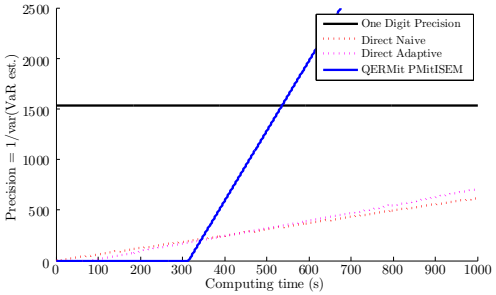
(b) ES, $H = 10$



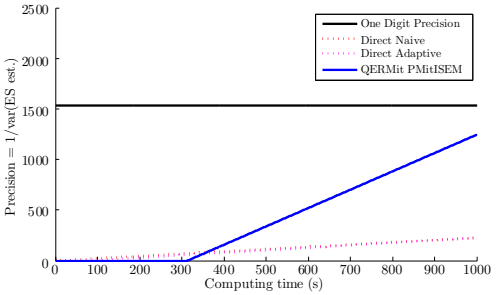
(c) VaR, $H = 20$



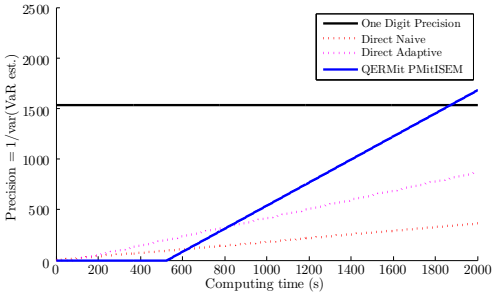
(d) ES, $H = 20$



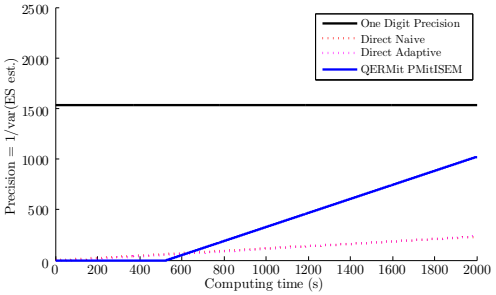
(e) VaR, $H = 40$



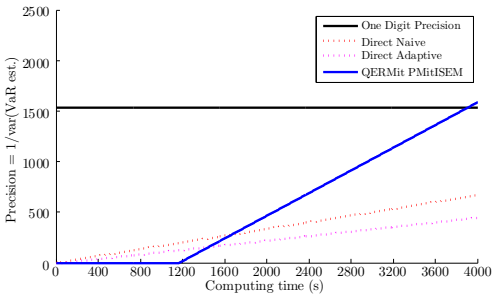
(f) ES, $H = 40$



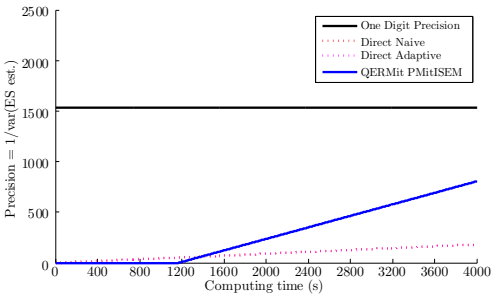
(g) VaR, $H = 100$



(h) ES, $H = 100$

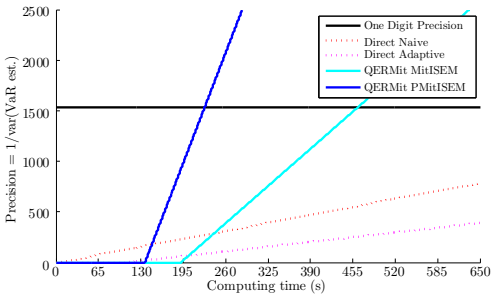


(i) VaR, $H = 250$

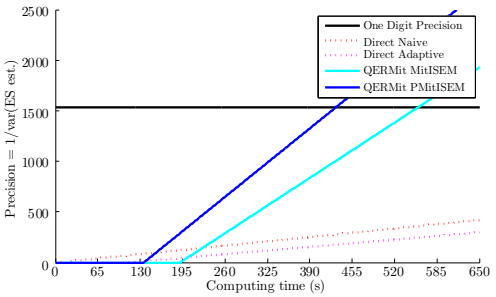


(j) ES, $H = 250$

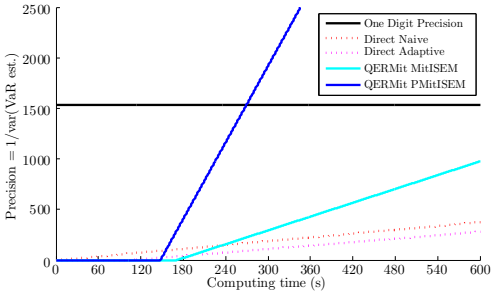
Figure D.1: Precision ($1/var$) of the predicted VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GARCH(1,1)-t** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



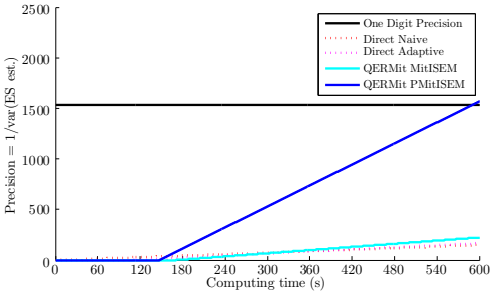
(a) VaR, $H = 10$



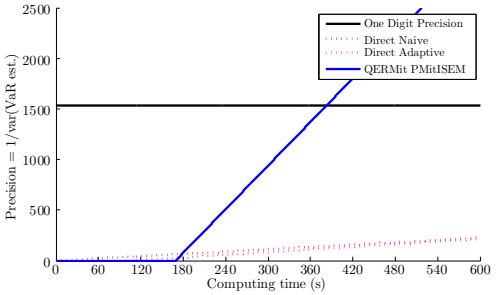
(b) ES, $H = 10$



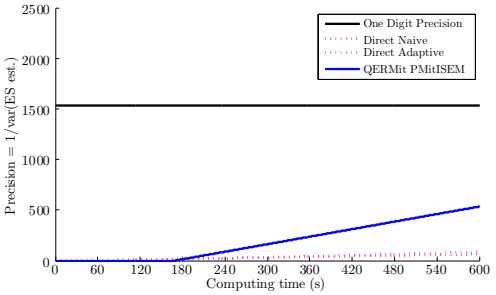
(c) VaR, $H = 20$



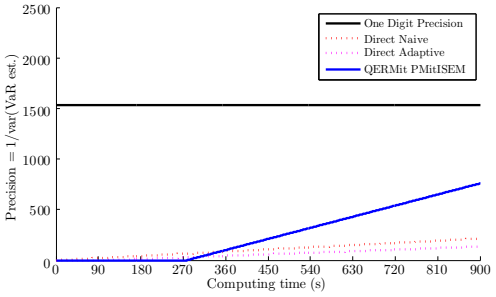
(d) ES, $H = 20$



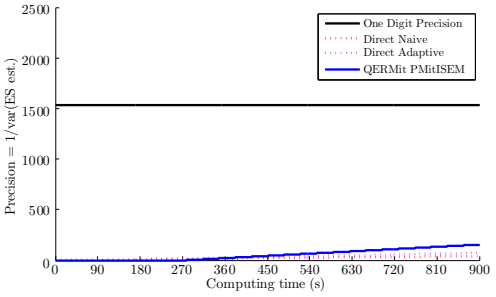
(e) VaR, $H = 40$



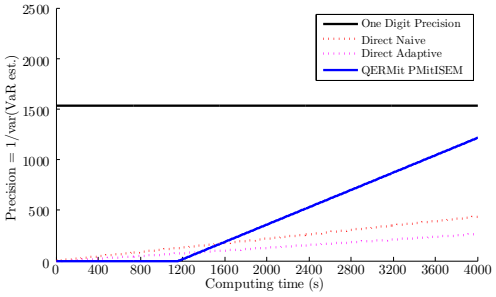
(f) ES, $H = 40$



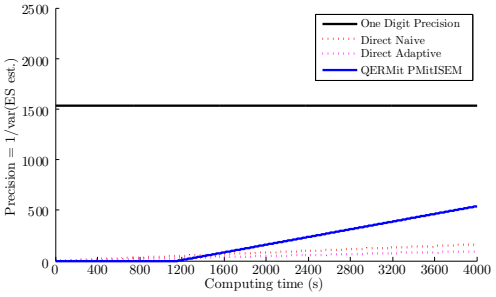
(g) VaR, $H = 100$



(h) ES, $H = 100$

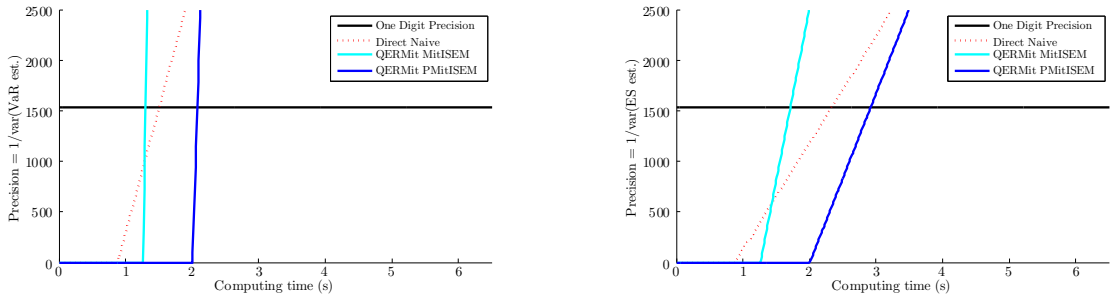


(i) VaR, $H = 250$



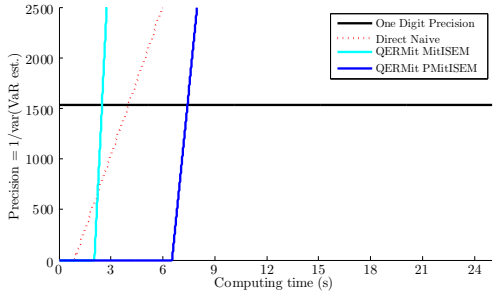
(j) ES, $H = 250$

Figure D.2: Precision ($1/var$) of the predicted VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the **GAS(1,1)-t** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.

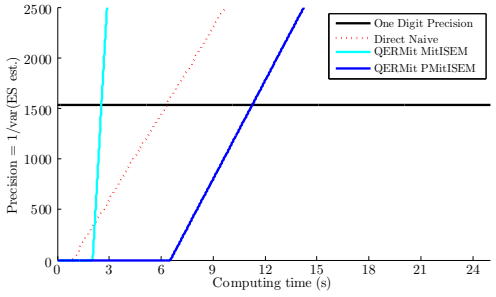


(a) VaR, $H = 10$

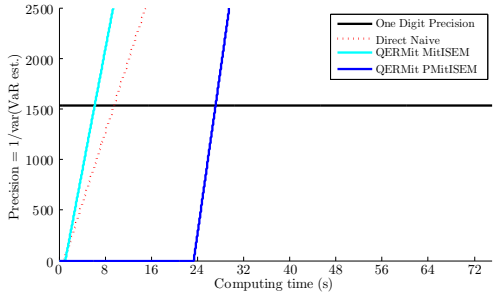
(b) ES, $H = 10$



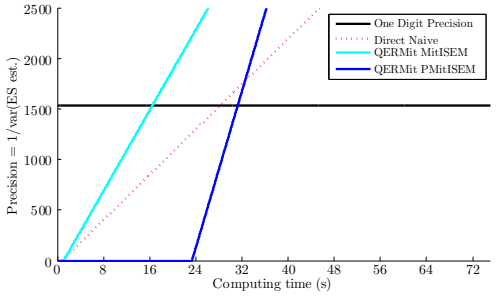
(c) VaR, $H = 20$



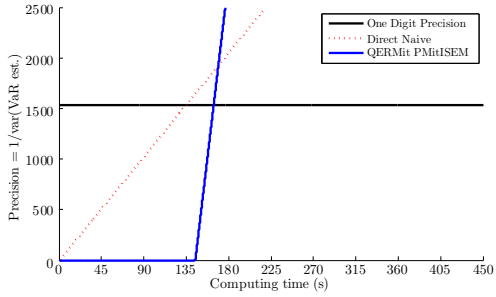
(d) ES, $H = 20$



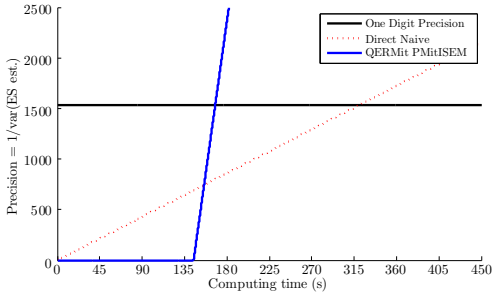
(e) VaR, $H = 40$



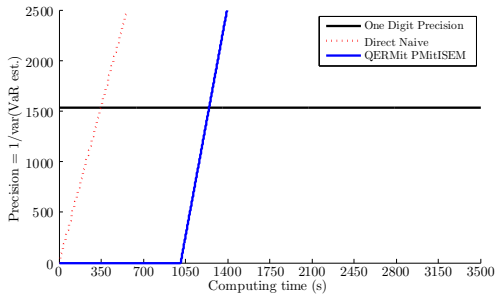
(f) ES, $H = 40$



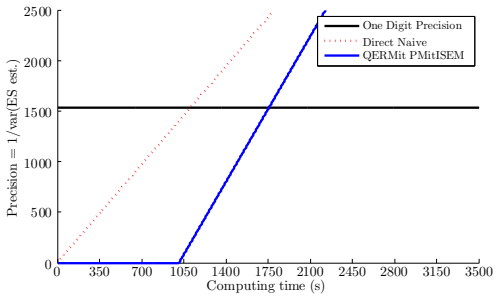
(g) VaR, $H = 100$



(h) ES, $H = 100$

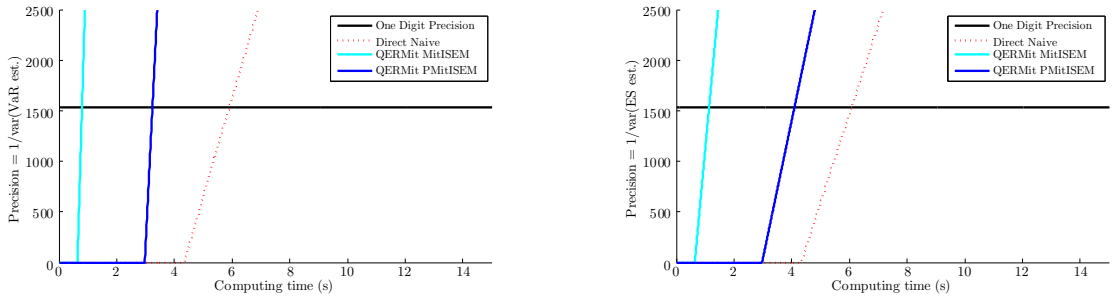


(i) VaR, $H = 250$



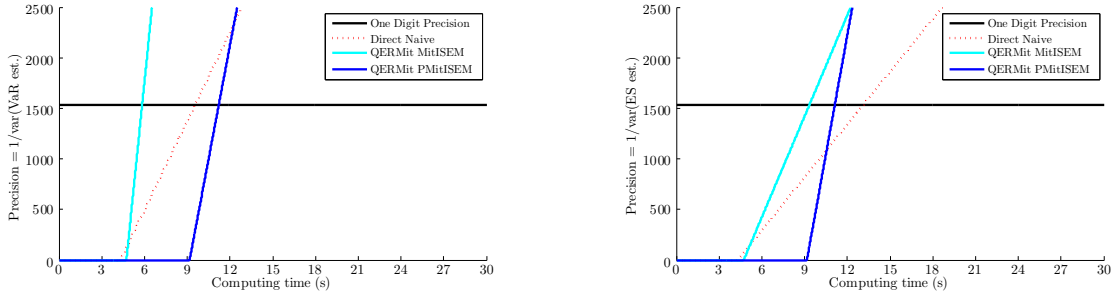
(j) ES, $H = 250$

Figure D.3: Precision ($1/var$) of the predicted VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the frequentist **GARCH(1,1)-t** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.



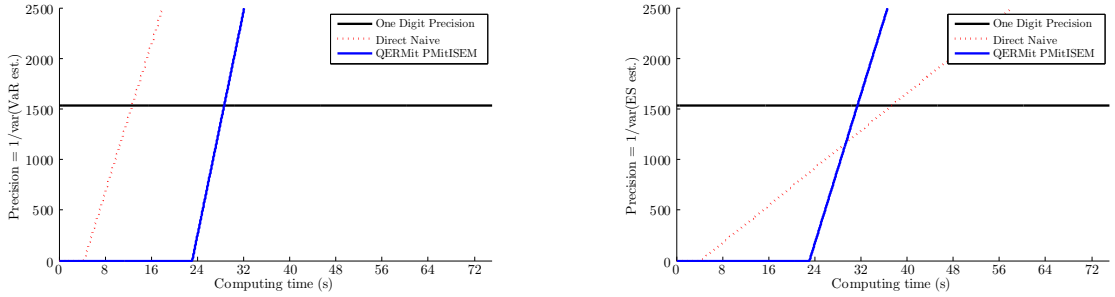
(a) VaR, $H = 10$

(b) ES, $H = 10$



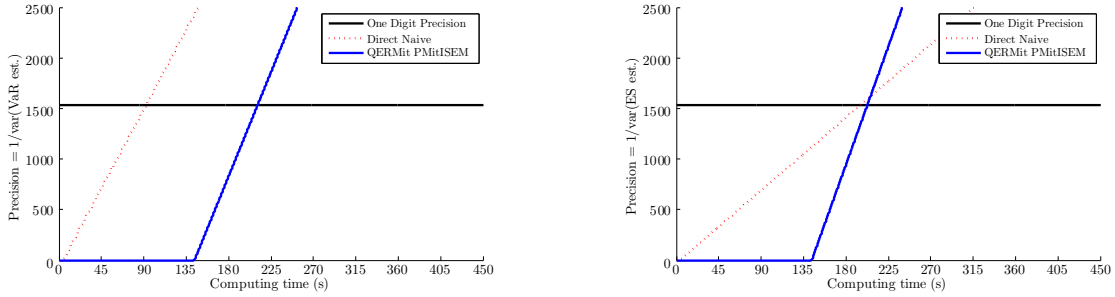
(c) VaR, $H = 20$

(d) ES, $H = 20$



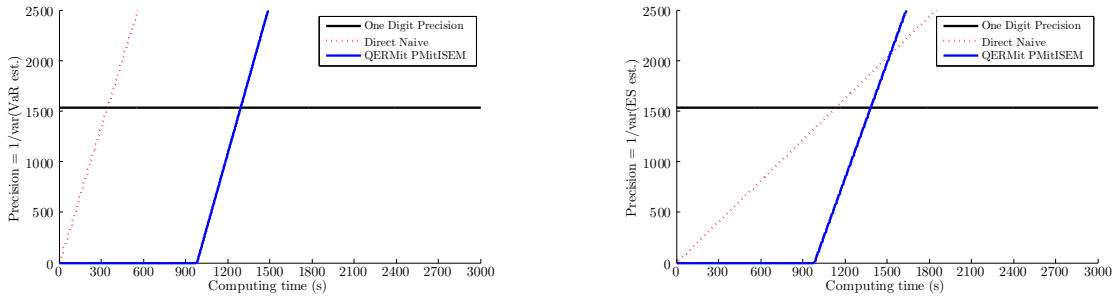
(e) VaR, $H = 40$

(f) ES, $H = 40$



(g) VaR, $H = 100$

(h) ES, $H = 10$



(i) VaR, $H = 25s0$

(j) ES, $H = 250$

Figure D.4: Precision ($1/var$) of the predicted VaR (left) and ES (right), as a function of the amount of computing time for different approaches, for the frequentist **GAS(1,1)-t** model, for different horizons. The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). A missing line for the MitISEM-based candidate corresponds to a situation when it was not possible to construct such a candidate.