TI 2008-092/4
Tinbergen Institute Discussion Paper

Bayesian Forecasting of Value at Risk and Expected Shortfall using Adaptive Importance Sampling

Lennart F. Hoogerheide
Herman K. van Dijk

Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, and Tinbergen Institute.

Tinbergen Institute

The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Bayesian Forecasting of Value at Risk and Expected Shortfall using Adaptive Importance Sampling*

Lennart F. Hoogerheide† & Herman K. van Dijk†

September 2008

Tinbergen Institute report 08-092/4

Abstract

An efficient and accurate approach is proposed for forecasting Value at Risk [VaR] and Expected Shortfall [ES] measures in a Bayesian framework. This consists of a new adaptive importance sampling method for Quantile Estimation via Rapid Mixture of t approximations [QERMit]. As a first step the optimal importance density is approximated, after which multi-step 'high loss' scenarios are efficiently generated. Numerical standard errors are compared in simple illustrations and in an empirical GARCH model with Student-t errors for daily S&P 500 returns. The results indicate that the proposed QERMit approach outperforms several alternative approaches in the sense of more accurate VaR and ES estimates given the same amount of computing time, or equivalently requiring less computing time for the same numerical accuracy.

Keywords: Value at Risk, Expected Shortfall, numerical standard error, numerical accuracy, importance sampling, mixture of Student-t distributions, variance reduction technique.

*A preliminary version of this paper was presented at the 2008 ESEM Conference in Milano. Helpful comments of several participants have led to substantial improvements. The authors further thank David Ardia for useful suggestions. The second author gratefully acknowledges financial assistance from the Netherlands Organization of Research (grant 400-07-703).

†Econometric and Tinbergen Institutes, Erasmus University Rotterdam, The Netherlands

1 Introduction

The issue that is considered in this paper is the efficient computation of accurate estimates of two risk measures, Value at Risk [VaR] and Expected Shortfall [ES], using simulation *given a chosen model*. There are several reasons why it is important to compute *accurate* VaR and ES estimates. An underestimation of risk could obviously cause immense problems for banks and other participants in financial markets (e.g. bankruptcy). On the other hand, an overestimation of risk may cause one to allocate too much capital as a cushion for risk exposures, having a negative effect on profits. Therefore, precise estimates of risk measures are obviously desirable. For simulation based estimates of VaR and ES there also several other issues that play a role. For 'backtesting' or model choice it is important that this model choice is based on the quality of the *model*, rather than the 'quality' of the *simulation run*. For example, one should not choose a model merely because simulation noise stemming from pseudo-random draws caused its historical VaR or ES estimates to be preferable. Next, risk measures should stay approximately constant when the actual risk level stays about constant over time. If changes of risk measures over time are merely caused by simulation noise, this leads to useless fluctuations in positions, leading to extra costs (e.g. transactions costs). Also for the choice between different risky investment strategies based on a risk-return-tradeoff it is important that the computed risk measures are accurate. Decision making on portfolios should not be misled by simulation noise. Moreover, the total volume of invested capital may obviously be huge, so that small percentage differences may correspond to huge amounts of money.

A typical disadvantage of computing simulation-based Value at Risk [VaR] and Expected Shortfall [ES] estimates with high precision is that this requires a huge amount of computing time. In practice, such computing times are often too long for 'real time' decision making. Then one typically faces the choice between a lower *numerical accuracy* - using a smaller number of draws or an approximating method - or a lower *'modeling accuracy'* using an alternative, computationally easier, typically less realistic model. In this paper we propose a simulation method that requires less computing time to reach a certain numerical accuracy, so that the latter choice between suboptimal alternatives may not be necessary.

The approaches for computing VaR and ES estimates can be divided into three groups (as indicated by McNeil and Frey (2000, p. 272)): non-parametric historical simulation, fully parametric methods based on an econometric model with explicit assumptions on volatility dynamics and conditional distribution, and methods based on extreme value

theory. In this paper we focus on the second method, although some ideas could be useful in the simulation-based approaches of the third method. We compute VaR and ES in a Bayesian framework: we consider the Bayesian predictive density. A specific focus is on the 99% quantile of a loss distribution for a 10-days ahead horizon. This particular VaR measure is accepted by the Basel Committee on Banking and Supervision of Banks for Internal Settlement (Basel Committee on Banking Supervision (1995)). The issues of model choice and 'backtesting' the VaR model or ES model are not *directly* addressed. However, as mentioned before, the numerical accuracy of the estimates can be *indirectly* important in the model choice or 'backtesting' procedure because simulation noise may misdirect the model selection process.

The contributions of this paper are as follows. First, we consider the numerical standard errors of VaR and ES estimates. Since VaR and ES are not simply unconditional expectations of (a function of) a random variable, the numerical standard errors do not directly fit within the importance sampling estimator's numerical standard error formula of Geweke (1989). We consider the optimal importance sampling density – that maximizes the numerical accuracy for a given number of draws – as derived by Geweke (1989) for the case of VaR estimation. Second, we propose a particular 'hybrid' mixture density that provides an approximation to the optimal importance density for VaR estimation. The proposed importance density is also useful – perhaps even more so – as an importance density for ES estimation. This 'hybrid' mixture approximation makes use of two mixtures of Student-t distributions as well as the distribution of future asset prices (or returns) given parameter values and historical asset prices (or returns). It is flexible so that it can provide a useful approximation in a wide range of situations. Further, it is easy to simulate from. Moreover, the main contribution of this paper is an iterative approach for constructing this 'hybrid' mixture approximation. It is automatic in the sense that it only requires a posterior density *kernel* - not the exact posterior density - and the distribution of future prices/returns given the parameters and historical prices/returns. We name the proposed two-step method, first constructing an approximation to the optimal importance density and subsequently using this in importance sampling, the Quantile Estimation via Rapid Mixture of t approximations [QERMit] approach. The QERMit procedure makes use of the Adaptive Mixture of t [AdMit] approach, see Hoogerheide et al. (2007), which constructs an approximating mixture of Student-t distributions given merely a kernel of a target density. Hoogerheide et al. (2007) apply the AdMit approach in order to approximate and simulate from a non-elliptical *posterior* of the *parameters*

in an Instrumental Variable [IV] regression model. In this paper we consider the joint distribution of *parameters and future returns* instead of merely the parameters. Moreover, our goal is not to approximate this distribution of parameters and future returns but to approximate the *optimal importance density* in which 'high loss' scenarios are generated more often, which is subsequently 'corrected' by giving these lower importance weights. Hence, the AdMit approach is merely one of the ingredients for the proposed QERMit approach.

There are four clear differences between this paper and the existing literature on importance sampling as a variance reduction technique for VaR estimation. First, typically the distribution of future returns is simply 'given', i.e. the exact density of future returns is known. We consider the Bayesian framework in which we assume that merely the exact density of future asset prices/returns *given the model parameters (and historical prices/returns)* and a *kernel* of the posterior density of the model parameters is known - as is typically the case in Bayesian inference. This has a huge impact on the optimal importance density. As will be described below, this means that the importance density should not only be focused on 'high loss' scenarios. The probability mass of the importance density should be divided 50%-50% over 'high loss' scenarios and 'common' scenarios. Second, the main distinction is that we consider a flexible mixture importance density for which we propose an automatic, iterative procedure to construct it. The speed of the construction procedure and the flexibility of its resulting importance density are the reason why it yields accurate and reliable estimates of VaR and/or ES in far less computing time than alternative approaches. Third, typically only the estimation of VaR is considered, whereas we also focus on ES estimation. The ES measure has several advantages over the VaR, as will be briefly discussed below. Fourth, the numerical accuracy of importance sampling procedures for VaR estimation is typically assessed by repeating *many simulations* and inspecting the standard deviation of the set of estimates. We also consider numerical standard errors, estimates of this standard deviation that are quickly and easily computed on the basis of *one simulation*. In practical situations one may often not have enough time to repeat a simulation experiment many times, so that the use of numerical standard errors may be a very convenient way to assess the numerical accuracy.

For example, Glasserman et al. (2000) specify a normal importance density based on a quadratic 'delta-gamma' approximation to the change in portfolio value. Glass (1999) uses a 'tilted' version of the returns distribution. They consider non-Bayesian applications, where the returns distribution of the assets within a portfolio is 'given', and do not

address estimation of the ES.

The outline of the paper is as follows. In section 2 we discuss the computation of numerical standard errors for VaR and ES estimates. Further we consider the optimal importance sampling density (due to Geweke (1989)), which minimizes the numerical standard errors (given a certain number of draws), for the case of VaR estimation. In section 3, we briefly reconsider the AdMit approach (Hoogerheide et al. (2007)). Section 4 describes the proposed QERMit method. In section 5 we illustrate the possible usefulness of the QERMit approach in an empirical example of estimating 99% VaR and ES in a GARCH model with Student-t innovations for S&P 500 log-returns. Section 6 concludes.

2 Computation of Value at Risk and Expected Shortfall using Importance Sampling: numerical standard errors and optimal importance distribution

2.1 Value at Risk [VaR] and Expected Shortfall [ES]

In literature, the VaR is referred to in several different manners. The quoted VaR is either a percentage or an amount of money, referring to either a future portfolio value or a future portfolio value in deviation from its expected value or current value. In this paper we refer to the $100\alpha\%$ VaR as the $100(1-\alpha)\%$ quantile of the percentage return's distribution and ES as the expected percentage return given that the loss exceeds the $100\alpha\%$ quantile. With these definitions VaR and ES are typically values between -100% and 0%.[1]

The VaR is a risk measure with several advantages: it is relatively easy to estimate and easy to explain to non-experts. The specific VaR measure of the 99% quantile for a horizon of two weeks - 10 trading days - is acceptable for the Basel Committee on Banking and Supervision of Banks for Internal Settlement (Basel Committee on Banking Supervision (1995)). This is motivated by the fear of a liquidity crisis where a financial institution might not be able to liquidate its holdings for a two weeks period. Even

[1]For certain derivatives, e.g. options or futures, it may not be natural or even possible to quote profit or loss as a certain percentage. It should be noted that the quality of our proposed method is not affected by which particular VaR or ES definition is used.

though the VaR has become a standard tool in financial risk management, it has several disadvantages. First, the VaR does not tell anything about the potential size of loss that exceeds the VaR level. This may lead to too risky investment strategies that optimize expected profit under the restriction that the VaR is not beyond a certain threshold, since the potential 'rare event' losses exceeding the VaR may be extreme. Second, the VaR is not a *coherent* measure, as indicated by Artzner et al. (1999). This results since the VaR lacks the property of *sub-additivity*. The ES has clear advantages over the VaR: it does say something about losses exceeding the VaR level, and the ES is a sub-additive, coherent measure. This property of sub-additivity means that the ES of a portfolio (firm) can not exceed the sum of the ES measures of its sub-portfolios (departments). Adding these individual ES measures yields a conservative risk measure for the whole portfolio (firm).[2] Because of these advantages of ES over VaR we not not only consider VaR but also ES in this paper. For a concise and clear discussion of the VaR and ES measures we also refer to Ardia (2008).

2.2 The 'direct' approach of Bayesian estimation of VaR or ES

As mentioned in the introduction, there are several approaches for computing VaR and ES estimates. In this paper, we focus on the Bayesian approach in an econometric model with explicit assumptions on volatility dynamics and conditional distribution. We use the following notation. The m-dimensional vector y_t consists of the returns (or asset prices) at time t.[3] Our data set on T historical returns is $y \equiv \{y_1, \ldots, y_T\}$. We consider τ-step ($\tau = 1, 2, \ldots$) ahead forecasting of VaR and ES, where we define the vector of future returns $y^* \equiv \{y_{T+1}, \ldots, y_{T+\tau}\}$. The model has a k-dimensional parameter vector θ. Finally, we have a (scalar valued) profit & loss function $PL(y^*)$ that is positive for profits, negative for losses.

[2] In order to see that the VaR measure is not *sub-additive*, consider the simple example of two independent assets, both with a 4% probability of becoming worthless in the next period and 96% probability that its value remains constant. Then the 95% VaR for the separate assets is zero, not providing any warning signal for risk, whereas the 95% VaR for the two assets together is non-zero.
The ES is *sub*-additive, as due to diversification the ES measure for a portfolio will typically be smaller than the sum of its sub-portfolios' ES measures.

[3] The examples in this paper consider integrated models for the S&P 500, i.e. GARCH type models for the S&P 500 log-returns (daily changes of the log-price) y_t. In the case of mean-reverting processes, e.g. electricity prices, one obviously uses the historical price process rather than merely the returns process in order to forecast future returns.

In order to estimate the τ-step ahead $100\alpha\%$ VaR or ES in a Bayesian framework, one can obviously use the following straightforward approach, that we will refer to as the *'direct approach'* of Bayesian VaR/ES estimation:

(Step 1) Simulate a set of draws θ^i $(i = 1, \ldots, n)$ from the posterior distribution, e.g. using Gibbs sampling (Geman and Geman (1984)) or the Metropolis-Hastings algorithm (Metropolis et al. (1953), Hastings (1970)).

(Step 2) Simulate corresponding future paths $y^{*i} \equiv \{y^i_{T+1}, \ldots, y^i_{T+\tau}\}$ $(i = 1, \ldots, n)$ from the model given parameter values θ^i and historical values $y \equiv \{y_1, \ldots, y_T\}$, i.e. from the density $p(y^*|\theta^i, y)$.

(Step 3) Order the values $PL(y^{*i})$ ascending as $PL^{(j)}$ $(j = 1, \ldots, n)$. The VaR and ES are then estimated as

$$\widehat{VaR}_{DA} \equiv PL^{(n(1-\alpha))} \tag{1}$$

and

$$\widehat{ES}_{DA} \equiv \frac{1}{n(1-\alpha)} \sum_{j=1}^{n(1-\alpha)} PL^{(n(1-\alpha))}, \tag{2}$$

the $(n(1-\alpha))$th sorted loss and the average of the first $(n(1-\alpha))$ sorted losses, respectively.

For example, one may generate 10000 profit/loss values, sort these ascending, and take the 100th sorted value as the 99% VaR estimate. In order to intuitively sketch that this 'direct approach' is not optimal, consider the simple example of the standard normal distribution with $PL \sim N(0,1)$. If we then estimate the 99% VaR by simulating 10000 standard normal variates, and taking the 100th sorted value, then this estimate is intuitively speaking 'based' on only 100 out of 10000 draws, see Figure 1. There is no specific focus on the 'high loss' subspace, the left tail. Roughly speaking, if we are only interested in the VaR or ES, then a large subset of the draws seems to be 'wasted' on the subspace that we are not particularly interested in. An alternative simulation approach that allows one to specifically focus on an *important* subspace is Importance Sampling.

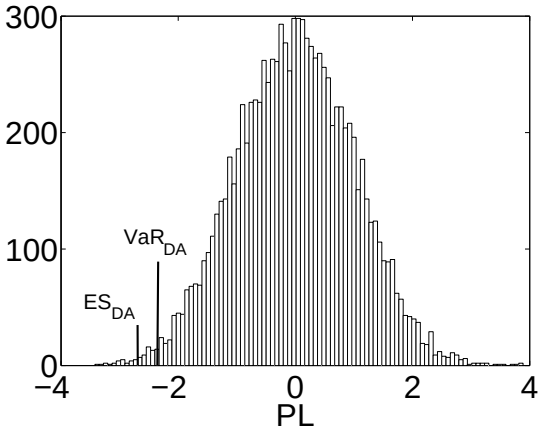


Figure 1: Example of standard normally distributed profit/loss PL: if VaR and ES are estimated by direct sampling using 10000 draws, these estimates particularly depend on only 100 out of 10000 draws.

2.3 Bayesian estimation of VaR or ES by Importance Sampling: numerical standard errors

In Importance Sampling [IS][4] the expectation $E[g(X)]$ of a certain function $g(.)$ of the random variable $X \in \mathrm{R}^r$ is estimated as

$$\widehat{E[g(X)]}_{IS} = \frac{\frac{1}{n}\sum_{i=1}^{n} w(\tilde{X}_i) g(\tilde{X}_i)}{\frac{1}{n}\sum_{j=1}^{n} w(\tilde{X}_j)} = \frac{\sum_{i=1}^{n} w(\tilde{X}_i) g(\tilde{X}_i)}{\sum_{j=1}^{n} w(\tilde{X}_j)}, \tag{3}$$

where $\tilde{X}_1, \ldots, \tilde{X}_n$ are independent realizations from the candidate distribution with density (= importance function) $q(x)$, and $w(\tilde{X}_1), \ldots, w(\tilde{X}_n)$ are the corresponding weights $w(\tilde{X}) = \frac{p(\tilde{X})}{q(\tilde{X})}$ with $p(x)$ a *kernel* of the target density $p^*(x)$ of X: $p(x) \propto p^*(x)$.[5]

[4] IS, see Hammersley and Handscomb (1964), has been introduced in Bayesian inference by Kloek and Van Dijk (1978) and is further developed by Van Dijk and Kloek (1980, 1984) and Geweke (1989).

[5] The consistency of the IS estimator in (3) is easily seen from

$$E[g(X)] = \int g(x) \frac{p(x)}{\int p(x)\,dx}\,dx = \frac{\int g(x)p(x)\,dx}{\int p(x)\,dx} = \frac{\int g(x)w(x)q(x)\,dx}{\int w(x)q(x)\,dx} = \frac{E[w(\tilde{X})g(\tilde{X})]}{E[w(\tilde{X})]}.$$

If we know the exact target density $p^*(x)$, then we also have

$$E[g(X)] = \int g(x)p^*(x)\,dx = \int g(x)\frac{p^*(x)}{q(x)}q(x)\,dx = E\left[\frac{p^*(\tilde{X})}{q(\tilde{X})}g(\tilde{X})\right],$$

so that we can use an alternative IS estimator of $E[g(X)]$:

$$\widehat{E[g(X)]}_{IS^*} = \frac{1}{n}\sum_{i=1}^{n} \frac{p^*(\tilde{X}_i)}{q(\tilde{X}_i)} g(\tilde{X}_i).$$

The IS estimator $\widehat{VaR}_{IS}$ of the $100\alpha\%$ VaR is computed by solving $\widehat{E[g(X)]}_{IS} = 1-\alpha$ with $g(X) = I\{PL(X) \leq \widehat{VaR}_{IS}\}$ (since $P[PL(X) \leq c] = E[I\{PL(X) \leq c\}]$). This amounts to sorting the profit/loss values of the candidate draws $PL(\tilde{X}_i)$ $(i = 1, \ldots, n)$ ascending as $PL(\tilde{X}^{(j)})$ $(j = 1, \ldots, n)$, and finding the value $PL(\tilde{X}^{(k)})$ such that $S_k = 1-\alpha$ where $S_k \equiv \sum_{j=1}^{k} \tilde{w}(\tilde{X}^{(j)})$ is the cumulative sum of scaled weights $\tilde{w}(\tilde{X}^{(j)}) \equiv \frac{w(\tilde{X}^{(j)})}{\sum_{i=1}^{n} w(\tilde{X}^{(i)})}$ (scaled to add to 1) corresponding to the ascending profit/loss values. In general there will be no $\tilde{X}^{(k)}$ such that $S_k = 1-\alpha$, so that one interpolates between the values of $PL(\tilde{X}^{(k)})$ and $PL(\tilde{X}^{(k+1)})$ where $PL(\tilde{X}^{(k+1)})$ is the smallest value with $S_{k+1} > 1-\alpha$.

The IS estimator $\widehat{ES}_{IS}$ of the $100\alpha\%$ ES is subsequently computed as $\widehat{ES}_{IS} = \frac{1}{k}\sum_{j=1}^{k} w^*(\tilde{X}^{(j)})\, PL(\tilde{X}^{(j)})$, the weighted average of the k values $PL(\tilde{X}^{(j)})$ $(j = 1, \ldots, k)$ with weights $w^*(\tilde{X}^{(j)}) \equiv \frac{w(\tilde{X}^{(j)})}{\sum_{i=1}^{k} w(\tilde{X}^{(i)})}$ (adding to 1).

Geweke (1989) provides formulas for the numerical accuracy of the IS estimator $\widehat{E[g(X)]}_{IS}$ in (3). See also Hoogerheide et al. (2008) for a discussion of the numerical accuracy of $\widehat{E[g(X)]}_{IS}$. Define

$$t_0 = \frac{1}{n}\sum_{i=1}^{n} w(\tilde{X}_i), \tag{4}$$

$$t_1 = \frac{1}{n}\sum_{i=1}^{n} w(\tilde{X}_i)\, g(\tilde{X}_i), \tag{5}$$

so that the importance sampling estimator can be written as $\widehat{E[g(\theta)]}_{IS} = t_1/t_0$. Using the delta method, the estimated variance $\hat{\sigma}^2_{\text{IS}}$ of $\widehat{E[g(X)]}_{IS} = t_1/t_0$ is given by

$$\begin{aligned}\hat{\sigma}^2_{\text{IS}} &= \left(\frac{\partial \widehat{E[g(X)]}_{IS}}{\partial t_0} \;\; \frac{\partial \widehat{E[g(X)]}_{IS}}{\partial t_1} \right) \begin{pmatrix} \hat{\text{var}}(t_0) & \hat{\text{cov}}(t_0, t_1) \\ \hat{\text{cov}}(t_0, t_1) & \hat{\text{var}}(t_1) \end{pmatrix} \begin{pmatrix} \frac{\partial \widehat{E[g(X)]}_{IS}}{\partial t_0} \\ \frac{\partial \widehat{E[g(X)]}_{IS}}{\partial t_1} \end{pmatrix} \\ &= \frac{t_1^2}{t_0^4}\hat{\text{var}}(t_0) + \frac{1}{t_0^2}\hat{\text{var}}(t_1) - 2\frac{t_1}{t_0^3}\hat{\text{cov}}(t_0, t_1), \end{aligned} \tag{6}$$

where

$$\hat{\text{var}}(t_0) = \frac{1}{n}\hat{\text{var}}(w(\tilde{X}_i)) = \frac{1}{n}\left(\left[\frac{1}{n}\sum_{i=1}^{n} w(\tilde{X}_i)^2 \right] - t_0^2 \right), \tag{7}$$

$$\hat{\text{var}}(t_1) = \frac{1}{n}\hat{\text{var}}(w(\tilde{X}_i)\, g(\tilde{X}_i)) = \frac{1}{n}\left(\left[\frac{1}{n}\sum_{i=1}^{n} w(\tilde{X}_i)^2 g(\tilde{X}_i)^2 \right] - t_1^2 \right), \tag{8}$$

$$\hat{\text{cov}}(t_0, t_1) = \frac{1}{n}\hat{\text{cov}}(w(\tilde{X}_i), w(\tilde{X}_i)\, g(\tilde{X}_i)) = \frac{1}{n}\left(\left[\frac{1}{n}\sum_{i=1}^{n} w(\tilde{X}_i)^2 g(\tilde{X}_i) \right] - t_0\, t_1 \right), \tag{9}$$

In the sequel we will use this formula to explain that for $\widehat{E[g(X)]}_{IS}$ and $\widehat{E[g(X)]}_{IS^*}$ different importance densities $q(x)$ are optimal.

and where t_0 and t_1 are evaluated at their realized values. It holds for large n and under mild regularity conditions reported by Geweke (1989) that $\widehat{E[g(X)]}_{IS}$ is approximately $\mathcal{N}(E[g(X)], \sigma^2_{\text{IS}})$ distributed. The accuracy of the estimate $\widehat{E[g(X)]}_{IS}$ for $E[g(X)]$ is reflected by the *numerical standard error* $\hat{\sigma}_{\text{IS}}$, and the 95% confidence interval for $E[g(\theta)]$ can be constructed as $(\widehat{E[g(X)]}_{IS} - 1.96\,\hat{\sigma}_{\text{IS}}, \widehat{E[g(X)]}_{IS} + 1.96\,\hat{\sigma}_{\text{IS}})$.[6]

The numerical standard error [NSE] $\hat{\sigma}_{\text{IS,VaR}}$ of the IS estimator of the VaR or ES does not directly follow from the NSE for $\widehat{E[g(X)]}_{IS}$, as both VaR and ES are not unconditional expectations $E[g(X)]$ for a random variable X of which we know the density kernel.[7] For the NSE of the VaR estimator, we make again use of the delta rule. We have

$$\begin{aligned}
\text{P}[PL(X) \leq \widehat{VaR}] &\approx \text{P}[PL(X) \leq VaR] \\
&\quad + \frac{\partial\, \text{P}[PL(X) \leq c]}{\partial\, c}\bigg|_{c=VaR} \left(\widehat{VaR} - VaR\right) \;\Rightarrow & (10) \\
1-\alpha &\approx \hat{\text{P}}[PL(X) \leq VaR] + \hat{p}_{_{PL}}(VaR)\left(\widehat{VaR} - VaR\right) \;\Rightarrow & (11) \\
\text{var}(\widehat{VaR}) &\approx \frac{\text{var}\left(\hat{\text{P}}[PL(X) \leq VaR]\right)}{(\hat{p}_{_{PL}}(VaR))^2} & (12)
\end{aligned}$$

where (11) results from (10) by substituting estimates for $\text{P}[PL(X) \leq \widehat{VaR}]$, $\text{P}[PL(X) \leq VaR]$ and $p_{_{PL}}(VaR)$, where $p_{_{PL}}(VaR)$ is the density of $PL(X)$ evaluated at VaR and $\hat{\text{P}}[PL(X) \leq \widehat{VaR}] = 1-\alpha$ since this equality defines $\widehat{VaR}$. Substituting the realized value of $\widehat{VaR}_{IS}$ for VaR into (12) and taking the square root yields the numerical standard error for $\widehat{VaR}_{IS}$:

$$\hat{\sigma}_{\text{IS,VaR}} = \frac{\hat{\sigma}_{\text{IS,P[PL}\leq\widehat{\text{VaR}}_{\text{IS}}]}}{\hat{p}_{_{PL}}(\widehat{VaR}_{IS})} \tag{13}$$

[6] If we know the exact target density $p^*(x)$, then we have

$$\text{var}\left(\widehat{E[g(X)]}_{\text{IS}^*}\right) = \frac{1}{n}\text{var}\left(\frac{p^*(\tilde{X})}{q(\tilde{X})}g(\tilde{X})\right) \;\Rightarrow\; \hat{\sigma}^2_{\text{IS}} = \frac{1}{n}\left(\left[\frac{1}{n}\sum_{i=1}^{n} w(\tilde{X}_i)^2 g(\tilde{X}_i)^2\right] - \left(\widehat{E[g(X)]}_{\text{IS}^*}\right)^2\right).$$

[7] Only if we would know the true value of the VaR with certainty, then the estimation of the ES would reduce to the 'standard' situation of IS estimation of the expectation of $PL(X)$ where X has target density kernel $p_{target}(x) \propto p(x)I\{PL(x) \leq VaR\}$. For an estimated $\widehat{VaR}$ value, the uncertainty on the ES estimator is larger than that. This uncertainty has two sources: (1) the variation of those draws $\tilde{X}_i$ with $PL(\tilde{X}_i) \leq VaR$ for $VaR = \widehat{VaR}$; and (2) the variation in $\widehat{VaR}$.

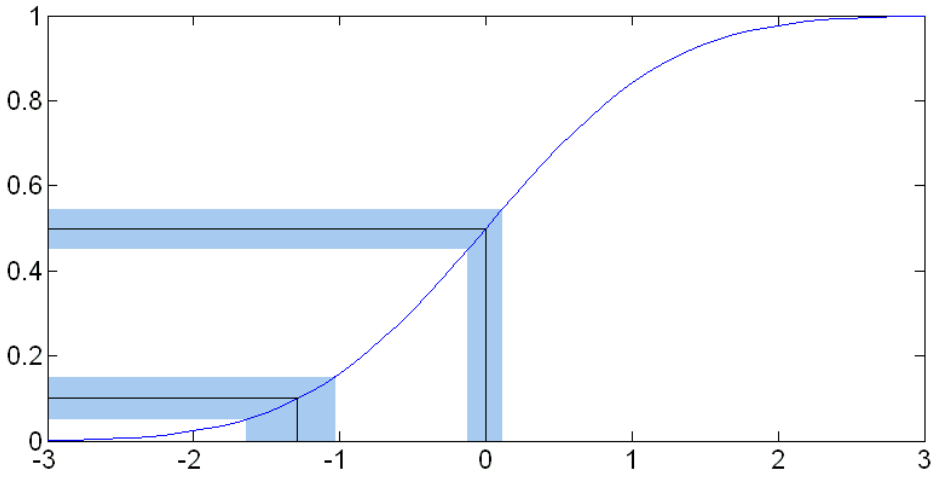


Figure 2: Illustration of the numerical standard error of the IS estimator for a VaR, a quantile of a profit/loss $PL(X)$ function of a random vector X. The uncertainty on $P[PL(X) \leq c]$ for $c = \widehat{VaR}_{IS}$ – on the vertical axis – is translated to the uncertainty on $\widehat{VaR}_{IS}$ – on the horizontal axis – by a factor $\frac{1}{p_{pl}(c)}$, the inverse of the density function that is the steepness of the displayed cumulative distribution function [CDF] of the profit/loss distribution.

The numerical standard error $\hat{\sigma}_{\mathrm{IS,P[PL}\leq\widehat{\mathrm{VaR}}_{\mathrm{IS}}]}$ for the IS estimator of the probability $\mathrm{P}[PL(X) \leq c]$ for $c = \widehat{VaR}_{IS}$ directly follows from (6) with $g(x) = I\{PL(x) \leq c\}$. Notice that in general we do not have an explicit formula for the density $p_{PL}(c)$ of $PL(X)$, but this is easily estimated by $\frac{Pr[PL(X)\leq c+\epsilon]-Pr[PL(X)\leq c-\epsilon]}{2\epsilon}$. One can compute this for several ϵ values, and use the ϵ that leads to the smallest estimate $\hat{p}_{pl(X)}(c)$, and hence the largest (conservative) value for $\hat{\sigma}_{IS,\widehat{VaR}}$.[8] Alternatively, one can use a kernel estimator of the profit/loss density at $c = \widehat{VaR}_{IS}$. Figure 2 provides an illustration of the numerical standard error for an IS estimator of a VaR, or more generally a quantile.

For the numerical standard error of the ES, we use that if the VaR would be known with certainty, we would be in a 'standard' situation of IS estimation of the expectation of a variable $PL(X)$ where X has the target density kernel $p_{target}(x) \propto p(x)I\{PL(x) \leq VaR\}$ for which the NSE $\hat{\sigma}_{IS,ES|VaR}$ and the (asymptotically valid) normal density are easily computed using (6). Since we do have the NSE $\hat{\sigma}_{IS,VaR}$ and the (asymptotically valid) normal density $\mathcal{N}\left(\widehat{VaR}_{IS}, \hat{\sigma}_{IS,VaR}\right)$ of the VaR estimator (as derived above), we can

[8] A convenient alternative is to compute $\frac{\mathrm{P}[PL(X)\leq c+\epsilon_1]-\mathrm{P}[PL(X)\leq c-\epsilon_2]}{\epsilon_1+\epsilon_2}$ for ϵ_1, ϵ_2 such that $\mathrm{P}[PL(X) \leq c+\epsilon_1] = (1-\alpha) + b\,\hat{\sigma}_{\mathrm{IS,P[PL}\leq\widehat{\mathrm{VaR}}_{\mathrm{IS}}]}$ and $\mathrm{P}[PL(X) \leq c-\epsilon_2] = (1-\alpha) - b\,\hat{\sigma}_{\mathrm{IS,P[PL}\leq\widehat{\mathrm{VaR}}_{\mathrm{IS}}]}$, e.g. for $b = 1, 2$.

proceed as follows to estimate the density for the ES estimator:

(1) Construct a grid of VaR values, e.g. on the interval $\left[\widehat{VaR}_{IS} - 4\hat{\sigma}_{IS,\widehat{VaR}}, \widehat{VaR}_{IS} + 4\hat{\sigma}_{IS,\widehat{VaR}}\right]$.

(2) For each VaR value on the grid evaluate the NSE $\hat{\sigma}_{IS,ES|VaR}$ of the ES estimator given the VaR value, and evaluate the (asymptotically valid) normal density $p(\widehat{ES}_{IS}|VaR)$ of the ES estimator on a grid.

(3) Estimate the ES estimator's density $p(\widehat{ES}_{IS})$ as the weighted average of the densities $p(\widehat{ES}_{IS}|VaR)$ in step (2) with weights from the estimated density of the VaR estimator $p(\widehat{VaR}_{IS})$.

The numerical accuracy of $\widehat{ES}_{IS}$ is now estimated by considering the 95% interval of the density $p(\widehat{ES}_{IS})$; the numerical standard error is obtained as its standard deviation. Figure 3 illustrates the procedure for estimation of the ES estimator's density in the case of $N(0,1)$ distributed profit/loss. The left panels show the case of direct sampling, where the density $p(\widehat{ES}_{IS}|VaR)$ has clearly higher variance for more negative, more extreme VaR values, resulting in a skewed density $p(\widehat{ES}_{IS})$. The reason is that for these extreme VaR values the estimate of ES given VaR is based on only few draws. The right panels show the case of IS with Student-t (10 degrees of freedom) importance sampling, where the density $p(\widehat{ES}_{IS}|VaR)$ has hardly a higher variance for more extreme VaR values. The Student-t distribution's fat tails assure that also for extreme VaR values the estimated ES is based on many draws. This example already reflects an advantage of IS (with an importance density having fatter tails than the target distribution) over a 'direct approach': IS results in lower NSE and especially less downward uncertainty on the ES – with lower risk of substantially underestimating risk.

2.4 Bayesian estimation of VaR or ES by Importance Sampling: the optimal importance density

The optimal importance distribution for IS estimation of $\bar{g} = E[g(X)]$ for a given target density $p(x)$ and function $g(x)$, which minimizes the numerical standard error for a given (large) number of draws, is given by Geweke (1989, Theorem 3). This optimal importance density has kernel $q_{opt}(x) \propto |g(x) - \bar{g}|\, p(x)$ (under the condition that $E[|g(x) - \bar{g}|]$ is finite). Geweke (1989) mentions three practical disadvantages of this optimal importance distribution. First, it is different for different functions $g(x)$. Second, a preliminary estimate of $\bar{g} = E[g(X)]$ is required. Third, methods for simulating from it would need to

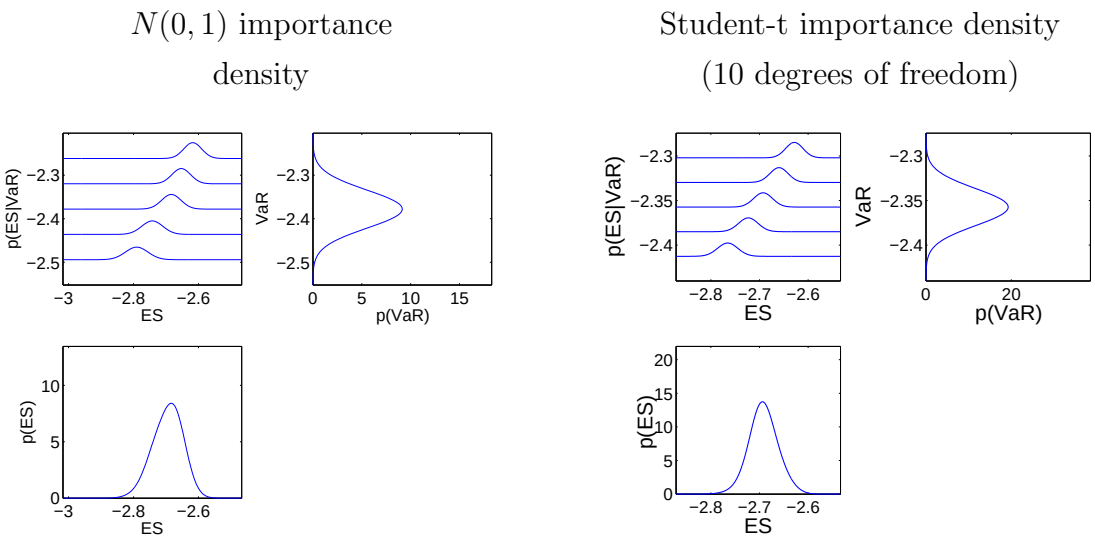


Figure 3: Example of standard normally distributed profit/loss PL: illustration of estimation of ES estimator's density using a $N(0,1)$ importance density (left) – corresponding to the case of direct sampling – or a Student-t importance density (right). The top-left panel gives the densities $p(\widehat{ES}_{IS}|VaR)$ for several VaR values. The top-right panel shows the density $p(\widehat{VaR}_{IS})$. The bottom panel gives the density $p(\widehat{ES}_{IS})$.

be devised. Geweke (1989) further notes that this result reflects that importance sampling densities with fatter tails may be more efficient than the target density itself, as is the case in the example above. In such cases the relative numerical efficiency [RNE], the ratio between (an estimate of) the variance of an estimator based on direct sampling and the IS estimator's estimated variance (with the same number of draws), exceeds 1.[9] An interesting result is the case where $g(x)$ is an indicator function $I\{x \in S\}$ for a subspace S, so that $E[g(X)] = P[X \in S] = \bar{p}$. Then the optimal importance density kernel is given by $q_{opt}(x) \propto (1-\bar{p})\, p(x)$ for $x \in S$ and $q_{opt}(x) \propto \bar{p}\, p(x)$ for $x \notin S$, so that half the draws should be made in S and half outside S, in proportion to the target kernel $p(x)$ in both

[9]The RNE is an indicator of the efficiency of the chosen importance function; if target and importance density coincide the RNE equals one, whereas a very poor importance density will have an RNE close to zero. The inverse of the RNE is known as the inefficiency factor [IF].

cases.[10]

From formula (13) it is seen that the NSE of the IS estimator for the $100\alpha\%$ VaR is proportional to the NSE of the IS estimator of $E[g(X)]$ with $g(x) = I\{x \in S\}$, where S is the subspace with $100(1-\alpha)\%$ lowest values $PL(X)$. Hence the optimal importance density for VaR estimation results from Geweke (1989): half the draws should be made in the 'high loss' subspace S and half the draws outside S, in proportion to the target kernel $p(x)$. Figure 4 shows the optimal importance density for IS estimation of the 99% VaR. Note the bimodality.[11] A mixture of Student-t distributions can approximate such shapes, see e.g. Hoogerheide et al. (2007). Figure 5 shows a mixture of three Student-t distributions providing a reasonable approximation.

The VaR estimation approach proposed in this paper - Quantile Estimation via Rapid Mixtures of t distributions [QERMit] - consists of two steps: (1) approximate the optimal importance density by a certain mixture density $\hat{q}_{opt}(.)$, where we must first compute a preliminary (less precise) estimate of the VaR; (2) apply IS using $\hat{q}_{opt}(.)$. Step (1) should be seen as an *'investment'* of computing time that will easily be *'profitable'*, since far fewer draws from the importance density are required in step (2).

The optimal importance density for IS estimation of the ES does not follow from Geweke (1989). We only mention that this will generally have fatter tails than the optimal importance density for VaR estimation, just like the optimal importance density for estimation of the mean has fatter tails than the target distribution itself (which is optimal for estimating the median). Since we anyway make use of a fat-tailed importance density

[10] This result differs from the case where the exact target density $p^*(x)$ is known. In that case

$$\text{var}\left(\widehat{E[g(X)]}_{IS^*}\right) = \frac{1}{n}\text{var}\left(\frac{p^*(\tilde{X})}{q(\tilde{X})} I\{\tilde{X} \in S\}\right).$$

is minimized by choosing $q^*_{opt}(x) \propto p^*(x) I\{x \in S\}$. In that case the IS estimator's variance is 0 since $\frac{p^*(x)}{q(x)} I\{x \in S\}$ is constant, equal to $E[g(X)]$. This explains why the IS approach for variance reduction of VaR estimation in a Bayesian framework, addressed in this paper, differs substantially from the non-Bayesian applications of e.g. Glass (1999) and Glasserman et al. (2000). In non-Bayesian applications one merely focuses on 'high loss' subspace whereas we focus 'half-half' on the 'high loss' subspace and the rest. Intuitively speaking, we divide our attention 'half-half' over accurately estimating numerator t_1 and denominator t_0, whereas non-Bayesians only need to focus on t_1.

[11] The optimal importance density can also have more than 2 modes. For example, if one shorts a straddle of options, one has high losses for both large decreases and increases of the underlying asset's price. The optimal importance density is trimodal. It is, especially in higher dimensions where one may not directly have a good 'overview' of the target distribution, important to use a flexible method such as the AdMit approach.

– being 'conservative' in the sense of assuring that our importance density does not 'miss' relevant parts of the parameter space – we simply reuse our approximation $\hat{q}_{opt}(.)$ to the optimal importance density for VaR estimation. In the examples this will be shown to work well.

In the extreme case of a Student-t profit-loss distribution with 2 degrees of freedom, the direct sampling estimator of the ES has no finite variance – just like the distribution itself. Whereas the IS estimator using as a Student-t importance density with 1 degree of freedom, a Cauchy density, does have a finite variance. Theoretically, the relative gain in precision from performing IS over direct simulation in estimation of the $100\alpha\%$ ES can therefore be infinite (for any $\alpha \in (0,1)$)!

On the other hand, for VaR estimation the relative gain of precision from IS over direct simulation (of the same number of *independent* draws from the target distribution) is limited (for a given $\alpha \in (0,1)$). From Geweke (1989, Theorem 3) we have that (for a large number of draws n) the variance of $\widehat{E[g(X)]}_{IS}$ with the optimal importance density $q_{opt}(x)$ is approximately $\sigma^2_{IS,opt} \approx \frac{1}{n} E[|g(x) - \bar{g}|]^2$. For $g(x) = I\{X \in S\}$ with $P[X \in S] = 1 - \alpha$ we have $\sigma^2_{IS,opt} \approx \frac{1}{n}[\alpha(1-\alpha) + (1-\alpha)\alpha]^2 = \frac{4}{n}\alpha^2(1-\alpha)^2$. For direct simulation (of *independent* draws) the variance of the estimator $\widehat{E[g(X)]}_{DS}$ results from the Binomial distribution: $\sigma^2_{DS} = \frac{1}{n}\alpha(1-\alpha)$. The gain from IS over direct simulation is therefore:

$$\frac{\sigma^2_{DS}}{\sigma^2_{IS,opt}} \approx \frac{1}{4\,\alpha\,(1-\alpha)}, \tag{14}$$

which is also the relative gain for the VaR estimator's precision (from formulas (12)-(13)). Figure 6 depicts formula (14). For $\alpha = 1/2$ formula (14) reduces to 1: for estimation of the median the optimal importance density is the target density itself. For $\alpha = 0.99$, the α value that is under specific focus in this paper, the relative gain in (14) is equal to 25.25. For $\alpha = 0.95$ and $\alpha = 0.995$ it is equal to 5.26 and 50.25, respectively. It is intuitively clear that the more extreme the quantile, the larger the potential gain is by focusing on the smaller subspace of interest using the IS method. The formula (14) gives an upper boundary for the (theoretical) RNE in IS based estimation of the $100\alpha\%$ VaR.[12] However, one should *not* interpret formula (14) as an upper boundary of the gain from the QERMit approach over the method that we name the 'direct approach', since the 'direct approach' typically yields *serially correlated* draws. If the serial correlation is high, due to non-elliptical shapes or simply due to high correlations between parameters in case of the Gibbs sampler, the relative gain can be much larger than the boundary of

[12] Quoted, *estimated* RNE values may however exceed this boundary due to estimation error.

formula (14). In such cases the RNE of the 'direct approach' may be far below 1.

First, we will briefly consider the Adaptive Mixture of t [AdMit] method which is an important ingredient in our QERMit approach. After that the QERMit approach will be discussed.

3 The Adaptive Mixture of t [AdMit] method

The AdMit approach consists of two steps. First, it constructs a mixture of Student-t distributions which approximates a target distribution of interest. The fitting procedure relies only on a kernel of the target density, so that the normalizing constant is not required. In a second step, this approximation is used as an importance function in importance sampling (or as a candidate density in the independence chain Metropolis-Hastings algorithm) to estimate characteristics of the target density. The estimation procedure is fully automatic and thus avoids the difficult task, especially for non-experts, of tuning a sampling algorithm. In a standard case of importance sampling the candidate density is unimodal. If the target distribution is multimodal then some draws may have huge importance weights or some modes may even be completely missed. Thus, an important problem is the choice of the importance density, especially when little is known a priori about the shape of the target density. The importance density should be close to the target density, and it is especially important that the tails of the candidate should not be thinner than those of the target. Hoogerheide et al. (2007) mention several reasons why mixtures of Student-t distributions are natural candidate densities. First, they can provide an accurate approximation to a wide variety of target densities, with substantial skewness and high kurtosis. Furthermore, they can deal with multi-modality and with non-elliptical shapes due to asymptotes. Second, this approximation can be constructed in a quick, iterative procedure and a mixture of Student-t distributions is easy to sample from. Third, the Student-t distribution has fatter tails than the normal distribution; especially if one specifies Student-t distributions with few degrees of freedom, the risk is small that the tails of the candidate are thinner than those of the target distribution. Finally, Zeevi and Meir (1997) showed that under certain conditions any density function may be approximated to arbitrary accuracy by a convex combination of basis densities; the mixture of Student-t distributions falls within their framework.

The AdMit approach determines the number of mixture components H, the mixing probabilities, the modes and scale matrices of the components in such a way that the

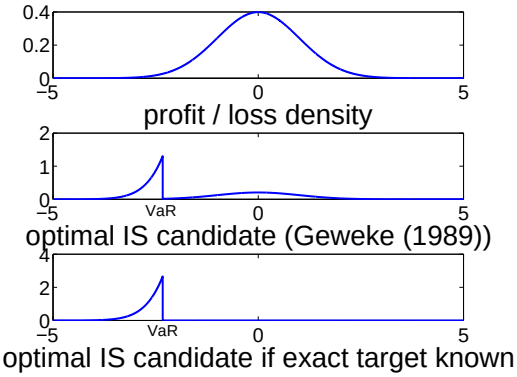


Figure 4: Standard normal profit/loss density and corresponding optimal importance density for IS estimation of 99% VaR in case with only the target density kernel *known and case with exact target density known (see also footnote 10).*

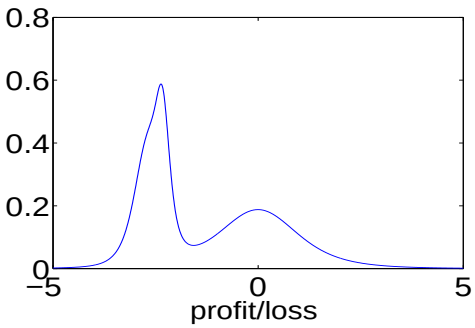


Figure 5: Mixture of three Student-t distributions providing an approximation to the optimal importance density.

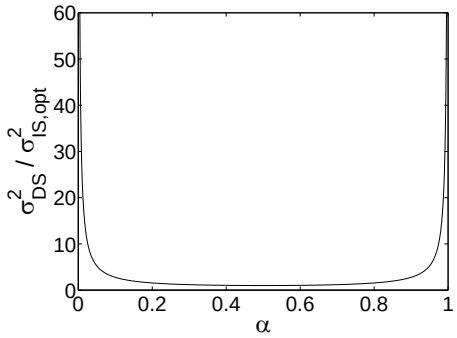


Figure 6: Relative gain in precision of estimation of the $100\alpha\%$ *VaR from IS with the optimal importance density over direct simulation (using the same number of* independent *draws).*

mixture density approximates the target density $p^*(\theta)$ of which we only know a kernel function $p(\theta)$ with $\theta \in R^k$. Typically, $p(\theta)$ will be a posterior density kernel for a vector of model parameters θ. The AdMit strategy consists of the following steps:

(0) Initialization: computation of the mode and scale matrix of the first component, and drawing a sample from this Student-t distribution;

(1) Iterate on the number of components: add a new component that covers a part of the space of θ where the previous mixture density was relatively small, as compared to $p(\theta)$;

(2) Optimization of the mixing probabilities;

(3) Drawing a sample from the new mixture;

(4) Evaluation of importance sampling weights: if the coefficient of variation, the standard deviation divided by the mean, of the weights has converged, then stop. Otherwise, go to step (1).

For more details we refer to Hoogerheide et al. (2007). The R package `AdMit` is available online (Ardia et al. (2008)).

Until now the AdMit approach has been applied to non-elliptical posterior distributions, where the reason for non-elliptical shapes is typically *local non-identification* of certain parameters. Examples are the IV model with weak instruments, or mixture models where one component has weight close to zero. In this paper the focus on the optimal importance density for VaR estimation gives rise to situations of Importance Sampling with non-elliptical target distributions, not only in the parameter space but also in the space of future price processes. Figure 4 already showed a bimodal target density in the case of normally distributed profit/loss.

4 Quantile Estimation via Rapid Mixtures of t approximations [QERMit]

The QERMit approach basically consists of two steps. First the optimal importance or candidate density of Geweke (1989) $q_{opt}(.)$ is approximated by a 'hybrid' mixture of densities $\hat{q}_{opt}(.)$. Second this candidate is used in Importance Sampling. In order to estimate the τ-step ahead $100\alpha\%$ VaR or ES the QERMit algorithm proceeds as follows:

(Step 1) Construct an approximation of the optimal importance density:

(Step 1a) Obtain a mixture of Student-t densities $q_{1,Mit}(\theta)$ that approximates the posterior density – given merely the posterior density kernel – using the AdMit approach.

(Step 1b) Simulate a set of draws θ^i $(i = 1, \ldots, n)$ from the posterior distribution using the independence chain MH algorithm with candidate $q_{1,Mit}(\theta)$. Simulate corresponding future paths $y^{*i} \equiv \{y^i_{T+1}, \ldots, y^i_{T+\tau}\}$ $(i = 1, \ldots, n)$ given parameter values θ^i and historical values $y \equiv \{y_1, \ldots, y_T\}$, i.e. from the density $p(y^*|\theta^i, y)$. Compute a preliminary estimate $\widehat{VaR}_{prelim}$ as the $100(1-\alpha)\%$ quantile of the profit-loss values $PL(y^{*i})$ $(i = 1, \ldots, n)$.

(Step 1c) Obtain a mixture of Student-t densities $q_{2,Mit}(\theta, y^*)$ that approximates the *conditional* joint density of parameters θ and future returns y^* given that $PL(y^*) < \widehat{VaR}_{prelim}$, using the AdMit approach.

(Step 2) Estimate the VaR and/or ES using Importance Sampling with the following mixture candidate density for θ, y^*:

$$\hat{q}_{opt}(\theta, y^*) = 0.5\ q_{1,Mit}(\theta)\ p(y^*|\theta, y) + 0.5\ q_{2,Mit}(\theta, y^*) \tag{15}$$

The reason for the particular term $q_{1,Mit}(\theta)\ p(y^*|\theta, y)$ in this candidate (15) is that the 50% of draws corresponding to the 'whole' distribution of (y^*, θ) can be generated more efficiently by using the density $p(y^*|\theta, y)$ that is specified by the model and approximating merely the posterior $q_{1,Mit}(\theta)$ than by approximating the joint distribution of (y^*, θ). This reduces the dimension of the approximation process, which has a positive effect on the computing time. In step 1b we actually compute a somewhat 'conservative', not-too-negative estimate $\widehat{VaR}_{prelim}$ of the VaR. For a too extreme, too negative $\widehat{VaR}_{prelim}$ may in step 1c yield an approximation of a distribution that covers not all of the 'high loss' region (with $PL < \widehat{VaR}$). This conservative $\widehat{VaR}_{prelim}$ can be based on its NSE, or simply by taking a somewhat higher value of α than the level of interest.[13]

The QERMit algorithm proceeds in an automatic fashion in the sense that it only requires the posterior kernel of θ, (evaluation and simulation from) the density of y^* given θ, and profit/loss as a function of y^* to be programmed. The generation of draws

[13]For this reason it does not make sense to use mixing probabilities $0.5/\alpha$ and $(\alpha - 0.5)/\alpha$ that would lead to an exact 50%-50% division of 'high loss' draws and other draws, instead of 0.5 and 0.5 in (15). Because $\widehat{VaR}_{prelim}$ is 'conservatively' chosen, anyway not entirely all of the candidate probability mass in $q_{2,Mit}(\theta, y^*)$ will be focused on the 'high loss' region.

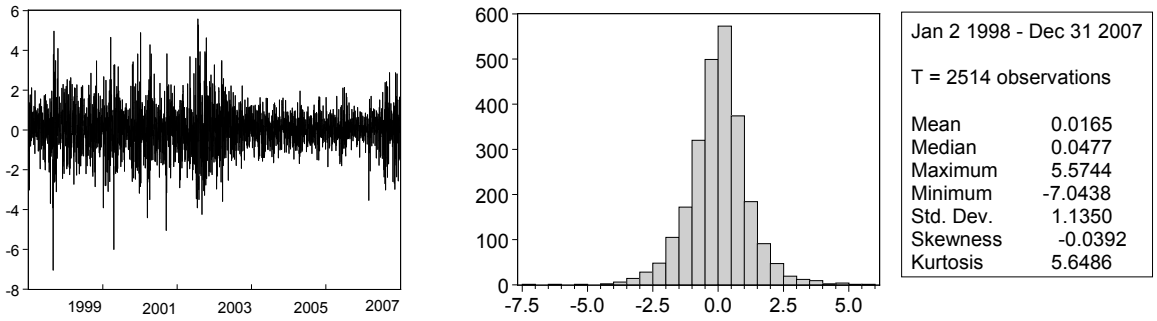


*Figure 7: S&P 500 log-returns (*100× *change of log-index): daily observations from 1998-2007.*

(θ^i, y^{*i}) requires only simulation from Student-t distributions and the model itself, which is performed easily and quickly. Notice that we focus on the distribution of (θ, y^*), whereas the loss only depends on y^*. The obvious reason is that we typically do not have the predictive density of the future path y^* as an explicit density kernel, so that we have to aim at (θ, y^*) of which we know the density kernel

$$p(\theta, y^*|y) \propto p(\theta|y)\, p(y^*|\theta, y) = \pi(\theta)\, p(y|\theta)\, p(y^*|\theta, y)$$

with prior density kernel $\pi(\theta)$.

We will now discuss the QERMit method in a simple, illustrative example of an ARCH(1) model. We consider the 1-day ahead 99% VaR and ES for the S&P500. That is, we assume that during 1 day one will keep a constant long position in the S&P500 index. We use daily observations y_t $(t = 1, \ldots, T)$ on log-return, 100x the change of the logarithm of the closing price, from January 2 1998 to April 14 2000. See Figure 7, in which April 14 2000 corresponds to the second negative 'shock' of approximately -6%. This particular day is chosen for illustrative purposes. We consider the ARCH model (Engle (1982)) for the demeaned series $\tilde{y}_t$:

$$\tilde{y}_t = \varepsilon_t (h_t)^{1/2} \tag{16}$$

$$\varepsilon_t \sim N(0, 1) \tag{17}$$

$$h_t = \alpha_0 + \alpha_1 \tilde{y}_{t-1}^2 \tag{18}$$

We further impose the variance targeting constraint $\alpha_0 = S^2 (1 - \alpha_1)$ with S^2 the sample variance of the y_t $(t = 1, \ldots, T)$, so that we have a model with merely 1 parameter α_1. We assume a flat prior on the interval $[0, 1)$.

Step 1a of the QERMit method is illustrated in Figure 8. The AdMit method constructs a mixture of t approximation to the posterior density - given merely its kernel. It starts with a Student-t distribution around the posterior mode $q_1(\alpha_1)$. After that it searches for the maximum of the weight function $w(\alpha_1) = p(\alpha_1)/q_1(\alpha_1)$, where a new Student-t component $q_2(\alpha_1)$ for the mixture distribution is specified. The mixing probabilities are chosen to minimize the coefficient of variation of the IS weights, yielding $q_{1,Mit}(\alpha_1) = 0.955q_1(\alpha_1) + 0.045q_2(\alpha_1)$, which in this case only provides a minor improvement – a slightly more skewed importance density – over the original Student-t density $q_1(\alpha_1)$.[14] Therefore, convergence is indicated after two steps. Note that we do not need a *perfect* approximation to the posterior kernel, which would generally require a huge amount of computing time. A *reasonably good* approximation is good enough. In this simple example QERMit step 1a took only 1.2 s[15].

The result of the QERMit method's step 1b is illustrated in Figure 9. We generate a set of draws α_1^i $(i = 1, \ldots, 10000)$ using the independence chain MH algorithm with candidate $q_{1,Mit}(\alpha_1)$, and simulate 10000 corresponding draws $\tilde{y}_{T+1}^i$ from the distribution $N(0, S^2 + \alpha_1^i(\tilde{y}_T^2 - S^2)) = N(0, 1.62 + 35.13\,\alpha_1^i)$ since $\tilde{y}_T = -6.06$. The 100th of the ascendingly sorted percentage loss values $PL(y^{*i}) = 100[\exp(y_{T+1}^i/100) - 1]$ — since y_{T+1} is 100x the log-return — or a 'conservatively' chosen less negative value, is then the preliminary VaR estimate $\widehat{VaR}_{prelim}$. In this simple example QERMit step 1b took only 3.4 s.

Figure 10 depicts QERMit step 1c. The top panels show the contour plots of the joint density of (α_1, y_{T+1}) and of $(\alpha_1, \varepsilon_{T+1})$, where it is indicated for which values the PL value falls below $\widehat{VaR}_{prelim}$. We will approximate the joint 'high loss' distribution of $(\alpha_1, \varepsilon_{T+1})$ rather than (α_1, y_{T+1}). The reason is that in general it is easier to approximate the 'high loss' distribution of (θ, ε^*), where ε^* is $\varepsilon^* \equiv \{\varepsilon_{T+1}, \ldots, \varepsilon_{T+\tau}\}$, by a mixture of Student-t distributions than the 'high loss' distribution of (θ, y^*). Especially in GARCH type models where the dependencies (of clustered volatility) between future values $y_{T+1}, \ldots, y_{T+\tau}$ are obviously much more complex than between the independent future values $\varepsilon_{T+1}, \ldots, \varepsilon_{T+\tau}$, it makes step 1c much faster. The 'high loss' subspace of parameters θ and future errors ε^* is somewhat more complex than for θ and y^*; for example, in Figure 10 the border line is described by $\varepsilon_{T+1} = c/\sqrt{1.62 + 35.13\,\alpha_1^i}$ instead of simply $y_{T+1} = c$ (for $c = 100\log(1+\widehat{VaR}_{prelim}/100)$). But it is still preferable to directly focusing on the parameters and future realizations y^*. The bottom panels show the contour plots of the joint 'high

[14] See Hoogerheide et al. (2007) for examples in which this improvement is huge. For the usefulness of the QERMit approach it is not necessary that the posterior has non-elliptical shapes.

[15] An Intel Centrino Duo Core processor was used.

loss' density of $(\alpha_1, \varepsilon_{T+1})$ and its mixture of t approximation. This illustrates that a two-component mixture can provide a useful approximation to the highly skewed shapes that are typically present in such tail distributions. In this simple example QERMit step 1c took only 2.0 s.

Figure 11 shows the result of QERMit step 1, a 'hybrid' mixture approximation $\hat{q}_{opt}(\alpha_1, \varepsilon_{T+1})$ to the optimal importance density $q_{opt}(\alpha_1, \varepsilon_{T+1})$. Table 1 shows the results of QERMit step 2, and compares the QERMit procedure to the 'direct' approach. For the 'direct' approach the series of 10000 profit/loss values is serially correlated, since we use the Metropolis-Hastings algorithm. Therefore, those numerical standard errors make use of the method of Andrews (1991), using a quadratic spectral (QS) kernel and pre-whitening as suggested by Andrews and Monahan (1992). Note the huge difference between the NSE's. The RNE for the 'direct approach' is somewhat smaller than 1 due to the serial correlation in the Metropolis-Hastings draws, whereas the RNE for the QERMit importance density is far above 1. In fact, it is not far from its theoretical boundary of 25.25 (for $\alpha = 0.99$). Notice that the fat-tailed approximation to the optimal importance density for VaR estimation works even better for ES estimation, with an even higher RNE. For a precision of 1 digit (with 95% confidence), i.e. $1.96\, NSE < 0.05$, we require far fewer draws and much less computing time using the QERMit approach than using the 'direct approach'. This is illustrated by Figure 12. In more complicated models the construction of a suitable importance density will obviously require more time. However, this bigger 'investment' of computing time may obviously still be profitable, possibly even more so, as the 'direct' approach will then also require more computing time. In the next section we consider a GARCH model with Student-t errors.

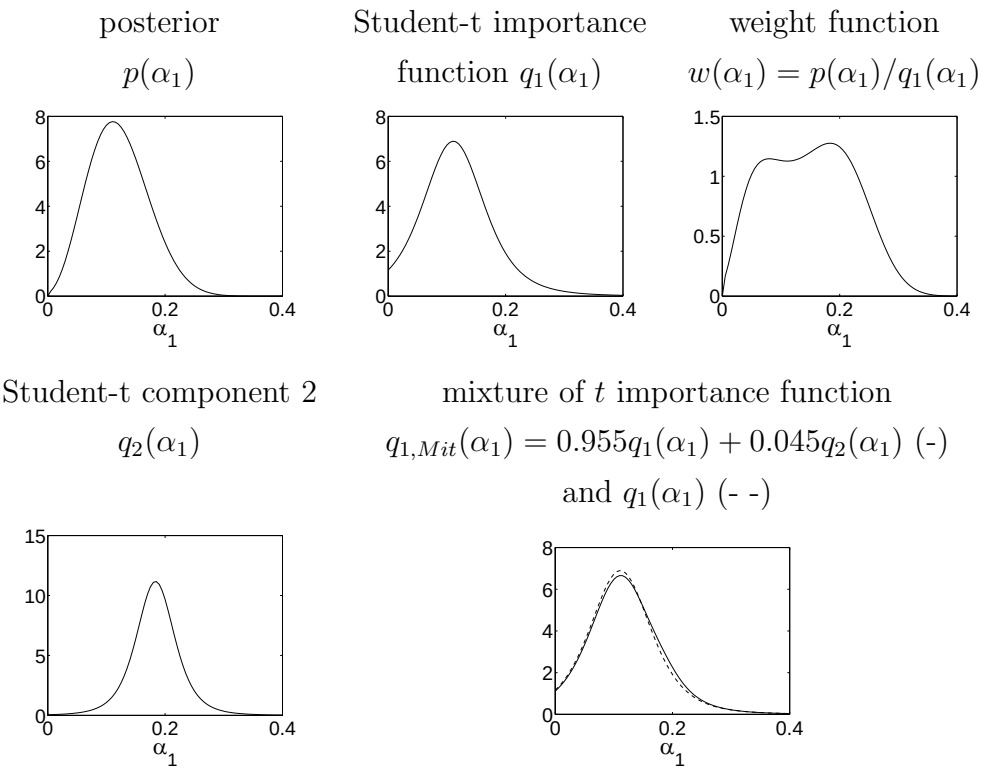


Figure 8: The QERMit method in an illustrative ARCH(1) model for S&P 500. Step 1a: the AdMit method iteratively constructs a mixture of t approximation $q_{1,Mit}(.)$ to the posterior density – given merely its kernel.

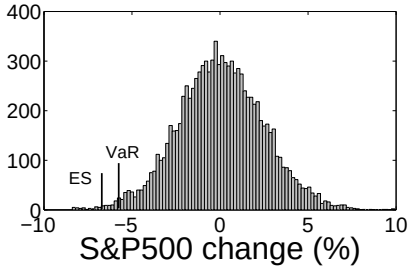


Figure 9: The QERMit method in an illustrative ARCH(1) model for S&P 500. Step 1b: obtain a preliminary estimate of the VaR.

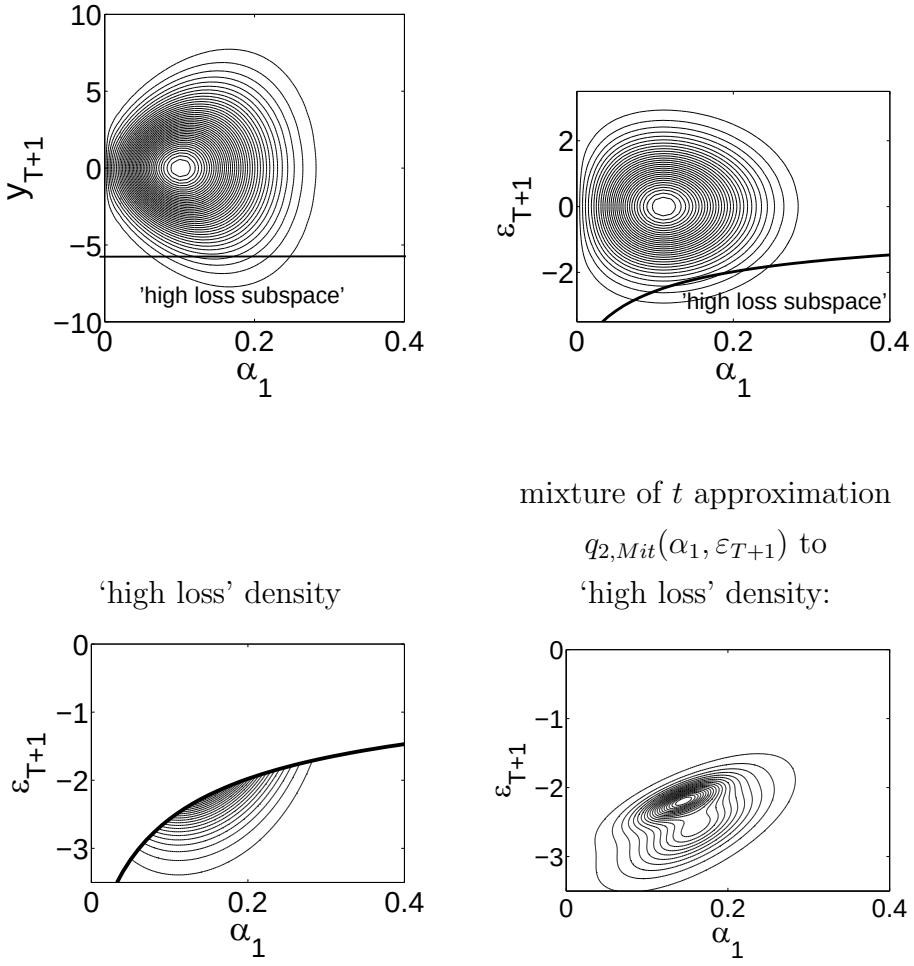


Figure 10: The QERMit method in an illustrative ARCH(1) model for S&P 500: Step 1c: the AdMit method constructs a mixture of t approximation $q_{2,Mit}(.)$ to the joint 'high loss' density of the parameters and the future errors.

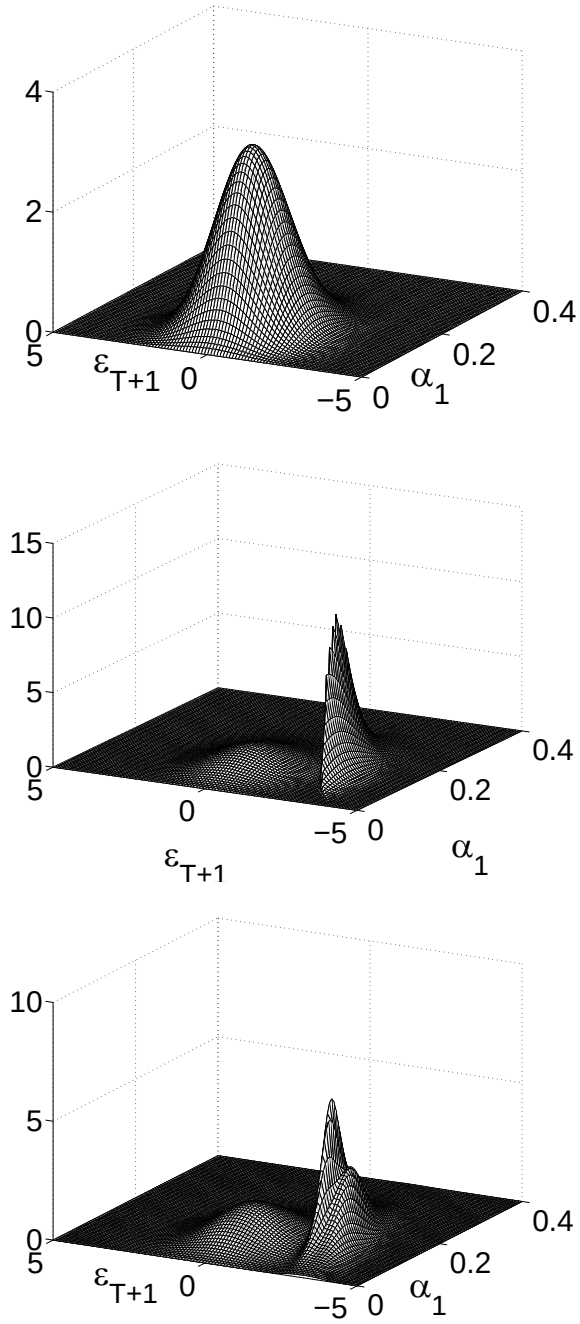


Figure 11: The QERMit method in an illustrative ARCH(1) model for S&P 500. Step 2: use the approximation $\hat{q}_{opt}(.)$ (bottom panel) to the optimal importance density $q_{opt}(.)$ (middle panel) for VaR or ES estimation. The top panel gives the joint density $p(\alpha_1, \varepsilon_{T+1}|y)$.

Table 1: Estimates of 1-day ahead 99% VaR and ES for S&P 500 in ARCH(1) model (for demeaned series under 'variance targeting' – given daily data of January 1 1998 - April 14 2000)

	'Direct' approach: Metropolis-Hastings (Student-t candidate) for parameter draws + direct sampling for future returns paths given parameter draws			**QERMit approach:** Adaptive Importance Sampling using a mixture approximation of the optimal candidate distribution		
	estimate	(NSE)	[RNE]	estimate	(NSE)	[RNE]
99% VaR	-5.744%	(0.099%)	[0.92]	-5.658%	(0.020%)	[22.1]
99% ES	-6.592%	(0.132%)	[0.86]	-6.566%	(0.024%)	[24.9]
total time	3.3 s			10.1 s		
time construction candidate				6.6 s		
time sampling	3.3 s			3.5 s		
draws	10000			10000		
time/draw	0.33 ms			0.35 ms		
required for % VaR estimate with 1 digit of precision (with 95% confidence):						
- number of draws	151216			6408		
- computing time	49.9 s			8.8 s		
required for % ES estimate with 1 digit of precision (with 95% confidence):						
- number of draws	268150			9036		
- computing time	88.5 s			9.8 s		

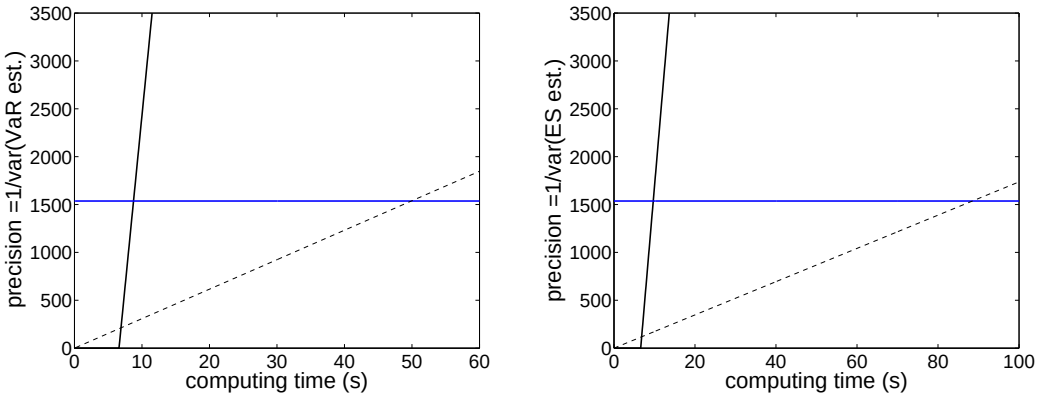


Figure 12: Precision (1/var) of estimated VaR and ES, as a function of the amount of computing time for 'direct' approach (- -) and QERMit approach (-). The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). Here the QERMit approach requires 6.6 seconds to construct an appropriate candidate density, and after that soon generates far more precise VaR and ES estimates.

5 Student-t GARCH model for S&P 500

In this section we consider the 10-day ahead 99% VaR and ES for the S&P500. We use $T = 2514$ daily observations y_t $(t = 1, \ldots, T)$ on log-return from January 2 1998 to December 31 2007. See Figure 7. We consider the GARCH model (Engle (1982), Bollerslev (1986)) with Student-t innovations:[16]

$$\begin{aligned} y_t &= \mu + u_t && (19)\\ u_t &= \varepsilon_t(\varrho h_t)^{1/2} && (20)\\ \varepsilon_t &\sim \text{Student-t}(\nu) && (21)\\ \varrho &\equiv \frac{\nu-2}{\nu} && (22)\\ h_t &= \alpha_0 + \alpha_1 u_{t-1}^2 + \beta h_{t-1} && (23) \end{aligned}$$

where Student-t(ν) is the standard Student-t distribution with ν degrees of freedom, with variance $\frac{\nu-2}{\nu}$. The scaling factor ϱ normalizes the variance of the Student-t distribution such that the innovation u_t has variance h_t. We specify flat priors for μ, α_0, α_1, β on

[16]We also considered the GJR model (Glosten et al. (1993)) with Student-t innovations. However, the results suggested a negative α_1 parameter for positive error values, suggesting that large positive shocks lead to a decrease in volatility as compared with modest positive innovations. This result may be considered counterintuitive and is a separate topic that does not fit within the scope of the current paper.

the parameter subspace with $\alpha_0 > 0$, $\alpha_1 \geq 0$, $\beta \geq 0$. These restrictions guarantee the conditional variance to be positive. For ν we use a proper yet uninformative Exponential prior for $\nu - 2$; the restriction $\nu > 2$ ensures that the conditional variance is finite.[17]

For the model (19)-(23) simulation results are in Table 2. Computing times refer to computations on an Intel Centrino Duo Core processor. The first MH approach uses a Student-t candidate distribution around the maximum likelihood estimator. The AdMit-MH approach in step 1a of the QERMit algorithm requires 16.1 s to construct a candidate distribution, which is a mixture of 2 Student-t distributions in this example. The AdMit-MH draws have a slightly higher acceptance rate and for all parameters but μ a somewhat lower serial correlation in the Markov chain of draws. The differences are however small, reflecting that the contours of the posterior are rather close to the elliptical shapes of the simple Student-t candidate. Figure 13 displays the estimated marginal posteriors from the AdMit-MH output.

We also considered the Griddy-Gibbs [GG] sampler (Ritter and Tanner (1992)). However, this GG approach requires 3734 seconds, i.e. over one hour, for generating a set of 1000 draws (using modest grids of merely 40 points). Further, for the GG approach the serial correlations are worse than for the MH methods, e.g. 0.93 and 0.95 for α_1 and β_1, respectively. Since we focus on the efficient computation of VaR and ES, we discard the GG sampler in the sequel of this paper.

Another alternative simulation method is to extend the approach of Nakatsuma (2000) for the case of Student-t innovations, see Ardia (2008). However, this 'MH within Gibbs' approach makes use of auxiliary candidate distributions that must be constructed in each step of the Gibbs sampler. For both (α_0, α_1) and β this requires two loops per draw, so that four loops occur within the loop over all draws. Summarizing, the extended version of the Nakatsuma (2000) approach is discarded for the same reason as the GG approach: it is much slower than the MH approaches.

We now compare the results of the 'direct' approach and the QERMit method. Figure 15 shows the estimated profit/loss density, the density of the percentage 10-day change in S&P500. Simulation results are in Table 3. In the QERMit approach the construction of the candidate distribution requires 103.8 seconds. This *'investment'* can again be considered quite *'profitable'* as the NSE of the VaR and ES estimators - both based on 10000 draws - are much smaller than the NSE of the estimators using the 'direct' approach. Suppose we want to compute estimates of the VaR and ES (in %) with a precision of 1

[17] Under a flat prior for ν the posterior would be improper, as for $\nu \to \infty$ the likelihood does not tend to 0, but to the likelihood under Gaussian innovations.

Table 2: Simulation results for the GARCH model with Student-t innovations (19)-(23) for S&P 500 log-returns: estimated posterior means, posterior standard deviations between ( ), and serial correlations in the Markov chains of draws between [ ].

	Metropolis-Hastings [MH] (candidate = Student-t)			Metropolis-Hastings [AdMit-MH] (candidate = mixture of 2 Student-t)		
	mean	(st.dev)	[s.c.]	mean	(st.dev)	[s.c.]
μ	0.0483	(0.0169)	[0.4322]	0.0489	(0.0177)	[0.4458]
α_0	0.0086	(0.0034)	[0.5463]	0.0080	(0.0033)	[0.5288]
α_1	0.0713	(0.0114)	[0.5081]	0.0697	(0.0108)	[0.4896]
β	0.9243	(0.0118)	[0.5157]	0.9262	(0.0114)	[0.5106]
ν	10.0953	(1.9717)	[0.6632]	9.8086	(1.6801)	[0.4791]
total time	47.2 s			65.4 s		
time construction candidate				16.1 s		
time sampling	47.2 s			49.3 s		
draws	10000			10000		
time/draw	4.7 ms			4.9 ms		
acceptance rate	53.9%			56.2%		

digit (with 95% confidence), i.e. $1.96NSE < 0.05$, so that we can quote e.g. -8.3% and -10.0% as *the* VaR and ES estimates from this model. In the 'direct' approach we would then require over 500000 draws (over 49 minutes) for the VaR and over 1000000 draws (over 89 minutes). However, in the QERMit approach we would require fewer than 60000 (or 75000) draws in fewer than 8 (10) minutes for the VaR (ES). Figure 14 illustrates that the investment of computing time in an appropriate candidate distribution is indeed very profitable if one desires estimates of VaR and ES with a reasonable precision.

Finally, notice that for the QERMit approach the RNE is much higher than 1, whereas for the 'direct' approach the RNE is somewhat below 1. The reason for the latter is again the serial correlation in the MH sequence of parameter draws.[18] The first phenomenon is in sharp contrast with the potential 'struggle' in importance sampling based Bayesian inference (for estimation of posterior moments of non-elliptical distributions) to have an RNE not too far below 1.

[18] One could consider to use only one in k draws, e.g. $k = 5$. However, this 'thinning' makes no sense in this application since generating a parameter draw (evaluating the posterior density kernel) takes certainly as much time as generating a path of 10 future log-returns. The quality of the draws, i.e. the RNE, would slightly increase, but the amount of computing time per draw would increase substantially.

Figure 13: Estimated marginal posterior distributions in model (19)-(23) for S&P 500 log-returns

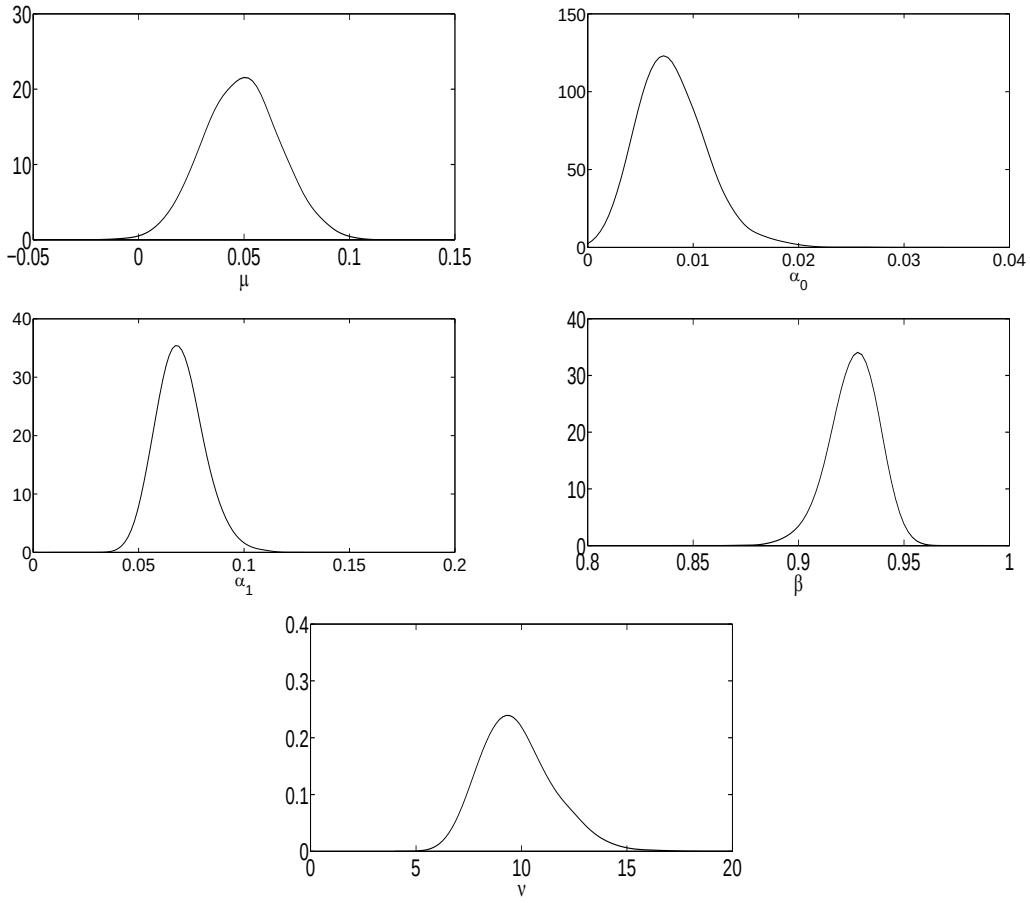


Table 3: Estimates of 10-day ahead 99% VaR and ES for S&P 500 in Student-t GARCH model (given daily data of January 1 1998 - December 2007)

	'Direct' approach: Metropolis-Hastings (Student-t candidate) for parameter draws + direct sampling for future returns paths given parameter draws			**QERMit approach:** Adaptive Importance Sampling using a mixture approximation of the optimal candidate distribution		
	estimate	(NSE)	[RNE]	estimate	(NSE)	[RNE]
99% VaR	-7.92%	(0.19%)	[0.76]	-8.27%	(0.06%)	[7.34]
99% ES	-9.51%	(0.26%)	[0.58]	-9.97%	(0.07%)	[8.11]
total time	51.9 s			165.9 s		
time construction candidate				103.8 s		
time sampling	51.9 s			62.1 s		
draws	10000			10000		
time/draw	5.2 ms			67.2 ms		
required for % VaR estimate with 1 digit of precision (with 95% confidence):						
- number of draws	567648			58498		
- computing time	2946 s (= 49 min. 6 s)			467 s (= 7 min. 47 s)		
required for % ES estimate with 1 digit of precision (with 95% confidence):						
- number of draws	1033980			74010		
- computing time	5366 s (= 89 min. 26 s)			563 s (= 9 min. 23 s)		

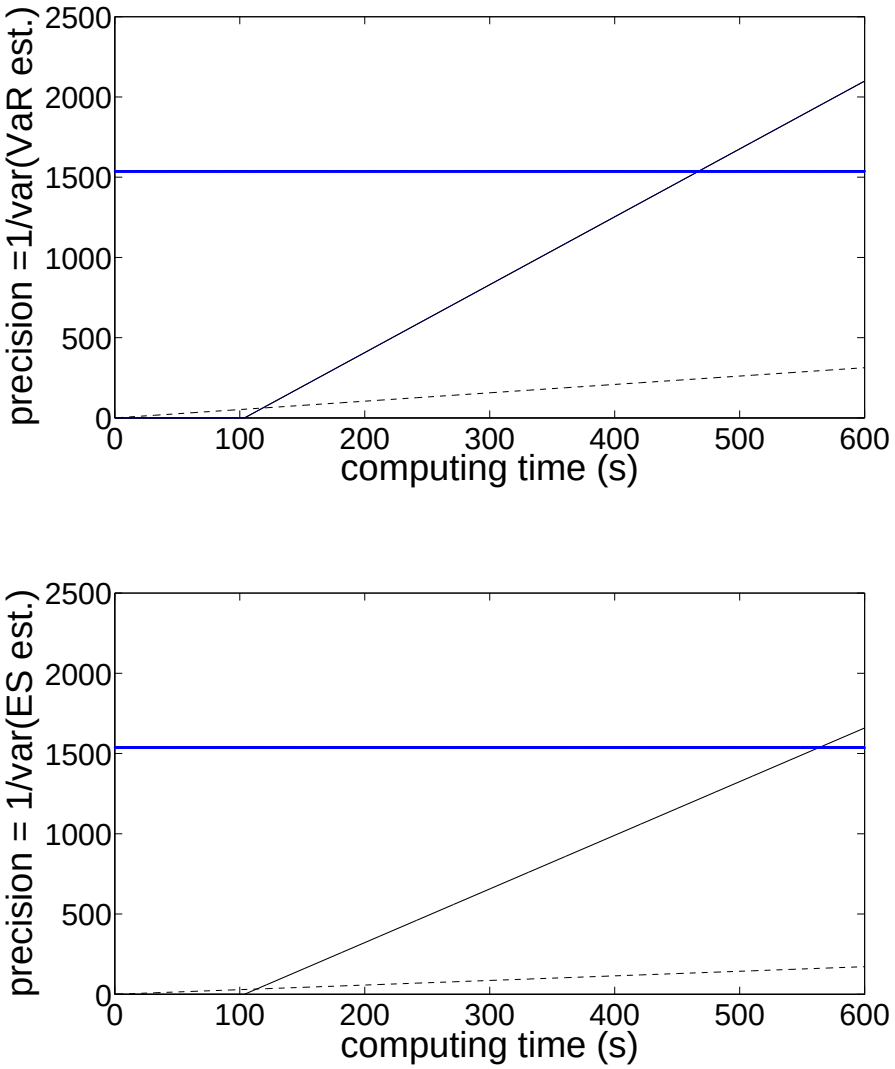
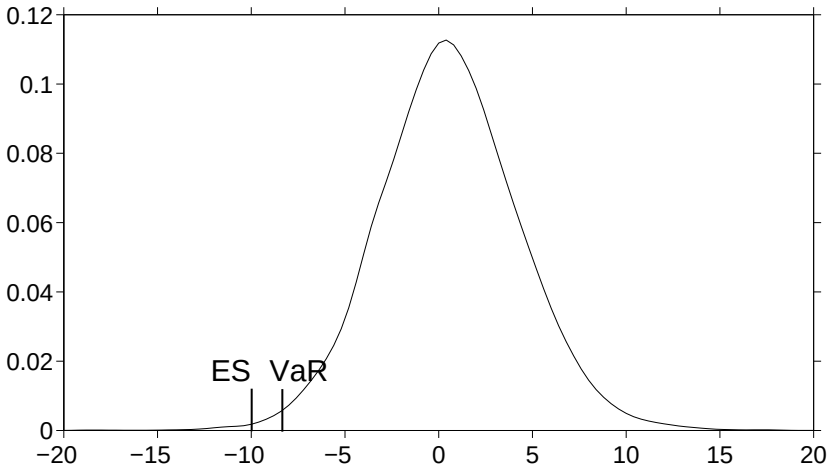


Figure 14: Precision (1/var) of estimated VaR and ES, as a function of the amount of computing time for 'direct' approach (- -) and QERMit approach (-). The horizontal line corresponds to a precision of 1 digit ($1.96NSE \leq 0.05$). Here the QERMit approach requires 103.8 seconds to construct an appropriate candidate density, and after that soon generates far more precise VaR and ES estimates.

Figure 15: Estimated profit/loss density: estimated density of 10-days % change in S&P 500 (for first 10 working days of January 2008) based on GARCH model with Student-t errors estimated on 1998-2007 data



6 Concluding remarks

We conclude that the proposed QERMit approach can yield far more accurate VaR and ES estimates given the same amount of computing time, or equivalently requiring less computing time for the same numerical accuracy. This enables 'real time' decision making on the basis of these risk measures in a simulation-based Bayesian framework based on results with a higher accuracy. In the case of 1-step ahead forecasting with a portfolio of several assets the proposed method can also be useful, as simulation of the future realizations is then typically also required. So, the sensible application of the QERMit method is not restricted to *multi-step* ahead forecasting of VaR and ES.

The examples in this paper only considered the case of a single asset, the S&P 500 index. In that sense, the application was 1-dimensional. However, the 10-days ahead forecasting of a single asset's price has similarities with 1-day ahead forecasting for a portfolio of 10 assets. Further, the subadditivity of the ES measure implies that ES measures of subportfolios may already be useful: adding these yields a conservative risk measure for a whole portfolio. Nonetheless, we intend to investigate portfolios of several assets and report on this in the near future. The application to portfolios of several assets whose returns' distributions are captured in a multivariate GARCH model or a copula is of interest. Having clearly different features than the S&P 500 index, an application to electricity prices would also be of interest.

As another topic for further research we mention the application of the approach for the efficient simulation-based computations in extreme value theory, e.g. efficient computations in the case of Pareto distributions.

References

[1] Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59(3), 817–858.

[2] Andrews, D.W.K., Monahan, J.C., 1992. An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica* 60(4), 953–966.

[3] Ardia, D. (2008). *Financial Risk Management with Bayesian Estimation of GARCH Models.* Lecture Notes in Economics and Mathematical Systems, Vol. 612. Springer.

[4] Ardia D., L.F. Hoogerheide and H.K. van Dijk (2008). The 'AdMit package: Adaptive Mixture of Student-t Distributions. R Foundation for Statistical Computing, URL http://cran.at.r-project.org/web/packages/AdMit/index.html.

[5] Artzner P., F. Delbaen, J.M. Eber, D. Heath (1999), "Coherent Measures of Risk", *Quantitative Finance* 9(3), 203−228.

[6] Basel Committee on Banking Supervision (1995). *An Internal Model-Based Approach to Market Risk Capital Requirements.* The Bank for International Settlements, Basel, Switzerland.

[7] Bollerslev T. (1986), "Generalized Autoregressive Conditional Heteroskedasticity". *Journal of Econometrics* 31(3), 307−327..

[8] Engle R.F. (1982), "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of the United Kingdom inflation". *Econometrica* 50(4), 987−1008.

[9] Geman S. and D. Geman (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721−741.

[10] Geweke J. (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration", *Econometrica*, 57, 1317−1339.

[11] Glass D. (1999), "Importance Sampling Applied to Value at Risk", Master of Science thesis, Department of Mathematics, Courant Institute of Mathematical Sciences, New York University.

[12] Glasserman P., Heidelberger P., Shahabuddin P. (2000), "Variance Reduction Techniques for Estimating Value-at-Risk", *Management Science*, Vol. 46, No. 10. (Oct., 2000), 1349−1364.

[13] Glosten L.R., R. Jaganathan and D.E. Runkle (1993), "On the relation between the expected value and the volatility of the nominal excess return on stocks", *Journal of Finance* 48(5), 1779−1801.

[14] Hammersley J.M. and D.C. Handscomb (1964), *Monte Carlo Methods*, first edition, Methuen, London.

[15] Hastings W.K. (1970), "Monte Carlo Sampling Methods using Markov Chains and their Applications", *Biometrika*, 57, 97−109.

[16] 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(1), 154−180.

[17] Hoogerheide, L.F., H.K. Van Dijk and R.D. Van Oest (2008), "Simulation Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances". Chapter in *Handbook of Computational Econometrics*, Wiley, forthcoming.

[18] Kloek T. and H.K. van Dijk (1978), "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo", *Econometrica*, 46, 1−20.

[19] 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 (3-4), 271−300.

[20] Metropolis N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller and E. Teller (1953), "Equation of State Calculations by Fast Computing Machines", *The Journal of Chemical Physics*, 21, 1087−1092.

[21] Nakatsuma, T. (2000). Bayesian analysis of Markov-GARCH models: A Markov Chain Sampling Approach. *Journal of Econometrics*, 95(1), 57-69.

[22] Ritter C. and M.A. Tanner (1992), "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler", *Journal of the American Statistical Association*, 87, 861−868.

[23] Van Dijk H.K. and T. Kloek (1980), "Further experience in Bayesian analysis using Monte Carlo integration", *Journal of Econometrics*, 14, 307−328.

[24] Van Dijk H.K. and T. Kloek (1984), "Experiments with some alternatives for simple importance sampling in Monte Carlo integration". In: Bernardo, J.M., Degroot, M., Lindley, D., Smith, A.F.M. (Eds.), *Bayesian Statistics*, Vol. 2. Amsterdam, North Holland.

[25] Zeevi A.J. and R. Meir (1997), "Density estimation through convex combinations of densities; approximation and estimation bounds", *Neural Networks*, 10, 99−106.