Available online at www.sciencedirect.com
ScienceDirect
International Journal of Forecasting 26 (2010) 231–247
international journal of forecasting
www.elsevier.com/locate/ijforecast

Bayesian forecasting of Value at Risk and Expected Shortfall using adaptive importance sampling

Lennart Hoogerheide[a,b,*], Herman K. van Dijk[a,b]

[a] *Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam, The Netherlands*

[b] *Tinbergen Institute, Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam, The Netherlands*

Abstract

An efficient and accurate approach is proposed for forecasting the Value at Risk (VaR) and Expected Shortfall (ES) measures in a Bayesian framework. This consists of a new adaptive importance sampling method for the Quick Evaluation of Risk using 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 alternative approaches, in the sense that it produces more accurate VaR and ES estimates given the same amount of computing time, or, equivalently, that it requires less computing time for the same numerical accuracy.

Keywords: Value at Risk; Expected Shortfall; Numerical standard error; Importance sampling; Mixture of Student-*t* distributions; Variance reduction technique

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 simulations, *given a chosen model*. There are several reasons why it is important to compute *accurate* VaR and ES estimates. An underestimation of the risk could obviously cause immense problems for banks and other participants in financial markets (e.g. bankruptcy). On the other hand, an overestimation of the risk may cause one to allocate too much capital as a cushion for risk exposure, which may have a negative effect on profits. Therefore, precise estimates of risk measures are obviously desirable. For simulation-based estimates of VaR and ES, there are also several other issues that play a role. For 'backtesting' or model choice, it is important that this model

* Corresponding author at: Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Burgemeester Oudlaan 50, Rotterdam, The Netherlands.
E-mail address: lhoogerheide@ese.eur.nl (L. Hoogerheide).

0169-2070/$ - see front matter
doi:10.1016/j.ijforecast.2010.01.007

choice be 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 the 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 in risk measures over time are merely caused by simulation noise, this leads to useless fluctuations in positions, leading to extra costs (e.g. transaction costs). Also, when choosing between different risky investment strategies based on a risk-return tradeoff, it is important that the computed risk measures be 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 VaR and 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, and 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 & Frey, 2000, p. 272): non-parametric historical simulation, fully parametric methods based on an econometric model with explicit assumptions on the volatility dynamics and conditional distribution, and methods based on extreme value theory. In this paper we focus on the second group, although some of the ideas could be useful in simulation-based approaches of the third group. 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-day-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 addressed *directly*. However, as was 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 fit directly within the importance sampling estimator's numerical standard error formula from 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 this optimal importance density. 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) for given parameter values and historical asset prices (or returns). It is flexible, so that it can provide useful approximations in a wide range of situations, and is easy to simulate from. The main contribution of this paper is the production of an iterative approach for constructing this 'hybrid' mixture approximation. Just like the traditional approach discussed below, it is 'automatic' in the sense that it only requires a posterior density *kernel* — rather than 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, which involves first constructing an approximation to the optimal importance density and subsequently using this for importance sampling estimation of the VaR or ES, the Quick Evaluation of Risk using Mixture of t approximations (QERMit). The QERMit procedure makes use of the Adaptive Mixture of t (AdMit) approach (see Hoogerheide, Kaashoek, & Van Dijk, 2007), which constructs an approximating mixture of Student-t distributions, given only 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, and subsequently 'corrected' by giving these lower importance weights. Hence, the AdMit approach is merely one of the ingredients for the proposed QERMit approach.

The outline of the paper is as follows. In Section 2 we discuss the computation of numerical standard errors for VaR and ES estimates. Furthermore, we consider the optimal importance sampling density (due to Geweke, 1989), which minimizes the numerical standard error (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. Finally, Section 6 concludes.

2. Bayesian estimation of Value at Risk and Expected Shortfall using importance sampling

In the literature, the VaR is referred to in several different ways. 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 to 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 it is 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 to 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 week period. Even though the VaR has become a standard tool in financial risk management, it has several disadvantages. First, the VaR does not give any information about the potential size of losses that exceed the VaR level. This may lead to overly risky investment strategies that optimize the 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 was indicated by Artzner, Delbaen, Eber, and Heath (1999); that is, it 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 it is a sub-additive, coherent measure. This property of sub-additivity means that the ES of a portfolio (firm) cannot exceed the sum of the ES measures of its subportfolios (departments). Adding these individual ES measures yields a conservative risk measure for the whole portfolio (firm). Because of these advantages of ES over VaR, we consider ES in this paper as well as VaR. For a concise and clear discussion of the VaR and ES measures, refer to Ardia (2008).

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

[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. The quality of our proposed method is not affected by which particular VaR or ES definition is used.

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

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

Step 2. Simulate corresponding future paths $y^{*i} \equiv \{y^i_{T+1}, \ldots, y^i_{T+h}\}$ $(i = 1, \ldots, n)$ 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

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

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

For example, one may generate 10,000 profit/loss values, sort these in ascending order, and take the 100th sorted value as the 99% VaR estimate. In order to intuitively show that this 'direct approach' is not optimal, consider the simple example of the standard normal distribution with $PL \sim \mathcal{N}(0, 1)$. If we then estimate the 99% VaR by simulating 10,000 standard normal variates and taking the 100th sorted value, then this estimate is mostly 'based' on only 100 out of 10,000 draws. 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 a 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.

In importance sampling (IS), due to Hammersley and Handscomb (1964) and introduced to Bayesian econometrics by Kloek and Van Dijk (1978), the expectation $E[g(X)]$ of a certain function $g(.)$ of the random variable $X \in \mathbb{R}^r$ 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 $w(\tilde{X}) = \frac{p(\tilde{X})}{q(\tilde{X})}$, where we only know a *kernel* $p(x)$ of the target density $p^*(x)$ of X: $p(x) \propto p^*(x)$. The consistency of the IS estimator in Eq. (3) is easily seen from

$$E[g(X)] = \frac{\int g(x)p(x)\,\mathrm{d}x}{\int p(x)\,\mathrm{d}x} = \frac{\int g(x)w(x)q(x)\,\mathrm{d}x}{\int w(x)q(x)\,\mathrm{d}x} = \frac{E[w(\tilde{X})g(\tilde{X})]}{E[w(\tilde{X})]}. \tag{4}$$

The IS estimator $\widehat{VaR}_{IS}$ of the $100\alpha\%$ VaR is computed by solving $\widehat{E[g(X)]}_{IS} = 1 - \alpha$ with indicator function $g(X) = I\{PL(X) \leq \widehat{VaR}_{IS}\}$, since $\Pr[PL(X) \leq c] = E[I\{PL(X) \leq c\}]$.[2] 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 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 S_{k+1} is the smallest value with $S_{k+1} > 1 - \alpha$.

[2] In our IS approach, X contains the model parameters θ and the future error terms of the process of asset returns (or prices), which together determine the future returns (or prices). This choice will be explained in Section 4.

The IS estimator $\widehat{ES}_{IS}$ of the $100\alpha\%$ ES is subsequently computed as

$$\widehat{ES}_{IS} = \sum_{j=1}^{k} w^*(\tilde{X}^{(j)})\, PL(\tilde{X}^{(j)}),$$

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).

2.1. Numerical standard errors

Geweke (1989) provides formulas for the numerical accuracy of the IS estimator $\widehat{E[g(X)]}_{IS}$ in Eq. (3). See also Hoogerheide, Van Dijk, and Van Oest (2009, chap. 7) for a discussion of the numerical accuracy of $\widehat{E[g(X)]}_{IS}$. It holds for large numbers of draws n and under the mild regularity conditions reported by Geweke (1989) that $\widehat{E[g(X)]}_{IS}$ has approximately the normal distribution $\mathcal{N}(E[g(X)], \sigma^2_{\text{IS}})$. 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 $(\widehat{E[g(X)]}_{IS} - 1.96\,\hat{\sigma}_{\text{IS}}, \widehat{E[g(X)]}_{IS} + 1.96\,\hat{\sigma}_{\text{IS}})$.

The numerical standard error (NSE) $\hat{\sigma}_{\text{IS,VaR}}$ of the IS estimator of the VaR or ES does not follow directly from the NSE for $\widehat{E[g(X)]}_{IS}$, as neither VaR nor ES are unconditional expectations $E[g(X)]$ for a random variable X of which we know the density kernel.[3] For the NSE of the VaR estimator, we make use of the delta rule. We have

$$\Pr[PL(X) \leq \widehat{VaR}] \approx \Pr[PL(X) \leq VaR] + \left.\frac{\partial \Pr[PL(X) \leq c]}{\partial c}\right|_{c=VaR} \left(\widehat{VaR} - VaR\right) \Rightarrow (5)$$

$$1 - \alpha \approx \widehat{\Pr}[PL(X) \leq VaR] + \hat{p}_{PL}(VaR)\left(\widehat{VaR} - VaR\right) \Rightarrow \quad (6)$$

[3] Only if we knew the true value of the VaR with certainty would the estimation of the ES reduce to the 'standard' situation of IS estimation of the expectation of $PL(X)$, where X has the 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 the draws $\tilde{X}_i$ with $PL(\tilde{X}_i) \leq VaR$ for $VaR = \widehat{VaR}$; and (2) the variation in $\widehat{VaR}$.

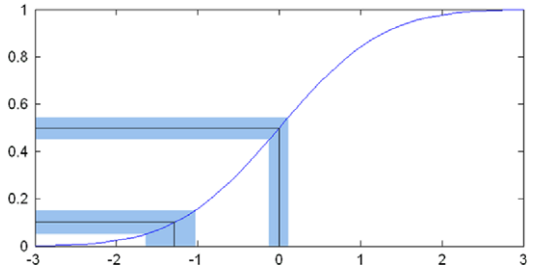


Fig. 1. Illustration of the numerical standard error of the IS estimator for a VaR, a quantile of a profit/loss $PL(X)$ function of a random vector X. The uncertainty on $\Pr[PL(X) \leq c]$ for $c = \widehat{VaR}_{IS}$ (on the vertical axis) is translated to the uncertainty on $\widehat{VaR}_{IS}$ (on 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.

$$\text{var}(\widehat{VaR}) \approx \frac{\text{var}\left(\widehat{\Pr}[PL(X) \leq VaR]\right)}{(\hat{p}_{PL}(VaR))^2}, \quad (7)$$

where Eq. (6) results from Eq. (5) by substituting estimates for $\Pr[PL(X) \leq \widehat{VaR}]$, $\Pr[PL(X) \leq VaR]$ and $p_{PL}(VaR)$, where $p_{PL}(VaR)$ is the density of $PL(X)$ evaluated at VaR and $\widehat{\Pr}[PL(X) \leq \widehat{VaR}] = 1-\alpha$, since this equality defines $\widehat{VaR}$. Substituting the realized value of $\widehat{VaR}_{IS}$ for VaR into Eq. (7) and taking the square root yields the numerical standard error for $\widehat{VaR}_{IS}$:

$$\hat{\sigma}_{\text{IS,VaR}} = \frac{\hat{\sigma}_{\text{IS,Pr}[PL \leq \widehat{VaR}_{IS}]}}{\hat{p}_{PL}(\widehat{VaR}_{IS})}. \quad (8)$$

The numerical standard error $\hat{\sigma}_{\text{IS,Pr}[PL \leq \widehat{VaR}_{IS}]}$ for the IS estimator of the probability $\Pr[PL(X) \leq c]$ for $c = \widehat{VaR}_{IS}$ follows directly from the NSE for $\widehat{E[g(X)]}_{IS}$ of Geweke (1989) with $g(x) = I\{PL(x) \leq c\}$. 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 $\hat{\sigma}_{IS,\widehat{VaR}}$. Alternatively, one can use a kernel estimator of the profit/loss density at $c = \widehat{VaR}_{IS}$. Fig. 1 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 the fact that if the VaR were 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 follow directly from Geweke (1989). Since we do have 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 proceed as follows to estimate the density for the ES estimator.

Step 1. Construct a grid of VaR values, e.g. on the interval $[\widehat{VaR}_{IS} - 4\hat{\sigma}_{IS,\widehat{VaR}}, \widehat{VaR}_{IS} + 4\hat{\sigma}_{IS,\widehat{VaR}}]$.
Step 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.
Step 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 standard error of $\widehat{ES}_{IS}$ is now obtained as the standard deviation of the estimated density $p(\widehat{ES}_{IS})$. Fig. 2 illustrates the procedure for the estimation of the ES estimator's density in the case of $\mathcal{N}(0, 1)$ distributed profit/loss. The left panels show the case of direct sampling, where the density $p(\widehat{ES}_{IS}|VaR)$ clearly has a higher variance for more negative, more extreme VaR values, resulting in a skewed density $p(\widehat{ES}_{IS})$. The reason for this 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 the Student-t importance density (with 10 degrees of freedom), where the density $p(\widehat{ES}_{IS}|VaR)$ has a variance which is hardly any higher for more extreme VaR values. The Student-t distribution's fat tails ensure that the estimated ES is based on many draws for extreme VaR values as well. This example already reflects one advantage of IS (with an importance density having fatter tails than the target distribution) over a 'direct approach': IS results in a lower NSE, and especially less downward uncertainty on the ES — with a correspondingly lower risk of substantially underestimating the risk.

2.2. Optimal importance density

The optimal importance distribution for the IS estimation of $\bar{g} \equiv E[g(X)]$ for a given kernel $p(x)$ of the 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 the 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 be devised. Geweke (1989) further notes that this result reflects the fact 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), which is the ratio of (an estimate of) the variance of an estimator based on direct sampling to the IS estimator's estimated variance (with the same number of draws), exceeds 1.[4] An interesting result is the case where $g(x)$ is an indicator function $I\{x \in S\}$ for a subspace S, so that $E[g(X)] = \Pr[X \in S] = \bar{p}$. Then the optimal importance density is given by

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

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

with c being a constant. Then $\int_{x\in S} q_{opt}(x)\mathrm{d}x = \int_{x\notin S} q_{opt}(x)\mathrm{d}x = 1/2$.[5] That is, half of the draws should be made in S and half outside S, in proportion to the target kernel $p(x)$ in both cases.

From Eq. (8) it can be 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

[4] The RNE is an indicator of the efficiency of the chosen importance function; if the target and importance densities 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).

[5] This is easily seen from $\int_{x\in S} q_{opt}(x)\mathrm{d}x + \int_{x\notin S} q_{opt}(x)\mathrm{d}x = 1$ and the equality of

$$\int_{x\in S} q_{opt}(x)\mathrm{d}x = c\,(1-\bar{p}) \int_{x\in S} p^*(x)\mathrm{d}x = c\,(1-\bar{p})\,\bar{p}$$

and

$$\int_{x\notin S} q_{opt}(x)\mathrm{d}x = c\,\bar{p} \int_{x\notin S} p^*(x)\mathrm{d}x = c\,\bar{p}\,(1-\bar{p}).$$

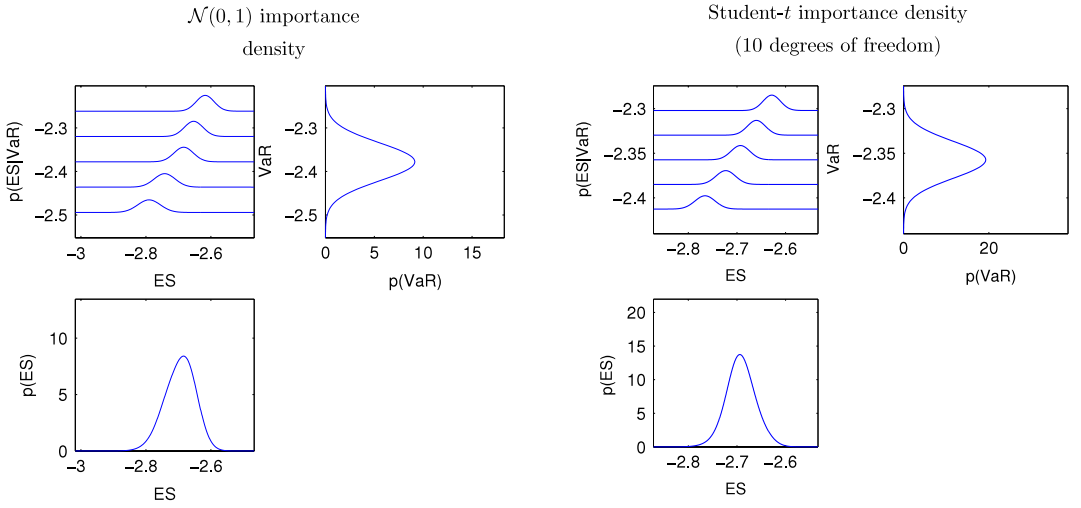


Fig. 2. Example of a standard normally distributed profit/loss PL: illustration of the estimation of the ES estimator's density using either a $\mathcal{N}(0,1)$ importance density (left) — corresponding to the case of direct sampling — or a Student-t importance density (right). The top-left panel gives 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 $p(\widehat{ES}_{IS})$.

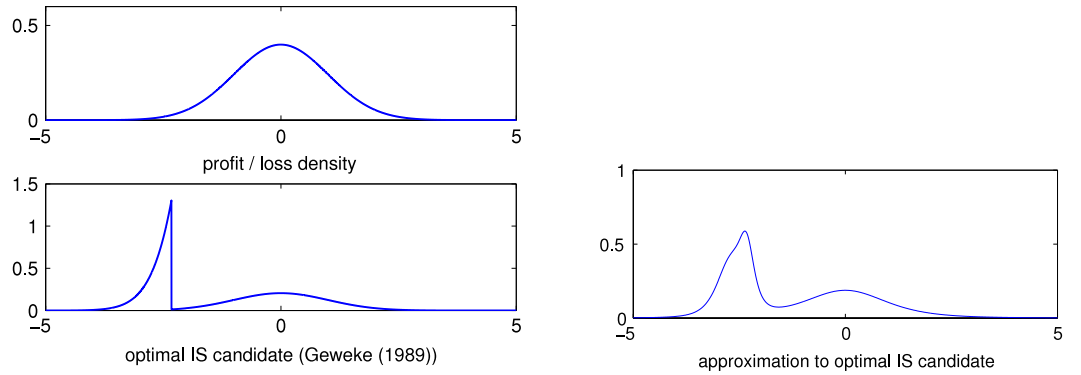


Fig. 3. Example of a standard normally distributed profit/loss PL (on the horizontal axis): the profit/loss density (top left) and the importance density for IS estimation of the 99% VaR that is optimal in the typical Bayesian case with only the target density *kernel* known (bottom left). Also given is a mixture of three Student-t distributions, providing an approximation to the optimal importance density (right).

$g(x) = I\{x \in S\}$, where S is the subspace with the $100(1-\alpha)\%$ lowest values $PL(X)$; hence the optimal importance density for VaR estimation results from Geweke (1989): half of the draws should be made in the 'high loss' subspace S and half of the draws outside S, in proportion to the target kernel $p(x)$. Intuitively, a precise numerator of Eq. (3) requires draws from S, whereas the denominator requires draws from the whole space of X. If the exact target density $p^*(x)$ were known, rather than merely a kernel $p(x)$, then the middle expression of the IS (Eq. (3)) would consist of only the numerator, so that it would be optimal to only simulate draws in the 'high loss' subspace S.

For the case of $\mathcal{N}(0,1)$ distributed profit/loss, the bottom left panel of Fig. 3 shows the optimal importance density for IS estimation of the 99% VaR. Note the bimodality.[6] A mixture of Student-t distributions can approximate such shapes, see e.g. Hoogerheide

[6] 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 large increases of the underlying asset's price, and the optimal importance density is then trimodal. Especially in higher dimensions, where one may not directly have a good 'overview' of the target distribution, it is important to use a flexible method such as the AdMit approach.

et al. (2007). The right panel of Fig. 3 shows a mixture of three Student-t distributions providing a reasonable approximation.

The VaR estimation approach proposed in this paper — Quick Evaluation of Risk using Mixture of t approximations (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; and (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 it will generally have fatter tails than the optimal importance density for VaR estimation, just as the optimal importance density for the estimation of the mean has fatter tails than the target distribution itself (which is optimal for estimating the median). Since we make use of a fat-tailed importance density in any case — being 'conservative' in the sense of ensuring that our importance density does not 'miss' relevant parts of the parameter space — we simply use our approximation $\hat{q}_{opt}(.)$ of 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 a Student-t importance density with 1 degree of freedom, a Cauchy density, does have a finite variance. This example shows that the relative gain in precision from performing IS over a direct simulation in the estimation of the $100\alpha\%$ ES can be infinite (for any $\alpha \in (0, 1)$)!

On the other hand, for VaR estimation the relative gain in precision from using IS rather than 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 know 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 $\Pr[X \in S] = 1 - \alpha$, 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 using IS rather than direct simulation is therefore

$$\frac{\sigma^2_{DS}}{\sigma^2_{IS,opt}} \approx \frac{1}{4\,\alpha\,(1-\alpha)}, \tag{9}$$

which is also the relative gain for the VaR estimator's precision (from Eqs. (7)–(8)). For $\alpha = 1/2$, Eq. (9) reduces to 1: for the estimation of the median, the optimal importance density is the target density itself. For $\alpha = 0.99$, the α value that is specific focussed on in this paper, the relative gain in Eq. (9) 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 by focusing on the smaller subspace of interest using the IS method. Eq. (9) gives an upper boundary for the (theoretical) RNE in IS based estimation of the $100\alpha\%$ VaR.[7] However, one should *not* interpret Eq. (9) 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 a low Metropolis-Hastings acceptance rate in the case of non-elliptical shapes or simply due to high correlations between parameters in the case of the Gibbs sampler, the relative gain can be much larger than the boundary of Eq. (9). In such cases, the RNE of the 'direct approach' may be far below 1.

3. The Adaptive Mixture of t (AdMit) method

In this section we briefly consider the AdMit approach, which is an important ingredient of our QERMit 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,

[7] This is an asymptotic result. For finite numbers of draws, the quoted *estimated* RNE values may exceed this boundary due to estimation error.

this approximation is used as an importance function in importance sampling (or as a candidate density in the independence chain Metropolis-Hastings algorithm) for estimating the 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 the 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 about the shape of the target density a priori. 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 accurate approximations of a wide variety of target densities, including 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 by 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; the risk that the tails of the candidate will be thinner than those of the target distribution is small, especially if one specifies Student-t distributions with few degrees of freedom. Finally, Zeevi and Meir (1997) showed that under certain conditions any density function may be approximated to an arbitrary level of accuracy using a convex combination of basis densities, and the mixture of Student-t distributions falls within their framework.

The AdMit approach determines the number of mixture components, the mixing probabilities, and the modes and scale matrices of the components in such a way that the mixture density approximates the target density $p^*(\theta)$, of which we only know a kernel function $p(\theta)$ with $\theta \in \mathbb{R}^k$. Typically, $p(\theta)$ will be a posterior density kernel for a vector of model parameters θ. The AdMit strategy consists of the following steps:

Step 0. Initialization: compute the mode and scale matrix of the first component, and draw a sample from this Student-t distribution;

Step 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, relative to $p(\theta)$;

Step 2. Optimize the mixing probabilities;

Step 3. Draw a sample from the new mixture;

Step 4. Evaluate the importance sampling weights: if the coefficient of variation of the weights, the standard deviation divided by the mean, has converged, then stop. Otherwise, return to step 1.

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

4. Quick Evaluation of Risk using Mixture 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 density is used for the importance sampling estimation of the VaR or ES. In order to estimate the h-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 only 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 $y^{*i} \equiv \{y^i_{T+1}, \ldots, y^i_{T+h}\}$ $(i = 1, \ldots, n)$ given the 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 $PL(y^*) \leq \widehat{VaR}_{prelim}$, using the AdMit approach.

Step 2. Estimate the VaR or ES using importance sampling with the following mixture candidate density for θ, y^*:

$$\hat{q}_{opt}(\theta, y^*) = 0.5\, q_{1,Mit}(\theta)\; p(y^*|\theta, y) + 0.5\, q_{2,Mit}(\theta, y^*), \quad (10)$$

where the weights 0.5 reflect the optimal 50%–50% division between 'high loss' draws and other draws.

The reason for the particular term $q_{1,Mit}(\theta)\; p(y^*|\theta, y)$ in this candidate (Eq. (10)) 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 only approximating the posterior of θ by $q_{1,Mit}(\theta)$ than by approximating the joint distribution of (y^*, θ). This reduces the dimensions 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, since an overly extreme, negative $\widehat{VaR}_{prelim}$ may yield an approximation of the distribution in step 1c that does not cover all of the 'high loss' region (with $PL \leq VaR$). This conservative $\widehat{VaR}_{prelim}$ can be based on its NSE, or simply take a somewhat higher value of α than the level of interest.[8]

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

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

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&P 500. That is, we assume that during a given day, one will keep a constant long position in the S&P 500 index. We use daily observations y_t $(t = 1, \ldots, T)$ on the log-return, 100 × the change in the logarithm of the closing price, from January 2, 1998, to April 14, 2000. See Fig. 4, 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 $\tilde{y}_t$:

$$\begin{aligned} \tilde{y}_t &= \varepsilon_t\, (h_t)^{1/2}, \\ \varepsilon_t &\sim \mathcal{N}(0, 1), \\ h_t &= \alpha_0 + \alpha_1 \tilde{y}_{t-1}^2. \end{aligned} \quad (11)$$

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

Step 1a of the QERMit method is illustrated in Fig. 5. The AdMit method constructs a mixture of Student-t distributions that approximates the posterior density, given only its kernel. It starts with a Student-t density around the posterior mode $q_1(\alpha_1)$, then 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, in this case yielding $q_{1,Mit}(\alpha_1) = 0.955\, q_1(\alpha_1) + 0.045\, q_2(\alpha_1)$, which only provides a minor improvement — a slightly more skewed importance density — over the original Student-t density $q_1(\alpha_1)$.[9] Therefore, convergence is achieved after two steps. Note that we do not need an (almost) *perfect* approximation to the posterior kernel, which would generally require a huge amount of computing time. A *reasonably good*

[8] 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 if $\widehat{VaR}_{prelim}$ and $q_{2,Mit}(\theta, y^*)$ were perfect, rather than 0.5 and 0.5, in Eq. (10). Since $\widehat{VaR}_{prelim}$ is chosen 'conservatively' and $q_{2,Mit}(\theta, y^*)$ is an approximation, not all of the candidate probability mass in $q_{2,Mit}(\theta, y^*)$ will be focused on the 'high loss' region in any case.

[9] See Hoogerheide et al. (2007) for examples in which this improvement is huge. For the QERMit approach to be useful, it is not necessary for the posterior to have non-elliptical shapes.

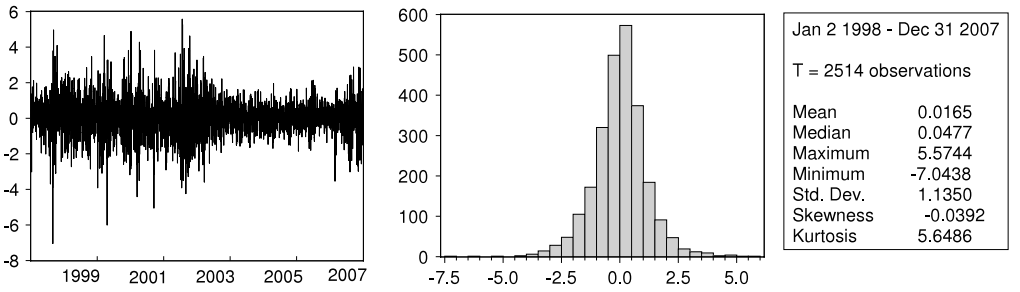


Fig. 4. S&P 500 log-returns (100 × change of log-index): daily observations from the period 1998–2007.

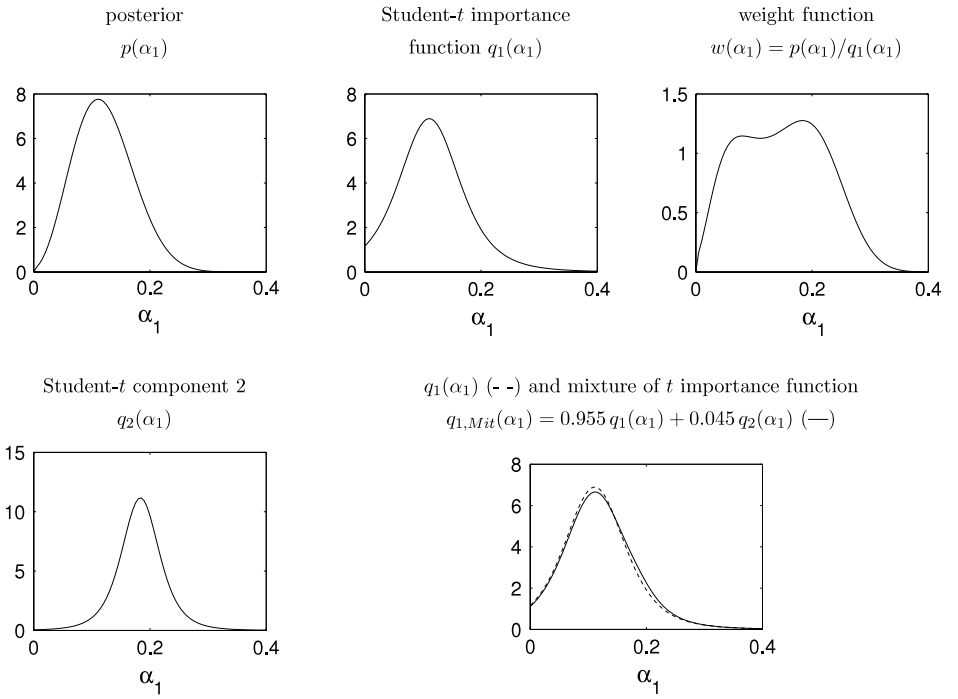


Fig. 5. The QERMit method in an illustrative ARCH(1) model for S&P 500. Step 1a: the AdMit method iteratively constructs a mixture of t approximation $q_{1,Mit}(.)$ to the posterior density, given only its kernel.

approximation is good enough. In this simple example, QERMit step 1a took only 1.2 s.[10]

In the QERMit method's step 1b, we generate a set of $n = 10{,}000$ draws α_1^i $(i = 1, \ldots, n)$ using the independence chain MH algorithm with candidate $q_{1,Mit}(\alpha_1)$, and simulate n corresponding draws $\tilde{y}_{T+1}^i$ from the distribution $\mathcal{N}(0, S^2 + \alpha_1^i(\tilde{y}_T^2 - S^2)) = \mathcal{N}(0, 1.62 + 35.13\,\alpha_1^i)$, 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 100 × 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.

The left panels of Fig. 6 depict QERMit step 1c. We approximate the joint 'high loss' distribution of $(\alpha_1, \varepsilon_{T+1})$ rather than (α_1, y_{T+1}). The reason for this is that in general it is easier to approximate the 'high loss' distribution of (θ, ε^*), where $\varepsilon^* \equiv$

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

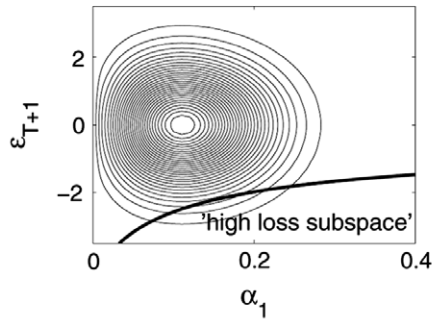
Table 1
Estimates of 1-day-ahead 99% VaR and ES for S&P 500 in the ARCH(1) model (for demeaned series under 'variance targeting', given daily data from January 1, 1998–April 14, 2000).

	'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
Time: - total		3.3 s			10.1 s	
- construction candidate					6.6 s	
- sampling		3.3 s			3.5 s	
Draws		10,000			10,000	
Time/draw		0.33 ms			0.35 ms	
	Required for % estimate with 1 digit of precision (with 95% confidence)					
For 99% VaR:						
- number of draws		151,216			6408	
- computing time		49.9 s			8.8 s	
For 99% ES:						
- number of draws		268,150			9036	
- computing time		88.5 s			9.8 s	

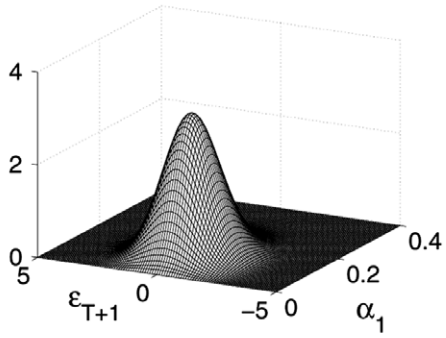
$\{\varepsilon_{T+1}, \ldots, \varepsilon_{T+h}\}$, by a mixture of Student-t distributions than the 'high loss' distribution of (θ, y^*). This makes step 1c much faster, especially in GARCH-type models where the dependencies (of the clustered volatility) between future values $y_{T+1}, \ldots, y_{T+h}$ are obviously much more complex than between the independent future values $\varepsilon_{T+1}, \ldots, \varepsilon_{T+h}$. The 'high loss' subspace of parameters θ and future errors ε^* is somewhat more complex than for θ and y^*. For example, in Fig. 6 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 focusing directly on the parameters and future realizations y^*. The middle and bottom left panels show the contour plots of the joint 'high loss' density of $(\alpha_1, \varepsilon_{T+1})$ and its mixture of t approximation. This shows that a two-component mixture can provide useful approximations of the highly skewed shapes that are typically present in such tail distributions. In this simple example, QERMit step 1c took only 2.0 s.

The right panels of Fig. 6 show 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 10,000 profit/loss values is serially correlated, since we use the Metropolis-Hastings algorithm. Therefore, the numerical standard errors make use of the method of Andrews (1991), using a quadratic spectral kernel and pre-whitening, as suggested by Andrews and Monahan (1992). Note the huge difference between the NSEs. 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'. 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.

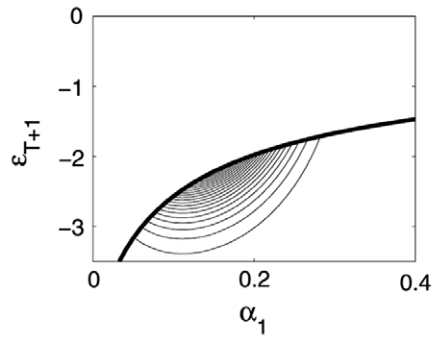
joint density $p(\alpha_1, \varepsilon_{T+1}|y)$:



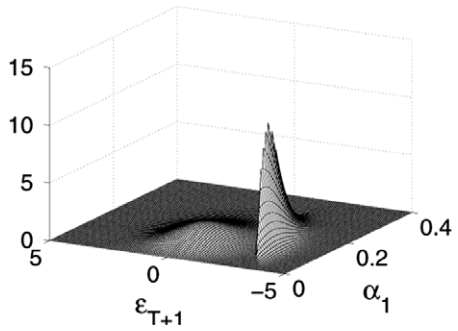
joint density $p(\alpha_1, \varepsilon_{T+1}|y)$:



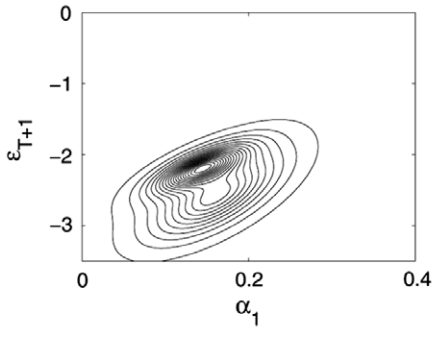
'high loss' density
$p(\alpha_1, \varepsilon_{T+1}|y, \varepsilon_{T+1}\sqrt{1.62 + 35.13\,\alpha_1} \leq c)$:



optimal importance
density $q_{opt}(\alpha_1, \varepsilon_{T+1})$:



mixture of t approximation $q_{2,Mit}(\alpha_1, \varepsilon_{T+1})$
to 'high loss' density:



approximation $\hat{q}_{opt}(\alpha_1, \varepsilon_{T+1})$
to optimal importance density:

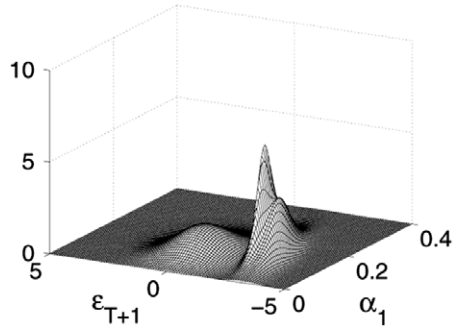


Fig. 6. The QERMit method in an illustrative ARCH(1) model for S&P 500. Step 1c: the AdMit method constructs a mixture of t approximation $q_{2,Mit}(.)$ to the joint 'high loss' density of the parameters and the future errors (contour plots in left panels). Step 2: use the approximation $\hat{q}_{opt}(.)$ (bottom right panel) to the optimal importance density $q_{opt}(.)$ (middle right panel) for VaR or ES estimation.

Table 2
Simulation results for the GARCH model with Student-t innovations (Eqs. (12)–(16)) for S&P 500 log-returns: estimated posterior means, posterior standard deviations and serial correlations in the Markov chains of draws.

	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		10,000			10,000	
Time/draw		4.7 ms			4.9 ms	
Acceptance rate		53.9%			56.2%	

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&P 500. We use $T = 2514$ daily observations y_t $(t = 1, \ldots, T)$ on log-returns from January 2, 1998, to December 31, 2007; see Fig. 4. We consider the GARCH model (Bollerslev, 1986; Engle, 1982) with Student-t innovations:[11]

$$y_t = \mu + u_t \tag{12}$$

$$u_t = \varepsilon_t (\varrho h_t)^{1/2} \tag{13}$$

$$\varepsilon_t \sim \text{Student-}t(\nu) \tag{14}$$

$$\varrho \equiv \frac{\nu - 2}{\nu} \tag{15}$$

$$h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + \beta h_{t-1} \tag{16}$$

where Student-$t(\nu)$ is the standard Student-t distribution with ν degrees of freedom, with variance $\frac{\nu}{\nu-2}$ for $\nu > 2$. 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, and β on the parameter subspace with $\alpha_0 > 0$, $0 \leq \alpha_1 \leq 1$, $0 \leq \beta \leq 1$. These restrictions guarantee that the conditional variance will be positive. We use a proper yet uninformative exponential prior distribution for $\nu - 2$; the restriction $\nu > 2$ ensures that the conditional variance is finite.[12]

For the model (12)–(16), 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 two 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. However, the differences are small, reflecting the fact that the contours of the posterior are rather close to the elliptical shapes of the simple Student-t candidate.

We now compare the results of the 'direct approach' and the QERMit method. Fig. 7 shows the estimated profit/loss density, the density of the percentage 10-day change in the S&P 500 index. Simulation results are in Table 3. In the QERMit approach the construction of the candidate distribu-

[11] We also considered the GJR model (Glosten, Jaganathan, & Runkle, 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 relative to modest positive innovations. This result may be considered counterintuitive, and is a separate topic which does not fit within the scope of the current paper.

[12] 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 — see Bauwens and Lubrano (1998).

Table 3
Estimates of 10-day-ahead 99% VaR and ES for the S&P 500 in the Student-t GARCH model (given daily data from the period January 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
Time: - total		51.9 s			165.9 s	
- construction candidate					103.8 s	
- sampling		51.9 s			62.1 s	
Draws		10,000			10,000	
Time/draw		5.2 ms			67.2 ms	
	Required for % estimate with 1 digit of precision (with 95% confidence):					
For 99% VaR:						
- number of draws		567,648			58,498	
- computing time		2946 s (= 49 min 6 s)			467 s (= 7 min 47 s)	
For 99% ES:						
- number of draws		1,033,980			74,010	
- computing time		5366 s (= 89 min 26 s)			563 s (= 9 min 23 s)	

tion requires 103.8 s. This '*investment*' can be considered quite '*profitable*', as the NSEs of the VaR and ES estimators — both based on 10,000 draws — are much smaller than the NSE of the estimators using the 'direct approach'. Suppose that we want to compute estimates of the VaR and ES (in %) with a precision of 1 digit (with 95% confidence), i.e. $1.96\,NSE \leq 0.05$, so that we can quote (for example) −8.3% and −10.0% as *the* VaR and ES estimates from this model. In the 'direct approach' we would then require over 500,000 draws (over 49 minutes) for the VaR and over 1,000,000 draws (over 89 minutes) for the ES. However, in the QERMit approach we would require fewer than 60,000 (75,000) draws (under 8 (10) min) for the VaR (ES). Fig. 8 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 level of precision.

Finally, notice that the RNE is much higher than 1 for the QERMit approach, whereas it is somewhat below 1 for the 'direct approach'. The reason for the latter is again the serial correlation in the MH sequence of parameter draws.[13] The first phenomenon is in sharp contrast to the potential 'struggle' in importance sampling based Bayesian inference (for the estimation

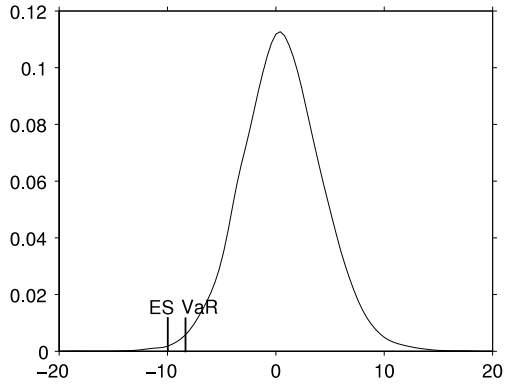


Fig. 7. Estimated profit/loss density: estimated density of 10 days % change in the S&P 500 (for the first 10 working days of January 2008) based on the GARCH model with Student-t errors estimated using 1998–2007 data.

[13] One could consider using 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 at least as much time as generating a path of 10 future log-returns. The quality of the draws, i.e. the RNE, would increase slightly, but the computing time required per draw would increase substantially.

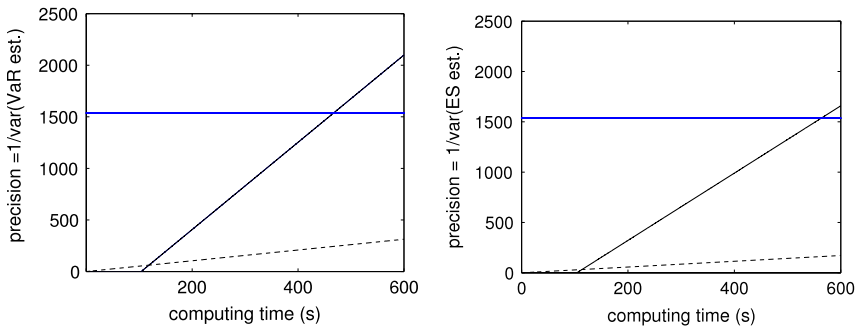


Fig. 8. Precision (1/var) of the estimated VaR and ES, as a function of the amount of computing time for the 'direct approach' (- -) and the QERMit approach (—). The horizontal line corresponds to a precision of 1 digit ($1.96\,NSE \leq 0.05$). Here the QERMit approach requires 103.8 s to construct an appropriate importance density, and after that soon generates far more precise VaR and ES estimates.

of posterior moments of non-elliptical distributions) to have an RNE not too far below 1.

A phenomenon that may cause problems for the QERMit approach is the so-called 'curse of dimensionality of importance sampling'. In this situation, the relative performance of the QERMit method may decrease as the prediction horizon h increases. In our example, the clear victory witnessed for $h = 10$ is still observed for $h = 20$. For very long horizons, e.g. $h = 100$, the preference for the QERMit approach vanishes. An adaptation of the QERMit algorithm, in which one adapts, in *real time*, a 'prefabricated' importance function that has been extensively trained on either historical data or one of a set of canonical examples, rather than starting from scratch, may provide better results in such situations. This is left as an issue for future research.

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, requires less computing time to achieve 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. The proposed method can also be useful in the case of 1-step-ahead forecasting with a portfolio of several assets, as the simulation of the future realizations is typically also required then. Thus, the sensible application of the QERMit method is not restricted to *multi-step*-ahead forecasting of VaR and ES.

One merit of the traditional 'direct approach' is that one can use the same algorithm for all risk levels (α) and prediction horizons (h) in which one is simultaneously interested. In the QERMit approach, a different importance function has to be designed for each α and h one wishes to consider. However, as an alternative one can also approximate the optimal importance function for the case with the largest relevant values of α and h, and use this importance function for smaller values of α and h as well. We intend to report on cases with multiple values of interest for α and h in the future.

The examples in this paper only consider the case of a single asset, the S&P 500 index. In that sense, the application is 1-dimensional. However, the 10-day-ahead forecasting of the price of a single asset 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 them yields a conservative risk measure for the whole portfolio. Nonetheless, we intend to investigate portfolios of several assets and report on it 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 also of interest. Having features which are clearly different to those of 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 efficient simulation-based computations in extreme value theory, e.g. efficient risk evaluation in the case of Pareto distributions.

Acknowledgements

A preliminary version of this paper was presented at the 2008 ESEM Conference in Milan, at the New York Camp Econometrics IV at Lake Placid, and at seminars at the European University Institute in Florence and the University of Pennsylvania. Helpful comments from several participants have led to substantial improvements. The authors further thank David Ardia, Luc Bauwens, Tomasz Woźniak and two anonymous referees for multiple useful suggestions. Some results were obtained during a students' research project at Erasmus University Rotterdam in 2008 by Jesper Boer, Rutger Brinkhuis and Jaap van Dam, supervised by Sander Scheerders of SNS REAAL and the first author. The second author gratefully acknowledges financial assistance from the Netherlands Organization of Research (grant 400-07-703).

References

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

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

Ardia, D. (2008). Financial risk management with Bayesian estimation of GARCH models. In *Lecture Notes in Economics and Mathematical Systems*: *Vol. 612*. Springer.

Ardia, D., Hoogerheide, L. F., & Van Dijk, H. K. (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.

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Quantitative Finance*, *9*(3), 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.

Bauwens, L., & Lubrano, M. (1998). Bayesian inference on GARCH models using the Gibbs sampler. *Econometrics Journal*, *1*, C23–C26.

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

Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of the United Kingdom inflation. *Econometrica*, *50*(4), 987–1008.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317–1339.

Glosten, L. R., Jaganathan, R., & Runkle, D. E. (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.

Hammersley, J. M., & Handscomb, D. C. (1964). *Monte Carlo methods* (1st ed.). London: Methuen.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Hoogerheide, L. F., Kaashoek, J. F., & Van Dijk, H. K. (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.

Hoogerheide, L. F., Van Dijk, H. K., & Van Oest, R. D. (2009). Simulation based Bayesian econometric inference: Principles and some recent computational advances. In *Handbook of computational econometrics* (pp. 215–280). Wiley.

Kloek, T., & Van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica*, *46*, 1–20.

McNeil, A. J., & Frey, R. (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.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087–1092.

Zeevi, A. J., & Meir, R. (1997). Density estimation through convex combinations of densities; Approximation and estimation bounds. *Neural Networks*, *10*, 99–106.

Lennart Hoogerheide is Assistant Professor of Econometrics at the Econometric Institute of Erasmus University Rotterdam. His research is mainly focused on flexible models and computational methods for the analysis of risk and treatment effects in finance and macroeconomics. His publications have appeared in international journals such as *Journal of Econometrics*, *International Journal of Forecasting*, *Journal of Forecasting*, and *Journal of Statistical Software*.

Herman K. van Dijk is Professor of Econometrics at the Econometric Institute of Erasmus University Rotterdam, and director of the Tinbergen Institute. His recent research has been related to simulation-based Bayesian inference, with applications in macroeconomics and finance. His publications have appeared in international journals such as *International Journal of Forecasting*, *Journal of Forecasting*, *Journal of Econometrics*, *Econometrica*, *European Economic Review*, *Journal of Applied Econometrics*, *Journal of Business and Economic Statistics*, and *Journal of Statistical Software*.