A Class of Adaptive EM-based Importance Sampling Algorithms for Efficient and Robust Posterior and Predictive Simulation

Lennart Hoogerheide*† Anne Opschoor* Herman K. van Dijk*

Abstract

A class of adaptive sampling methods is introduced for efficient posterior and predictive simulation. The proposed methods are robust in the sense that they can handle target distributions that exhibit non-elliptical shapes such as multimodality and skewness. The basic method makes use of sequences of importance weighted Expectation Maximization steps in order to efficiently construct a mixture of Student-t densities that approximates accurately the target distribution – typically a posterior distribution, of which we only require a kernel – in the sense that the Kullback-Leibler divergence between target and mixture is minimized. We label this approach *Mixture of t by Importance Sampling and Expectation Maximization* (MitISEM). The constructed mixture is used as a candidate density for quick and reliable application of either Importance Sampling (IS) or the Metropolis-Hastings (MH) method. The MitISEM algorithm performs well in exploring non-elliptical shapes of posterior and predictive distributions, in estimating predictive likelihoods and forecasting Value at Risk, for several examples of statistical and econometric models. We also introduce three extensions of the basic MitISEM approach. First, we propose a method for applying MitISEM in a *sequential* manner, so that the candidate distribution for posterior simulation is cleverly updated when new data become available. Our results show that the computational effort reduces enormously, while the quality of the approximation remains almost unchanged. This sequential approach can be combined with a tempering approach, which facilitates the simulation from densities with multiple modes that are far apart. Second, we introduce a *permutation-augmented* MitISEM approach, for importance sampling from posterior distributions in mixture models without the requirement of imposing identification restrictions on the model's mixture regimes' parameters. Third, we propose a *partial* MitISEM approach, which aims at approximating the marginal and conditional posterior distributions of subsets of model parameters, rather than the joint. This division can substantially reduce the dimension of the approximation problem, which facilitates the application of adaptive importance sampling for posterior simulation in more complex models with larger numbers of parameters. Our results indicate that the proposed methods can substantially reduce the computational burden in econometric models like mixture GARCH models and a mixture instrumental variables model.

*Econometric and Tinbergen Institutes, Erasmus University Rotterdam, The Netherlands
†Corresponding author, e-mail address: lhoogerheide@ese.eur.nl.

Keywords: mixture of Student-t distributions, importance sampling, Kullback-Leibler divergence, Expectation Maximization, Metropolis-Hastings algorithm, predictive likelihoods, Mixture GARCH models, Value at Risk.

1 Introduction

Since a few decades there is considerable interest in Bayesian analysis using computer generated pseudo random draws from the posterior and predictive distribution. Markov Chain Monte Carlo (MCMC) techniques are useful for this purpose and a popular MCMC technique is the Metropolis-Hastings algorithm, developed by Metropolis et al. (1953) and generalized by Hastings (1970). Several updates of this sampler are proposed in the literature, especially the idea of adapting the proposal distribution given sampled draws.

Monte Carlo procedures based on Importance Sampling (IS), see Hammersley and Handscomb (1964), are an alternative. This idea has been introduced in Bayesian inference by Kloek and Van Dijk (1978) and is further developed by Van Dijk and Kloek (1980, 1984) and, in particular, by Geweke (1989). According to Cappé et al. (2008), there exists renewed interest in Importance Sampling. This is due to its relatively simple properties which allow for the development of parallel implementation. The increased popularity of Importance Sampling goes jointly with the development of multiple core machines and computer clusters.

In this paper we specify a class of adaptive sampling methods for efficient and reliable posterior and predictive simulation. The proposed methods are robust in the sense that they can handle target distributions that exhibit non-elliptical shapes such as multimodality and skewness. These methods are especially useful for posteriors where the convergence of alternative simulation methods is slow or even doubtful, such as high serial correlation in Gibbs sequences that may be caused by large numbers of latent variables or non-elliptical shapes. Importance Sampling and Gibbs sampling are not necessarily substitutes: given that diagnostic checks can never fully guarantee that results have converged to the true values (that is, that convergence has been reached and that no errors have been made in the derivations and code), the use of both simulation methods that have completely different theory and implementation can be a useful validity check. Further, an appropriate candidate distribution can be used to draw initial values for multiple Gibbs sequences, whereas a sample of Gibbs draws can be used to obtain initial values for the mean and covariance matrix in the process of constructing an approximating candidate distribution. Our proposed methods make use of the novel *Mixture of t by Importance Sampling and Expectation Maximization* (MitISEM) approach. This approach uses sequences of importance weighted

steps in an Expectation Maximization algorithm in order to relatively quickly construct a mixture of Student-t densities, which is used as an efficient and reliable candidate density for Importance Sampling (IS) or the Metropolis-Hastings (MH) method. Next to assessing possibly non-elliptical posterior distributions, MitISEM is particulary useful for accurately estimating marginal and predictive likelihoods via IS.

Apart from specifying the basic approach of MitISEM, we introduce three extensions. First, we propose a method for applying MitISEM in a sequential manner, so that the candidate distribution for posterior simulation is cleverly updated when new data become available. Our results show that the computational effort reduces enormously, while the quality of the approximation remains almost unchanged, as compared with an 'ad hoc' procedure in which the construction of the MitISEM candidate is performed 'from scratch' at every moment in time. This sequential approach can be combined with a tempering approach, which facilitates the simulation from densities with multiple modes that are far apart. The proposed tempering method moves sequentially from a tempered target density kernel, the target density kernel to the power of a positive number that is smaller than 1, towards the real target density kernel. The tempered target distribution is more diffuse and hence the probability of detecting far-away modes is higher. This tempering idea is used in the Equi-Energy sampler, developed by Kou, Zhou and Wong (2006).

Second, we introduce a permutation-augmented MitISEM approach, for importance sampling from posterior distributions in mixture models without the requirement of imposing *a priori* identification restrictions on the mixture components' parameters. As discussed by Geweke (2007), the mixture model likelihood function is invariant with respect to permutation of the components of the mixture model. If functions of interest are permutation sensitive, as in classification applications, then interpretation of the likelihood function requires valid inequality constraints. If functions of interest are permutation invariant, as in prediction applications, then there are no such problems of interpretation. Geweke (2007) proposes the permutation-augmented Gibbs sampler, which can be considered as an extension of the random permutation sampler of Frühwirth-Schnatter (2001). The practical implementation of the idea of the permutation-augmented Gibbs sampler is that one simulates a Gibbs sequence with total disregard for label switching or the prior's labeling restrictions. Only after that and only if functions of interest are permutation sensitive, then one simply permutes the Gibbs sampler's output so as to satisfy the labeling restrictions. We propose a method of permutation-augmented IS, for which we extend the MitISEM approach to construct an approximation to the unrestricted posterior, taking into account the permutation structure. If m is the number of components of the mixture model, then the addition of a

Student-t component to the candidate implies an addition of the $m!$ equivalent permutations. Thereby, we construct a mixture of mixtures of $m!$ Student-t components, where the restriction is imposed that the $m!$ permutations have equal candidate density. Intuitively stated, we help the basic MitISEM approach by 'telling' it about the invariance with respect to permutations. It should be noted that this invariance with respect to permutations is not the only possible cause of non-elliptical shapes in a mixture model's posterior. For example, if the probability of one of the model's components tends to zero, the local non-identification of the component's other parameters causes ridge shapes.

Third, we propose a partial MitISEM approach, which aims at approximating the marginal and conditional posterior distributions of subsets of model parameters, rather than the joint. This division can substantially reduce the dimension of the approximation problem, which facilitates the application of adaptive importance sampling for posterior simulation in more complex models with larger numbers of parameters. Approximating the joint posterior density kernel with a mixture of Student-t distributions allows for a huge flexibility of shapes. However, rarely all of this flexibility is required. It is typically enough to use mixtures of Student-t distributions for the dependence *within* subsets of the parameters. We can often divide the parameters into subsets, where the dependence *between* different subsets is less complicated. Our partial MitISEM approach divides the model parameters into ordered subsets, where the conditional candidate distributions' means are linear combinations of (functions of) the parameters in previous subsets. The conditional candidate distributions' covariances can also be made to depend on the parameters in previous subsets, by allowing the probabilities of the mixture components of the conditional candidate distribution to differ for different ranges of values for functions of the parameters in previous subsets. The partial MitISEM approach is a way to provide a usable approximation to the posterior, while preventing problems such as numerical issues with specifying huge covariance matrices for a joint candidate distribution – problems that have led researchers to conclude that IS necessarily suffers from a 'curse of dimensionality'.

Several approaches of adaptive sampling using mixtures exist in the literature. Keith et al. (2008) developed adaptive independence samplers by minimizing the Kullback-Leibler (KL) divergence in order to provide the best candidate density, which consists of a mixture of Gaussian densities. The minimization of the KL-divergence is done by applying the EM algorithm of Dempster et al. (1977) and the number of mixture components is selected through information criteria like AIC (Akaike (1974)), BIC (Schwarz (1978)) or DIC (Gelman et al. (2003)). Our basic approach is a 'bottom up' procedure that starts with one Student-t distribution (instead of a Gaussian distribution) and Student-t components are

added iteratively until a certain stop criterion is met. We emphasize that the IS-weighted version of the EM algorithm is applied in order to use all candidate draws without requiring the Metropolis-Hastings algorithm to transform the candidate draws into a set of posterior draws. The advantages are that we do not require a burn-in sample, that the use of all candidate draws helps to prevent numerical problems with estimating candidate covariance matrices – also draws with relatively small, but positive importance weights are helpful for this purpose – and that the use of all candidate draws may lead to a better approximation. Cappé et al. (2008) and Cornuet et al. (2009) also use IS-weights in the EM algorithm with a mixture of Student-t densities as candidate density. Cappé et al. (2008) developed the M-PMC (Mixture Population Monte Carlo) algorithm, which is an adaptive algorithm that iteratively updates both the weights and component parameters of a mixture importance sampling density. An important difference between Cappé et al. (2008) (and also Cornuet et al. (2009)) and the present paper is the choice of the number of mixture components and the starting values of the candidate mixture's Student-t components' means and covariances in the EM optimization procedure. Regarding the first issue, in earlier papers the number of mixture components is chosen a priori, where we let the algorithm choose the required number of components. Second, we choose the starting values based on the draws that correspond to the highest IS-weights for the previous mixture of Student-t candidate in the algorithm, where Cappé et al. (2008) do not provide a strategy for choosing starting values. Although the EM procedure is guaranteed to converge to a *local* optimum, the choice of the starting values may still be crucial, given that the KL divergence between target and candidate (as a function of the candidate mixture's means, covariances, degrees of freedom and component weights) is a highly non-elliptical, multimodal function. Moreover, we provide extensions (sequential, tempered, permutation-augmented and partial MitISEM) that facilitate simulation for specific applications and for particular statistical and econometric models.

A final remark considering the literature regards the Adaptive Mixture of t (AdMit) approach of Hoogerheide, Kaashoek and Van Dijk (2007). Whereas the idea behind AdMit and MitISEM is the same, i.e. iteratively constructing an approximation of a target distribution by a mixture of Student-t distributions, there are three substantial differences. First, AdMit aims at minimizing the variance of the IS estimator directly, whereas MitISEM aims at this goal indirectly by minimizing the Kullback-Leibler divergence. As a result, AdMit optimizes the mixture component weights using a non-linear optimization procedure that requires considerable computational effort. Second, in the AdMit method, means and covariance matrices of the candidate components are chosen heuristically and are never updated

when additional components are added to the mixture, whereas in MitISEM all mixture parameters are optimized jointly by means of the relatively quick EM algorithm. This implies a large reduction of the computing time in the approximation procedure, and is expected to lead to a better candidate in most applications. Third, AdMit requires the joint target density kernel, whereas MitISEM requires candidate draws and importance weights. This implies that AdMit can not be applied partially to the marginal and conditional posterior distributions of subsets of parameters, whereas we propose a partial MitISEM approach. One relative advantage of the AdMit approach is the step in which the importance weight function is maximized with respect to the parameter vector, which may lead to finding relevant areas of the parameter space that were 'missed' by all draws from the previous candidate. We intend to investigate the use of such an AdMit step within MitISEM in further research.

The outline of this paper is as follows. In section 2 we introduce the MitISEM method. Section 2 also provides three subsections of applications in which MitISEM is used for estimating posterior moments, forecasting Value at Risk, and estimating model probabilities. Section 3 introduces the sequential MitISEM method, and includes a subsection on the tempering method. Section 4 introduces the partial MitISEM method. Section 5 concludes. The appendix provides the derivations of the IS-weighted EM methods.

2 Mixture of t by Importance Sampling and Expectation Maximization (MitISEM)

If one uses Importance Sampling or the Metropolis-Hastings algorithm to conduct posterior analysis, a key issue is to find a candidate density which approximates the target distribution. This can be quite difficult if the target density is not elliptical. This paper proposes to specify the candidate distribution as a mixture of Student-t distributions. According to Hoogerheide et al. (2007), the usage of mixtures of Student-t distributions has several advantages. First, they can provide an accurate approximation to a wide variety of target densities. For example, they can exhibit substantial skewness or irregularly curved contours such as multimodality. Zeevi and Meir (1997) show 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 densities falls within their framework. Second, simulation from the Student-t distribution and evaluation of the Student-t density are performed easily and efficiently. Third, Student-t distributions have fatter tails than normal distributions, which reduces the risk that the tails of the candidate density are thinner than those of the

target distribution. Fourth, a mixture of t approximation to a target distribution can be constructed in a quick, automatic, reliable manner by our novel procedure.

We will use the notation $f(\theta)$ for the target density kernel of θ, the k-dimensional vector of interest. $f(\theta)$ is typically a posterior density kernel, but it can also be a density kernel of observable variables or a density kernel of both parameters and observable variables. $g(\theta)$ is the candidate density, a mixture of H Student-t densities:

$$g(\theta) = g(\theta|\zeta) = \sum_{h=1}^{H} \eta_h \; t_k(\theta|\mu_h, \Sigma_h, \nu_h), \tag{1}$$

where ζ is the set of modes μ_h, scale matrices Σ_h, degrees of freedom ν_h, and mixing probabilities η_h $(h = 1, \ldots, H)$ of the k-dimensional Student-t components with density:

$$t_k(\theta|\mu_h, \Sigma_h, \nu_h) = \frac{\Gamma\left(\frac{\nu_h+k}{2}\right)}{\Gamma\left(\frac{\nu_h}{2}\right)(\pi\nu_h)^{k/2}} |\Sigma_h|^{-1/2} \left(1 + \frac{(\theta-\mu_h)'\Sigma_h^{-1}(\theta-\mu_h)}{\nu_h}\right)^{-(k+\nu_h)/2}. \tag{2}$$

Here Σ_h is positive definite, $\eta_h \geq 0$ and $\sum_{h=1}^{H} \eta_h = 1$. We further restrict ν_h such that $\nu_h \geq 1$.

First, assume that the number of components H is given. In the sequel of this section we will propose a 'bottom up' procedure that starts with one Student-t distribution and which iteratively adds Student-t components until a certain stop criterion is met. The aim is to choose the candidate mixture density $g(\theta)$ in such a way that it provides a good approximation of the target density $\tilde{f}(\theta)$ of which $f(\theta)$ is a kernel. We do this by choosing ζ such that it minimizes the Kullback-Leibler divergence (or Cross-entropy distance) (Kullback and Leibler (1951)), which is defined as

$$\mathcal{D}_1(\tilde{f} \to g) = \int \tilde{f}(\theta) \log \frac{\tilde{f}(\theta)}{g(\theta|\zeta)} \, d\theta. \tag{3}$$

This is obviously equivalent with minimizing

$$\mathcal{D}_1(f \to g) = \int f(\theta) \log \frac{f(\theta)}{g(\theta|\zeta)} \, d\theta. \tag{4}$$

as long as the same kernel f of the target density $\tilde{f}$ is used throughout the minimization. Since

$$\mathcal{D}_1(f \to g) = \int f(\theta) \log \frac{f(\theta)}{g(\theta|\zeta)} \, d\theta = \int f(\theta) \log f(\theta) \, d\theta - \int f(\theta) \log g(\theta|\zeta) \, d\theta, \tag{5}$$

where only the second term on the right-hand side of (5) depends on ζ, this amounts to maximizing

$$\int f(\theta) \log g(\theta|\zeta) \, d\theta = E_{\theta\sim f(\theta)}[\log g(\theta|\zeta)] = \tag{6}$$

$$\int g_0(\theta) \frac{f(\theta)}{g_0(\theta)} \log g(\theta|\zeta) \, d\theta = E_{\theta\sim g_0(\theta)}\left[\frac{f(\theta)}{g_0(\theta)} \log g(\theta|\zeta)\right], \tag{7}$$

where $g_0(\theta)$ is a given candidate density that has been obtained in a previous step. For $H = 1$ the density $g_0(\theta)$ is an initial candidate distribution, such as a Student-t distribution around the posterior mode with scale matrix equal to minus the inverse Hessian of the log-posterior at the mode, or an adapted version thereof. For $H \geq 2$, g_0 is a mixture of $H - 1$ Student-t components, that has been obtained in the previous step of the 'bottom up' construction procedure.

We use an Expectation-Maximization (EM) algorithm for minimizing the stochastic counterpart of (7) in order to find

$$\zeta^* = \arg\max_{\zeta} \frac{1}{N} \sum_{i=1}^{N} W^i \log g(\theta^i|\zeta) \qquad \text{with} \qquad W^i = \frac{f(\theta^i)}{g_0(\theta^i)},$$

where θ^i $(i = 1, 2, \ldots, N)$ are independent draws from g_0. Note that both the θ^i and W^i are given the during the optimization; θ^i and W^i $(i = 1, 2, \ldots, N)$ do not depend on ζ. We emphasize that the importance weighted version of the EM algorithm is applied, rather than minimizing the stochastic counterpart of (6) by a 'regular' EM algorithm, in order to use all candidate draws without requiring the Metropolis-Hastings algorithm to transform the candidate draws into a set of posterior draws. This has three advantages. First, we do not require a burn-in sample. Second, the use of all candidate draws θ^i $(i = 1, 2, \ldots, N)$ helps to prevent numerical problems with estimating candidate covariance matrices; also draws with relatively small, but positive importance weights are helpful for this purpose. Third, the use of all candidate draws may lead to a better approximation.

The EM algorithm (Dempster et al. (1977)) is based on the idea that a complex model for some observable 'data' θ with parameters ζ can be formulated in a simpler form with latent data $\tilde{\theta}$ in addition to θ and ζ. If the latent data $\tilde{\theta}$ were observed, the computation of the Maximum Likelihood estimator of θ would be relatively straightforward. Each iteration L of the EM algorithm consists of two (iterative) steps, the Expectation and Maximization step. The first (Expectation) step takes the expectation of the log-likelihood function with respect to the latent data $\tilde{\theta}$ (given the parameter values $\zeta^{(L-1)}$ from the previous iteration). The second (Maximization) step maximizes this expected log-likelihood with respect to the parameters.

In our situation we maximize the *weighted* log-likelihood

$$\frac{1}{N} \sum_{i=1}^{N} W^i \log g(\theta^i|\zeta)$$

where $g(.|\zeta)$ is the mixture of Student-t densities (1). The mixture of Student-t densities (1) for θ^i is equivalent with the specification

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

where z^i is a latent H-dimensional vector indicating from which Student-t component the observation θ^i stems: if θ^i stems from component h, then $z_h^i = 1$, $z_j^i = 0$ for $j \neq h$; $\Pr[z^i = e_h] = \eta_h$ with e_h the h-th column of the identity matrix; w_h^i has the Inverse-Gamma distribution $IG(\nu_h/2, \nu_h/2)$. For a more extensive explanation of this continuous scale mixing representation of (mixtures of) Student-t distributions we refer to Peel and McLachlan (2000). Here we have latent 'data' $\tilde{\theta}^i$ $(i = 1, \ldots, N)$

$$\tilde{\theta}^i = \{z_h^i, w_h^i | h = 1, \ldots, H\}$$

and

$$\begin{aligned}
\log p(\theta^i, w^i, z^i|\zeta) &= \log p(\theta^i|w^i, z^i, \zeta) + \log p(w^i|\zeta) + \log p(z^i|\zeta) \\
&= \sum_{h=1}^{H} z_h^i \ \log\left[\mathrm{pdf}_{N(\mu_h, w_h^i \Sigma_h)}(\theta^i)\right] + \\
&\quad \sum_{h=1}^{H} \log \mathrm{pdf}_{IG(\nu_h/2, \nu_h/2)}(w_h^i) + \sum_{h=1}^{H} z_h^i \log(\eta_h) \\
&= \sum_{h=1}^{H} z_h^i \ \left\{-\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_h| - \frac{k}{2}\log(w_h^i) - \frac{1}{2}\frac{(\theta^i - \mu_h)'(\Sigma_h)^{-1}(\theta^i - \mu_h)}{w_h^i}\right\} \\
&\quad + \sum_{h=1}^{H} \left\{\frac{\nu_h}{2}\log\left(\frac{\nu_h}{2}\right) - \left(\frac{\nu_h}{2} - 1\right)\log(w_h^i) - \frac{\nu_h}{2}\ \frac{1}{w_h^i} - \log\left(\Gamma\left(\frac{\nu_h}{2}\right)\right)\right\} \\
&\quad + \sum_{h=1}^{H} z_h^i \log(\eta_h) \qquad (8)
\end{aligned}$$

where w^i and z^i are *a priori* independent. The expressions of the latent variables w^i and z^i that appear in terms which also involve the parameters ζ to be optimized are z_h^i, $\frac{z_h^i}{w_h^i}$, $\log w_h^i$, and $\frac{1}{w_h^i}$. The conditional expectations given θ^i and $\zeta = \zeta^{(L-1)}$, the optimal parameters in the previous EM iteration, are as follows:

$$\tilde{z}_h^i \equiv E\left[z_h^i \left| \theta^i, \zeta = \zeta^{(L-1)}\right.\right] = \frac{t(\theta^i|\mu_h, \Sigma_h, \nu_h)\ \eta_h}{\sum_{j=1}^{H} t(\theta^i|\mu_j, \Sigma_j, \nu_j)\ \eta_j}, \qquad (9)$$

$$\widetilde{z/w}_h^i \equiv E\left[\left.\frac{z_h^i}{w_h^i}\right| \theta^i, \zeta = \zeta^{(L-1)}\right] = \tilde{z}_h^i\ \frac{k + \nu_h}{\rho_h^i + \nu_h}, \qquad (10)$$

$$\begin{aligned}
\xi_h^i &\equiv E\left[\log w_h^i \left| \theta^i, \zeta = \zeta^{(L-1)}\right.\right] = \\
&= \left[\log\left(\frac{\rho_h^i + \nu_h}{2}\right) - \psi\left(\frac{k + \nu_h}{2}\right)\right]\tilde{z}_h^i + \left[\log\left(\frac{\nu_h}{2}\right) - \psi\left(\frac{\nu_h}{2}\right)\right](1 - \tilde{z}_h^i), \qquad (11)
\end{aligned}$$

$$\delta_h^i \equiv E\left[\left.\frac{1}{w_h^i}\right| \theta^i, \zeta = \zeta^{(L-1)}\right] = \frac{k + \nu_h}{\rho_h^i + \nu_h}\ \tilde{z}_h^i + (1 - \tilde{z}_h^i), \qquad (12)$$

with $\rho_h^i = (\theta^i - \mu_h)'\Sigma_h^{-1}(\theta^i - \mu_h)$, $\psi(.)$ the digamma function (the derivative of the logarithm

of the gamma function $\log\Gamma(.)$), and all parameters $\mu_h, \Sigma_h, \nu_h, \eta_h$ elements of $\zeta^{(L-1)}$. For the derivations of these expectations we refer to the appendix.

Define $\log\tilde{p}(\theta^i, w^i, z^i|\zeta)$ as the result of substituting the expectations (9)-(12) into $\log p(\theta^i, w^i, z^i|\zeta)$ in (8). The Maximization step amounts to computing the ζ that maximizes

$$\zeta^{(L)} = \arg\max_{\zeta} \frac{1}{N}\sum_{i=1}^{N} W^i \log\tilde{p}(\theta^i, w^i, z^i|\zeta).$$

Using the analogy with Maximum Likelihood estimation for the Seemingly Unrelated Regression model with Gaussian errors (for the k elements of θ^i) and the same 'regressor' (a constant term) in each equation, in which case the Ordinary Least Squares (OLS) estimator provides the Maximum Likelihood Estimator, and Maximum Likelihood estimation for the multinomial distribution, it is easily derived that $\zeta^{(L)}$ consists of:

$$\mu_h^{(L)} = \left[\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i\right]^{-1}\left[\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i \ \theta^i\right], \tag{13}$$

$$\hat{\Sigma}_h^{(L)} = \frac{\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i \ (\theta^i - \mu_h^{(L)})(\theta^i - \mu_h^{(L)})'}{\sum_{i=1}^{N} W_i \ \tilde{z}_h^i}, \tag{14}$$

$$\eta_h^{(L)} = \frac{\sum_{i=1}^{N} W_i \ \tilde{z}_h^i}{\sum_{i=1}^{N} W_i}. \tag{15}$$

Further, $\nu_h^{(L)}$ is solved from the first order condition of ν_h:

$$-\psi(\nu_h/2) + \log(\nu_h/2) + 1 - \frac{\sum_{i=1}^{N} W_i \ \xi_h^i}{\sum_{i=1}^{N} W_i} - \frac{\sum_{i=1}^{N} W_i \ \delta_h^i}{\sum_{i=1}^{N} W_i} = 0. \tag{16}$$

Cappé et al. (2008) only update the expectations and covariance structures of the Student-t distributions and not the degrees of freedom, because there is no closed-form solution for the latter. We propose to optimize also the degrees of freedom parameter ν_h during the EM procedure for three reasons. First, the larger flexibility may lead to a better approximation of the target distribution. Second, solving ν_h from (16) requires only a one-dimensional root finder, which requires little computation time. Moreover, $1 - \frac{\sum_{i=1}^{N} W_i \ \xi_h^i}{\sum_{i=1}^{N} W_i} - \frac{\sum_{i=1}^{N} W_i \ \delta_h^i}{\sum_{i=1}^{N} W_i}$ is constant with respect to ν_h, so that it only has to be evaluated once in the process of solving the equation. Third, the resulting values of ν_h ($h = 1, \ldots, H$) may provide information on the shape of the target distribution (e.g. whether the kurtosis is small, moderate or large).

We now discuss two remaining issues: (1) how to choose the number of components H; (2) how to specify the initial values in the EM algorithm. In order to deal with both issues,

we use a 'bottom up' procedure that starts with one Student-t distribution and which iteratively adds Student-t components until a certain stop criterion is met:

Algorithm 1. *The MitISEM approach for obtaining an approximation to a target density:*

(0) **Initialization**: Simulate draws $\theta_1, \ldots, \theta_N$ from the naive proposal density g_{naive} where g_{naive} denotes a Student-t distribution with mode and scale matrix equal to the target distribution's mode and minus the inverse Hessian of the log-target density kernel evaluated at the mode.

(1) **Adaptation**: Estimate the target distribution's mean and covariance matrix using IS with the draws $\theta_1, \ldots, \theta_N$ from g_{naive}. Use these estimates as the mode and scale matrix of Student-t distribution $g_{adaptive}$. Draw a sample $\theta_1, \ldots, \theta_N$ from this adaptive Student-t distribution $g_0 = g_{adaptive}$, and compute the IS weights for this sample.

(2) Apply the **IS-weighted EM algorithm** given the latest IS weights and the drawn sample of step 1. The output consists of the new candidate density g with optimized ζ, the set of $\mu_h, \Sigma_h, \nu_h, \eta_h$ $(h = 1, \ldots, H)$. Draw a new sample $\theta_1, \ldots, \theta_N$ from this proposal density and compute corresponding IS weights.

(3) **Iterate on the number of mixture components**: Given the current mixture of H components with corresponding μ_h, Σ_h, ν_h and η_h $(h = 1, \ldots, H)$, take $x\%$ of the sample $\theta_1, \ldots, \theta_N$ that correspond to the highest IS weights. Construct with these draws and IS weights a new mode μ_{H+1} and scale matrix Σ_{H+1} which are starting values for the additional component in de mixture candidate density. The reason behind this choice is that the new component is meant to cover a region of the parameter space in which the previous candidate mixture had relatively too little probability mass. Starting values for η_{H+1} and ν_{H+1} are at each iteration set at 0.10 and 1, respectively. Obvious starting values for μ_h, Σ_h and ν_h $(h = 1, \ldots, H)$ are the optimal values in the mixture of H components, while η_h is 0.90 times the previously optimal value. Given the latest IS weights and the drawn sample from the current mixture of H components, apply the IS-weighted EM algorithm to optimize *each* mixture component μ_h, Σ_h, ν_h and η_h with $h = 1, \ldots, H + 1$. Draw a new sample from the mixture of $H + 1$ components and compute corresponding IS weights.

(4) **Evaluate the IS weights** by computing the Coefficient of Variation (C.o.V.), i.e. the standard deviation of the IS weights divided by their mean. Stop the algorithm when this coefficient has converged. Otherwise return to step 3.

Step (1) can be seen as an intermediate step which quickly tries to improve the initial candidate distribution g_0, before calling the IS-weighted EM algorithm. If during the EM algorithm, a scale matrix Σ_h of a Student-t component (with very small weight η_h) becomes (nearly) singular, then this h-th component is removed from the mixture. We emphasize that in the iteration on the number of mixture components, the EM algorithm is applied to optimize *all* components. This is a qualitative improvement compared to the AdMit approach of Hoogerheide et al. (2007), which fixes the Student-t densities once they are formed.

There are still two strategic issues to be discussed about the MitISEM algorithm. The first issue relates to the following question: what is an efficient simulation method? Is this a simulation method that, given a certain amount of computing time, provides an estimate of a quantity of interest with the highest possible precision? Or is this a simulation method that, given a certain required precision, needs the shortest computing time. The optimal number of Student-t components may depend on the available computing time or the required precision. The more computing time is available, or the higher the required precision, the more rewarding a large 'investment' in an accurate approximation may be. Moreover, in order to choose the optimal number of Student-t components, we need to know the quantity of interest. That is, for a particular quantity of interest and a particular desired precision (or available amount of computing time), one could attempt to compute an optimal allocation of computing time over the construction of the candidate and the subsequential use in IS or the MH algorithm. We intend to investigate this issue in future research. In the current paper, we propose a heuristic procedure that continues adding Student-t components until the approximation's quality 'hardly' improves. We define the latter as a relative change in the C.o.V. of the IS weights that is smaller than 10%.

We discuss examples in which the posterior distribution is itself approximated, which seems a reasonable choice when we are interested in quantities such as the posterior mean, median or covariance. For the specific application of multi-step-ahead forecasting Value at Risk (VaR), we approximate the optimal importance density of Geweke (1989). In the latter case, one may monitor the Numerical Standard Error (NSE) of the estimated VaR, as an alternative to the C.o.V. of IS weights.

Second, although the EM procedure is guaranteed to converge to a *local* optimum – the (weighted) log-likelihood is a non-decreasing function of the number of EM iterations – the choice of the starting values may still be crucial, given that the KL divergence between target and candidate (as a function of the candidate mixture's means, covariances, degrees of freedom and component weights) is a highly non-elliptical, multimodal function.

MitISEM uses $x\%$ of the sample $\theta_1, \ldots, \theta_N$ that correspond to the highest IS weights, in order to compute starting values for the mode μ_{H+1} and scale matrix Σ_{H+1} of the additional component in de mixture candidate density. The optimal choice of $x\%$ depends on the particular target distribution and the current candidate mixture of H Student-t components. Therefore, we apply the EM algorithm with three different starting values (based on 1%, 5% or 10% of the draws $\theta_1, \ldots, \theta_N$), and continue the algorithm with the resulting mixture of $H+1$ Student-t components that yields the lowest C.o.V. value of the IS weights among the three approaches.

The results in the present paper suggest that the current implementation of MitISEM is successful at constructing approximations that are useful candidate distributions. It should be stressed that we do *not* require the globally optimal candidate distribution: it suffices to have a 'good' approximation that makes a trade-off between the computing time of constructing a candidate distribution and the efficiency during the subsequential simulation.

In the following subsections the MitISEM approach is applied in mixture GARCH models, for the estimation of posterior moments, mult-step-ahead prediction of Value at Risk, and the analysis of model probabilities.

2.1 Application I: analysis of a non-elliptical posterior distribution in a mixture GARCH(1,1) model

In this subsection the MitISEM approach is applied to the two-component Gaussian Mixture GARCH (1,1) model of Ausín and Galeano (2007). For the Bayesian estimation of this model, Ausín and Galeano (2007) propose a Griddy-Gibbs sampler (Ritter and Tanner (1992)), since the recursive structure of the likelihood in GARCH-type models implies that a regular Gibbs sampling approach is not feasible. However, the Griddy-Gibbs sampler is known to be very slow. We use the MH sampler and IS with a candidate density resulting from the MitISEM algorithm, and compare the performance of the MitISEM candidate density with a naive and an adaptive Student-t candidate density.

The two-component Gaussian mixture GARCH(1,1) model for the returns y_t ($t = 1, 2, \ldots, T$) is given by

$$y_t = \mu + \sqrt{h_t}\,\varepsilon_t, \tag{17}$$

$$h_t = \omega + \alpha(y_{t-1} - \mu)^2 + \beta h_{t-1}, \tag{18}$$

$$\varepsilon_t \sim \begin{cases} N(0, \sigma^2) & \text{with probability } \rho, \\ N(0, \sigma^2/\lambda) & \text{with probability } 1-\rho, \end{cases} \tag{19}$$

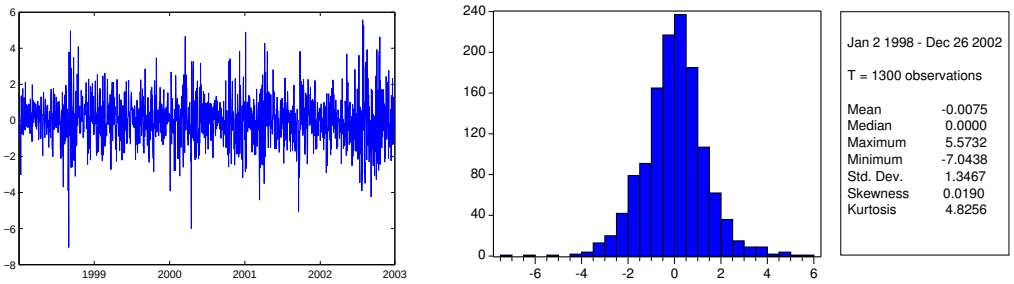

*Figure 1: S&P 500 log-returns (*100 × *change of log-index): daily observations from* 1998 − 2002.

with h_t the conditional variance of y_t given the information set $I_{t-1} = \{y_{t-1}, y_{t-2}, y_{t-3}, \ldots\}$. In addition, $0 < \lambda < 1$, and $\sigma^2 \equiv 1/(\rho + (1-\rho)/\lambda)$ so that $\text{var}(\varepsilon_t) = 1$; h_0 is treated as a known constant. We restrict $\omega > 0, \alpha \geq 0$ and $\beta \geq 0$ to ensure positivity of h_t. We follow Ausín and Galeano (2007) by imposing the prior restriction $0.5 < \lambda < 1$, so that it is ensured that the state with smaller variance has larger probability than the state with larger variance. We follow Ausín and Galeano (2007) also in specifying flat priors for the model parameters. Moreover, we truncate ω and μ such that these have proper (non-informative) priors. For the parameter vector $\theta = (\rho, \lambda, \mu, \omega, \gamma, \alpha, \beta)$ of dimension $k = 7$ we have a uniform prior on $[0.5, 1] \times [0, 1] \times [-1, 1] \times [0, 1] \times [-1, 1] \times [0, 1] \times [0, 1]$ with $\alpha + \beta < 1$ which implies covariance stationarity of h_t.

The returns y_t are taken from the *S&P* 500 index. From this index we use daily observations y_t $(t = 1, \ldots, T)$ on the log return (100 times the change of the logarithm of the closing price) from January 2 1998 to December 26 2002. We chose this pre-crisis period, since the performance of the model was better than during the recent crisis. Therefore, this period is a plausible choice for this illustrative example. Figure 1 shows the returns and their corresponding descriptive statistics. This shows clearly some stylized facts of equity returns' distributions: non-normality (excess kurtosis) and volatility clustering.

Posterior means of the model parameters are estimated by using IS and the independence chain MH algorithm. In more detail we use three candidate distributions based on Student-t densities: the mixture of Student-t densities resulting from the MitISEM algorithm, an 'adaptive' Student-t distribution and a 'naive' Student-t distribution. The adaptive candidate is in fact the distribution that is produced in step 1 of the MitISEM algorithm, whereas the 'naive' density simply uses the mode and the scale matrix estimated from the Hessian.

The top left panel of Figure 2 shows the non-elliptical shapes of the posterior density. Contour lines are plotted for (λ, ρ), where the remaining parameters are fixed at posterior means (estimated by IS). A non-identification issue arises if $\rho \to 1$, because in this

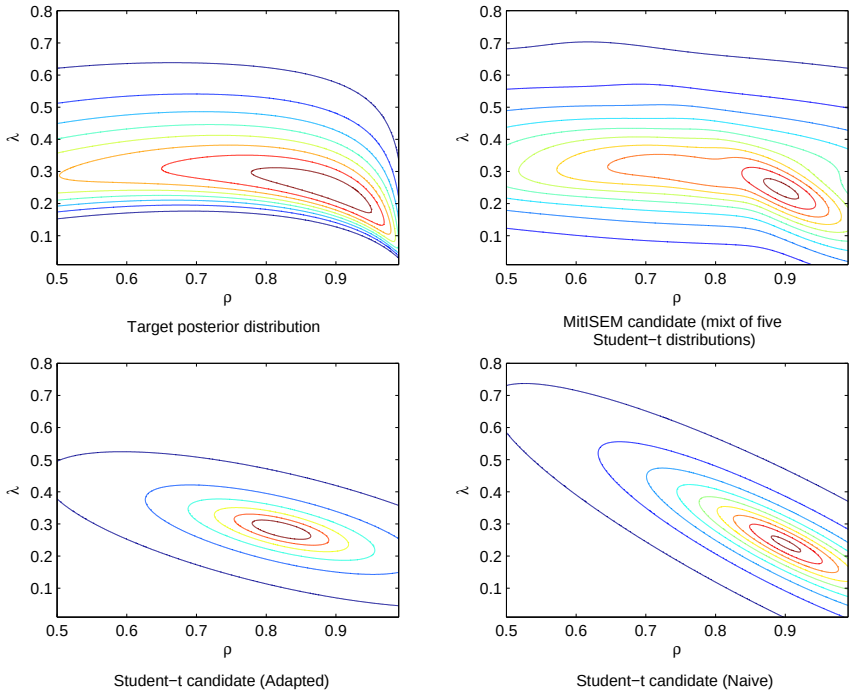

Figure 2: Contour plots of the Gaussian Mixture GARCH (1,1) model applied to S&P 500 data. All panels show plots of the conditional (posterior/candidate) density of (ρ, λ) *given* $(\mu, \omega, \alpha, \beta)$ *equal to the posterior mean (estimated by IS). The top panels depict the conditional posterior density and the candidate density contours resulting from MitISEM. The bottom panels show contours of the 'adaptive' and 'naive' candidate densities.*

case λ becomes unidentified since the model does not contain a regime with larger variance anymore. Then $1/\lambda$, the ratio of the large and small variance, can take a wide range of values. The remaining panels show contours of the candidate density implied by MitISEM, the 'adaptive' and the 'naive' candidate distribution. MitISEM has produced a candidate density that covers all the relevant (non-elliptically shaped) areas of the posterior target distribution, whereas the adaptive and naive candidates may 'miss' relevant areas, for example around points $(\rho = 0.5, \lambda = 0.2)$, $(\rho = 0.9, \lambda = 0.5)$ or $(\rho = 0.99, \lambda = 0.01)$.

Table 1 shows posterior means estimated by the IS and MH algorithms. For both methods, we simulate 10000 draws. For the MH approach, we take an burn-in sample of 1000 draws. Numerical standard errors (NSE) for IS and the MH algorithm are obtained by repeating the procedure 100 times. The main result from Table 1 is that MitISEM clearly outperforms the other candidate densities, irrespective whether IS or the MH method is used, since the NSE values are (much) smaller than the corresponding values implied by the Adaptive and Naive candidate densities. Regarding IS, an additional column is included

in the table: here we combine MitISEM-based IS with the variance reduction technique of antithetic sampling, where we simulate half the number of draws from the MitISEM candidate, and for each draw the 'mirror image' within the Student-t component (at the other side of the candidate Student-t component's mode) is added. For some parameters this leads to an improvement of the NSE, indicating that the combination of MitISEM-based IS with well-known variance reduction techniques such as antithetic sampling may be worthwhile. However, for ρ and λ no improvement is observed, reflecting that, roughly stated, the other side of the Student-t component may still be in a nearby subdomain of the whole parameter space. The latter phenomenon makes the effect of antithetic sampling much less clear than under symmetric candidate distributions. We leave the combination of MitISEM with variance reduction techniques such as antithetic sampling and control variates as a topic for further research.

We end this subsection with a remark on the computing time. Given the candidate density, the IS or MH method using the MitISEM candidate costs hardly more computing time than under a Student-t candidate. That is, the difference in computing time between evaluating and simulating from a mixture of Student-t and a Student-t density is small, as compared with the computing time required for the evaluation of the target density kernel. However, the construction of the MitISEM candidate necessarily requires more computing time than the naive and adaptive Student-t candidates, since the computations for these Student-t candidates are merely the initial steps of the MitISEM procedure.

Here, the construction of the MitISEM candidate took less than a minute (on a common laptop processor), whereas the simulation of 10000 draws requires approximately 6 seconds. From this it is clear that the MitISEM approach is especially useful if one desires estimates with a high precision. In the next subsection we will consider the Bayesian estimation of Value at Risk, where we will take a closer look at the computing time.

Table 1: Estimated posterior means and NSE's, obtained by using three different candidate densities for IS and the independence chain MH method. NSE values of the IS method and the MH-algorithm are obtained by repeating the procedure 100 times. Maximum Likelihood estimates are provided in the first panel of the table.

		Independence chain MH estimates					
		MitISEM		Adaptive		Naive	
	ML est.	mean	NSE · 100	mean	NSE · 100	mean	NSE · 100
ρ	0.92	0.81	0.20	0.82	0.97	0.82	6.21
λ	0.23	0.28	0.09	0.28	0.36	0.29	1.35
μ	0.04	0.04	0.07	0.04	0.22	0.04	0.95
ω	0.06	0.09	0.12	0.08	0.27	0.09	1.31
α	0.07	0.08	0.04	0.08	0.15	0.08	0.45
β	0.90	0.87	0.06	0.87	0.23	0.86	1.05
	IS estimates						
	MitISEM		MitISEM antithetic		Adaptive		Naive	
	mean	NSE · 100	mean	NSE · 100	mean	NSE · 100	mean	NSE · 100
ρ	0.79	0.16	0.79	0.17	0.79	0.74	0.79	3.44
λ	0.28	0.08	0.28	0.08	0.28	0.15	0.28	0.66
μ	0.04	0.04	0.04	0.02	0.04	0.09	0.04	0.37
ω	0.09	0.07	0.09	0.05	0.09	0.12	0.09	0.57
α	0.08	0.03	0.08	0.02	0.08	0.05	0.08	0.28
β	0.86	0.06	0.87	0.04	0.86	0.11	0.86	0.51

2.2 Application II: efficient Bayesian forecasting of Value at Risk

In the previous subsection we illustrated that local non-identification of model parameters can cause non-elliptical shapes of the target distribution. In this subsection we will illustrate that aiming at the optimal importance density for a particular (tail-related) quantity of interest may be another cause for non-elliptical shapes of the target distribution.

A basic Bayesian procedure to multi-step-ahead prediction of Value at Risk (VaR) is as follows. Given draws of the posterior density, obtained by for example an independence chain MH algorithm, one simulates possible future paths of the returns and takes the quantile of interest of the simulated future returns. We label this procedure the 'direct approach'. Hoogerheide and Van Dijk (2010) propose an indirect way to compute the multi-step-ahead VaR. They developed the approach of Quick Evaluation of Risk using Mixture of t approximations (QERMit), where first the optimal importance density, derived by Geweke (1989), $q_{opt}(\cdot)$ of future returns and model parameters is approximated by a 'hybrid' mixture of densities $\hat{q}_{opt}(\cdot)$. After that, this approximation $\hat{q}_{opt}(\cdot)$ is used as a candidate density in

Importance Sampling.

The optimal importance distribution has 50% of the future returns below the VaR and 50% above the VaR; that is, 50% of the draws should consist of high losses. Therefore the optimal importance density $q_{opt}(\cdot)$ is typically multimodal, even if the posterior is elliptically shaped (as is the case in the Student-t GARCH model in this subsection), since it has one mode near the mode of the future paths' distribution (and the posterior mode) and at least one mode in the 'high loss region'. We refer to Hoogerheide and Van Dijk (2010) for more details.

Following Hoogerheide and Van Dijk (2010), the step-by-step procedure to estimate the τ-step ahead 100 α% VaR by the QERMit approach is as follows [1]:

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

(Step 1a) Use the MitISEM algorithm to obtain a mixture of Student-t densities $q_{1,Mit}(\theta)$ that approximates the posterior density.

(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 $y^{*i} \equiv \{y^i_{T+1}, \ldots, y^i_{T+\tau}\}$ $(i = 1, \ldots, N)$ from the model given parameter values θ^i and historical values $y \equiv \{y_1, \ldots, y_T\}$, i.e. from the density $p(y^*|\theta^i, y)$. Compute a preliminary estimate $\widehat{VaR}_{prelim}$ as the 100 α% quantile of the profit/loss values $PL(y^{*i})(i = 1, \ldots, N)$.

(Step 1c) Use again the MitISEM algorithm to 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 $PL(y^*) < \widehat{VaR}_{prelim}$.

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

$$\hat{q}_{opt}(\theta, y^*) = 0.5\ q_{1,Mit}(\theta)p(y^*|\theta^i, y) + 0.5\ q_{2,Mit}(\theta, y^*). \tag{20}$$

The first term in the candidate (20) is caused by the fact that 50% of the draws corresponding to the 'whole' distribution of (θ, y^*) can be generated more efficiently by using the density $p(y^*|\theta^i, y)$ that is specified by the model and approximating merely the posterior $q_{1,Mit}(\theta)$

[1] There is obviously a crucial difference between the method of Hoogerheide and Van Dijk (2010) and the method described in this paper: the mixture of Student-t densities is obtained by AdMit in Hoogerheide and Van Dijk (2010), whereas we obviously use the MitISEM algorithm

than by approximating the joint distribution of (θ, y^*)[2]. Further, the profit/loss function equals simply the sum of all returns y^{*i} in this paper.

We apply the QERMit approach by considering the 10-day ahead 99% VaR forecast for the $S\&P$ 500 index. For estimation we use the same pre-crisis data as in the previous subsection. We use the GARCH model (Engle (1982), Bollerslev (1986)) with Student-t innovations:

$$\begin{aligned} y_t &= \mu + u_t && (21)\\ u_t &= \varepsilon_t(\rho h_t)^{1/2} && (22)\\ \varepsilon_t &= \text{Student-}t(\nu) && (23)\\ \rho &\equiv \frac{\nu-2}{\nu} && (24)\\ h_t &= \alpha_0 + \alpha_1 u_{t-1}^2 + \beta h_{t-1} && (25) \end{aligned}$$

with Student-$t(\nu)$ the standard Student-t distribution with ν degrees of freedom and variance $\frac{\nu-2}{\nu}$. The reasons for choosing this GARCH(1,1) model with Student-t errors are that it is a popular model among practitioners and moreover that its posterior is elliptically shaped, so that our example illustrates that flexible candidate distributions can also be useful in cases with elliptically shaped posteriors. Non-informative priors are specified for all parameters; a proper non-informative prior is used for ν to avoid an improper posterior density; see Bauwens and Lubrano (1998). The factor ρ ensures that h_t is the conditional variance of y_t.

We now compare the results of the QERMit method with the 'direct approach' explained at the start of this subsection. Table 2 shows simulation results. The 'investment' of computing time into the construction of a candidate density for IS in case of the QERMit approach is obviously larger than for the direct approach. However, this is 'profitable' as the NSE of the estimated VaR – based on 10000 draws – is much smaller than the NSE of the estimator using the direct approach. As the table shows, if one wants to compute an estimate of the VaR with a precision of 1 digit with 95% confidence, (1.96 NSE < 0.05) one needs four times more draws in the 'direct approach' than using the QERMit approach. This corresponds to almost eight minutes for the first approach, whereas QERMit needs only three minutes for the same precision. That is, the computational gain of QERMit is equal to 2.64 (= 477/181). However, when one requires a higher precision this ratio will tend to 4.11 (= 452/110), since the 'investing time' of constructing the candidate will become relatively negligible. To summarize, if one needs a precise Bayesian forecast of a

[2]For small values of 100 $(1-\alpha)$% (like the 1% or 5% percentile), the 'whole' distribution is close to the part of the distribution that does *not* correspond to high losses. Therefore we simulate 50% of the draws from the 'whole' distribution.

multi-step-ahead VaR, then the investment of computing time in an appropriate candidate distribution – (20) with two mixtures of Student-t distributions constructed by MitISEM – is very profitable, as also shown by Figure 3.

Table 2: Estimates of 10-day ahead 99 % VaR forecast for S&P500 based on the Student-t GARCH model. Daily data are used from 1998 - 2002.

	'Direct' MitISEM approach MH-algorithm (mixt of Student-t cand) for parameter draws + direct sampling of future returns paths given parameter draws		**QERMit approach** Adaptive Importance Sampling using a mixture approximation of the optimal candidate distribution	
	estimate	(NSE)	estimate	(NSE)
99 % VaR	-10.62%	0.24%	-10.89%	0.12%
total time	30.3 s		76.4 s	
time construction candidate	25 s		71 s	
time sampling	5.3 s		5.4 s	
draws	10000		10000	
required for % VaR estimate with 1 digit of precision (with 95 % confidence)				
- number of draws	852948		203574	
- computing time	477 s (= 7 min. 57s)		181 s (= 3 min. 1s)	

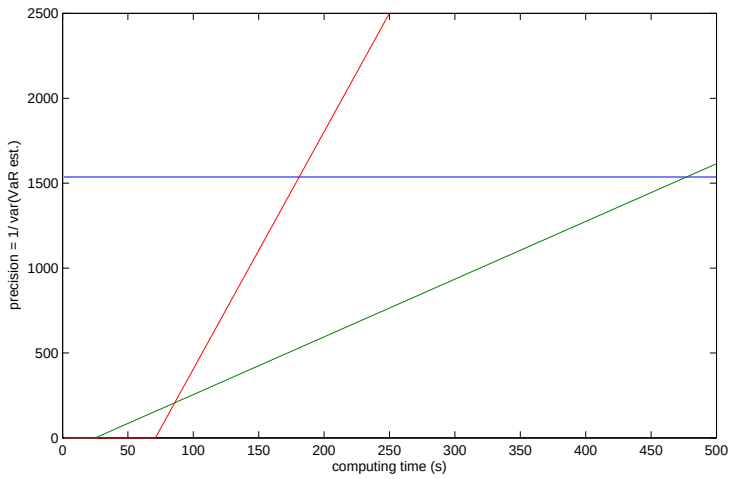

Figure 3: Precision $(1/var)$ *of estimated VaR as a function of the amount of computing time for the 'direct approach' (green line), and the QERMit approach (steepest, red line). The horizontal blue line corresponds to a precision of 1 digit* (1.96 $NSE \leq 0.05$).

2.3 Application III: accurate estimation of posterior model probabilities in case of non-elliptically shaped posteriors

In this subsection we compare the posterior model probabilities of two extensions of the Gaussian Mixture GARCH (1,1) model (17)-(19), the Gaussian Mixture GJR GARCH(1,1) and the Gaussian Mixture EGARCH(1,1) model. In these models, equation (18) is replaced by the GJR specification proposed by Glosten, Jaganathan and Runkle (1993)

$$h_t = \omega + \alpha(y_{t-1} - \mu)^2 + \gamma(y_{t-1} - \mu)^2 I[y_t - \mu < 0] + \beta h_{t-1}, \tag{26}$$

or by the EGARCH specification introduced by Nelson (1990)

$$\log(h_t) = \omega + \gamma \frac{y_{t-1} - \mu}{\sqrt{h_{t-1}}} + \alpha \left(\frac{|y_{t-1} - \mu|}{\sqrt{h_{t-1}}} - \frac{E|y_{t-1} - \mu|}{\sqrt{h_{t-1}}} \right) + \beta \log(h_{t-1}). \tag{27}$$

Both models aim at capturing the 'leverage-effect', i.e. that an unexpected negative shock in the asset price boosts volatility more up than a positive shock of the same magnitude. This effect is discovered by Black (1976) and confirmed by findings of Nelson (1990) and Schwert (1990).

We have no a priori preference for one particular model, so that the posterior odds ratio is equal to the Bayes factor, the ratio of the marginal likelihoods of both models, whereas the marginal likelihood of model M_1 is given by

$$p(y|M_1) = \int p(y|\theta_1, M_1) p(\theta_1|M_1) d\theta_1 \tag{28}$$

where $p(y|\theta_1, M_1)$ is the likelihood of the model and $p(\theta_1|M_1)$ the exact prior density of the parameters θ_1 in model M_1. However, since we use flat priors for the parameters of both models, we can not directly use marginal likelihoods, due to Bartlett's paradox (Bartlett (1957)). In order to get reasonable model probabilities, we compute the *predictive* likelihood of both models. Eklund and Karlsson (2007) show that the sensitivity of model probabilities to the prior choice can be handled using predictive likelihoods and summarize alternative ways to specify and calculate the predictive likelihood. We compute the predictive likelihood as follows. By splitting the data $y = (y_1, \dots y_T)$ into $y^* = (y_1, \dots y_m)$ and $\tilde{y} = (y_{m+1}, \dots y_T)$, the predictive likelihood of model M_1 is given by:

$$p(\tilde{y}|y^*, M_1) = \int p(\tilde{y}|\theta_1, y^*, M_1) p(\theta_1|y^*, M_1) d\theta_1, \tag{29}$$

which is actually the marginal likelihood if we consider $\tilde{y}$ as 'the data' and $p(\theta_1|y^*, M_1)$, the exact posterior density after observing y^*, as the prior. Using Bayes' rule for this exact posterior density $p(\theta_1|y^*, M_1)$ and substituting into (29) yields

$$p(\tilde{y}|y^*, M_1) = \frac{\int p(y|\theta_1, M_1) p(\theta_1|M_1) d\theta_1}{\int p(y^*|\theta_1, M_1) p(\theta_1|, M_1) d\theta_1}. \tag{30}$$

Hence this predictive likelihood is simply the ratio of the marginal likelihood for all observations over the marginal likelihood for the first part of the data.

We estimate these marginal likelihoods by IS, where we compare the performance of the candidate density resulting from MitISEM with the 'adaptive' Student-t density. The computation of a *predictive* likelihood may be yet another reason why one needs an approximation of a non-elliptically shaped target distribution. For the posterior after observing only the first subset of data y^* may 'suffer' more from local non-identification of model parameters than the posterior based on the whole data set y. Roughly stated, the subset of data y^* may not contain strong enough information to 'keep the posterior away from difficult areas (e.g., ridges due to local non-identification) of the parameter space.

In this application, we use again the $S\&P$ 500 data and repeat the simulation-based computation of the predictive likelihoods, Bayes factors and model probabilities 100 times. The first 600 observations are regarded as the 'training sample' $y^* = (y_1, \dots y_m)$. Table 3 shows simulation results. Two main findings arise from the table. First, the NSE values suggest that MitISEM produces far more precise estimates of predictive likelihoods and hence model probabilities. Second, an even more important result is that there is a sizeable difference between the means of the estimated predictive likelihoods from both approaches (over the 100 repetitions). The reason is arguably that the adaptive Student-t candidate density misses an important subdomain of the parameter space. The considerable number of Student-t components in the mixture approximations (between 3 and 6) suggests the presence of rather non-elliptical shapes. In future research we will investigate this difference in more detail. In any case, the example stresses that the specification of an appropriate candidate density may be relevant for estimating model probabilities, and hence for model choice or Bayesian Model Averaging.

Table 3: Model comparison between a Gaussian Mixture GJR-GARCH and a Gaussian Mixture EGARCH model. Means and corresponding NSE values are based on 100 simulation runs. Predictive likelihoods are computed by IS with adaptive Student-t and MitISEM candidate densities.

		Mixt GJR-GARCH		Mixt EGARCH	
MitISEM results					
# components	Training sample	3		5	
in candidate	Full Sample	6		4	
Predictive Likelihood					
		$10^{234}\cdot$ mean	$10^{236}\cdot$ NSE	$10^{234}\cdot$ mean	$10^{236}\cdot$ NSE
	MitISEM	1.70	2.96	7.03	8.46
	Adaptive	1.76	8.42	5.86	22.51
Bayes Factors and Model Probabilities					
		mean	$10^{3}\cdot$ NSE	mean	$10^{3}\cdot$ NSE
Bayes Factor	MitISEM	0.24	4.90		
	Adaptive	0.30	19.10		
Model prob	MitISEM	0.19	3.18		
	Adaptive	0.23	11.18		

3 Sequential MitISEM

In this section, we propose a method for applying MitISEM in a sequential manner, so that the candidate distribution for posterior simulation is cleverly updated when new data become available. Our results show that the computational effort reduces enormously, while the quality of the approximation remains almost unchanged, as compared with an 'ad hoc' procedure in which the construction of the MitISEM candidate is performed 'from scratch' at every moment in time. In the next subsection we show how this sequential approach can be combined with a tempering approach, which facilitates the simulation from densities with multiple modes that are far apart.

The previous sections showed that, although the IS-weighted EM steps are relatively efficient, the construction of an appropriate candidate distribution may still require considerable computing time. After all, it requires evaluations of the target density kernel. This may seem a serious disadvantage if one requires multiple estimates over time, for example daily Bayesian forecasts. However, the idea behind the procedure in this section is that the posterior for data $y_{1:T+1} = \{y_1, \ldots, y_T, y_{T+1}\}$ is typically not so different from the posterior for data $y_{1:T} = \{y_1, \ldots, y_T\}$. Therefore, one can 'recycle' the same candidate distribution. At many moments, the candidate distribution can simply be reused. Further, if the candidate distribution needs to be updated, i.e. if its quality falls below a certain level, then

we still do not require to start from scratch. It may suffice to perform an update using the IS-weighted EM algorithm, keeping the number H of Student-t components the same. Only if the resulting quality is still below a desired level, then we start the MitISEM procedure, adding components until convergence has been reached.

Suppose that at time $T + \tau$ ($\tau = 1, 2, \ldots$) we want to analyze the posterior based on data $y_{1:T+\tau} = \{y_1, \ldots, y_{T+\tau}\}$, and that time T was the last time when we had to update the candidate density. That is, the current candidate distribution has been estimated using the data $y_{1:T}$. Then at time $T + \tau$ we perform the following algorithm:

Algorithm 2. *The Sequential MitISEM approach for obtaining a candidate density for the posterior density for data $y_{1:T+\tau}$:*

(1) Compute *C.o.V.(no update)*, the C.o.V. value that is based on the posterior density kernel for data $y_{1:T+\tau}$ and the current candidate density.

(2) Compare *C.o.V.(no update)* with *C.o.V.(T)*, the C.o.V. value of the last time when the candidate was updated. If the change is below a certain threshold (10%), stop. Otherwise go to step (3).

(3) Run the IS-weighted EM algorithm with the current mixture of H Student-t densities as starting values. Sample from the new distribution (with the same number of components H) and compute IS weights and the corresponding C.o.V. value *C.o.V.(only EM update)*. Since the IS-weighted EM algorithm updates all mixture components, it can easily perform a useful shift of the candidate density.

(4) Judge the value of *C.o.V.(only EM update)*. If the change of quality is below a certain threshold (10%), stop. Otherwise go to step (5).

(5) Iterate on the number of components until the C.o.V. value has converged.

When a particular Student-t component gets a minimal weight, then the practical relevance is negligible. In such a case we delete the Student-t component from the mixture. So, the number of Student-t components is not monotonically increasing over time. In step (2) we compare *C.o.V.(no update)* with *C.o.V.(T)* rather than the C.o.V. for the posterior at time $y_{T+\tau-1}$, since in the latter case a series of small increases of the C.o.V. may eventually lead to a much worse candidate density, without the algorithm ever being 'alarmed' to update the candidate.

We apply the Sequential MitISEM algorithm to the Gaussian Mixture EGARCH model with the S&P 500 data. We estimate the model on the first 1300 observations and recycle the obtained candidate density by adding iteratively one observation of the forecast sample to the existing sample. At each time $t = 1301, \ldots, 1350$, the predictive likelihood is computed. The training sample y^* (for the marginal likelihood in the denominator of the predictive likelihood) consists of 500 observations, and is remained fixed.

We compare the Sequential MitISEM approach with the 'ad hoc MitISEM approach', whichs run the MitISEM algorithm from scratch at each time $t = 1300, \ldots, 1350$. The comparison is twofold. First we compare the computating time that is involved with both methods. Second the quality of the estimates of the predictive likelihood is compared. In order to fulfill the second comparison measure, we repeat the calculation of the predictive likelihoods 100 times and compute the NSE as the standard deviation over the repetitions.

Table (4) compares both methods in computational effort and provides more details about the results of the Sequential MitISEM algorithm. During the forecast sample, the constructed candidate density is adapted only one time (step (3)). In all other cases, it was not necessary in our strategy to adapt the candidate density.

To emphasize that the number of times the candidate density is left unchanged is not a result of coincidence, we have run the Sequential MitISEM approach for a different data set and a different model. We have considered the Gaussian Mixture GJR-GARCH and Gaussian Mixture-EGARCH model, applied to daily log-returns for the SMI-index (1992-1998), data used by Ausín and Galeano (2007). Likewise, we iteratively add one observation of the forecast sample to the starting sample $\tilde{y} = (y_1, \ldots y_{1000})$. The forecast sample is denoted by $y_\tau = y_{1001}, \ldots, y_{1858}$, hence 858 times the candidate density is updated, extended or left unchanged. Table (5) shows that for both models in almost 90% of the cases the current candidate density is recycled, i.e. no adaption or extension is required.

We now turn back to the application of this section. Using the Sequential MitISEM algorithm implies a huge computational advantage, as it is more than 45 times faster than the 'ad hoc MitISEM method'. The Sequential MitISEM algorithm is visualized in Figure (4). The blue line represents *C.o.V.(T)*, the Coefficient of Variation that is used in step (2) for comparison, whereas the green line denotes *C.o.V.(no update)*. Finally the red line gives an impression of the quality of the 'ad hoc MitISEM approach': the average C.o.V. value of the 'ad hoc MitISEM approach' over the same period. When the dataset includes the 25th observation of the forecast sample, the new C.o.V . value is relatively too high. In this case the candidate density is updated which is shown by the upward shift of the blue line, representing the new value of *C.o.V.(T)* (and the new moment T of the latest update). The

Table 4: Results of the Sequential MitISEM algorithm, applied to a Gaussian Mixture EGARCH model, compared with the 'ad hoc MitISEM method', which simply runs the MitISEM algorithm from scratch on each sample $(y_{1:t})$ $t = (1301, \ldots, 1350)$. *The number of times adapted denotes the case when the candidate is only updated, using IS-weighted EM, while the number of components is held constant. When the candidate is adapted and extended, the number of components increases. Reusing the candidate density implies that the same candidate density is held, hence no updating occurs.*

	Sequential MitISEM	Adhoc MitISEM
Sequential MitISEM		
# adapted	1	
# adapted and extended	0	
# reused	48	
Computational effort		
Construct 50 candidate densities over period $(1300 - 1349)$	117 s	5602 s

figure suggests that the quality of Sequential MitISEM is approximately the same as the 'ad hoc MitISEM approach', since the difference in C.o.V. values is quite small. (Note that the y-axis corresponds to merely the interval $[0.66, 0.84]$.)

An additional indication is given by Figure 5, which shows the mean of 100 predictive likelihoods with 95% confidence bounds. Since the blue and red asterisks lie most of the time in both confidence intervals, we suggest again that the quality of the Sequential MitISEM algorithm is of the same order as the 'ad hoc MitISEM approach'. We further note that the same procedure can be used if one makes use of a *moving window* instead of the *expanding window* of data that we use. To conclude this subsection, Sequential MitISEM is far more efficient compared to a 'ad hoc approach' as it produces approximately the same quality of candidate distributions for predictive likelihood estimation with considerably less computational effort.

Table 5: Results of the Sequential MitISEM algorithm, applied to a Gaussian Mixture EGARCH model and Gaussian Mixture GJR-GARCH model. The date consist of daily log-returns from the SMI-index from July 1991 until August 1998 (data of Ausín and Galeano (2007)). The models are estimated on the first 1000 obervations and recycled after iteratively adding one observation of the forecast sample $(t = 1001, \ldots, 1859)$ *to the existing sample. The number of times adapted denotes the case when the candidate is only updated, using IS-weighted EM, while the number of components is held constant. When the candidate is adapted and extended, the number of components increases. Reusing the candidate density implies that the same candidate density is held, hence no updating occurs.*

	Mixture EGARCH(1,1)	**Mixture GJR-GARCH(1,1)**
# adapted	56	38
# adapted and extended	45	33
# reused	757	787

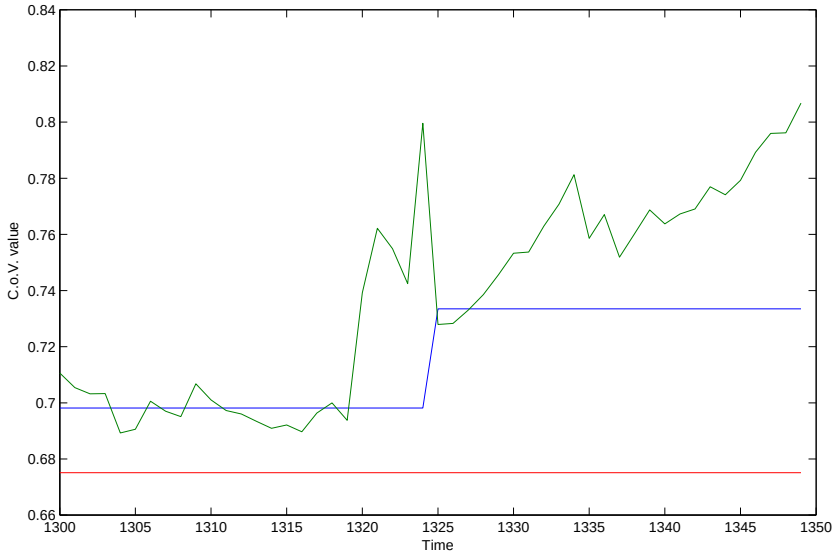

Figure 4: The blue line represents C.o.V.(T)*, the Coefficient of Variation that is used for comparison in step (2) of the Sequential MitISEM approach, whereas the green line denotes* C.o.V.(no update)*. Finally the red line gives an impression of the quality of the 'ad hoc MitISEM approach': the average C.o.V. value of the 'ad hoc MitISEM approach' over the same period. When the dataset includes the 25th observation of the forecast sample, the new C.o.V . value is relatively too high. In this case the candidate density is updated which is shown by the upward shift of the blue line, representing the new value of* C.o.V.(T) *(and the new moment* T *of the latest update).*

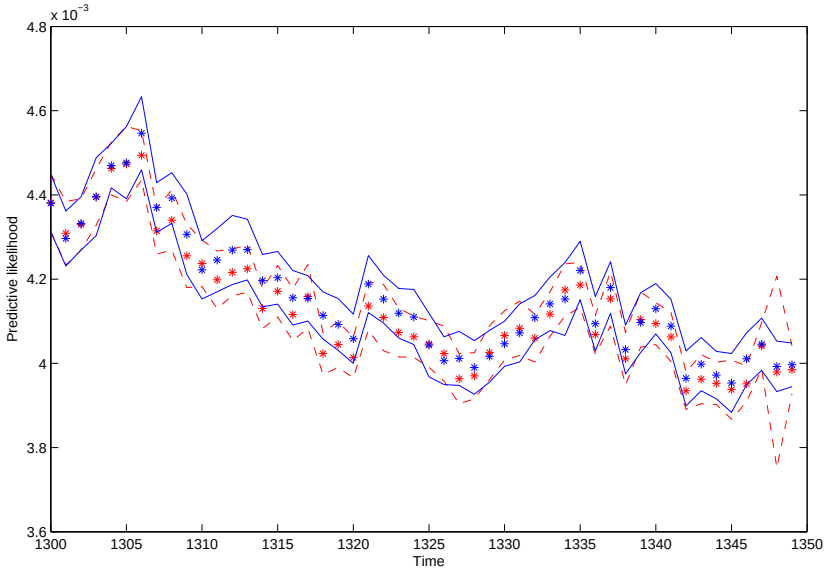

Figure 5: Predictive likelihood estimates based on a Gaussian Mixture EGARCH model. The asterisks show at each time the mean of 100 predictive likelihoods; the red and blue line correspond with 95% confidence bounds (estimated from the 100 repetitions). The red asterisks and confidence bounds are based the 'ad hoc MitISEM approach', where each day the MitISEM approach is applied from scratch. The blue asterisks and confidence bounds are based on the Sequential MitISEM algorithm.

3.1 Tempered MitISEM

Although the MitISEM approach can approximate multimodal target distributions, it may occur in extreme cases that the modes of a target distribution are so wide apart that one or more of the modes are 'missed'. To decrease the probability that distant modes are 'missed', one can combine MitISEM with a tempering approach. The proposed tempering method moves sequentially from a tempered target density kernel, the target density kernel to the power of a positive number that is smaller than 1, towards the real target density kernel. The tempered target distribution is more diffuse, roughly stated 'more uniform', and hence the probability of detecting far-away modes is higher. The tempering idea is used in the Equi-Energy sampler, developed by Kou, Zhou and Wong (2006).

We apply the tempering approach in the following way as a Sequential MitISEM algorithm. Given a target kernel $f(\theta)$, we temper this kernel by raising it to the power $(1/P_0)$ with $P_0 > 1$, i.e. $f(\theta)^{1/P_0}$. The MitISEM algorithm is applied to this tempered kernel $f(\theta)^{1/P_0}$. The resulting mixture of Student-t densities is used as input for the updated tempered target kernel, say $f(\theta)^{1/P_1}$, with $1 \leq P_1 < P_0$. This approach is repeated by decreasing P_n $(n = 0, 1, 2, \ldots, \tilde{n})$ iteratively to $P_{\tilde{n}} = 1$, corresponding to the real target kernel. Many

possible choices can be made on the number of iterations and the distance between the P_n. We follow Kou, Zhou and Wong (2006), and take equidistant steps of $\log(P_n)$. We label this approach the Tempered MitISEM procedure.

We apply the Tempered MitISEM approach to the same highly multimodal density that is used by Kou, Zhou and Wong (2006): the two-dimensional normal mixture:

$$f(x) = \sum_{i=1}^{20} \frac{w_i}{2\pi\sigma_i^2} \exp\left(-\frac{1}{2\sigma_i^2}(x-\mu_i)'(x-\mu_i) \right) \tag{31}$$

where $\sigma_1 = \ldots = \sigma_{20} = 0.1, w_1 = \ldots = w_{20} = 0.05$, and the 20 mean vectors

$$\begin{aligned}(\mu_1, \mu_2, \ldots, \mu_{20}) \quad = \quad & \begin{pmatrix} 2.18 & 8.67 & 4.24 & 8.41 & 3.93 & 3.25 & 1.70 \\ 5.76 & 9.59 & 8.48 & 1.68 & 8.82 & 3.47 & 0.50 \end{pmatrix} \\ & \begin{pmatrix} 4.59 & 6.91 & 6.87 & 5.41 & 2.70 & 4.98 & 1.14 \\ 5.60 & 5.81 & 5.40 & 2.65 & 7.88 & 3.70 & 2.39 \end{pmatrix} \\ & \begin{pmatrix} 8.33 & 4.93 & 1.83 & 2.26 & 5.54 & 1.69 \\ 9.50 & 1.50 & 0.09 & 0.31 & 6.86 & 8.11 \end{pmatrix}. \end{aligned} \tag{32}$$

Since most local modes are 15 standard deviations away from the nearest one, this mixture distribution is a good test for our approach. We compare three methods. First the Tempered MitISEM approach is used. In more detail, we choose $P_0 = 5$ and move sequentially in five steps to $P_5 = 1$ with equally (log) spaced intervals. Second, we apply the MitISEM algorithm to the real target density, hence no tempering approach is used. The final sampler is an ordinary Student-t distribution with adapted mode and scale matrix.

Figures 6 and 7 and Table 6 show simulation results from these three methods. First of all, panel (A^*) of Figure 6 suggests that the 'adaptive' Student-t density produces poor results. In other words, one really needs advanced samplers to handle multimodal target kernels. Second, the MitISEM approach without tempering is a serious improvement, as the C.o.V. value decreases substantially from 23 to 0.77. The MitISEM algorithm is able to detect most of the modes, however by comparing panel (C^*) to panel (D^*) of Figure 6, which represents simulated draws from the target density, not all modes are covered. The mode around $(8.41, 1.68)$ is missed by MitISEM. This reflects that if the mode lies to far away from the remaining modes, MitISEM may not be able to detect this important subdomain of the target density. Finally, the 'tempered MitISEM' approach is shown in Figure 7. From panel A to E, candidate draws are shown for the target $p(\theta)^{1/P}$, where P is equally log-spaced from 5 to 1. The importance of sequentially lowering the value of P_n lies in the fact that first the global area of interest is captured. Then a lower P_n in the subsequent panels shows an increasing precision of the local modes. In the end, the improvement of

'tempered MitISEM' over the basic MitISEM algorithm is clearly illustrated in panel (E), since all 20 modes are covered. The quality of the final candidate density is also confirmed by Table 6, as the *C.o.V.* value drops further from 0.77 to 0.43.

Table 6: Results of simulation from the two-dimensional normal mixture (31) by three different candidates: an (adaptive) Student-t density, and mixtures of Student-t densities with and without tempering.

	Adaptive t	**MitISEM**	**Tempered MitISEM**
Number of components in candidate mixture	1	14	16
C.o.V. of IS weights	21.57	0.78	0.43

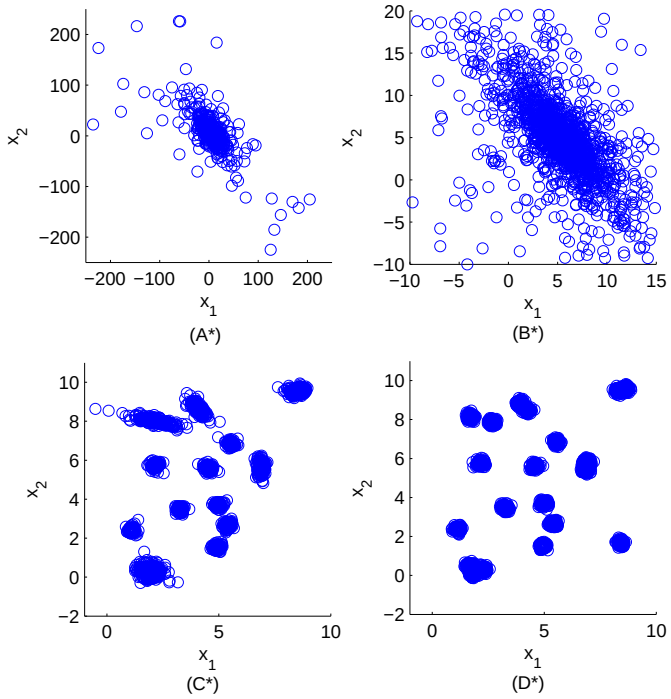

Figure 6: Samples generated by the Adaptive Student-t density (panel (A) and (B*)) and the MitISEM algorithm (panel (C*)). panel(D*) shows draws simulated from the real target distribution given in (31).*

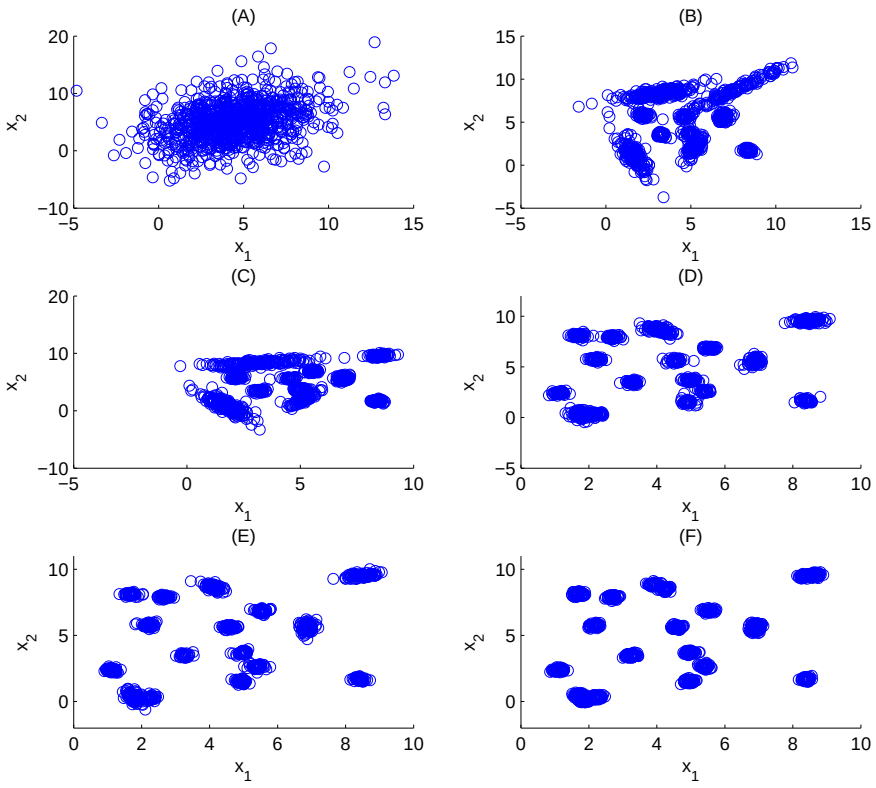

Figure 7: Samples generated form each step of the Sequential MitISEM algorithm. Starting from panel (A) to (E), P (in the target kernel $f(\theta)^{1/P}$) is equally log-spaced from 5 to 1. Panel (F) shows draws simulated from the real target distribution given in (31).

4 Permutation-augmented MitISEM

In this section, we introduce a permutation-augmented MitISEM approach, for importance sampling (or the MH algorithm) from posterior distributions in mixture models without the requirement of imposing *a priori* identification restrictions on the mixture components' parameters. As discussed by Geweke (2007), the mixture model likelihood function is invariant with respect to permutation of the components of the mixture. If functions of interest are permutation sensitive, as in classification applications, then interpretation of the likelihood function requires valid inequality constraints. If functions of interest are permutation invariant, as in prediction applications, then there are no such problems of interpretation. Geweke (2007) proposes the permutation-augmented Gibbs sampler, which can be considered as an extension of the random permutation sampler of Frühwirth-Schnatter (2001). The practical implementation of the idea of the permutation-augmented Gibbs sampler is that one simulates a Gibbs sequence with total disregard for label switching or the prior's labeling restrictions. Only after that and only if functions of interest are permutation sensitive, then one simply permutes the Gibbs sampler's output so as to satisfy the labeling restrictions. We propose a method of permutation-augmented IS, for which we extend the MitISEM approach to construct an approximation to the unrestricted posterior, taking into account the permutation structure. If m is the number of components of the mixture model, then the addition of a Student-t component to the candidate implies an addition of the $m!$ equivalent permutations. Thereby, we construct a mixture of mixtures of $m!$ Student-t components, where the restriction is imposed that the $m!$ permutations have equal candidate density. Intuitively stated, we help the basic MitISEM approach by 'telling' it about the invariance with respect to permutations. It should be noted that this invariance with respect to permutations is not the only possible cause of non-elliptical shapes in a mixture model's posterior. For example, if the probability of one of the model's components tends to zero, the local non-identification of the component's other parameters causes ridge shapes.

To illustrate our permutation-augmented method, we consider mixtures of m normal distributions. We assume that y_t are independently distributed with

$$y_t \sim N(\mu_j, \sigma_j^2) \qquad \text{if } z_{tj} = 1 \qquad (t = 1, \ldots, T; j = 1, \ldots, m),$$

where $z_t = (z_{t1}, \ldots, z_{tJ})'$ is a vector of latent 0/1 variables of which exactly one of the m elements is equal to 1, where

$$\Pr[z_{tj} = 1] = \pi_j \qquad (t = 1, \ldots, T; j = 1, \ldots, m).$$

Define $y = (y_1, \ldots, y_T)'$ and $z = \{z_1, \ldots, z_T\}$. Then the likelihood is given by:

$$p(y|\theta) = \prod_{t=1}^{T} \left\{ \sum_{j=1}^{m} \pi_j \left[(2\pi)^{-1/2} \sigma_j^{-1} \exp\left(-\frac{1}{2\sigma_j^2} (y_t - \mu_j)^2 \right) \right\} \right]. \tag{33}$$

with $\theta = (\mu_1, \ldots, \mu_m, \sigma_1, \ldots, \sigma_m, \pi_1, \ldots, \pi_{m-1})$, where $\pi_m \equiv 1 - \sum_{j=1}^{m-1} \pi_j$. We use proper non-informative priors for all parameters θ: truncated uniform priors for μ_j and $\log \sigma_j$ and $(\pi_1, \ldots, \pi_{m-1}, \pi_m) \sim Dirichlet(1, 1, \ldots, 1)$.

First, we consider the simple case of $m = 2$ with $\mu_1 = \mu_2 = 0$, so that $\theta = (\sigma_1, \sigma_2, \pi_1)$. We simulate 250 observations from this model with true values $\theta = (\sigma_1, \sigma_2, \pi_1) = (1,\ 2,\ 0.8)$. The top panel of Figure 8 shows the shapes of the unrestricted posterior distribution. In addition to the multimodality due to the absence of identification restrictions, the distribution 'per mode' is also non-elliptical in the sense of 'curved contours'.

The bimodal shapes reflect that the model with parameter values $(\sigma_1, \sigma_2, \pi_1)$ and the permuted version $(\sigma_2, \sigma_1, 1 - \pi_1)$ are obviously equivalent. We will use the subscript c to denote the permutations of the original vector θ. In the case of $m = 2$ components with $m! = 2$ permutations, we use $\theta_{c=1}$ for the original parameter vector, and $\theta_{c=2}$ for the permuted version. For the model with $m = 3$ and $\mu_1 = \mu_2 = \mu_3 = 0$, we have $\theta = (\sigma_1, \sigma_2, \sigma_3, \pi_1, \pi_2)$. Here we have $m! = 6$ permutations θ_c $(c = 1, \ldots, m!)$. For an explanation of our notation θ_c we refer to Table 7. During the permutation-augmented algorithm we also make use of the inverse permutation $\theta_{inv(c)}$, defined such that $(\theta_{inv(c)})_c = (\theta_c)_{inv(c)} = \theta$. In the case of $m = 2$ regimes, $\theta_{inv(c)} = \theta_c$; there are only two options, leaving θ the same or switching the two regimes, where applying the same operation twice always returns the original θ. The case of $m = 3$ regimes is somewhat less straightforward; there are two permutations that require a different permutation to return to the original θ. Table 7 provides the details.

The basic idea of the permutation-augmented MitISEM approach is the same as the basic, 'plain vanilla' MitISEM. However, there are subtle differences in the IS-weighted EM algorithm. Instead of H Student-t components h $(h = 1, \ldots, H)$, the candidate distribution now consists of $H \cdot m!$ Student-t components (h, c) $(h = 1, \ldots, H; c = 1, \ldots, m!)$, where for each Student-t component (h, c) $\mu_{h,c}$, $\Sigma_{h,c}$ are permuted versions of $\mu_h = \mu_{h,1}$ and $\Sigma_h = \Sigma_{h,1}$; further we have $\nu_{h,c} = \nu_h$ and $\eta_{h,c} = \eta_h / m!$. Instead of (9)-(12), the conditional expectations of the latent variables given θ^i and $\zeta = \zeta^{(L-1)}$, the optimal parameters in the previous EM

iteration, are given by:

$$\tilde{z}^i_{h,c} \equiv E\left[z^i_{h,c}\left|\theta^i, \zeta = \zeta^{(L-1)}\right.\right] = \frac{t(\theta^i|\mu_{h,c}, \Sigma_{h,c}, \nu_h)\ \eta_h}{\sum_{j=1}^{J}\sum_{l=1}^{m!} t(\theta^i|\mu_{j,l}, \Sigma_{j,l}, \eta_j)\ \eta_j}. \tag{34}$$

$$\widetilde{z/w}^i_{h,c} \equiv E\left[z^i_{h,c}\frac{1}{w^i_h}\middle|\theta^i, \zeta = \zeta^{(L-1)}\right] = \tilde{z}^i_{h,c}\ \frac{k+\nu_h}{\rho^i_{h,c}+\nu_h}. \tag{35}$$

$$\begin{aligned}\xi^i_h &\equiv E\left[\log w^i_h\left|\theta^i, \zeta = \zeta^{(L-1)}\right.\right] = \\ &= \sum_{c=1}^{m!}\left\{\left[\log\left(\frac{\rho^i_{h,c}+\nu_h}{2}\right) - \psi\left(\frac{k+\nu_h}{2}\right)\right]\tilde{z}^i_{h,c}\right\} \\ &\quad + \left[\log\left(\frac{\nu_h}{2}\right) - \psi\left(\frac{\nu_h}{2}\right)\right]\left(1 - \sum_{c=1}^{m!}\tilde{z}^i_{h,c}\right),\end{aligned} \tag{36}$$

$$\begin{aligned}\delta^i_h &\equiv E\left[\frac{1}{w^i_h}\middle|\theta^i, \zeta = \zeta^{(L-1)}\right] \\ &= \sum_{c=1}^{m!}\frac{k+\nu_h}{\rho^i_{h,c}+\nu_h}\ \tilde{z}^i_{h,c} + \left(1 - \sum_{c=1}^{m!}\tilde{z}^i_{h,c}\right).\end{aligned} \tag{37}$$

with $\rho^i_{h,c} = (\theta^i - \mu_{h,c})'\Sigma^{-1}_{h,c}(\theta^i - \mu_{h,c})$, and all parameters $\mu_{h,c}, \Sigma_{h,c}, \nu_h, \eta_h$ elements of $\zeta^{(L-1)}$. Instead of (13)-(15), the expressions of the Maximization step are given by:

$$\mu^{(L)}_h = \left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i\ \widetilde{z/w}^i_{h,c}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i\ \widetilde{z/w}^i_{h,c}\ \theta^i_{inv(c)}\right], \tag{38}$$

$$\hat{\Sigma}^{(L)}_h = \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i\ \widetilde{z/w}^i_{h,c}\ (\theta^i_{inv(c)} - \mu^{(L)}_h)(\theta^i_{inv(c)} - \mu^{(L)}_h)'}{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i\ \tilde{z}^i_{h,c}}, \tag{39}$$

$$\eta^{(L)}_h = \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i\ \tilde{z}^i_{h,c}}{\sum_{i=1}^{N} W_i}, \tag{40}$$

whereas the equation of the first order condition for ν_h remains (16). For the derivations we refer to the appendix.

We apply the permutation-augmented MitISEM approach to the posterior distribution in the top panel of Figure 8, resulting in a mixture of $5{\cdot}2$ Student-t distributions shown in the bottom panel of Figure 8. We use this candidate in the IS and MH methods to estimate the standard deviation of y_t $(t = 1, \ldots, T)$, $\sigma = \sqrt{\sum_{j=1}^{m}\pi_j\ (\sigma_j^2 + \mu_j^2) - \mu^2}$ with $\mu = \sum_{j=1}^{m}\pi_j\ \mu_j$. This quantity is clearly not permutation-sensitive, so that we do not require identification restrictions. The results are in the first row of Table 8. The C.o.V. of the IS weights and the high MH acceptance rate reflect the accuracy of the MitISEM approximation. Table 9 shows the results of Gibbs sampling (with data augmentation), which requires more computing time to reach the same accuracy. If we would desire a higher level of precision, then the difference in computing time would be enormous, since simulating 10000 extra draws requires much more time in the Gibbs sampler.

At this point, we must address a disadvantage of the permutation-augmented MitISEM approach. The number of expectations of latent variables $\tilde{z}_{h,c}^i$ and $\widetilde{z/w}_{h,c}^i$ in (34) and (35) that need to be computed, increases with the factorial $m!$ of the number of regimes in the model. This implies that we should only apply the permutation-augmented MitISEM approach with a 'limited' value of m. The second, third and fourth row of Tables 8 and 9 show that the permutation-augmented MitISEM approach is at least feasible (and useful) for $m = 2$, $m = 3$ and $m = 4$ regimes (with $2! = 2$, $3! = 6$ and $4! = 24$). For each setting, we simulated 250 observations, applied the permutation-augmented MitISEM approach, and compared the results of IS with the Gibbs sampler. Again, the Gibbs sampler requires more computing time to reach the same (or worse) accuracy. Since the increase from $4! = 24$ to $5! = 120$ is obviously huge, the permutation-augmented MitISEM algorithm may have its practical limit at $m = 4$.

It should be noted that the permutation-augmented MitISEM approach outperforms the Gibbs sampler, even though the latter does not suffer from a large serial correlation in the Gibbs sequence (the first order serial correlation is at each instance below 0.30), which may be a problem in other settings. Further, the IS approach has the advantage that an estimate of the marginal likelihood is immediately available as the average of the IS weights, whereas for the Gibbs sampler the method of Chib (1995) would require additional reduced runs.

In the next section, we will consider an empirical example involving an extended version of the permutation-augmented MitISEM algorithm for a subset of the parameters, where the candidate Student-t components' means are allowed to depend on the draw of a different subset of parameters.

We now explain why we do not use the permutation-augmented MitISEM approach for the *mixture* GARCH models in previous sections. First, suppose that we relax the identification restriction $0 < \lambda < 1$. Then the models with parameter values $(\mu, \omega, \alpha, \beta, \lambda, \rho)$ and $(\mu, \omega, \alpha, \beta, 1/\lambda, 1 - \rho)$ would be equivalent. Suppose we have a mixture of Student-t distributions that approximates the distribution of $(\mu, \omega, \alpha, \beta, \lambda, \rho)$. Then we do not have a mixture of Student-t distributions that approximates the distribution of $(\mu, \omega, \alpha, \beta, 1/\lambda, 1 - \rho)$. That is, if (X, Y) has a bivariate Student-t distribution, then $(1/X, Y)$ does not have a bivariate Student-t distribution. This reflects that the permutation-augmented MitISEM approach should only be used if equivalent parameter vectors are linear combinations of each other, which is typically the case. The mixture GARCH model is an exception. Second, we follow Ausín and Galeano (2007) in imposing an informative prior that incorporates the restriction that the regime with the smallest variance has the highest probability. Therefore, we have a 'real' restriction, not merely identification restrictions.

Also if equivalent parameter vectors are linear combinations of each other, one could still impose identification restrictions and apply the basic MitISEM approach to the restricted posterior. However, permutation-augmented MitISEM is typically more efficient for several reasons: (i) during the construction of the candidate we make use of the a priori knowledge on the permutation-invariant structure; (ii) no draws are rejected that do not satisfy the identification restrictions; if we desire to compute permutation-sensitive quantities of interest, then these draws are simply permuted such that they do satisfy the identification restrictions. (iii) imposing identification restrictions may itself lead to more irregular shapes of the target distribution.

The use of the tempered MitISEM approach would not make sense here, since we know a priori the permutation-invariant structure, so that we know all modes as soon as we find one mode. The tempered MitISEM approach needs only to be used when we are confronted with multimodality having an 'unknown structure'.

Finally, we note that also in mixture models with more than 4 regimes the permutation-augmented MitISEM approach can be useful. Although in such cases we can not proceed without any identification restrictions, we can still use permutation-augmented MitISEM to reduce the number of identification restrictions. For example, in a mixture of 6 normal distributions, we can impose that the first and last have the smallest and largest variance (or mean), whereas the 4 middle regimes are left unrestricted. This may still have the same positive effect on the computing time and the quality of the candidate.

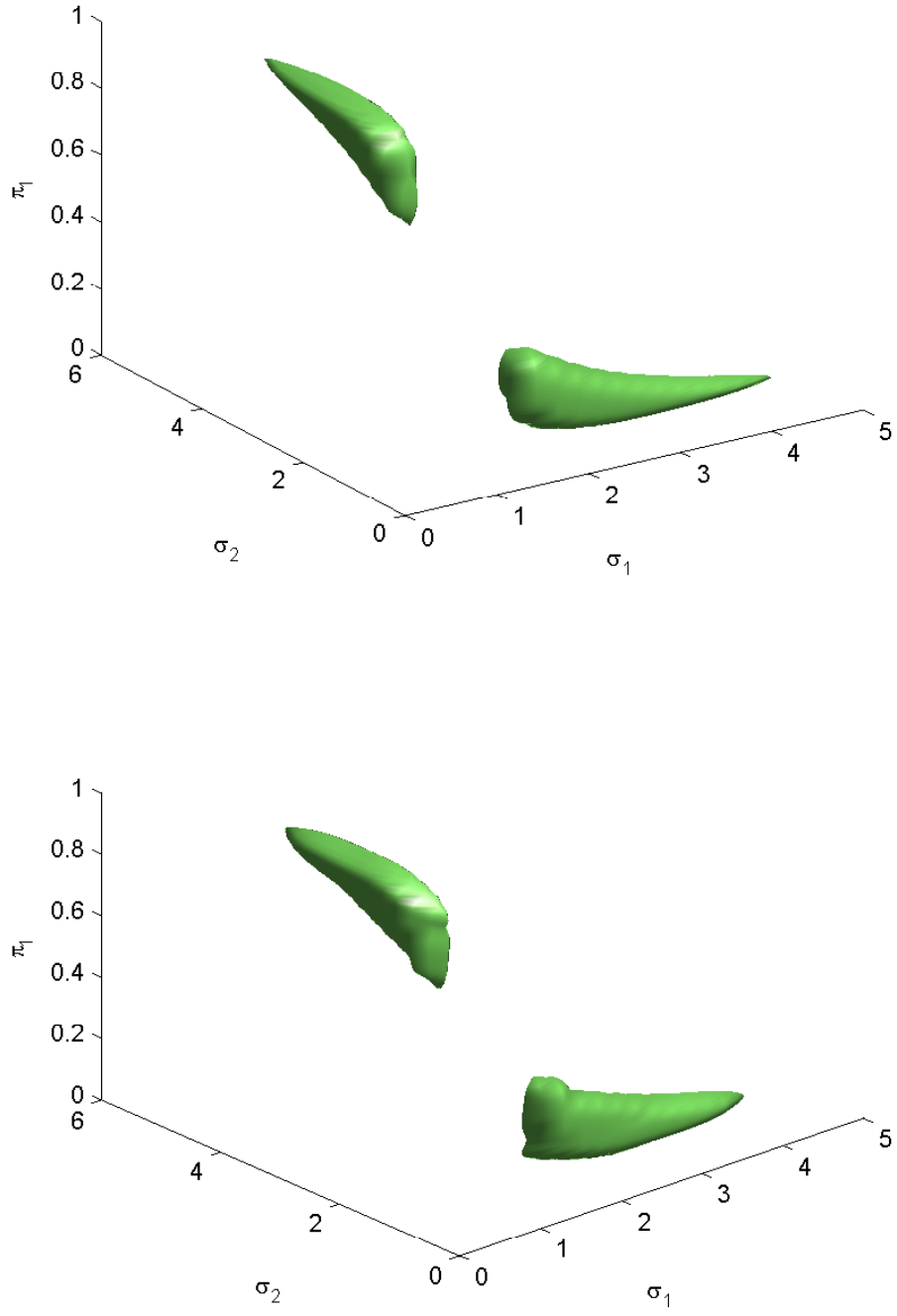

Figure 8: Mixture of two normal distributions: Highest Posterior Density credible region of $\theta = (\sigma_1, \sigma_2, \pi_1)$ *(top) and 'Highest Candidate Density region' for mixture of* $2 \cdot 5$ *Student-t candidate distribution (bottom), constructed by permutation-augmented MitISEM algorithm.*

Table 7: Explanation of notation for permutation θ_c and inverse permutation $\theta_{inv(c)}$ in mixture models with $m = 2$ and $m = 3$ regimes with parameter vector θ. The examples that are referred to are the mixtures of normal distributions with $\mu_j = 0$ $(j = 1, \ldots, m)$

Mixture model with $m = 2$ components and $m! = 2$ permutations:					
c	permutation	θ_c in example	inverse permutation	$\theta_{inv(c)}$ in example	inv(c)
1	(1,2) → (1,2)	$(\sigma_1, \sigma_2, \pi_1)$	(1,2) → (1,2)	$(\sigma_1, \sigma_2, \pi_1)$	1
2	(1,2) → (2,1)	$(\sigma_2, \sigma_1, 1 - \pi_1)$	(1,2) → (2,1)	$(\sigma_2, \sigma_1, 1 - \pi_1)$	2

Mixture model with $m = 3$ components and $m! = 6$ permutations:					
c	permutation	θ_c in example	inverse permutation	$\theta_{inv(c)}$ in example	inv(c)
1	(1,2,3) → (1,2,3)	$(\sigma_1, \sigma_2, \sigma_3, \pi_1, \pi_2)$	(1,2,3) → (1,2,3)	$(\sigma_1, \sigma_2, \sigma_3, \pi_1, \pi_2)$	1
2	(1,2,3) → (1,3,2)	$(\sigma_1, \sigma_3, \sigma_2, \pi_1, 1 - \pi_1 - \pi_2)$	(1,2,3) → (1,3,2)	$(\sigma_1, \sigma_3, \sigma_2, \pi_1, 1 - \pi_1 - \pi_2)$	2
3	(1,2,3) → (2,1,3)	$(\sigma_2, \sigma_1, \sigma_3, \pi_2, \pi_1)$	(1,2,3) → (2,1,3)	$(\sigma_2, \sigma_1, \sigma_3, \pi_2, \pi_1)$	3
4	(1,2,3) → (2,3,1)	$(\sigma_2, \sigma_3, \sigma_1, \pi_2, 1 - \pi_1 - \pi_2)$	(1,2,3) → (3,1,2)	$(\sigma_3, \sigma_1, \sigma_2, 1 - \pi_1 - \pi_2, \pi_1)$	5
5	(1,2,3) → (3,1,2)	$(\sigma_3, \sigma_1, \sigma_2, 1 - \pi_1 - \pi_2, \pi_1)$	(1,2,3) → (2,3,1)	$(\sigma_2, \sigma_3, \sigma_1, \pi_2, 1 - \pi_1 - \pi_2)$	4
6	(1,2,3) → (3,2,1)	$(\sigma_3, \sigma_2, \sigma_1, 1 - \pi_1 - \pi_2, \pi_2)$	(1,2,3) → (3,2,1)	$(\sigma_3, \sigma_2, \sigma_1, 1 - \pi_1 - \pi_2, \pi_2)$	6

Table 8: Simulation results for IS and the MH algorithm, using the candidate distribution resulting from the permutation-augmented MitISEM procedure, for posterior simulation in mixture models with normally distributed regimes

	posterior mean of σ	NSE	time for construction of MitISEM candidate (in s)	time for simulating 10000 draws	C.o.V. of IS weights	MH accep-tance rate	number of t components in MitISEM
$m = 2$ ($\mu_j = 0$)	1.2360	0.0009	40.48	0.72	0.36	0.84	5
$m = 2$ (μ_j in model)	4.9339	0.0009	19.63	0.79	0.30	0.83	3
$m = 3$ (μ_j in model)	7.4978	0.0014	23.20	1.24	0.47	0.74	2
$m = 4$ (μ_j in model)	10.8300	0.0031	75.08	1.89	0.61	0.67	2

Table 9: Simulation results for Gibbs sampling (with data augmentation) for posterior simulation in mixture models with normally distributed regimes

	posterior mean of σ	NSE	time for simulating 10000 draws + 1000 burn-in)
$m = 2$ ($\mu_j = 0$)	1.2358	0.0009	42.65
$m = 2$ (μ_j in model)	4.9330	0.0009	46.75
$m = 3$ (μ_j in model)	7.4963	0.0021	66.11
$m = 4$ (μ_j in model)	10.8267	0.0029	84.54

5 Partial MitISEM

In this section, we propose a partial MitISEM approach, which aims at approximating the marginal and conditional posterior distributions of subsets of model parameters, rather than the joint. This division can substantially reduce the dimension of the approximation problem, which facilitates the application of adaptive importance sampling for posterior simulation in more complex models with larger numbers of parameters. Approximating the joint posterior density kernel with a mixture of Student-t distributions allows for a huge flexibility of shapes. However, rarely all of this flexibility is required. It is typically enough to use mixtures of Student-t distributions for the dependence *within* subsets of the parameters. We can often divide the parameters into subsets, where the dependence *between* different subsets is less complicated. Our partial MitISEM approach is to divide the model parameters into ordered subsets, where the conditional candidate distributions' means are linear combinations of (functions of) the parameters in previous subsets. The conditional candidate distributions' covariances can also be made to depend on the parameters in previous subsets, by allowing the probabilities of the mixture components of the conditional candidate distribution to differ for different ranges of values for functions of the parameters in previous subsets. We will analyze this extension, which still fits within the framework of the IS-weighted EM algorithm, in future research. The partial MitISEM approach is a way to provide a usable approximation to the posterior, while preventing problems such as numerical issues with specifying huge covariance matrices for a joint candidate distribution – problems that have led researchers to conclude that IS necessarily suffers from a 'curse of dimensionality'.

Intuitively, the idea behind the basic MitISEM approach is as follows. First, the asymptotic normal distribution $N(\theta_{mode}, -H(\theta_{mode})^{-1})$, with θ_{mode} the mode of the target distribution, and $H(\theta_{mode})$ the Hessian of the log-target distribution at the mode, is replaced by a Student-t distribution $t(\theta_{mode}, -H(\theta_{mode})^{-1}, \nu)$ with low degrees of freedom ν to have fat tails. Second, $t(\theta_{mode}, -H(\theta_{mode})^{-1}, \nu)$ is replace by a mixture of Student-t distributions with optimized modes, scale matrices, degrees of freedom and weights, to have more flexibility of the candidate's shapes.

The intuitive idea behind the partial MitISEM approach is as follows. Divide the set of parameters θ into two subsets θ_1 and θ_2. The asymptotic normal distribution $\theta \sim N(\mu = \theta_{mode}, \Sigma = -H(\theta_{mode})^{-1})$ is equivalent with

$$\theta_1 \sim N(\mu_1, \Sigma_{11}) \tag{41}$$

$$\theta_2|\theta_1 \sim N(\mu_2 + \Sigma_{22}^{-1}\Sigma_{21}(\theta_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}). \tag{42}$$

with

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

In the partial MitISEM approach we replace both normal distributions of (41) and (42) by mixtures of Student-t distributions, with optimized scale matrices, degrees of freedom and weights of the marginal candidate of θ_1 and conditional candidate of θ_2 given θ_1. For θ_1 we further optimize the Student-t components' modes, such that this reduces to the basic MitISEM method. For the conditional candidate of θ_2 we use a slightly different IS-weighted EM algorithm in which *coefficients* are optimized. That is, we basically replace μ_2 and $\Sigma_{22}^{-1}\Sigma_{21}$ by optimized coefficients that are allowed to differ between the Student-t components. Moreover, the conditional means are allowed to be a linear combination of *non-linear* functions of θ_1 (and the given data set).

Suppose we have S subsets of parameters θ_s ($s = 1, \ldots, S$). Then the partial MitISEM approach constructs one marginal candidate distribution of θ_1, and $S-1$ conditional candidate distributions (θ_2 given θ_1; θ_3 given θ_1, θ_2; $\ldots$; θ_S given $\theta_1, ..., \theta_{S-1}$), by iteratively adding Student-t components until for all subsets the latest addition has not caused a substantial improvement of the candidate, as an approximation to the target. For the marginal distribution of θ_1 we use the basic IS-weighted EM algorithm. However, for the conditional distribution of θ_s ($k_s \times 1$) given $\theta_1, \ldots, \theta_{s-1}$ we use an extended version where $\mu_h = \beta_h X$ ($h = 1, \ldots, H$), with β_h a $k_s \times r$ matrix and X an $r \times 1$ vector (of which the elements are functions of $\theta_1, \ldots, \theta_{s-1}$ (and the given data)). To obtain the appropriate Expectation and Maximization steps in the IS-weighted EM algorithm, one substitutes $\mu_h = \mu_h^i = \beta_h X^i$. Moreover, (13) is replaced by

$$\beta_h^{(L)'} = \left[\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i \ X_i \, X_i' \right]^{-1} \left[\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i \ X_i \, \theta^{i'} \right], \tag{43}$$

or in case of the permutation-augmented MitISEM approach (38) is replaced by

$$\beta_h^{(L)'} = \left[\sum_{i=1}^{N} \sum_{c=1}^{m!} W_i \ \widetilde{z/w}_{h,c}^i \ X_i \, X_i' \right]^{-1} \left[\sum_{i=1}^{N} \sum_{c=1}^{m!} W_i \ \widetilde{z/w}_{h,c}^i \ X_i \, \theta_{inv(c)}^{i'} \right]. \tag{44}$$

We apply the partial MitISEM approach to an instrumental variables model in which the distribution of the error terms is a mixture of two normal distributions. We use quarter of birth as an instrumental variable for education. The data are from Angrist and Krueger (1991): 8933 observations on individuals of the state of Kentucky, the state in which the

instrument is the strongest (or the 'least weak'), in the sense that the multiple F-test of the first stage regression has the smallest (significant) p-value.

The dependent variable y_t is the log of weekly income of individual t, the possibly endogenous regressor x_t is the number of years of education, z_t consists of three dummies indicating quarter of birth (the first quarter being the reference category). The structural form of the model is:

$$y_t = x_t\beta + \varepsilon_t \tag{45}$$

$$x_t = z_t\gamma + v_t \tag{46}$$

with

$$(\varepsilon_t, v_t)' \sim N(0, \Sigma_j) \qquad \text{if } Z_{tj} = 1 \qquad (t = 1, \ldots, T; j = 1, 2),$$

and

$$\Pr[Z_{tj} = 1] = \pi_j \qquad (t = 1, \ldots, T; j = 1, 2).$$

The restricted reduced form is:

$$y_t = z_t\gamma\beta + v_{1t} \tag{47}$$

$$x_t = z_t\gamma + v_t \tag{48}$$

with $v_{1t} = v_t\beta + \varepsilon_t$; here

$$(v_{1t}, v_t)' \sim N(0, \Omega_j) \qquad \text{if } Z_{tj} = 1 \qquad (t = 1, \ldots, T; j = 1, 2).$$

We specify proper non-informative priors.

We consider the 11-dimensional vector of the restricted reduced form's parameters $\theta = (\beta, \gamma, \omega_{1,11}, \omega_{1,12}, \omega_{1,22}, \omega_{2,11}, \omega_{2,12}, \omega_{2,22}, \pi_1)$, with $\omega_{l,ij}$ the element (i,j) of Ω_l.

The reason for simulating the elements of the reduced form matrices Ω_j $(j = 1, 2)$, rather than the structural form matrices Σ_j, is that we divide θ into two subsets $\theta_1 = (\beta, \gamma)$ $(k_1 = 4)$ and $\theta_2 = (\omega_{1,11}, \omega_{1,12}, \omega_{1,22}, \omega_{2,11}, \omega_{2,12}, \omega_{2,22}, \pi_1)$ $(k_2 = 7)$. The relationship between (β, γ) and Ω_j $(j = 1, 2)$ is 'simpler' than the relationship between (β, γ) and Σ_j $(j = 1, 2)$, where

$$\Sigma_j = \begin{pmatrix} \omega_{j,11} + \omega_{j,22}\beta^2 - 2\omega_{j,12}\beta & \omega_{j,12} - \omega_{j,22}\beta \\ \\ \omega_{j,12} - \omega_{j,22}\beta & \omega_{j,22} \end{pmatrix}$$

depends on β. In other words, in the restricted reduced form β only appears in γ β; this product is always identified, even if $\gamma \to 0$. So, even if $\gamma \to 0$, we would not have 'problems' with the posterior distribution of the Ω_j $(j = 1, 2)$. For $\gamma \to 0$ we are faced with the well-known case of local non-identification of β.

For the covariance matrices Ω_j $(j = 1, 2)$ we have local non-identification for $\pi_j \to 0$. Therefore, multiple parameters may exhibit irregular, non-elliptical posterior contours. However, we can approximate the posterior shapes of θ_1 and θ_2 separately, since these two issues of possible non-identification are not strongly related.

For θ_2 we use the permutation-augmented MitISEM algorithm with $\mu^i_{h,c} = \beta_{h,c} X^i$, where X^i consists of a constant and the elements of the sample covariance matrix of the restricted reduced form's 'residuals' $y_t - z_t\gamma\beta$ and $x_t - z_t\gamma$ $(t = 1, \ldots, T)$ for given values of (β, γ).

The posterior mean of β is estimated as 0.0432, with a posterior standard deviation of 0.0254. The 95% posterior interval is estimated as $[-0.0095, 0.0921]$. For comparison, in the IV model with 1 normal regime, the posterior mean of β is estimated as 0.0983, with a posterior standard deviation of 0.0362. For this 1-regime model, the 95% posterior interval is estimated as $[0.0325, 0.1740]$, not including 0. This huge difference stresses the importance of taking into account the non-normality of the data in case of weak instruments. Figure 9 shows the log-income data with substantial negative skewness (due to the logarithmic transformation of some low wages) and large kurtosis.

We also applied the Gibbs sampler (without Rao-Blackwellization), which provides similar but less accurate estimates given the same amount of computing time. We do not use Rao-Blackwellization for two reasons. First, we could also extend the MitISEM-IS approach, adding one step of simulating latent variables and performing Rao-Blackwellization, a possibility that we will investigate in future research. Without this extension, the comparison would not be fair. Second, for different quantities of interest, such as the effect of education on *wage* rather than *log-wage* (for a particular amount of extra education), Rao-Blackwellization would not be feasible.

In future research we will investigate the performance of the partial MitISEM approach in larger models (with larger numbers of parameters).

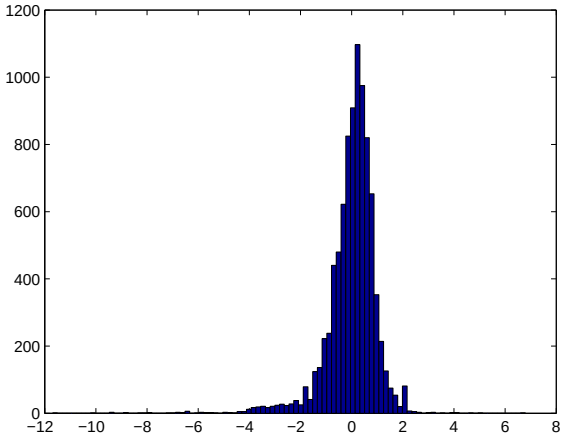

Figure 9: Log-income data (in deviation from mean, scaled by standard deviation) for state of Kentucky.

6 Concluding remarks

We introduced a new class of adaptive sampling methods for efficient and reliable posterior and predictive simulation. Multiple examples have shown the possible relevance of the novel methods, as a substitute for worse candidate distributions in Importance Sampling or the Metropolis-Hastings algorithm, or as a substitute or complement (e.g., as a validity check for estimated posterior moments or marginal likelihoods) for Gibbs sampling.

In future research we intend to investigate further extensions of the methods, such as the combination of MitISEM with variance reduction techniques such as antithetic sampling and control variates, the incorporation of an AdMit-step in the MitISEM method ('AdMit within MitISEM'), or the implementation of Rao-Blackwellization in the MitISEM procedure ('Rao-Blackwellization within MitISEM'). Further, we think that the applications of partial MitISEM to more complicated models (with a larger number of parameters) is of particular interest. The practical applicability and usefulness of adaptive importance sampling methods may be substantially increased by the partial MitISEM approach and extensions thereof.

References

[1] Akaike H. (1974), "A new look at the statistical model identification", *IEEE Transactions by Automatic Control*, 19, 716−723.

[2] Angrist, J.D. , Krueger, A.B. (1991), "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics* 106, 979−1014.

[3] Ausín M.C. and P. Galeano (2007), "Bayesian estimation of the mixture GARCH model", *Computational Statistics & Data Analysis*, 51, 2636−2652.

[4] Bartlett M.S. (1957), "A comment on D.V. Lindley's statistical paradox", *Biometrika*, 45, 533−534.

[5] Bauwens L. and Lubrano M. (1998), "Bayesian inference on GARCH models using the Gibbs sampler", *Econometrics Journal* 1, C23−C46.

[6] Black F. (1976), "Studies of Stock Prices Volatility Changes", *Proceeding from the American Statistical Association, Business and Economics Statistics Section*, 177−181.

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

[8] Cappé O., R. Douc, A. Guillin, J.M. Marin and C.P. Robert (2008), "Adaptive Importance Sampling in general mixture classes", *Statistics and Computing*, 18, 447−459.

[9] Chib S. (1995), "Marginal likelihood from the Gibbs output" *Journal of the American Statistical Association* 90(432), 1313−1321.

[10] Cornuet J.M., J.M. Marin, A. Mira and C.P. Robert (2009), "Adaptive Multiple Importance Sampling", Working Paper.

[11] Dempster A.P., N.M. Laird and D.B. Rubin (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society Series B* (Methodological), 39(1), 1−38.

[12] Eklund, J and S. Karlsson (2007), "Forecast combination and model averaging using predictive measures", *Econometric Reviews*, 26, 329−363.

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

[14] Frühwirth−Schnatter S. (2001), "Markov chain Monte Carlo estimation of classical and dynamic switching models", *Journal of the American Statistical Association*, 96, 194−209.

[15] Gelman A., J.B. Carlin, H.S. Stern and D.B. Rubin (2003), "Bayesian Data Analysis", 2nd edition. Chapman and Hall, London.

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

[17] Geweke J. (2007), "Interpretation and inference in mixture models: Simple MCMC works", *Computational Statistics & Data Analysis*, 51, 3529−3550.

[18] 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.

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

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

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

[22] Hoogerheide L.F., H.K. Van Dijk and R.D. Van Oest (2009), "Simulation Based Bayesian Econometric Inference: Principles and Some Recent Computational Advances". Chapter 7 in *Handbook of Computational Econometrics*, 215−280. Wiley, in press.

[23] 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.

[24] Hu W. (2005), "Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk." Dissertation at the Florida State University, College of Arts and Sciences.

[25] Keith J.M., D.P. Kroese and G.Y. Sofronov (2008), "Adaptive Independence Samplers", *Statistics and Computing*, 18, 409−420.

[26] 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.

[27] Kou S.C., Q. Zhou and W.H. Wong (2006), "Equi-energy sampler with applications in statistical inference and statistical mechanics", *The Annals of Statistics*, 34, 1581−1619.

[28] Kullback S. and R.A. Leibler (1951), "On information and sufficiency", *The Annals of Mathematical Statistics*, 22, 79−86.

[29] 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.

[30] Nelson D.B. (1991), "Conditional Heteroskedasticity in Asset Returns: A new Approach", *Econometrica*, 59, 347−370.

[31] Peel D. and G.J. McLachlan, (2000), "Robust Mixture Modelling using the t Distribution", *Statistics and Computing*, 10, 339−348.

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

[33] Schwarz G. (1978), "Estimating the dimension of a model", *Annals of Statistics*, 6, 461−464.

[34] Schwert G.W. (1990), "Why Does Stock Market Volatility Change over Time?", *Journal of Finance*, 44, 1115−1154.

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

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

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

A Derivation of the IS-weighted EM algorithm for mixtures of Student-t distributions

This appendix provides the derivation of the most general IS-weighted EM algorithm that is considered in this paper: the permutation-augmented algorithm in a mixture model of m components, in which the modes $\mu_{h,c}$ ($k\times 1$) of the candidate mixture's Student-t components are linear combinations $\mu_{h,c} = \beta_{h,c}X$ (with $\beta_{h,c}$ $k \times r$ and X $r \times 1$) where X consists of (functions of) parameters in previous subsets (plus typically a constant term). For the 'plain vanilla' algorithm, that is used in the basic MitISEM approach, one simply sets $m = 1$ (deleting the permutation-related subscripts c and $inv(c)$ at all variables), $X = 1$ ($r = 1$) and $\beta_{h,c} = \mu_h$.

The candidate density $g(\theta)$ is a mixture of $H \cdot m!$ Student-t densities ($h = 1, \ldots, H; c = 1, \ldots, m!$):

$$g(\theta) = g(\theta|\zeta) = \sum_{h=1}^{H} \eta_{h,c} \sum_{c=1}^{m!} t_k(\theta|\beta_{h,c}X, \Sigma_{h,c}, \nu_h), \tag{49}$$

where ζ is the set of coefficients $\beta_{h,c}$, scale matrices $\Sigma_{h,c}$, degrees of freedom ν_h, and mixing probabilities $\eta_{h,c}$ of the k-dimensional Student-t components with density:

$$t_k(\theta|\beta_{h,c}X, \Sigma_{h,c}, \nu_h) = \frac{\Gamma\left(\frac{\nu_h+k}{2}\right)}{\Gamma\left(\frac{\nu_h}{2}\right)(\pi\nu_h)^{k/2}} |\Sigma_{h,c}|^{-1/2} \left(1 + \frac{(\theta - \beta_{h,c}X)'\Sigma_{h,c}^{-1}(\theta - \beta_{h,c}X)}{\nu_h}\right)^{-(k+\nu_h)/2}. \tag{50}$$

Here $\Sigma_{h,c}$ is positive definite, $\nu_h \geq 1$, $\eta_h \geq 0$ and $\sum_{h=1}^{H} \eta_h = 1$. Moreover, in order to have a permutation-invariant candidate the mixing probabilities satisfy $\eta_{h,c} = \frac{\eta_h}{m!}$.

In our situation we maximize the *weighted* log-likelihood

$$\frac{1}{N} \sum_{i=1}^{N} W^i \log g(\theta^i|\zeta)$$

where $g(.|\zeta)$ is the mixture of Student-t densities (49).

The mixture of Student-t densities (49) for θ^i is equivalent with the specification

$$\theta^i \sim N(\beta_{h,c}X^i, w_h^i\Sigma_{h,c}) \qquad \text{if} \qquad z_{h,c}^i = 1,$$

where z^i is a set of $H \cdot m!$ latent variables indicating from which Student-t component, and from which permutation thereof, the observation θ^i stems: if θ^i stems from component h and permutation c, then $z_{h,c}^i = 1$, $z_{j,l}^i = 0$ for $(j,l) \neq (h,c)$; $\Pr[z_{h,c} = 1] = \eta_{h,c}$; w_h^i has the Inverse-Gamma distribution $IG(\nu_h/2, \nu_h/2)$. For a more extensive explanation of this continuous scale mixing representation of (mixtures of) Student-t distributions we refer to

Peel and McLachlan (2000). Here we have latent 'data' $\tilde{\theta}^i$ $(i = 1, \ldots, N)$

$$\tilde{\theta}^i = \{z^i_{h,c}, w^i_h | h = 1, \ldots, H;\ c = 1, \ldots, m!\}$$

and

$$\begin{aligned}
\log p(\theta^i, w^i, z^i|\zeta) &= \log p(\theta^i|w^i, z^i, \zeta) + \log p(w^i|\zeta) + \log p(z^i|\zeta) \\
&= \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\ \log\left[\mathrm{pdf}_{N(\beta_{h,c}X^i, w^i_h\Sigma_{h,c})}(\theta^i)\right] + \\
&\quad \sum_{h=1}^{H} \log \mathrm{pdf}_{IG(\nu_h/2,\nu_h/2)}(w^i_h) + \sum_{h=1}^{H} z^i_{h,c} \log\left(\frac{\eta_h}{m!}\right) \\
&= \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\ \left\{-\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{h,c}| - \frac{k}{2}\log(w^i_h)\right. \\
&\quad \left. -\frac{1}{2}\frac{(\theta^i - \beta_{h,c}X^i)'(\Sigma_{h,c})^{-1}(\theta^i - \beta_{h,c}X^i)}{w^i_h}\right\} \\
&\quad + \sum_{h=1}^{H}\left\{\frac{\nu_h}{2}\log\left(\frac{\nu_h}{2}\right) - \left(\frac{\nu_h}{2} - 1\right)\log(w^i_h) - \frac{\nu_h}{2}\ \frac{1}{w^i_h} - \log\left(\Gamma\left(\frac{\nu_h}{2}\right)\right)\right\} \\
&\quad + \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\log\left(\frac{\eta_h}{m!}\right) \\
&= \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\ \left\{-\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_h| - \frac{k}{2}\log(w^i_h)\right. \\
&\quad \left. -\frac{1}{2}\frac{(\theta^i_{inv(c)} - X^i\beta_h)'(\Sigma_h)^{-1}(\theta^i_{inv(c)} - X^i\beta_h)}{w^i_h}\right\} \\
&\quad + \sum_{h=1}^{H}\left\{\frac{\nu_h}{2}\log\left(\frac{\nu_h}{2}\right) - \left(\frac{\nu_h}{2} - 1\right)\log(w^i_h) - \frac{\nu_h}{2}\ \frac{1}{w^i_h} - \log\left(\Gamma\left(\frac{\nu_h}{2}\right)\right)\right\} \\
&\quad + \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\log\left(\frac{\eta_h}{m!}\right), \qquad (51)
\end{aligned}$$

where w^i and z^i are *a priori* independent, and where $inv(c)$ is the inverse of the permutation c. That is, applying permutation c and permutation $inv(c)$ subsequentially yields the original vector or matrix.

The expressions of the latent variables w^i and z^i that appear in terms which also involve the parameters ζ to be optimized are $z^i_{h,c}$, $\frac{z^i_{h,c}}{w^i_h}$, $\log w^i_h$, and $\frac{1}{w^i_h}$. Therefore, we derive the conditional expectations of $z^i_{h,c}$, $\frac{z^i_{h,c}}{w^i_h}$, $\log w^i_h$, and $\frac{1}{w^i_h}$ given θ^i and $\zeta = \zeta^{(L-1)}$, the optimal parameters in the previous EM iteration:

(1) **Expectation of $z^i_{h,c}$**: in order to speed up the convergence of the (IS weighted) EM algorithm we compute the expectation

$$\tilde{z}^i_{h,c} \equiv E\left[z^i_{h,c}\left|\theta^i, \zeta = \zeta^{(L-1)}\right.\right] = \Pr[z^i_{h,c} = 1|\theta^i, \zeta = \zeta^{(L-1)}]$$

not given w^i_h; that is, w^i_h is integrated out:

$$\begin{aligned} p(\theta^i, z^i|\zeta) &= \prod_{h=1}^{H}\prod_{c=1}^{m!}\left[p(\theta^i|z^i_{h,c}=1,\zeta)\ \Pr[z^i_{h,c}=1|\zeta]\right]^{z^i_{h,c}} \\ &= \prod_{h=1}^{H}\prod_{m=1}^{m!}\left[t(\theta^i|\beta_{h,c}X^i, \Sigma_{h,c}, \nu_h)\ \frac{\eta_h}{m!}\right]^{z^i_{h,c}}, \end{aligned}$$

which is a kernel of a probability function of a multinomial distribution for the set of $z^i_{h,c}$ $(h = 1, \ldots, H; c = 1, \ldots, m!)$ given θ^i and ζ, with probabilities $\Pr[z^i_{h,c} = 1|\theta^i, \zeta = \zeta^{(L-1)}]$ equal to

$$\tilde{z}^i_{h,c} \equiv E\left[z^i_{h,c}\left|\theta^i, \zeta = \zeta^{(L-1)}\right.\right] = \frac{t(\theta^i|\beta_{h,c}X^i, \Sigma_{h,c}, \nu_h)\ \eta_h}{\sum_{j=1}^{J}\sum_{l=1}^{m!} t(\theta^i|\beta_{j,l}X^i, \Sigma_{j,l}, \eta_j)\ \eta_j}. \tag{52}$$

(2) **Expectation of $\frac{z^i_{h,c}}{w^i_h}$:**

$$\begin{aligned} \widetilde{z/w}^i_{h,c} \equiv E\left[z^i_{h,c}\frac{1}{w^i_h}\middle|\theta^i, \zeta = \zeta^{(L-1)}\right] &= \Pr[z^i_{h,c} = 1|\theta^i, \zeta = \zeta^{(L-1)}]\ \times \\ &\quad E\left[\frac{1}{w^i_h}\middle| z^i_{h,c} = 1, \theta^i, \zeta = \zeta^{(L-1)}\right]. \end{aligned}$$

Given $z^i_{h,c} = 1$, i.e. given that θ^i stems from permutation c of Student-t component h, the situation reduces to the case of the EM algorithm for a Student-t distribution without mixtures (see Hu (2005) for an extensive explanation):

$$E\left[\frac{1}{w^i_h}\middle| z^i_{h,c} = 1, \theta^i, \zeta\right] = \frac{k + \nu_h}{\rho^i_{h,c} + \nu_h}.$$

with

$$\rho^i_{h,c} = (\theta^i - \beta_{h,c}X)'\Sigma^{-1}_{h,c}(\theta^i - \beta_{h,c}X).$$

Therefore we have

$$\widetilde{z/w}^i_{h,c} \equiv E\left[z^i_{h,c}\frac{1}{w^i_h}\middle|\theta^i, \zeta = \zeta^{(L-1)}\right] = \tilde{z}^i_{h,c}\ \frac{k + \nu_h}{\rho^i_{h,c} + \nu_h}. \tag{53}$$

(3) **Expectation of $\log w_h^i$:**

$$\begin{aligned}
\xi_h^i &\equiv E\left[\log w_h^i \left| \theta^i, \zeta = \zeta^{(L-1)} \right.\right] = \\
&= \sum_{c=1}^{m!} E\left[\log w_h^i \left| z_{h,c} = 1, \theta^i, \zeta = \zeta^{(L-1)} \right.\right] \Pr[z_{h,c}^i = 1 | \theta^i, \zeta = \zeta^{(L-1)}] \\
&\quad + E\left[\log w_h^i \left| z_{h,c}^i = 0 \ \forall c,\ \theta^i, \zeta = \zeta^{(L-1)} \right.\right] \Pr[z_{h,c}^i = 0 \ \forall c | \theta^i, \zeta = \zeta^{(L-1)}] \\
&= \sum_{c=1}^{m!} \left\{ \left[\log\left(\frac{\rho_{h,c}^i + \nu_h}{2}\right) - \psi\left(\frac{k+\nu_h}{2}\right) \right] \tilde{z}_{h,c}^i \right\} \\
&\quad + \left[\log\left(\frac{\nu_h}{2}\right) - \psi\left(\frac{\nu_h}{2}\right) \right] \left(1 - \sum_{c=1}^{m!} \tilde{z}_{h,c}^i \right), \qquad (54)
\end{aligned}$$

where $\psi(.)$ is the digamma function (the derivative of the logarithm of the gamma function $\log\Gamma(.)$), and where we again used that given $z_{h,c} = 1$ the situation reduces to the case of the EM algorithm for a Student-t distribution without mixtures (see Hu (2005) for an extensive explanation). For $z_{h,c}^i = 0 \ \forall c$, the conditional distribution of w_h^i given θ^i, ζ is the distribution given only ζ (since the observation θ^i does not depend on w_h^i) which is Inverse-Gamma $IG(\nu_h/2, \nu_h/2)$:

$$E\left[\log w_h^i \left| z_{h,c}^i = 0 \ \forall c,\ \theta^i, \zeta = \zeta^{(L-1)} \right.\right] \ = \ \log\left(\frac{\nu_h}{2}\right) - \psi\left(\frac{\nu_h}{2}\right).$$

(4) **Expectation of $\frac{1}{w_h^i}$:**

$$\begin{aligned}
\delta_h^i &\equiv E\left[\frac{1}{w_h^i} \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] \\
&= \sum_{c=1}^{m!} E\left[\frac{1}{w_h^i} \middle| z_{h,c}^i = 1, \theta^i, \zeta = \zeta^{(L-1)} \right] \Pr[z_{h,c}^i = 1 | \theta^i, \zeta = \zeta^{(L-1)}] \\
&\quad + E\left[\frac{1}{w_h^i} \middle| z_{h,c}^i = 0 \ \forall c,\ \theta^i, \zeta = \zeta^{(L-1)} \right] \Pr[z_{h,c}^i = 0 \ \forall c | \theta^i, \zeta = \zeta^{(L-1)}] \\
&= \sum_{c=1}^{m!} \frac{k+\nu_h}{\rho_{h,c}^i + \nu_h} \ \tilde{z}_{h,c}^i + \left(1 - \sum_{c=1}^{m!} \tilde{z}_{h,c}^i \right). \qquad (55)
\end{aligned}$$

where if $z_{h,c}^i = 0 \ \forall c$, $1/w_h^i$ has the $Gamma(\nu_j/2, \nu_j/2)$ distribution with

$$E[1/w_h^i | z_{h,c}^i = 0 \ \forall c, \theta^i, \zeta = \zeta^{(L-1)}] = 1.$$

Define $\log\tilde{p}(\theta^i, w^i, z^i|\zeta)$ as the result of substituting the expectations (52)-(55) into $\log p(\theta^i, w^i, z^i|\zeta)$ in (51). The Maximization step amounts to computing the ζ that maximizes

$$\zeta^{(L)} = \arg\max_{\zeta} \frac{1}{N} \sum_{i=1}^{N} W^i \log\tilde{p}(\theta^i, w^i, z^i|\zeta).$$

Using the analogy with Maximum Likelihood estimation for the Seemingly Unrelated Regression model with Gaussian errors (for the k elements of θ^i) and the same r 'regressors' X^i in each equation, in which case the Ordinary Least Squares (OLS) estimator provides the Maximum Likelihood Estimator, and with Maximum Likelihood estimation for the multinomial distribution, it is easily derived that $\zeta^{(L)}$ consists of:

$$\beta_h^{(L)\prime} = \left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \widetilde{z/w}^i_{h,c}\, X_i\, X_i'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \widetilde{z/w}^i_{h,c}\, X_i\, \theta^{i\prime}_{inv(c)}\right], \tag{56}$$

$$\hat{\Sigma}_h^{(L)} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \widetilde{z/w}^i_{h,c}\,(\theta^i_{inv(c)} - \beta_h^{(L)}X^i)(\theta^i_{inv(c)} - \beta_h^{(L)}X^i)'}{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \tilde{z}^i_{h,c}}, \tag{57}$$

$$\eta_h^{(L)} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i\, \tilde{z}^i_{h,c}}{\sum_{i=1}^{N} W_i}. \tag{58}$$

Further, $\nu_h^{(L)}$ is solved from the first order condition of ν_h:

$$-\psi(\nu_h/2) + \log(\nu_h/2) + 1 - \frac{\sum_{i=1}^{N} W_i\, \xi_h^i}{\sum_{i=1}^{N} W_i} - \frac{\sum_{i=1}^{N} W_i\, \delta_h^i}{\sum_{i=1}^{N} W_i} = 0 \tag{59}$$

using a procedure for one-dimensional root finding.