A Class of Adaptive Importance Sampling Weighted EM 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. 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. This is useful for importance or Metropolis-Hastings 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 joint distribution by estimating a product of marginal and conditional distributions. 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 DCC or mixture GARCH models and a mixture instrumental variables model.

Keywords: mixture of Student-t distributions, importance sampling, Kullback-Leibler divergence, Expectation Maximization, Metropolis-Hastings algorithm, predictive likelihood, DCC GARCH, mixture GARCH, instrumental variables.

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

[*] Department of Econometrics and Tinbergen Institute, Vrije Universiteit Amsterdam, The Netherlands

[†] Corresponding author, e-mail address: l.f.hoogerheide@vu.nl.

[‡] Econometric and Tinbergen Institutes, Erasmus University Rotterdam, The Netherlands

[§] Econometric Institute and Tinbergen Institute, Erasmus University Rotterdam and Vrije Universiteit Amsterdam, The Netherlands

is further developed by Van Dijk and Kloek (1980, 1984) and, in particular, by Geweke (1989). Cappé et al. (2008) discuss that 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. The idea of tempering was introduced by Geyer (1991), see also Hukushima and Nemoto (1996).

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 joint distribution by estimating a product of marginal and conditional distributions. This division can substantially reduce the dimension of the approximation problem, which facilitates the application of adaptive importance or Metropolis-Hastings 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 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. 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 different strand of literature is the use of adaptive MCMC algorithms where the parameters of the candidate density are automatically tuned during the sampling procedure. Learning about candidate density parameter values leading to more efficient sampling while maintaining the ergodicity property for asymptotic convergence is, of course, important. Roberts and Rosenthal (2009) consider an adaptive random walk Metropolis sampler and Giordani and Kohn (2010) use a mixture of densities in their adaptive independent MH sampler. We differ from these authors by using a two-stage approach. Using the Kullback-Leibler distance function, we fit during a first stage of *preliminary adaptation* a flexible candidate to the target with the IS-weighted EM algorithm. In the second stage we insert the obtained candidate in a 'standard', non-adaptive IS or MH algorithm. So, in terms of Giordani and Kohn (2010), we do not perform *strict adaptation*; our second phase of non-adaptive IS or MH ensures that the simulation output converges to the correct distribution. In section 2.1, we compare the efficiency of our approach with those from Roberts and Rosenthal (2009) and Giordani and Kohn (2010) in the context of a DCC-GARCH model with 11 parameters. The results indicate that our approach compares favorably with these alternative adaptive MCMC schemes, but we emphasize that a systematic study of the relevant merits of alternative sampling schemes for a variety of target density shapes is a topic of great interest, which is however beyond the scope of the present study.

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, and we show applications in a multivariate GARCH model with 11 parameters, and a (Wishart) posterior density kernel of up to 36 parameters in an inverse covariance matrix. Section 3 introduces the sequential MitISEM method, and includes a subsection on the tempering method. In section 4 the permutation-augmented MitISEM approach is proposed. Section 5 introduces the partial MitISEM method. Section 6 concludes. The appendix provides the derivations of the IS-weighted EM methods, and discusses the alternative simulation methods of Roberts and Rosenthal (2009) and Giordani and Kohn (2010).

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. As discussed by 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 $(h = 1, \dots, H)$ 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

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

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

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 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-density

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

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 the continuous scale mixing representation of Student-t distributions we refer to Rubin (1983) and to Lange, Little and Taylor (1989) who consider the more general situation with unknown degrees of freedom. For 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 the so-called data-augmented density is given by

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

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:

$$\begin{aligned}
\tilde{z}_h^i &\equiv E\left[z_h^i \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] = \frac{t(\theta^i|\mu_h, \Sigma_h, \nu_h)\ \eta_h}{\sum_{j=1}^{H} t(\theta^i|\mu_j, \Sigma_j, \nu_j)\ \eta_j}, \qquad (9) \\
\widetilde{z/w}_h^i &\equiv E\left[\frac{z_h^i}{w_h^i} \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] = \tilde{z}_h^i \ \frac{k+\nu_h}{\rho_h^i + \nu_h}, \qquad (10) \\
\xi_h^i &\equiv E\left[\log w_h^i \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] = \\
&= \left[\log\left(\frac{\rho_h^i+\nu_h}{2}\right) - \psi\left(\frac{k+\nu_h}{2}\right) \right] \tilde{z}_h^i + \left[\log\left(\frac{\nu_h}{2}\right) - \psi\left(\frac{\nu_h}{2}\right) \right] (1-\tilde{z}_h^i), \qquad (11) \\
\delta_h^i &\equiv E\left[\frac{1}{w_h^i} \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] = \frac{k+\nu_h}{\rho_h^i+\nu_h}\ \tilde{z}_h^i + (1-\tilde{z}_h^i), \qquad (12)
\end{aligned}$$

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:

$$\begin{aligned}
\mu_h^{(L)} &= \left[\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i \right]^{-1} \left[\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i \ \theta^i \right], \qquad (13) \\
\hat{\Sigma}_h^{(L)} &= \frac{\sum_{i=1}^{N} W_i \ \widetilde{z/w}_h^i \ (\theta^i - \mu_h^{(L)})(\theta^i - \mu_h^{(L)})'}{\sum_{i=1}^{N} W_i \ \tilde{z}_h^i}, \qquad (14) \\
\eta_h^{(L)} &= \frac{\sum_{i=1}^{N} W_i \ \tilde{z}_h^i}{\sum_{i=1}^{N} W_i}. \qquad (15)
\end{aligned}$$

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 $(h = 1, \dots, H)$ 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 *basic MitISEM* approach for obtaining an approximation to a target density:

(0) **Initialization**: Simulate draws $\theta^1, \dots, \theta^N$ 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 $\theta^1, \dots, \theta^N$ from g_{naive}. Use these estimates as the mode and scale matrix of Student-t distribution $g_{adaptive}$. Draw a sample $\theta^1, \dots, \theta^N$ 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 $\mu_h, \Sigma_h, \nu_h, \eta_h$ $(h = 1, \dots, H)$. Draw a new sample $\theta^1, \dots, \theta^N$ 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 $(h = 1, \dots, H)$, take $x\%$ of the sample $\theta^1, \dots, \theta^N$ 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 the 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 $(h = 1, \dots, H)$ 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 $(h = 1, \dots, H + 1)$. 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 of Value at Risk (VaR), it is arguably wise to approximate the optimal importance density of Geweke (1989), see Hoogerheide and Van Dijk (2010). 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 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 the 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.

2.1 Application: Bayesian analysis of the DCC-GARCH model

In this subsection the MitISEM approach is applied to the popular Dynamic Conditional Correlation (DCC) GARCH model of Engle (2002). This multivariate GARCH model allows the conditional correlation between multiple time series to be time-varying, whereas it allows flexible GARCH specifications for the univariate processes. For the Bayesian estimation of this model, a regular Gibbs sampling approach is not feasible due to the recursive structure of GARCH models. One could apply a Griddy-Gibbs sampler (Ritter and Tanner (1992)), but this sampler is known to be relatively very slow. We use the MH sampler with a candidate density resulting from the MitISEM algorithm, and compare the performance of the MitISEM candidate density with two samplers from the literature: the adaptive Metropolis (AM) sampler of Roberts and Rosenthal (2009) and the adaptive independent Metropolis-Hastings (AIMH) sampler of Giordani and Kohn (2010).

In our example, the d-dimensional vector y_t $(t = 1, \ldots, T)$ consists of (demeaned) returns of asset prices and is supposed to follow the following conditional distribution:

$$y_t | \mathcal{I}_{t-1} \sim N(0, H_t) \tag{17}$$

with $\mathcal{I}_{t-1}$ the information set at time $t-1$ and H_t representing the time-varying conditional covariance matrix of the returns. Decomposing H_t into conditional variances and correlations, H_t can be written as

$$H_t = D_t R_t D_t, \tag{18}$$

where D_t represents a $d \times d$ time-varying diagonal matrix containing the square root of the conditional variances $h_{i,t}$ $(i = 1, \ldots, d)$ of the asset returns $y_{i,t}$. The d conditional variances follows a unit GARCH process, which reads as

$$h_{i,t} = \omega_i + \alpha_i y_{i,t-1}^2 + \beta_i h_{i,t-1} \qquad (i = 1, \ldots, d) \tag{19}$$

with the usual restrictions $\omega_i \geq 0, \alpha_i \geq 0$ and $\beta_i \geq 0$ in order to ensure positive values of the conditional variance. To ensure covariance stationarity of y_t, one must impose $\alpha_i + \beta_i \leq 1$.

Regarding the correlations, Engle (2002) suggests a dynamic process

$$Q_t = (1 - A - B)\,\bar{Q} + A\,(z_{t-1} z_{t-1}') + B\,Q_{t-1} \tag{20}$$

with scalars A and B satisfying $A \geq 0, B \geq 0$ and $A + B \leq 1$, and with $\bar{Q}$ representing the unconditional correlation matrix of the standardized residuals $z_t = D_t^{-1} y_t$. This matrix $\bar{Q}$ is estimated by the sample correlation matrix $\frac{1}{T}\sum_{t=1}^{T} \hat{z_t}\hat{z_t}'$. The stated conditions plus a positive definite initial matrix Q_0 guarantee a positive definite matrix Q_t. The matrix Q_t has to be rescaled however to produce a valid time-varying *correlation* matrix (with diagonal elements 1):

$$R_t = (I \circ Q_t)^{-1/2} Q_t (I \circ Q_t)^{-1/2}, \tag{21}$$

where '$\circ$' denotes the Hadamard product. We take uniform priors on the parameter vector θ, which consists of 11 parameters (3 times a univariate GARCH process plus the Dynamic Correlation process). For α_i, β_i, A and B, we are restricted to [0,1] plus the covariance stationarity restrictions. For the remaining three parameters ω_i we use a truncated uniform prior $[0, P]$ in order to have proper non-informative priors. When using truncated uniform priors, many draws may fall outside the feasible prior region for a naive Student-t candidate distribution. An advantage of the MitISEM algorithm is that it produces a rather close approximation to the posterior (including the 0 level outside the 'allowed range') that will have almost all probability mass inside the feasible region.

We take returns from three indices: the MSCI World, the MSCI Emerging Markets and the Barclays Global Bond Index. The first series is a stock market index of 1500 world stocks of 23 different developed countries, maintained by Morgan Stanley Capital International (MSCI inc.). The Emerging Markets Index is a market capitalization index that consists of indices in 26 emerging economies. The third series is often used to represent investment grade bonds being traded in the United States. From these indices we use daily observations on log return (100 times the change of the logarithm of the closing price) from January 3 2000 to November 3 2003.

Posterior means of the model parameters are estimated by using the independence chain MH algorithm with the candidate density produced by MitISEM. We compare the performance of MitISEM to the AM sampler of Roberts and Rosenthal (2009) and the AIMH sampler of Giordani and Kohn (2010). The first sampler is based on a mixture of two multivariate normal distributions, where both covariance matrices are multiplied by factors that aim at an optimal random walk proposal distribution and a high acceptance rate of the local sampler. The second sampler consists of a mixture of normal densities, which are estimated by the k-harmonic means clustering algorithm instead of the EM-algorithm that is used in MitISEM. See the appendix for a brief description of both samplers.

Table 1 shows posterior means estimated by the MH algorithms. For all methods, we simulate 20,000 draws after a burn-in sample of 5000 draws. Numerical standard errors (NSE) are obtained by using the integrated autocorrelation time (IACT),

$$IACT = 1 + 2\sum_{\tau=1}^{\infty} \rho_\theta(\tau), \tag{22}$$

where we truncate this sum of τ-th order autocorrelations $\rho_\theta(\tau)$ at $\tau_{max} = 50$. Hence the variance of the sample mean $\hat{\theta}_{mean}$ after N iterations of the MCMC algorithm is equal to:

$$\text{var}(\hat{\theta}_{mean}) = \sigma_\theta^2/N \times IACT. \tag{23}$$

The main result from Table 1 is that MitISEM outperforms the competing algorithms in this DCC-GARCH application, since the NSE values of most parameters are smaller than the corresponding values of AIMH and AM. MitISEM combines a higher acceptance rate with lower first order autocorrelations than the competing algorithms. MitISEM does require more computing time (on an AMD Athlon$^{\text{tm}}$ II X2 B24 processor) than AM, but if we give AM the same computing time (generating more draws), then its NSEs only drop approximately 10%, so that these are still worse than those of MitISEM. If we compare the AIMH and AM algorithms, then the acceptance rate of AM is higher than the rate of AIMH, but the high serial correlation of the AM draws increases the IACT, causing higher NSE values.

Note that the relative quality of the AIMH and AM algorithms, as compared with MitISEM, may improve for parameter spaces with higher dimension. In such cases a comparison of AIMH and AM with the basic MitISEM approach and the partial MitISEM method of section 5 would be particularly interesting. A systematic study of the relevant merits of alternative sampling schemes for a variety of target density shapes and dimensions is a topic of great interest, which is however beyond the scope of the present study. In any case, we expect that no algorithm will dominate in all applications. Moreover, given that diagnostic checks can never fully guarantee that simulation results have converged to the true values, the use of multiple simulation methods can be a quite useful validity check.

Table 1: Application of the basic MitISEM algorithm, the AIMH sampler of Giordani and Kohn (2010), and the AM sampler of Roberts and Rosenthal (2009) to posterior in DCC-GARCH model with 11 parameters: estimated posterior means, corresponding numerical standard errors (NSEs) and first order autocorrelations of draws $\rho_\theta(1)$. Results are based on 20,000 draws after a burn-in period of 5,000 draws. We report 100 times the NSE values which are obtained by equation (23).

	MitISEM			AIMH (GK)			AM (RR)		
	mean	NSE · 100	$\rho_\theta(1)$	mean	NSE · 100	$\rho_\theta(1)$	mean	NSE · 100	$\rho_\theta(1)$
ω_1	0.056	0.032	0.499	0.058	0.085	0.846	0.059	0.096	0.959
α_1	0.091	0.028	0.458	0.092	0.064	0.823	0.092	0.081	0.959
β_1	0.858	0.048	0.478	0.855	0.119	0.837	0.854	0.143	0.960
ω_2	0.143	0.254	0.696	0.134	0.268	0.860	0.137	0.329	0.959
α_2	0.107	0.086	0.609	0.105	0.102	0.842	0.103	0.117	0.956
β_2	0.770	0.288	0.694	0.780	0.309	0.857	0.778	0.371	0.957
ω_3	0.005	0.008	0.702	0.005	0.006	0.828	0.005	0.008	0.954
α_3	0.092	0.076	0.578	0.090	0.078	0.811	0.090	0.102	0.952
β_3	0.744	0.303	0.691	0.754	0.221	0.820	0.752	0.294	0.954
A	0.019	0.009	0.499	0.020	0.015	0.823	0.019	0.019	0.955
B	0.972	0.020	0.572	0.971	0.024	0.830	0.972	0.032	0.956
Acceptance rate			54%			19%			29%
Computational time (in seconds)									
constructing candidate			1800						
MH-sampler			740			4100			2140

2.2 Application: Wishart posterior distribution of inverse covariance matrix

In this subsection we show an example to analyze the MitISEM algorithm for different numbers of parameters k. Suppose we have i.i.d. observations y_t $(t = 1, 2, \ldots, T)$, with the $(d \times 1)$ vector $y_t \sim N(0, \Sigma)$. We specify the diffuse prior $p(\Sigma) \propto |\Sigma|^{-\frac{d+1}{2}}$, so that the posterior of Σ is Inverted Wishart; the posterior of $\Psi \equiv \Sigma^{-1}$ is Wishart with mean the inverse of the sample covariance matrix of y_t $(t = 1, 2, \ldots, T)$ and degrees of freedom the number of observations T. Suppose $T = 250$ and the sample covariance matrix is $\rho\iota\iota' + (1-\rho)I$ with $\rho = 0.5$ and ι the $(T \times 1)$ vector of ones, i.e., all variances are 1 and all covariances are $\rho = 0.5$. Of course, we could simulate directly from this Wishart distribution. However, we use this example to illustrate the quality of the MitISEM algorithm for different numbers of parameters, the $k = \frac{1}{2}d(d+1)$ elements of the symmetric $d \times d$ matrix Ψ $(\psi_{ij};\ i = 1, \ldots, d; j = 1, \ldots, d; i \leq j)$. Table 2 gives simulation results for $d = 1, 2, \ldots 8$, where $d = 8$ implies $k = 36$ parameters. The Relative Numerical Efficiency (RNE) is the ratio of the variance of a direct sampling estimator and the variance of the MitISEM IS estimator (with the same number of draws); see Geweke (1989). This can be used to compute the Effective Sample Size (ESS), see Liu (2001). The mean, min and max RNE indicate the mean, minimum and maximum of the $k = \frac{1}{2}d(d+1)$ RNEs for the estimated posterior means of the parameters ψ_{ij} $(i = 1, \ldots, d; j = 1, \ldots, d; i \leq j)$. For small numbers of parameters k, the MitISEM approach produces an extremely good candidate distribution (RNEs close to 1, MH acceptance rates above 80%). For larger numbers of parameters up to 36, the computing time increases, whereas the quality decreases, although the latter is still rather high. In such dimensions, a RNE of 0.378 or a MH acceptance rate of 40.9% can still be considered fine. For comparison, for the case of 36 parameters, the RNEs of IS with the 'naive' candidate distribution around the posterior mode (with scale matrix equal to minus the inverse Hessian of the log-posterior at the mode) have mean 0.0469 (falling in the interval [0.0382, 0.0514]), and the MH acceptance rate is 8.5%. From the ratio of the RNEs we see that approximately 8 times more draws would be required from the naive candidate to reach the same precision as MitISEM; that is, assuming that no relevant parts of the parameter space are 'missed' by the naive candidate, otherwise the naive results would be biased. Summarizing, this example illustrates

Table 2: Application of basic MitISEM to posterior (Wishart) density kernel for $k = \frac{1}{2}d(d+1)$ parameters in symmetric $d \times d$ inverse covariance matrix $\Psi \equiv \Sigma^{-1}$ for different dimensions d: simulation results where the mean, min and max RNE indicate the mean, minimum and maximum of the $k = \frac{1}{2}d(d+1)$ Relative Numerical Efficiencies for the estimated posterior means of the parameters ψ_{ij} $(i = 1, \dots, d; j = 1, \dots, d; i \leq j)$.

		Basic MitISEM					computing time (s)
		Importance Sampling				MH	
d	k	C.o.V. IS weights	mean RNE	min RNE	max RNE	acc. rate	
1	1	0.130	0.946	0.946	0.946	0.939	39.1
2	3	0.227	0.925	0.921	0.927	0.884	46.0
3	6	0.351	0.864	0.856	0.871	0.806	48.8
4	10	0.491	0.783	0.774	0.791	0.724	53.7
5	15	0.643	0.686	0.680	0.692	0.639	61.5
6	21	0.805	0.588	0.573	0.595	0.561	72.3
7	28	1.002	0.481	0.470	0.494	0.486	96.8
8	36	1.240	0.378	0.365	0.389	0.409	117.7

that MitISEM can produce useful results for moderate dimensions, at least up to 35 parameters.

For larger numbers of parameters, one may split the parameters into subsets, and construct the joint candidate distribution by estimating a product of marginal and conditionals. This *partial* MitISEM approach will be introduced in Section 5. Alternatively, one may consider the methods of Roberts and Rosenthal(2009) or Giordani and Kohn (2010).

For the efficient evaluation of Value-at-Risk or Expected Shortfall, we may specifically focus on higher values of variance and covariance parameters; see Hoogerheide and Van Dijk (2010). In such a case, no direct sampling is possible. One may also consider mixtures of Wishart or Inverted Wishart distributions, instead of Student-t distributions, which is left as a topic for further research.

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. For sequential Monte Carlo methods, we refer to Liu and Chen (1998), Doucet et al. (2001), and Chopin (2002). The latter explicitly takes into account that a candidate proposal will not be updated until the sequential weights become very variable.

The previous section showed that, although the IS-weighted EM steps are relatively efficient, the construction of an appropriate candidate distribution may still require considerable computing time. 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 $y_{1:T+1} = \{y_1, \dots, y_T, y_{T+1}\}$ is typically not so different from the posterior for data $y_{1:T} = \{y_1, \dots, y_T\}$. 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 $y_{1:T+\tau}$ ($\tau = 1, 2, \ldots$):

(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 univariate two-component Gaussian Mixture EGARCH model, which is given by:

$$y_t = \mu + \sqrt{h_t}\,\varepsilon_t, \tag{24}$$

$$\log(h_t) = \omega + \gamma \frac{y_{t-1}-\mu}{\sqrt{h_{t-1}}} + \alpha \left(\frac{|y_{t-1}-\mu|}{\sqrt{h_{t-1}}} - \frac{E|y_{t-1}-\mu|}{\sqrt{h_{t-1}}} \right) + \beta \log(h_{t-1}), \tag{25}$$

$$\varepsilon_t \sim \begin{cases} N(0,\sigma^2) & \text{with probability } \rho \\ N(0,\sigma^2/\lambda) & \text{with probability } 1-\rho \end{cases}, \tag{26}$$

with h_t the conditional variance of y_t given the information set $\mathcal{I}_{t-1} = \{y_{t-1}, y_{t-2}, y_{t-3}, \ldots\}$. See Nelson (1991) for the original (one-component) EGARCH model. 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 $|\beta| \leq 1$ to ensure covariance stationarity of h_t and impose 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. Moreover, we truncate μ, ω, α and γ 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 $[0.5, 1] \times [0, 1] \times [-1, 1] \times [-1, 1] \times [-1, 1] \times [0, 1] \times [0, 1]$.

The returns y_t are taken from the *S&P* 500 index. From this index we use daily observations y_t $(t = 1, \dots, T)$ on the log return (100 times the change of the logarithm of the closing price) from January 2 1998 to March 6 2003 (1350 observations).

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 $t = 1301, \dots, 1350$, we compute the predictive likelihood, see Gelfand and Dey (1994) and Eklund and Karlsson (2007), who provide an overview of several approaches including the fractional Bayes factor of O'Hagan(1995) and the intrinsic Bayes factor of Berger and Pericchi (1996). In principle, the marginal likelihood exists for this flat, proper prior. However, if we would perform a model comparison, then we could make the marginal likelihood for the mixture EGARCH model as low as we want, for example, by increasing the prior range for the parameter μ or ω. For the predictive likelihood this does not hold, since the effect of a lower (exact) prior density due to a wider prior range for μ or ω drops out of the ratio in (28) below. For other adaptive sampling methods for estimating marginal and predictive likelihoods we refer to Frühwirth-Schnatter and Wagner (2008) and Pitt et al. (2010).

The predictive likelihood, a useful quantity in Bayesian inference for model comparison, is computed 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 is given by:

$$p(\tilde{y}|y^*, M) = \int p(\tilde{y}|\theta, y^*, M) p(\theta|y^*, M) d\theta, \tag{27}$$

which is actually the marginal likelihood if we consider $\tilde{y}$ as 'the data' and $p(\theta|y^*, M)$, the exact posterior density after observing y^*, as 'the prior'. Using Bayes' rule for this exact posterior density $p(\theta|y^*, M)$ and substituting into (27) yields

$$p(\tilde{y}|y^*, M) = \frac{\int p(y|\theta, M) p(\theta|M) d\theta}{\int p(y^*|\theta, M) p(\theta|M) d\theta}, \tag{28}$$

where $p(y|\theta, M)$ is the likelihood of the model M and $p(\theta|M)$ the prior density of the parameters θ in the model. 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. In our example, the training sample y^* (for the marginal likelihood in the denominator of the predictive likelihood) consists of 500 observations, and remains fixed.

We compare the Sequential MitISEM approach with the 'ad hoc MitISEM approach', which runs the MitISEM algorithm from scratch at each time $t = 1301, \dots, 1350$. The comparison is twofold. First we compare the computing time that is required by 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 3 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.

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 the left panel of Figure 1. 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

Table 3: Application of Sequential MitISEM and 'ad hoc MitISEM' (which simply runs the MitISEM algorithm from scratch on each sample $y_{1:t}$ $(t = 1301, \ldots, 1350)$*) to a Gaussian Mixture EGARCH model. 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 steps		
# adapted	1	
# adapted and extended	0	
# reused	48	
Computational effort		
Construct 50 candidate densities over period (1301 − 1350)	117 s	5602 s

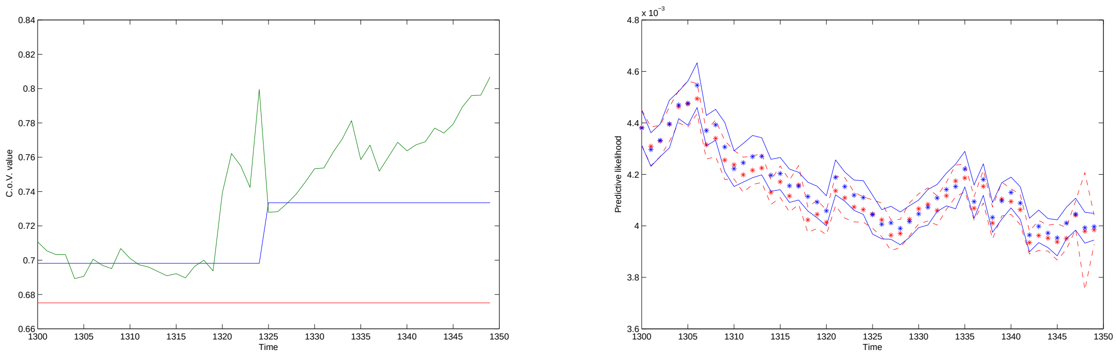

Figure 1: Illustration of the Sequential MitISEM approach for predictive likelihood estimation in a Gaussian Mixture EGARCH model. Left panel: The blue line represents C.o.V.(T)*, the Coefficient of Variation that is used for comparison in step (2) of the Sequential MitISEM approach, whereas the green 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).*
Right Panel: Predictive likelihood estimates. The asterisks show at each time the mean of 100 predictive likelihoods; the red and blue lines correspond with 95% confidence bounds. The red asterisks and confidence bounds are based on 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.

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 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 the right panel of Figure 1 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 an 'ad hoc approach' as it produces approximately the same quality of candidate distributions for predictive likelihood estimation with considerably less computational effort.

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 idea of tempering was introduced by Geyer (1991), see also Hukushima and Nemoto (1996). The tempering idea is also 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 density 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:

Algorithm 2*. The *Tempered MitISEM* approach for obtaining an approximation to a multimodal target density with kernel $f(\theta)$: Apply the Sequential MitISEM algorithm to $f(\theta)^{1/P_n}$ $(n = 0, 1, 2, \ldots, \tilde{n})$ with P_n monotonically decreasing to $P_{\tilde{n}} = 1$.

To illustrate the Tempered MitISEM approach, we apply it to the same highly multimodal density that is used by Kou, Zhou and Wong (2006), a two-dimensional normal mixture for $\theta = (x_1, x_2)'$ with 20 modes that are relatively very far apart. 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 apply the MitISEM algorithm to the tempered target density. That is, we start with a 'naive' Student-t distribution around one of the modes, with scale matrix equal to minus the inverse Hessian of the log-density. We use this 'naive' Student-t distribution as a candidate in IS to obtain a first estimate of the mean and covariance matrix of the target distribution. We then continue with an 'adaptive' Student-t distribution with mode and scale matrix given by the first estimates of the target distribution's mean and covariance matrix. After that, the usual steps 2-4 of Algorithm 1 in Section 2 are conducted. Given a candidate density, we move sequentially in five steps to $P_5 = 1$ with equally (log) spaced intervals. The second method applies the basic MitISEM algorithm to the real target density. Here no tempering approach is used. The final method is the aforementioned 'adaptive' candidate density, which is the Student-t distribution with adapted mode and scale matrix. That is, for the 'adaptive' candidate density we perform only step (0) and step (1) of the original MitISEM algorithm.

Figure 2 and Table 4 show simulation results from these three methods. All figures are based on 10,000 simulated draws. Panels (A^*) and (B^*) of Figure 2 show simulated draws from the adaptive candidate density, where panel (B^*) is similar to panel (A^*) but zoomed in on a closer interval. These panels plus the huge C.o.V. of IS weights in the table suggest that the 'adaptive' Student-t density produces poor results. In other words, one really needs advanced samplers to handle multimodal target kernels. Second, the basic 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

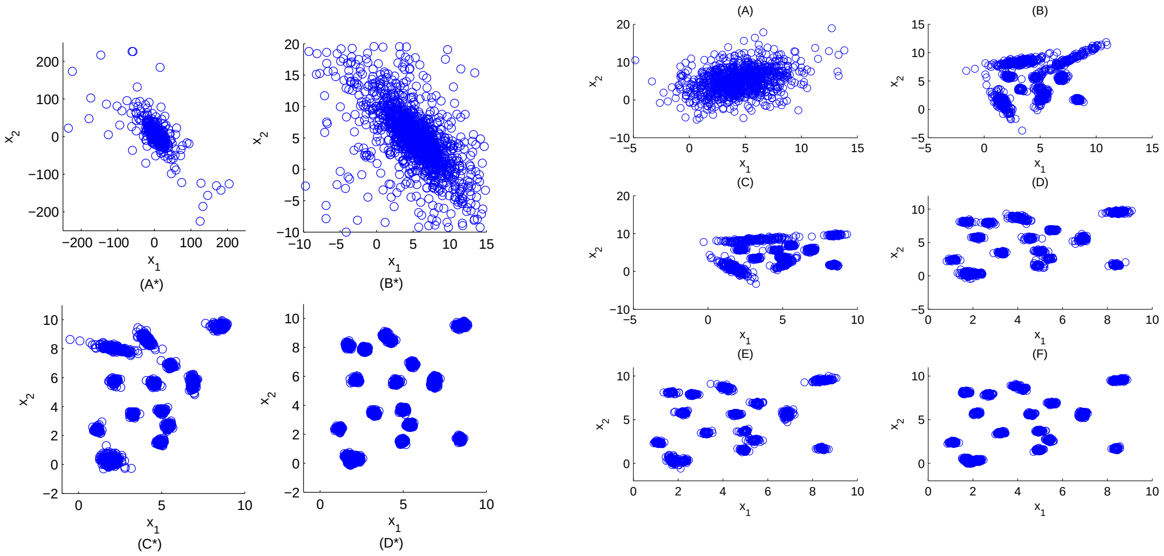

Figure 2: Application to bivariate multimodal distribution of Kou, Zhou and Wong (2006). Left: Panel (A) and (B*) denote samples generated by the Adaptive Student-t density. These panels represent the same draws, but panel (B*) focuses on a smaller interval. Panel (C*) shows draws resulting from applying MitISEM to the original target density. Panel (D*) shows draws simulated from the real target distribution.*
Right: Samples generated from each step of the Tempered MitISEM algorithm. Starting from panel (A) to (E), simulated draws are shown from the candidate density that is produced by applying MitISEM to the target density $f(\theta)^{1/P}$, with P equally log-spaced from 5 to 1. Panel (F) shows draws simulated from the real target distribution.

comparing panel (C^*) - simulated draws from the candidate density that is produced by MitISEM - to panel (D^*) of Figure 2, which represents simulated draws from the real 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 too 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 the right-hand panels of Figure 2. From panel (A) to (E), simulated candidate draws from the resulting candidate density of MitISEM applied to the target density $p(\theta)^{1/P}$ are shown, 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 basic MitISEM is clearly illustrated in panel (E), since all 20 modes are covered. The quality of the final candidate density is also confirmed by Table 4, as the $C.o.V.$ value drops further from 0.77 to 0.43. We stress that the reported numbers of Student-t components are not chosen beforehand by the user; these are automatically found by the basic and tempered MitISEM methods.

Table 4: Results of simulation from the two-dimensional normal mixture of Kou, Zhou and Wong (2006) using three different candidates: an (adaptive) Student-t density, and mixtures of Student-t densities from basic and tempered MitISEM. The number of components of (Tempered) MitISEM and the corresponding C.o.V. of IS weights correspond with the last iteration of the MitISEM algorithm, as described in Algorithms 1 and 2.*

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

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 scalar 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:

$$p(y|\theta) = \prod_{t=1}^{T} \left\{ \sum_{j=1}^{m} \pi_j \left[(2\pi)^{-1/2} \sigma_j^{-1} \exp\left(-\frac{1}{2\sigma_j^2} (y_t - \mu_j)^2 \right) \right\} \right] \tag{29}$$

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 $\theta = (\sigma_1, \sigma_2, \pi_1) = (1,\ \sqrt{2},\ 0.8)$. The left panel of Figure 3 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 highly 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

Table 5: 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

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 5. 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 5 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:

$$\tilde{z}^i_{h,c} \equiv E\left[z^i_{h,c} \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] = \frac{t(\theta^i|\mu_{h,c}, \Sigma_{h,c}, \nu_h)\, \eta_h}{\sum_{j=1}^{J} \sum_{l=1}^{m!} t(\theta^i|\mu_{j,l}, \Sigma_{j,l}, \eta_j)\, \eta_j}. \tag{30}$$

$$\widetilde{z/w}^i_{h,c} \equiv E\left[z^i_{h,c} \frac{1}{w^i_h} \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] = \tilde{z}^i_{h,c}\, \frac{k + \nu_h}{\rho^i_{h,c} + \nu_h}. \tag{31}$$

$$\begin{aligned} \xi^i_h &\equiv E\left[\log w^i_h \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] = \\ &= \sum_{c=1}^{m!} \left\{ \left[\log\left(\frac{\rho^i_{h,c} + \nu_h}{2} \right) - \psi\left(\frac{k + \nu_h}{2} \right) \right] \tilde{z}^i_{h,c} \right\} \\ &\quad + \left[\log\left(\frac{\nu_h}{2} \right) - \psi\left(\frac{\nu_h}{2} \right) \right] \left(1 - \sum_{c=1}^{m!} \tilde{z}^i_{h,c} \right), \end{aligned} \tag{32}$$

$$\begin{aligned} \delta^i_h &\equiv E\left[\frac{1}{w^i_h} \middle| \theta^i, \zeta = \zeta^{(L-1)} \right] \\ &= \sum_{c=1}^{m!} \frac{k + \nu_h}{\rho^i_{h,c} + \nu_h}\, \tilde{z}^i_{h,c} + \left(1 - \sum_{c=1}^{m!} \tilde{z}^i_{h,c} \right). \end{aligned} \tag{33}$$

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:

$$\mu_h^{(L)} = \left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \widetilde{z/w}^{\,i}_{h,c}\right]^{-1} \left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \widetilde{z/w}^{\,i}_{h,c}\ \theta^i_{inv(c)}\right], \tag{34}$$

$$\hat{\Sigma}_h^{(L)} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \widetilde{z/w}^{\,i}_{h,c}\ (\theta^i_{inv(c)} - \mu_h^{(L)})(\theta^i_{inv(c)} - \mu_h^{(L)})'}{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \tilde{z}^i_{h,c}}, \tag{35}$$

$$\eta_h^{(L)} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W_i \ \tilde{z}^i_{h,c}}{\sum_{i=1}^{N} W_i}, \tag{36}$$

whereas the equation of the first order condition for ν_h remains (16). For the derivations we refer to the appendix. The permutation-augmented MitISEM algorithm is briefly summarized as:

Algorithm 3. The *permutation-augmented MitISEM* approach for obtaining an approximation to an unrestricted posterior distribution in a mixture model: Apply the basic MitISEM algorithm with formulas (9)-(15) replaced by (30)-(36) to the unrestricted posterior density kernel.

We apply the permutation-augmented MitISEM approach to the posterior distribution in the left panel of Figure 3, resulting in a mixture of $2 \cdot 7$ Student-t distributions shown in the middle panel of Figure 3. 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 6. The low C.o.V. of the IS weights and the high MH acceptance rate reflect the accuracy of the permutation-augmented MitISEM approximation.

To stress the advantage of the permutation-augmented MitISEM algorithm over the basic MitISEM method, we compare their performance in two simple examples. First, for the posterior in the left panel of Figure 3 the basic MitISEM approximation is given in the right panel of Figure 3. The approximation is slightly worse than for permutation-augmented MitISEM, which is clear from the second row of Table 6. After more computing time the quality of the basic MitISEM candidate is somewhat worse. Second, we consider a simple case of a posterior for 250 simulated observations from the model with $m = 2$, $\mu_1 = \mu_2 = 0$, and $\theta = (\sigma_1, \sigma_2, \pi_1) = (1,\ 5,\ 0.8)$. This posterior is shown in the first panel of Figure 4. The permutation-augmented MitISEM method quickly constructs a close approximation, shown in the second panel of Figure 4. The third row of Table 6 confirms the high quality of the resulting candidate. On the other hand, the basic MitISEM approach yields a candidate, shown in the third panel of Figure 4, that completely misses one of the two modes. Apparently, the distance between the two modes is too large for basic MitISEM. Only if we make use of tempering, then both modes are found; the candidate from tempered MitISEM is shown in the fourth panel of Figure 4. Here tempered MitISEM starts with 10,000 draws from a Student-t approximation of $f(\theta)^{1/50}$ around the posterior mode, with $f(\theta)$ the posterior density kernel. The tempered MitISEM method requires substantially more computing time than permutation-augmented MitISEM, while it also results in a somewhat worse candidate; see the fourth row of Table 6. For the candidate resulting from basic MitISEM, the missed mode does not necessarily cause a problem in a permutation-augmented approach, as long as the second mode is also completely missed during the second stage of using the candidate for IS or MH. However, if accidentally one or more draws are generated near the missed mode, then these will have huge IS weight. IS or MH estimates will then have huge variance. Summarizing, for the computing time and accuracy, it is profitable that the a priori knowledge on the likelihood's invariance is incorporated within the permutation-augmented MitISEM method.

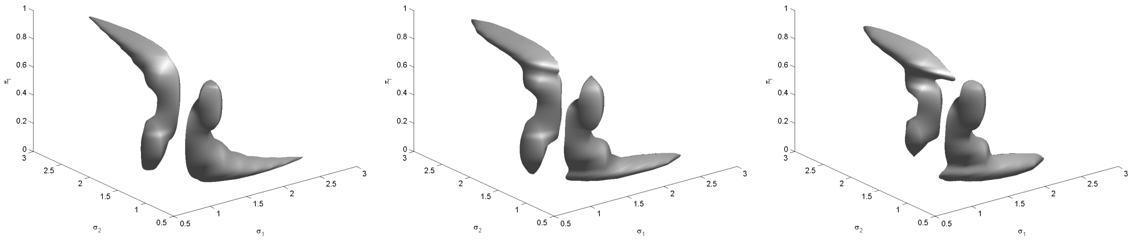

Figure 3: Application of permutation-augmented and basic MitISEM to posterior in mixture model with 2 normally distributed regimes: Highest Posterior Density credible region of $\theta = (\sigma_1, \sigma_2, \pi_1)$ (left); 'Highest Candidate Density region' for mixture of $2 \cdot 7$ Student-t candidate distribution, constructed by permutation-augmented MitISEM (middle); 'Highest Candidate Density region' for mixture of 9 Student-t candidate distribution, constructed by basic MitISEM (right).

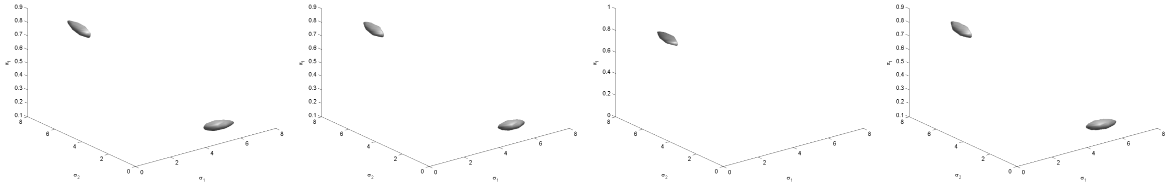

Figure 4: Application of permutation-augmented, basic and tempered MitISEM to posterior in mixture model with 2 normally distributed regimes: Highest Posterior Density credible region of $\theta = (\sigma_1, \sigma_2, \pi_1)$ [first]; 'Highest Candidate Density region' for mixture of $2 \cdot 4$ Student-t candidate distribution, constructed by permutation-augmented MitISEM [second]; 'Highest Candidate Density region' for mixture of 3 Student-t candidate distribution, constructed by basic MitISEM [third]; 'Highest Candidate Density region' for mixture of 7 Student-t candidate distribution, constructed by tempered MitISEM [fourth].

Table 6: Application of permutation-augmented and basic MitISEM to posteriors in mixture models with 2 normally distributed regimes: simulation results for IS using 10,000 draws from the candidate distribution resulting from the permutation-augmented and basic MitISEM procedures

	posterior mean of σ	NSE · 100	time (in s) for construction of MitISEM candidate	time (in s) for simulating 10000 draws	C.o.V. of IS weights	number of t components in MitISEM
DGP ($m = 2$ and $\mu_j = 0$): $(\sigma_1, \sigma_2, \pi_1) = (1, \sqrt{2}, 0.8)$						
permutation-augmented MitISEM	1.0397	0.0683	73.26	0.73	0.58	$2 \cdot 7$
basic MitISEM	1.0395	0.0853	85.47	0.72	0.87	9
DGP ($m = 2$ and $\mu_j = 0$): $(\sigma_1, \sigma_2, \pi_1) = (1, 5, 0.8)$						
permutation-augmented MitISEM	2.4340	0.2869	26.65	0.70	0.24	$2 \cdot 4$
tempered MitISEM	2.4347	0.3293	104.27	0.70	0.31	7

To further stress the relevance of the (permutation-augmented) MitISEM method, we note that for the posterior in Figure 3 a 'naive' Student-t candidate distribution (around posterior mode with scale matrix equal to minus the inverse Hessian of the log-posterior at the mode) would yield a RNE of 0.0009 for the estimated posterior mean of σ, whereas the permutation-augmented MitISEM gives a RNE of 0.7109. The naive sampler would require approximately 800 times more draws than permutation-augmented MitISEM to reach the same accuracy, whereas permutation-augmented MitISEM only requires 1.4 times the number of draws of a hypothetical direct simulation method.

At this point, we must address a disadvantage of the permutation-augmented MitISEM approach. The number of expectations of latent variables $\tilde{z}^i_{h,c}$ and $\widetilde{z/w}^{\,i}_{h,c}$ in (30) and (31) that need to be computed,

Table 7: Application of permutation-augmented MitISEM and Gibbs sampler to posteriors in mixture models with m normally distributed regimes: simulation results for IS and the MH algorithm, using the candidate distribution resulting from the permutation-augmented MitISEM procedure, and for the permutation-augmented Gibbs sampler of Geweke (2007)

	IS or MH with permutation-augmented MitISEM						Gibbs sampling		
	posterior mean of σ	NSE	time (in s) for construction of candidate	time (in s) for simulating 10000 draws	C.o.V. of IS weights	MH acceptance rate	posterior mean of σ	NSE	time (in s) for simulating 10000 draws (+1000 burn-in)
$m = 2$	4.9339	0.0009	19.63	0.79	0.30	0.83	4.9330	0.0009	46.75
$m = 3$	7.4978	0.0014	23.20	1.24	0.47	0.74	7.4963	0.0021	66.11
$m = 4$	10.8300	0.0031	75.08	1.89	0.61	0.67	10.8267	0.0029	84.54

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. Table 7 shows 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 to the model including μ_j ($j = 1, \ldots, m$), and compared the results of IS with the permutation-augmented Gibbs sampler of Geweke (2007). The permutation-augmented Gibbs sampler requires more computing time to reach the same (or worse) accuracy. If we would desire a higher level of precision, then the difference in computing time would be enormous, since simulating $10,000$ extra draws requires much more time in the Gibbs sampler. 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.

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. A tempered permutation-augmented MitISEM approach may be useful then.

5 Partial MitISEM

In this section, we propose a partial MitISEM approach, which aims at approximating the joint posterior *indirectly*, by approximating the product of marginal and conditional posterior distributions of subsets of model parameters. This division can substantially reduce the dimension of the approximation problem, which facilitates the application of adaptive importance or Metropolis-Hastings 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 are mixtures of Student-t distributions with component modes that are linear combinations of (functions of) the parameters in previous subsets. In this way the mode, covariance and shape of the conditional candidate are allowed to depend on 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'.

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 density 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 replaced 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 partial MitISEM approach is based on the following idea. 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

$$\begin{aligned} \theta_1 &\sim N(\mu_1, \Sigma_{11}), && (37)\\ \theta_2|\theta_1 &\sim N(\mu_2 + \Sigma_{22}^{-1}\Sigma_{21}(\theta_1 - \mu_1), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}), && (38) \end{aligned}$$

with

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

In the partial MitISEM approach we replace both normal distributions of (37) and (38) 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. 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)). It is important that X contains a constant, a column of ones. If not, the conditional candidate density would be too restrictive. 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

$$\beta_h^{(L)'} = \left[\sum_{i=1}^{N} W^i \;\; \widetilde{z/w}_h^{\,i} \, X^i \, X^{i'} \right]^{-1} \left[\sum_{i=1}^{N} W^i \;\; \widetilde{z/w}_h^{\,i} \, X^i \, \theta^{i'} \right], \qquad (39)$$

or in case of the permutation-augmented MitISEM approach (34) is replaced by

$$\beta_h^{(L)'} = \left[\sum_{i=1}^{N} \sum_{c=1}^{m!} W^i \ \widetilde{z/w}^{\,i}_{h,c} \, X^i \, X^{i'} \right]^{-1} \left[\sum_{i=1}^{N} \sum_{c=1}^{m!} W^i \ \widetilde{z/w}^{\,i}_{h,c} \, X^i \, \theta^{i'}_{inv(c)} \right]. \qquad (40)$$

For the derivations we refer to the appendix.

We propose the following ‘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 4. The *partial MitISEM* approach for obtaining an approximation to a target density with kernel $f(\theta)$, where θ is decomposed in subsets $\theta_{(s)}$ ($s = 1, \ldots, S$):

(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 $W_0^i \equiv f(\theta^i)/g_0(\theta^i)$ $(i = 1, 2, \ldots, N)$ for this sample.

(2) Do for subset $\theta_{(s)}$ $(s = 1, \ldots, S)$ — approximating the marginal target distribution of θ_1 and conditional target distribution of $\theta_{(s)}$ given $\theta_{(1)}, \ldots, \theta_{(s-1)}$ $(s = 2, \ldots, S)$:

(2a) Apply the **IS-weighted EM algorithm** given the drawn sample from g_0 and the corresponding IS weights W_0^i. The output consists of the new marginal or conditional candidate density $g_{(s)}$ of $\theta_{(s)}$ with optimized $\zeta_{(s)}$, the set of $\mu_{(s),h}$ or $\beta_{(s),h}$, $\Sigma_{(s),h}$, $\nu_{(s),h}$, $\eta_{(s),h}$ $(h = 1, \ldots, H_{(s)})$.

(2b) **Iterate on the number of mixture components $H_{(s)}$ of the candidate $g_{(s)}$ for subset $\theta_{(s)}$**: Take $x\%$ of the sample $\theta^1, \ldots, \theta^N$ from g_0 that correspond to the highest IS weights for the *current* candidate $W^i_{current} \equiv f(\theta^i)/\prod_{s=1}^{S} g(\theta^i_{(s)}|\theta^i_{(1)}, \ldots, \theta^i_{(s-1)})$ $(i = 1, 2, \ldots, N)$, where all marginal and conditional candidate densities are the latest that have been obtained during the algorithm. Construct with these draws a new mode $\mu_{(s),H_{(s)}+1}$ and scale matrix $\Sigma_{(s),H_{(s)}+1}$ which are starting values for the additional component in the mixture candidate density. Starting values for $\eta_{(s),H_{(s)}+1}$ and $\nu_{(s),H_{(s)}+1}$ are at each iteration set at 0.10 and 1, respectively. Obvious starting values for $\mu_{(s),h}$, $\Sigma_{(s),h}$ and $\nu_{(s),h}$ $(h = 1, \ldots, H_{(s)})$ are the optimal values in the mixture of $H_{(s)}$ components, while $\eta_{(s),h}$ is 0.90 times the previously optimal value. Given the drawn sample from g_0 and the corresponding IS weights W_0^i, apply the IS-weighted EM algorithm to optimize parameters $\mu_{(s),h}$ or $\beta_{(s),h}$, $\Sigma_{(s),h}$, $\nu_{(s),h}$ and $\eta_{(s),h}$ with $h = 1, \ldots, H_s + 1$.

(2c) **Evaluate the IS weights** by estimating the Coefficient of Variation (C.o.V.) for the latest candidate, where the expected IS weight and the expected squared IS weight for the latest candidate are estimated using IS with the draws from g_0. Stop the iteration s for subset $\theta_{(s)}$ when this coefficient has converged. Otherwise return to step 2b.

(3) Simulate a new set of draws $\theta^1, \ldots, \theta^N$ from the latest candidate, sampling from marginal and conditional candidate densities for $s = 1, \ldots, S$. Update all marginal and conditional candidate distributions, applying the IS-weighted EM algorithms with the latest set of draws and corresponding IS weights (with initial values for the IS-weighted EM algorithms given by the latest marginal and conditional candidates, and keeping the numbers of components constant). If this leads to a change of

the C.o.V. that is smaller than the 'tolerance' of 10%, then stop. Otherwise, call the latest candidate density g_0, the corresponding weights W_0^i, and return to step 2.

The idea of step 3 is that if the C.o.V. of IS weights improves substantially when the marginal and conditional candidates are re-estimated using a new, better set of candidate draws, then the quality of the approximation has not yet converged. The expected IS weight $E_{\theta\sim g(\theta)}\left[\frac{f(\theta)}{g(\theta)}\right]$ and expected squared IS weight $E_{\theta\sim g(\theta)}\left[\frac{f(\theta)^2}{g(\theta)^2}\right]$ for the latest joint candidate $g(\theta)$ (the product of the latest marginal and conditional candidates), which are required for the evaluation of the C.o.V., are simply estimated as the sample mean of $\frac{f(\theta)}{g_0(\theta)}$ and $\frac{f(\theta)^2}{g(\theta)g_0(\theta)}$ with θ from g_0, since

$$\begin{aligned} E_{\theta\sim g(\theta)}\left[\frac{f(\theta)}{g(\theta)}\right] &= \int \frac{f(\theta)}{g(\theta)}\,g(\theta)\,d\theta = \int \frac{f(\theta)}{g_0(\theta)}\,g_0(\theta)\,d\theta = E_{\theta\sim g_0(\theta)}\left[\frac{f(\theta)}{g_0(\theta)}\right], \\ E_{\theta\sim g(\theta)}\left[\frac{f(\theta)^2}{g(\theta)^2}\right] &= \int \frac{f(\theta)^2}{g(\theta)}\,d\theta = \int \frac{f(\theta)^2}{g(\theta)\,g_0(\theta)}\,g_0(\theta)\,d\theta = E_{\theta\sim g_0(\theta)}\left[\frac{f(\theta)^2}{g(\theta)g_0(\theta)}\right]. \end{aligned}$$

Note that we only require the evaluation of the *joint* target density kernel, no marginal or conditional target densities need to be evaluated. Further, this joint target density kernel is typically evaluated less often than in the basic MitISEM approach, since not each addition of a Student-t component to one of the marginal or conditional candidates implies a new set of candidate draws for which target and candidate density must be evaluated. There are two reasons for this choice. First, otherwise the computing time would be much longer; the number of target density evaluations would increase quadratically with an increasing number of subsets of parameters. Second, since the parameter subsets are smaller than the whole parameter vector, their marginal and conditional distributions are typically much easier to approximate, as the distributions have smaller dimension (and may have 'easier' shapes as well). That is, fewer Student-t components are required, and each Student-t component has fewer candidate parameters. Therefore it is much more efficient to work with 'older' candidate draws and IS weights in the IS-weighted EM steps. Step 3 takes care that the 'older' draws are not too bad for obtaining a good approximation to the target density.

Further, also only the marginal or conditional candidate density for subset $\theta_{(s)}$ that has just been updated needs to be evaluated in step 2c. The candidate densities for the other subsets have been evaluated in previous steps. The number of Student-t components $H_{(s)}$ may (and typically will) differ between the subsets $\theta_{(s)}$; typically a low number of 2 or 3 components suffices for most or even all subsets.

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 male 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 in 1979, 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), where each variable is taken in deviation from its sample mean for the individual's year of birth (1930-1939). The structural form of the model is:

$$y_t = x_t\beta + \varepsilon_t \tag{41}$$

$$x_t = z_t\gamma + v_t \tag{42}$$

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:

$$\begin{aligned} y_t &= z_t \gamma \beta + v_{1t} && (43) \\ x_t &= z_t \gamma + v_t && (44) \end{aligned}$$

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

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

As in the mixture EGARCH model, we assume that the state with smaller variance has larger probability than the state with smaller variance: $\pi_1 > 0.5$ and $\omega_{1,11} < \omega_{2,11}$ (with $\omega_{l,ij}$ the element (i, j) of Ω_l). Further, we specify proper non-informative priors. We consider the 11-dimensional vector of the restricted reduced form's parameters

$$\theta = (\beta, \gamma, \omega_{1,11}, \omega_{1,12}, \omega_{1,22}, \omega_{2,11}, \omega_{2,12}, \omega_{2,22}, \pi_1)'.$$

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

$$\theta_{(1)} = (\beta, \gamma')' \quad (k_1 = 4) \qquad \text{and} \qquad \theta_{(2)} = (\omega_{1,11}, \omega_{1,12}, \omega_{1,22}, \omega_{2,11}, \omega_{2,12}, \omega_{2,22}, \pi_1)' \quad (k_2 = 7).$$

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 the product $\gamma \cdot \beta$; 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 $\theta_{(1)}$ and $\theta_{(2)}$ separately, since these two issues of possible non-identification are not strongly related. The parameters in each subset do not become unidentified for particular parameter values in the other subset.

For $\theta_{(2)}$ the conditional candidate is specified as a mixture of Student-t distributions with modes given by $\mu^i_{h,c} = \beta_{h,c} X^i$, where X^i consists of a constant and the elements of the sample covariance matrix of the restricted reduced form's 'residuals' $y_t - z_t \gamma^i \beta^i$ and $x_t - z_t \gamma^i$ $(t = 1, \ldots, T)$ for given values of $\theta^i_{(1)} = (\beta^i, \gamma^i)$. There are two reasons for choosing this ordering of the subsets $\theta_{(1)}$ and $\theta_{(2)}$. First, the covariance matrix of the restricted reduced form's 'residuals' provides a concise summary of the effect of β and γ on Ω_l $(l = 1, 2)$. Second, this 'residual' covariance matrix arguably affects mainly the conditional *mean* of Ω_l $(l = 1, 2)$. On the other hand, Ω_l $(l = 1, 2)$ would affect mainly the *(co)variance* of β and γ, a dependence that may be somewhat more difficult to approximate. This may require more Student-t components in the conditional candidate, and hence more computing time.

Note that for an optimal selection of the subsets, some understanding of the model and its posterior distribution are required, and that the ordering of the subsets should preferably be chosen in a careful manner. For example, in a cointegration model one could group the parameters occurring in the matrix

Table 8: Application of basic and partial MitISEM to posterior in mixture IV model: estimated posterior means and corresponding numerical standard errors (NSEs), obtained by Importance Sampling with candidate distributions resulting from basic MitISEM and partial MitISEM approaches. We report 100 times the NSE values.

	basic MitISEM		partial MitISEM	
	mean	NSE · 100	mean	NSE · 100
β	0.046	0.035	0.046	0.031
γ_1	0.109	0.115	0.110	0.112
γ_2	0.363	0.120	0.365	0.116
γ_3	0.528	0.125	0.529	0.116
$\omega_{1,11}$	0.215	0.005	0.215	0.005
$\omega_{1,12}$	0.615	0.025	0.615	0.024
$\omega_{1,22}$	11.629	0.237	11.628	0.228
$\omega_{2,11}$	3.111	0.242	3.111	0.236
$\omega_{2,12}$	1.734	0.347	1.735	0.353
$\omega_{2,22}$	21.025	1.610	21.020	1.524
π_1	0.906	0.007	0.906	0.007
C.o.V. of IS weights		0.589		0.529
Computational time (in seconds)				
constructing candidate		238		139
IS using candidate		36		35

having a reduced rank restriction as the first subset, whereas the other parameters may typically have 'easy' conditional posterior distributions. Alternatively, a preliminary diagnostic analysis using MitISEM may be used when the model structure does not give a clear indication.

Table 8 shows the results of basic and partial MitISEM for 11,000 draws. The basic MitISEM approach produces a good candidate distribution leading to a low C.o.V. of the IS weights of 0.589 and low NSEs; for the computation of NSEs for IS estimates see e.g. Hoogerheide et al. (2009). The Metropolis-Hastings algorithm (with a burnin of 1,000 draws) yields similar results to IS with a high acceptance rate of 68.7%. However, the partial MitISEM provides a somewhat better candidate distribution with lower NSEs and C.o.V. of the IS weights, and requires substantially less computing time than the basic MitISEM approach. For this candidate, the Metropolis-Hastings algorithm again yields similar results to IS with a high acceptance rate of 71.8%. This example involves 11 parameters and 2 subsets. If the numbers of parameters and subsets increase, the relative performance (in terms of computing time and quality of the candidate) of the partial MitISEM method, as compared with the basic MitISEM approach, may improve even much further.

If we know the full conditional distribution for one of the subsets, then we may consider this subset as the last subset $\theta_{(S)}$ and use the true full conditional distribution (instead of approximating it by mixtures of Student-t distributions). This obviously reduces the computing time and improves the quality of the candidate. For example, in a random effects model one can apply partial MitISEM in which the random effects are the last subset of parameters, and in which the true conditional posterior distribution of the random effects is used. Alternatively, if possible, one may analytically integrate out the random effects from the data-augmented posterior, and subsequently apply the basic MitISEM approach to the posterior density kernel of the remaining parameters (or partial MitISEM to subsets of the remaining parameters).

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 or Metropolis-Hastings sampling methods may be substantially increased by the partial MitISEM approach and extensions thereof. Finally, we will investigate the application of sequential MitISEM (or an adapted version) to build sequentially candidate densities in time-series state-space models.

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

[4] Berger, J.O. and L.R. Pericchi (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction" *Journal of the American Statistical Association*, 91(433), 109−122.

[5] 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.

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

[7] Chopin, N. (2002), "A sequential particle filter method for static models" *Biometrika*, 89 (3), 539−551.

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

[9] 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.

[10] Doucet, A., De Freitas, N., and N.J. Gordon (2001), "Sequential Monte Carlo Methods in Practice". New York: Springer-Verlag.

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

[12] Engle R.F. (2002), "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models", *Journal of Business & Economic Statistics* 20, 339−350.

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

[14] Frühwirth-Schnatter, S. and H. Wagner (2008), "Marginal likelihoods for non-Gaussian models using auxiliary mixture sampling", *Computational Statistics & Data Analysis*, 52, 4608−4624.

[15] Gelfand, A.E. and D.K. Dey (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations", *Journal of the Royal Statistical Society Series B*, 56(3), 501−514.

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

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

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

[19] Geyer, C.J. (1991), "Markov chain Monte Carlo maximum likelihood", *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E.M. Keramidas, ed.), 156163.

[20] Giordani P. and R. Kohn (2009), "Appendix to 'Adaptive Independent Metropolis-Hastings by Fast Estimation of Mixtures of Normals'," available at website of the *Journal of Computational and Graphical Statistics.*

[21] Giordani P. and R. Kohn (2010), "Adaptive Independent Metropolis-Hastings by Fast Estimation of Mixtures of Normals", *Journal of Computational and Graphical Statistics*, 19, 243−259.

[22] Haario H., E. Saksman and J. Tamminen (2001), "An Adaptive Metropolis Algorithm", *Bernoulli*, 7, 223−242.

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] Hukushima, K. and K. Nemoto (1996), "Exchange Monte Carlo and application to spin glass simulations", Journal of the Physical Society of Japan, 65, 1604−1608.

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

[31] 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.

[32] 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.

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

[34] Lange, K.L., Little, R.J.A. and J.M.G. Taylor (1989), "Robust statistical modeling using the t distribution", *Journal of the American Statistical Association* 84, 881−896.

[35] Liu, J.S. (2001), "Monte Carlo strategies in scientific computing", Springer, New York.

[36] Liu, J.S. and R. Chen (1998), "Sequential Monte Carlo methods for dynamic systems", *Journal of the American Statistical Association* 93, 10321044.

[37] 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.

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

[39] O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparison", *Journal of the Royal Statistical Society Series B*, 57(1), 99−138.

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

[41] Pitt, M.K., Silva, R.S., Giordani, P. and R. Kohn (2010), "Auxiliary particle filtering within adaptive Metropolis-Hastings sampling", http://arxiv.org/abs/1006.1914.

[42] 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.

[43] Roberts G.O. and J.S. Rosenthal (2001), “Optimal Scaling for Various Metropolis-Hastings Algorithms”, *Statistical Science*, 16, 351−367.

[44] Roberts G.O. and J.S. Rosenthal (2009), “Examples of Adaptive MCMC”, *Journal of Computational and Graphical Statistics*, 18, 349−367.

[45] Rubin, D.B. (1983), “Iteratively reweighted least squares”, in: *Encyclopedia of Statistical Sciences*, Vol. 4, 272−275. John Wiley, New York.

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

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

[48] 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.

[49] 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 $(h = 1, \ldots, H; c = 1, \ldots, m!)$:

$$g(\theta) = g(\theta|\zeta) = \sum_{h=1}^{H} \eta_{h,c} \sum_{c=1}^{m!} t_k(\theta|\beta_{h,c}X, \Sigma_{h,c}, \nu_h), \tag{45}$$

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:

$$t_k(\theta|\beta_{h,c}X, \Sigma_{h,c}, \nu_h) = \frac{\Gamma\left(\frac{\nu_h+k}{2}\right)}{\Gamma\left(\frac{\nu_h}{2}\right)(\pi\nu_h)^{k/2}} |\Sigma_{h,c}|^{-1/2} \left(1 + \frac{(\theta - \beta_{h,c}X)'\Sigma_{h,c}^{-1}(\theta - \beta_{h,c}X)}{\nu_h}\right)^{-(k+\nu_h)/2}. \tag{46}$$

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 (45).

The mixture of Student-t densities (45) 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, $z^i_{j,l} = 0$ for $(j,l) \neq (h,c)$; $\Pr[z_{h,c} = 1] = \eta_{h,c}$; w^i_h 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 the so-called data-augmented density is given by

$$\begin{aligned}
\log p(\theta^i, w^i, z^i|\zeta) &= \log p(\theta^i|w^i, z^i, \zeta) + \log p(w^i|\zeta) + \log p(z^i|\zeta) \\
&= \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\ \log\left[\mathrm{pdf}_{N(\beta_{h,c}X^i, w^i_h \Sigma_{h,c})}(\theta^i)\right] + \\
&\quad \sum_{h=1}^{H} \log \mathrm{pdf}_{IG(\nu_h/2,\nu_h/2)}(w^i_h) + \sum_{h=1}^{H} z^i_{h,c} \log\left(\frac{\eta_h}{m!}\right) \\
&= \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\ \left\{-\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{h,c}| - \frac{k}{2}\log(w^i_h)\right. \\
&\quad \left. -\frac{1}{2}\frac{(\theta^i - \beta_{h,c}X^i)'(\Sigma_{h,c})^{-1}(\theta^i - \beta_{h,c}X^i)}{w^i_h}\right\} \\
&\quad + \sum_{h=1}^{H}\left\{\frac{\nu_h}{2}\log\left(\frac{\nu_h}{2}\right) - \left(\frac{\nu_h}{2} - 1\right)\log(w^i_h) - \frac{\nu_h}{2}\frac{1}{w^i_h} - \log\left(\Gamma\left(\frac{\nu_h}{2}\right)\right)\right\} \\
&\quad + \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\log\left(\frac{\eta_h}{m!}\right) \\
&= \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\ \left\{-\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_h| - \frac{k}{2}\log(w^i_h)\right. \\
&\quad \left. -\frac{1}{2}\frac{(\theta^i_{inv(c)} - X^i\beta_h)'(\Sigma_h)^{-1}(\theta^i_{inv(c)} - X^i\beta_h)}{w^i_h}\right\} \\
&\quad + \sum_{h=1}^{H}\left\{\frac{\nu_h}{2}\log\left(\frac{\nu_h}{2}\right) - \left(\frac{\nu_h}{2} - 1\right)\log(w^i_h) - \frac{\nu_h}{2}\frac{1}{w^i_h} - \log\left(\Gamma\left(\frac{\nu_h}{2}\right)\right)\right\} \\
&\quad + \sum_{h=1}^{H}\sum_{c=1}^{m!} z^i_{h,c}\log\left(\frac{\eta_h}{m!}\right),
\end{aligned} \tag{47}$$

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

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

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 $z^i_{h,c}$ ($h = 1,\dots,H; c = 1,\dots,m!$) given θ^i and ζ, with probabilities $\Pr[z^i_{h,c} = 1|\theta^i, \zeta = \zeta^{(L-1)}]$ equal to

$$\tilde{z}^i_{h,c} \equiv E\left[z^i_{h,c}\middle|\theta^i, \zeta = \zeta^{(L-1)}\right] = \frac{t(\theta^i|\beta_{h,c}X^i, \Sigma_{h,c}, \nu_h)\ \eta_h}{\sum_{j=1}^{J}\sum_{l=1}^{m!} t(\theta^i|\beta_{j,l}X^i, \Sigma_{j,l}, \eta_j)\ \eta_j}. \tag{48}$$

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

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

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

$$\widetilde{z/w}^i_{h,c} \equiv E\left[z^i_{h,c}\frac{1}{w^i_h}\middle|\theta^i, \zeta = \zeta^{(L-1)}\right] = \tilde{z}^i_{h,c}\ \frac{k+\nu_h}{\rho^i_{h,c}+\nu_h}. \tag{49}$$

(3) **Expectation of $\log w^i_h$:**

$$\begin{aligned}\xi^i_h &\equiv E\left[\log w^i_h \middle|\theta^i, \zeta = \zeta^{(L-1)}\right] = \\ &= \sum_{c=1}^{m!} E\left[\log w^i_h \middle| z_{h,c} = 1, \theta^i, \zeta = \zeta^{(L-1)}\right]\ \Pr[z^i_{h,c} = 1|\theta^i, \zeta = \zeta^{(L-1)}] \\ &\quad + E\left[\log w^i_h \middle| z^i_{h,c} = 0\ \forall c,\ \theta^i, \zeta = \zeta^{(L-1)}\right]\ \Pr[z^i_{h,c} = 0\ \forall c|\theta^i, \zeta = \zeta^{(L-1)}] \\ &= \sum_{c=1}^{m!}\left\{\left[\log\left(\frac{\rho^i_{h,c}+\nu_h}{2}\right) - \psi\left(\frac{k+\nu_h}{2}\right)\right]\tilde{z}^i_{h,c}\right\} \\ &\quad + \left[\log\left(\frac{\nu_h}{2}\right) - \psi\left(\frac{\nu_h}{2}\right)\right]\left(1 - \sum_{c=1}^{m!}\tilde{z}^i_{h,c}\right),\end{aligned} \tag{50}$$

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 $z^i_{h,c} = 0\ \forall c$, the conditional distribution of w^i_h given θ^i, ζ is the distribution given only ζ (since the observation θ^i does not depend on w^i_h) which is Inverse-Gamma $IG(\nu_h/2, \nu_h/2)$:

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

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

$$\begin{aligned}
\delta_h^i &\equiv E\left[\left.\frac{1}{w_h^i}\right|\theta^i, \zeta = \zeta^{(L-1)}\right] \\
&= \sum_{c=1}^{m!} E\left[\left.\frac{1}{w_h^i}\right| z_{h,c}^i = 1, \theta^i, \zeta = \zeta^{(L-1)}\right] \Pr[z_{h,c}^i = 1|\theta^i, \zeta = \zeta^{(L-1)}] \\
&\quad + E\left[\left.\frac{1}{w_h^i}\right| z_{h,c}^i = 0 \;\forall c,\; \theta^i, \zeta = \zeta^{(L-1)}\right] \Pr[z_{h,c}^i = 0 \;\forall c|\theta^i, \zeta = \zeta^{(L-1)}] \\
&= \sum_{c=1}^{m!} \frac{k+\nu_h}{\rho_{h,c}^i + \nu_h}\, \tilde{z}_{h,c}^i + \left(1 - \sum_{c=1}^{m!} \tilde{z}_{h,c}^i\right). \qquad (51)
\end{aligned}$$

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 (48)-(51) into $\log p(\theta^i, w^i, z^i|\zeta)$ in (47). 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:

$$\begin{aligned}
\beta_h^{(L)\prime} &= \left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W^i\; \widetilde{z/w}_{h,c}^{\,i}\; X^i\, X^{i\prime}\right]^{-1} \left[\sum_{i=1}^{N}\sum_{c=1}^{m!} W^i\; \widetilde{z/w}_{h,c}^{\,i}\; X^i\; \theta_{inv(c)}^{i\prime}\right], \qquad (52) \\
\hat{\Sigma}_h^{(L)} &= \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W^i\; \widetilde{z/w}_{h,c}^{\,i}\; (\theta_{inv(c)}^i - \beta_h^{(L)} X^i)(\theta_{inv(c)}^i - \beta_h^{(L)} X^i)'}{\sum_{i=1}^{N}\sum_{c=1}^{m!} W^i\; \tilde{z}_{h,c}^i}, \qquad (53) \\
\eta_h^{(L)} &= \frac{\sum_{i=1}^{N}\sum_{c=1}^{m!} W^i\; \tilde{z}_{h,c}^i}{\sum_{i=1}^{N} W^i}. \qquad (54)
\end{aligned}$$

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

$$-\psi(\nu_h/2) + \log(\nu_h/2) + 1 - \frac{\sum_{i=1}^{N} W^i\, \xi_h^i}{\sum_{i=1}^{N} W^i} - \frac{\sum_{i=1}^{N} W^i\, \delta_h^i}{\sum_{i=1}^{N} W^i} = 0 \qquad (55)$$

using a procedure for one-dimensional root finding.

B Two alternative adaptive simulation methods

B.1 Adaptive Metropolis (AM) of Roberts and Rosenthal (2009)

The Adaptive Metropolis (AM) sampler is proposed on page 3 of Roberts and Rosenthal (2009). It is a version of the AM algorithm of Haario, Saksman and Tamminen (2001). Defining V as the covariance matrix of the Laplace approximation to the posterior at the posterior mode, at iteration j the proposal distribution is given by:

$$\begin{aligned}
q_j(\theta^c, .) &= N(\theta^c, (0.1)^2 V/k) \qquad \text{if} \quad j < 5k, \\
q_j(\theta^c, .) &= (1-\beta)\, N(\theta^c, (2.38)^2 \Sigma_j/k) + \beta\, N(\theta^c, (0.1)^2 V/k) \qquad \text{if} \quad j \geq 5k, \qquad (56)
\end{aligned}$$

with θ^c the current value of θ, k the dimension of θ, $\beta = 0.05$ and Σ_j the current empirical estimate of the covariance matrix of the target distribution based on the iterations thus far. The scalar 0.1 tries to achieve a high acceptance rate by moving the sampler locally. From previous literature, (see Roberts and Rosenthal (2001)) it is known that the proposal $N(\theta^c, (2.38)^2\Sigma/k)$ is optimal in a particular large-dimensional context. In the original setting, Roberts and Rosenthal (2009) propose I_k instead of the covariance matrix V when $j < 5k$. Here we follow Giordani and Kohn (2010) and replace the identity matrix by V.

B.2 Adaptive Independence Metropolis-Hastings (AIMH) sampler of Giordani and Kohn (2010)

Giordani and Kohn (2010) propose a mixture with four terms as candidate density in their adaptive independent Metropolis-Hastings approach. Starting with the general form, the candidate density at iteration j is given by

$$q_j(\theta, \lambda_n) = \omega_1\, g_0(\theta) + (1 - \omega_1)\, g_j(\theta; \lambda_j), \tag{57}$$

where λ_j denotes a parameter vector that evolves over time. The density $g_0(\theta)$ is constant and given by a mixture of the form $g_0(\theta) = 0.6\,\phi_0(\theta) + 0.4\,\tilde{\phi}_0(\theta)$ where $\phi_0(\theta)$ is a mixture of normals, initialized at iteration $j = 1$ by a Laplace expansion in which case it is a multivariate normal. The density $\tilde{\phi}_0(\theta)$ is a mixture of normals with similar parameters as $\phi_0(\theta)$, however the covariance matrices are multiplied by a factor k_1. Omitting the parameter vector λ_j, the density $g_j(\theta)$ is given by:

$$g_j(\theta) = \omega_2^*\, g_j^*(\theta) + (1 - \omega_2^*)\, \tilde{g}_j^*(\theta), \tag{58}$$

where both $g_j^*(\theta)$ and $\tilde{g}_j^*(\theta)$ are mixtures of normals with again the same parameters, except that the covariance matrices of the last density are multiplied by a scalar k_2. These parameters are estimated by the k-harmonic means clustering algorithm, see Giordani and Kohn (2009) for a discussion of this k-harmonic means clustering algorithm.