Importance sampling from posterior distributions using copula-based approximations

Petros Dellaportas*and Mike G. Tsionas†

August 30, 2017

We provide generic approximations to k-dimensional posterior distributions through an importance sampling strategy. The importance function is a product of k univariate of Student-t densities and a k-dimensional beta-Liouville density. The parameters of the densities and the number of components in the mixtures are adaptively optimized along the Monte Carlo sampling. For challenging high dimensional latent Gaussian models we propose a nested importance function approximation. We apply the techniques to a range of econometric models that have appeared in the literature, and we document their satisfactory performance relative to the alternatives.

Key Words: Bayesian Analysis; beta - Liouville distribution; GARCH; EGARCH; Simultaneous equation model; vector autoregressive.

JEL Classification: C11,C13

*Department of Statistical Science , University College London, Gower Street, London WC1E 6BT, UK and Department of Statistics, Athens University of Economics and Business, Greece. p.dellaportas@ucl.ac.uk

†Lancaster University Management School, Lancaster, LA1 4YX, UK and Department of Economics, Athens University of Economics and Business, Greece. m.tsionas@lancaster.ac.uk

1 Introduction

The impressive growth of Bayesian econometric applications in the last two decades has been clearly due to the driving force of computational methods. The need to evaluate high dimensional integrals for all practically important or interesting applications was acknowledged very early and the very first numerical integration strategies for the implementation of the Bayesian paradigm were based on importance sampling simulations introduced in the seminal paper by Kloek and van Dijk (1978). Although the major advances in Bayesian econometrics has been compelling in terms of Markov Chain Monte Carlo (MCMC) and Sequential Monte Carlo methods, the original goal of drawing Monte Carlo based independent samples from a high-dimensional posterior density in a computationally straightforward way has attracted a lot of research interest in the past 35 years.

The importance sampling scheme we consider is the following. Suppose that interest lies in approximating expectations Ef of some measurable function $f(\theta)$ with respect to a (target) un-normalised posterior density $p(\theta|Y) \propto L(\theta;Y)p(\theta) \equiv \pi(\theta|Y)$, $\theta \in \Theta$ is a continuous parameter vector in $\Re^k$, Y denotes the data, $L(\theta;Y)$ denotes the likelihood function and $p(\theta)$ denotes the prior density. The vast majority of the information required in any Bayesian analysis problem can be expressed as an expectation which requires evaluation of at most k-dimensional integrals. These expectations can be approximated by first sampling I i.i.d. samples θ_i, $i = 1, \ldots, I$ from a density $q(\theta;\boldsymbol{\alpha})$ indexed by a parameter vector $\boldsymbol{\alpha}$ and then using the approximation

$$\hat{E}f = \frac{\sum_{i=1}^n w_i f(\theta_i)}{\sum_{i=1}^n w_i}, \quad w_i = \frac{p(\theta_i|Y)}{q(\theta_i;\boldsymbol{\alpha})}, \; i = 1, \ldots I, \tag{1}$$

where $n^{-1}\sum_{i=1}^n w_i$ is an approximation of the marginal likelihood of the data $\mathcal{M}(Y) = \int_\Theta L(\theta;Y)p(\theta)d\theta$.

The strong law of large numbers guarantees that $\hat{E}f \to Ef$ almost surely and a central limit theorem yields that $\sqrt{n}(\hat{E}f - Ef)$ is asymptotically normal with zero mean and variance σ^2 equal to the expected value, with respect to $p(\theta|Y)$, of $p(\theta|Y)^2 q(\theta;\boldsymbol{\alpha})^{-2}(f(\theta) - Ef)^2$; see Geweke et al. (1989). Following the article by Kloek and van Dijk (1978), the development of importance sampling simulation strategies in Bayesian econometrics was enhanced by the work of Van Dijk and Kloek (1980) and Geweke (1988); Geweke et al. (1989). Evans (1991) was the first to suggest the idea of adaptive importance sampling in which the parameters $\boldsymbol{\alpha}$ can be adapted along with Monte Carlo sampling. The idea has been turned out to be popular, see for example West (1993), Oh and Berger (1993), Givens and Raftery (1996), Bauwens et al. (2004), Richard and Zhang (2007), Hoogerheide et al. (2007), Cappé et al. (2008), Ardia et al. (2009). It is well known that the variance of the consistent and asymptotically normal estimator $\hat{E}f$ depends crucially on how well $q(\theta;\boldsymbol{\alpha})$ approximates $p(\theta|Y)$. Therefore, the key problem considered in these papers is the construction of off-the-shelve algorithms for posterior integration that automatically update $q(\theta;\boldsymbol{\alpha})$ so that it approximates $p(\theta|Y)$. In recent years, this problem has been more popular because of the immediate application of such methods to sequential Monte Carlo algorithms, see for example Cornebise et al. (2008, 2014).

Since $q(\theta;\boldsymbol{\alpha})$ is typically multidimensional and needs to be sampled efficiently, a common choice is a mixture of multivariate normal or Student-t densities. Adaptation of $\boldsymbol{\alpha}$, which here denotes number of components, means, covariance matrices and mixture proportions, is based on either minimisation of the chi-square distance (Kong et al., 1994; Liu, 2008) or the Kullback-Leibler divergence (Cappé et al., 2008) between $q(\theta;\boldsymbol{\alpha})$ and $p(\theta|Y)$. The fact that adaptation is being performed in a high dimensional space with often multi-modal posterior densities makes this adaptation process often problematic.

Our basic idea is based on Sklar (1959) celebrated theorem which states that the density $p(\theta|Y)$ can be written as a product of k univariate marginal densities and a k-dimensional copula density for which the marginal probability of each variable is a uniform density on $(0, 1)$. Research directions in this area have been focused on optimal choice of copulas. Simulation methods to generate random variables from a given copula have been derived to aid Monte Carlo checking of certain estimator properties. Random number generation methods consist of the conditional inverse

method in which samples from conditional densities are generated recursively and the methods by Marshall and Olkin (1988) and McNeil (2008). Although these algorithms are in general inefficient for high dimensional copulas, there are recent examples of fast sampling algorithms, see for example Smith and Maneesoonthorn (2016) and Oh and Patton (2017). We exploit Sklar's theorem by constructing an adaptive importance sampling strategy based on a proposal density $q(\theta;\boldsymbol{\alpha})$ which approximates the k marginal densities and the k-dimensional multivariate copula density by probability density functions that are easy to sample from. The marginal densities are approximated with finite mixtures of univariate Student-t distributions and the copula density with finite mixture of multivariate beta-Liouville distributions. The parameters $\boldsymbol{\alpha}$ of all mixture densities, including the number of components in each finite mixture, are updated adaptively along with the Monte Carlo sampling by minimising the chi-square distance between the target and proposal densities. We document that this approximation performs well in a series of Bayesian econometrics problems and it is easy to craft in practice.

Our sampling method can be applied with various variants in challenging high-dimensional Bayesian inference problems. We illustrate that in the popular family of latent Gaussian models, adoption of sequential, nested approximations are adequate to construct an importance sampling function that samples efficiently the required posterior distribution. The flexibility of our sampling strategy is successfully tested in a series of challenging, new high-dimensional vector autoregressive models with time-varying parameters and multivariate stochastic volatility.

The rest of the paper proceeds as follows. Section 2 describes our methodology and Section 3 illustrates it to a collection of popular Bayesian inference problems. In Section 4 we present variants of the basic methodology that can handle latent Gaussian models together with prediction and sequential updating inferences. We conclude with a short discussion in Section 5.

2 Construction of the importance sampling

2.1 Specification of the importance function

We follow Bauwens et al. (2004) and start by applying an initial transformation as follows. If the initial parameter is denoted by $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$, it is transformed to a new parameter vector $\theta = (\rho, \eta) \in \Re \times \{\eta \in \Re^{k-1} : \eta'\eta < 1\}$ using the transformation

$$\rho = \text{sgn}(\vartheta_k)\sqrt{\vartheta'\vartheta}, \tag{2}$$

$$\eta_j = \vartheta_j \rho^{-1},\ j = 1, \ldots, k-1. \tag{3}$$

The Jacobian of the transformation is $\rho^{k-1}(1-\eta^{'}\eta)^{-1/2}$. Bauwens et al. (2004) have proposed efficient MCMC schemes based on the Metropolis-Hastings algorithm to sample ρ and η by introducing the class of adaptive radial-based direction sampling methods to sample from a posterior distribution which may be non-elliptical.

We construct the importance function $q(\theta;\boldsymbol{\alpha})$ by exploiting Sklar (1959) theorem which states that any posterior density $p(\theta|Y)$ can be written as

$$p(\theta|Y) = \prod_{j=1}^{k} p_j(\theta_j) \cdot c(u_1, \ldots, u_k), \tag{4}$$

where $p_j(\theta_j)$ denotes the marginal density of the j-th element of θ, $u_j = P_j(\theta_j) = \int_{-\infty}^{\theta_j} p_j(\phi)d\phi$ and $c(u_1, \ldots, u_k)$ represents a copula density. Our suggested proposal density $q(\theta;\boldsymbol{\alpha})$ is constructed as follows. First, we choose $\tilde{p}_j(\theta_j)$ to be flexible univariate densities that can capture many shapes of the marginals $p_j(\theta_j)$ and let

$$u_j = \tilde{P}_j(\theta_j) = \int_{-\infty}^{\theta_j} \tilde{p}_j(\phi)d\phi. \tag{5}$$

We now choose $\tilde{c}(u_1,\ldots,u_k)$, $0<u_j<1$, $j=1,\ldots,k$, to be a k-dimensional density that can be sampled efficiently and define the importance function as

$$q(\theta;\boldsymbol{\alpha}) = \tilde{c}(\tilde{P}_1(\theta_1),\ldots,\tilde{P}_k(\theta_k))\prod_{j=1}^{k}\tilde{p}_j(\theta_j). \tag{6}$$

Thus, samples from $q(\theta;\boldsymbol{\alpha})$ can be obtained by first sampling $u_1,\ldots,u_k$ from $\tilde{c}(u_1,\ldots,u_k)$ and then obtaining θ_j from (5) by inverting $\tilde{P}_j$. It is necessary to choose densities $\tilde{p}_j$ of known form, such as normal or Student-t, so that numerical inversion of $\tilde{P}_j$ can be performed through commonly available software. We emphasize here that the crucial advantage between our sampling strategy against existing adaptive Monte Carlo methods for generating samples from $p(\theta|Y)$ is that the hard problem of fitting the high-dimensional dependence induced by the posterior density is achieved by approximating the copula density and all marginal densities rather than the posterior density itself. Our method involves an approximation in $\{u\in\Re_+^k : 0\leq u_j\leq 1, j=1,\ldots,k\}$ and k univariate approximations in $\mathcal{R}$ which are much easier than one approximation over $\Re^k$.

We propose approximating each marginal density by a mixture of G_j Student-t univariate densities with parameters $\boldsymbol{\alpha}_{pj}=(G_j,\sigma_{jg}^2,\pi_{jg},\mu_{jg},\nu_{jg})$, $\sum_{g=1}^{G_j}\pi_{jg}=1$, $0<\pi_{jg}<1$, $\mu_{jg}\in\mathcal{R}$, $\nu_{jg},\sigma_{jg}\in\mathcal{R}^+$:

$$\tilde{p}_j(\theta_j) = \sum_{g=1}^{G_j}\pi_{jg}\frac{\Gamma(\frac{\nu_{jg}+1}{2})}{\Gamma(\frac{\nu_{jg}}{2})(\nu_{jg}\pi\sigma_{jg}^2)^{1/2}}\left(1+\frac{(\theta_j-\mu_{jg})^2}{\nu_{jg}\sigma_{jg}^2}\right)^{-(\nu_{jg}+1)/2}, j=1,\ldots,k. \tag{7}$$

This choice is by no means unique. Any flexible family of univariate densities could have been chosen and we do not believe that the performance of our proposed methodology would be affected.

A key methodological aspect in our sampling strategy is the choice of the importance function that approximates the copula density. An obvious choice is a finite mixture of Gaussian or t-copulas which are based on the corresponding elliptical multivariate distributions. However, the estimation of these copula densities requires estimation of $k(k+1)/2$ parameters of covariance matrices which increases quadratically with k and renders the estimation process cumbersome. Other choices such as, for example, the random Bernstein polynomial copula density, see Burda and Prokhorov (2014), which is a mixture of a product of beta densities is hard to sample from and its evaluation is computationally very expensive.

We propose the use of a finite mixture of beta-Liouville densities which have $k+2$ parameters and sampling from them is straightforward. The beta-Liouville density is a product of a Dirichlet density with parameters α_i and a beta density g with parameters a and b, see (Fang et al., 1990, p. 147). It is written as

$$p_L(u_1,\ldots,u_k;\alpha_1,\ldots,\alpha_k,a,b) = \prod_{i=1}^{k}\frac{u_i^{\alpha_i-1}}{\Gamma(\alpha_i)}\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{(\sum_{i=1}^{k}u_i)^{\sum_{i=1}^{k}\alpha_i-1}}g_{a,b}(\sum_{i=1}^{k}u_i), \tag{8}$$

where $0\leq u_i\leq 1$. It is a generalisation of the Dirichlet distribution since its covariance elements can be, unlike the Dirichlet distribution, positive or negative. Since we need to approximate $c(u_1,\ldots,u_k)$ we need $\sum_{i=1}^{k}u_i\leq k$ so we choose $g_{a,b}$ to be a beta density of the first kind with parameters a and b, see McDonald and Xu (1995):

$$g_{a,b}(r) = \frac{1}{B(a,b)}k^{1-a-b}r^{a-1}(k-r)^{b-1}, 0<r<k.$$

The induced copula function of (8) inherits the dependence structure of the scale mixture representation of Liouville densities and its survival copula belongs to the family of Liouville copulas introduced by McNeil and Nešlehová (2010). Like the Liouville copulas, the copula function induced by (8) does not have an explicit form and can be only written with respect to the Williamson's transform. There are recent results about the tail behaviour of Liouville

copulas and that of their corresponding survival counterparts such as the one we propose through (8), see Belzile and Nešlehová (2017) and Hua (2016). Clearly, this tail behaviour depends on the interaction between the tail behaviour of the Dirichlet density induced by the parameters α_i and the beta density $g_{a,b}$. One clear advantage of this copula is that it has richer tail behaviour than that of the Archimedean copulas derived by setting all Dirichlet parameters $\alpha_i = 1$ which assumes symmetric (exchangeable) dependence. In general, unlike other symmetric copulas such as, for example, Gaussian or Student-t copulas, the copula function induced by (8) will have the ability to better adapt to more complex posterior shapes because it is able to capture asymmetric, non-exchangeable dependence. An enrichment of the association structures captured by copula functions is achieved via mixtures of copulas, see for example Arakelian and Karlis (2014), so our copula density is taken to be a finite mixture of beta-Liouville densities:

$$\tilde{c}(u_1, \ldots, u_k) = \sum_{g=1}^{G_c} \pi_{cg} p_L(u_{g1}, \ldots, u_{gk}; \alpha_{g1}, \ldots, \alpha_{gk}, a_g, b_g) \tag{9}$$

with parameters $\boldsymbol{\alpha}_c = (G_c, \pi_{cg}, \alpha_{g1}, \ldots, \alpha_{gk}, a_g, b_g)$, $\sum_{g=1}^{G_c} \pi_{cg} = 1$, $0 < \pi_{cg} < 1$, $\alpha_{g1}, \ldots, \alpha_{gk}, a_g, b_g \in \mathcal{R}^+$. Thus, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_c, \boldsymbol{\alpha}_{pj}, j = 1, \ldots, k)$.

To generate random drawings from (8) we use the following construction, see (Fang et al., 1990, p. 146). Suppose $w_i \sim Be(\sum_{j=1}^{i} \alpha_j, \alpha_i)$, are mutually independent and independent of $r = \sum_{i=1}^{k} u_i$. Then the required draw is $u = r(\prod_{i=1}^{k-1} w_i, (1-w_1)\prod_{i=2}^{k-2} w_i, \ldots, 1-w_{k-1})$. The inversion of $\tilde{P}_j$ is achieved numerically as follows. The distribution function of the Student-t density is available through standard statistical packages, so $\tilde{P}_j(\theta_j)$ is available through (7). Then, the required θ_j is obtained by solving the optimisation problem $\theta_j = \arg\min_x(\tilde{P}_j(x) - u_j)^2$.

2.2 Adaptation of the importance function

We use the general methodology of adaptive importance sampling which is based on the following steps. For a given $\boldsymbol{\alpha}$ we sample θ_i, $i = 1, \ldots, I$ from $q(\theta; \boldsymbol{\alpha})$ and we compute the un-normalised weights w_i in (1). Adaptation refers to the way $\boldsymbol{\alpha}$ is being estimated adaptively from the sample θ_i so that it approximates the posterior density of interest $p(\theta|Y)$. The most often used criterion is the chi-squared distance between $p(\theta|Y)$ and $q(\theta; \boldsymbol{\alpha})$ defined as

$$\mathcal{W}(\boldsymbol{\alpha}) = \int_{\Theta} \frac{p(\theta|Y)}{q(\theta; \boldsymbol{\alpha})} p(\theta|Y) d\theta - 1 = \int_{\Theta} \left\{ \frac{p(\theta|Y)}{q(\theta; \boldsymbol{\alpha})} \right\}^2 q(\theta; \boldsymbol{\alpha}) d\theta - 1.$$

Geweke et al. (1989) argued that this is a reasonable objective function to minimise and the criterion has been used extensively since, see for example Ardia et al. (2009). In fact $\mathcal{W}(\boldsymbol{\alpha})$ is just the variance of the weight function w_i defined in (1) under the proposal density $q(\theta; \boldsymbol{\alpha})$ and can be readily estimated by computing the squared coefficient of variation of the un-normalised weights

$$\tilde{\mathcal{W}}(\boldsymbol{\alpha}) = \frac{I \sum_{i=1}^{I} w_i^2}{(\sum_{i=1}^{I} w_i)^2} - 1.$$

Note that $\tilde{\mathcal{W}}(\boldsymbol{\alpha})$ is related to the efficient sample size (ESS) which is often used to measure the overall efficiency of the importance function since it represents the number of i.i.d samples equivalent to the number of importance sampling drawings, see Kong et al. (1994):

$$I^{-1} ESS = (1 + \tilde{\mathcal{W}}(\boldsymbol{\alpha}))^{-1}. \tag{10}$$

The chi-squared distance is not the only criterion that can be used in our proposed sampling strategy. In one of our examples we also use, for comparison purposes, the relative numerical efficiency introduced by Geweke et al. (1989) to measure how well an adaptive importance sampling density is tailored to the target density. This quantity

is just the ratio $var(q(\theta;\boldsymbol{\alpha}))/\sigma^2$ and is interpreted as the ratio of number of replications required to achieve any specified numerical standard error using the adaptive importance sampling density, to the number required using the posterior as an importance sampling density. The Kullback-Leibler divergence which is central in cross-entropy methodology is also another alternative, see for example Rubinstein and Kroese (2013); we have not explored this criterion here.

Minimization of $\tilde{\mathcal{W}}(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ can be performed using widely available conjugate gradient algorithms with numerical derivatives. We use subroutine `tn` from package `opt` in `netlib`, a truncated Newton algorithm due to S. G. Nash which is efficient when the number of variables is large. We choose not to update all vector of parameters $\boldsymbol{\alpha}$ simultaneously but instead we update the subset $\boldsymbol{\alpha}^{'}$ which denotes all elements of $\boldsymbol{\alpha}$ except the number of components G_j and G_c. The algorithm proceeds by starting with one mixture component for each marginal and the beta-Liouville density and if a chosen desired optimisation criterion is not satisfied we add one component in each sampling density; see Algorithm 1. This heuristic adaptation strategy has been suggested by Hoogerheide et al. (2012) who discuss the different merits of an algorithm that continues adding mixture components until the quality of the approximation does not improve against alternatives which may require more computing time.

`Start with` $G_j = G_c = 1,\ j = 1, \ldots, k$; `Fix` ϵ.
while *the relative change of* $\tilde{\mathcal{W}}(\boldsymbol{\alpha})$ *is greater than* ε **do**
  `for all` j `set` $G_j = G_j + 1$; `set` $G_c = G_c + 1$; `Minimize` $\tilde{\mathcal{W}}(\boldsymbol{\alpha}^{'})$;
end

Algorithm 1: The adaptive importance sampling algorithm

Algorithm 1 does not necessarily reach a global minimum of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}^{'})$. This issue is very important when an inference problem requires a probabilistic description of the marginals or the copula with a parsimonious model based on finite mixture of densities. However, our goal here is to construct an efficient importance function so the key criterion is the ESS and we have found that Algorithm 1 obtains values of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}) < 1$ in all our real data examples. The trade-off between searching for optimum values or just increasing the number of mixtures depends on whether one would like to adopt a black-box or a more elaborate, adaptive optimisation algorithm. Algorithm 1 needs only one tuning parameter, ϵ. In the extended and challenging examples we present in the following Sections we have found that $\epsilon = 0.01$ works very well and the algorithm converges with at most three components.

The ESS reported in our illustrative examples is based on the resulting number of mixture components and it does not take into account the numerical effort to construct the importance function through the iterations of Algorithm 1. Depending on the cost to evaluate the posterior kernel, this effort might render the efficiency of Algorithm 1 questionable. To address this issue, and for a more direct comparison with other methods, we also report $\tilde{\mathcal{W}}(\boldsymbol{\alpha}^{'})$ obtained by fitting directly five components $G_j = G_c = 5,\ j = 1, \ldots, k$; see Algorithm 2.

`Start with` $G_j = G_c = 5,\ j = 1, \ldots, k$. `Minimize` $\tilde{\mathcal{W}}(\boldsymbol{\alpha}^{'})$.

Algorithm 2: The non-adaptive importance sampling algorithm

Algorithm 2 may unnecessarily use more components for the importance function and it is more probable that it will converge to a local minimum. But there is considerable improvement in numerical efficiency compared to Algorithm 1 and it serves as a yardstick for comparing ESS against other Monte Carlo algorithms. In the Appendix, we report results from this strategy in all our illustrative examples. We have found that although the resulting importance function is not as good as the one derived by Algorithm 1, the overall ESS is satisfactory.

During the first iteration of both Algorithms, the minimisation of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}^{'})$, requires initial values. In all our examples we used some plausible, naive initial values, taken as follows. The Student-t densities were initialised at zero mean, unit scale and five degrees of freedom, the parameters of beta-Liouville densities were set to $\alpha_j = 0.5$, $j = 1, \ldots, k$, and $a = b = 1$ and the five mixing parameters for Algorithm 2 were taken to be equal to 0.2

When the number of components is increased, our optimisation strategy in Algorithm 1 exploits the current optimal values exactly as described in Algorithm 1 of Hoogerheide et al. (2012): we propose the new mixture component in the region in which the current importance sampling weights are larger and we keep all other parameter estimates equal to the current values. An alternative that provided identical results in all our examples is to split the component which has the largest weights by sampling new parameters and mixing probabilities exactly as Richardson and Green (1997) proposed split moves in their reversible jump algorithm for finite mixtures of normals.

In all the examples of the following Sections, the estimation of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ in each iteration of Algorithms 1 and 2 was based on samples of size $I_2 = 100,000$. We did not experiment with the values of I_2, as we consider it as a preliminary stage to construct the importance function. For a direct comparison with other methods such as MCMC with respect to function evaluations, one may replace the sample size I with $I + I_2$ and make a direct comparison of ESS obtained by Algorithm 2. We discuss this issue further in Section 5.

3 Empirical applications

3.1 Incomplete simultaneous equation model

We follow closely Hoogerheide et al. (2007) and consider the following possibly over-identified instrumental variables model, also known as the incomplete simultaneous equations model or errors in variables model, see Zellner et al. (1988):

$$y_1 = y_2\beta + \varepsilon$$
$$y_2 = X\pi^* + v$$

where y_1 and y_2 are $T \times 1$ observation vectors, X is a $T \times k$ matrix of weakly exogenous variables, β is a scalar structural parameter of interest, π^* is a $k \times 1$ vector of reduced form parameters and ε, v are $T \times 1$ vectors of error terms such that their corresponding T elements follow a bivariate normal distribution with zero mean and covariance matrix Σ. Assume that the prior density is non-informative and has the form

$$p(\beta, \pi^*, \Sigma) \propto |\Sigma|^{-h/2}$$

and we set $h = 3$. After integrating out Σ we obtain

$$p(\beta, \pi^*|Y, X) \propto \left\{ \det \begin{bmatrix} (y_1 - y_2\beta)'(y_1 - y_2\beta) & (y_1 - y_2\beta)'(y_2 - X\pi^*) \\ (y_2 - X\pi^*)'(y_1 - y_2\beta) & (y_2 - X\pi^*)'y_2 - X\pi^*) \end{bmatrix} \right\}^{-T/2} \tag{11}$$

which is a bivariate density that is a challenging case for our method since it may show highly non-elliptical shapes when instruments are weak, see Drèze (1976, 1977) and Kleibergen and Van Dijk (1994, 1998). For $\pi^* = 0$ it is well known from Kleibergen and Van Dijk (1994, 1998) that the posterior kernel is improper, although it can be made proper by restricting β and π^* to certain bounded regions. We simulate data as in Hoogerheide et al. (2007) and we evaluate the posterior density in (11) as follows. We set $k = 1$, $T = 100$, $\beta = 0$, $\sigma_{11} = \sigma_{22} = 1$, $\pi^* = 0, 0.1$ or 1 and $\rho = 0, 0.1$ or 1 where ρ is the correlation deduced from Σ between the error terms ϵ and v. Thus, we have nine combinations resulted from the 3×3 values of π^* and ρ that represent three different cases of identification, or quality of instruments, expressed via π^*, and three cases of endogeneity expressed through ρ. The matrix X was filled with independent draws from a standard normal density. Hoogerheide et al. (2007) used their `AdMit` procedure to construct a Type 3 neural network approximation, a mixture of 15 Student-t distributions, and used a million drawings from an algorithm based on importance sampling and Metropolis-Hastings. We did not use the

exact same artificial data as in Hoogerheide et al. (2007) so slight sampling errors may have occurred.

The convergence behaviour of Algorithm 1 is reported in Table 1 whereas final parameter estimates, based on three components, are shown in Table 2. Direct comparison with the results of Hoogerheide et al. (2007) based on one million drawings, reported in Table 2 in their paper, is illustrated in Table 3. The results of the two methods are very similar, indicating that our importance function captured very well the non-elliptical contour of the posterior. Our ESS, expressed through $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ is slightly better than that of Hoogerheide et al. (2007). Algorithm 1 was based on $I = 10^7$ after adaptation and the posterior standard errors are based on 20 independent replications of our algorithm.

Number of components	$\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$	$\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ without transformation
1	85.18	97.12
2	7.43	12.44
3	0.75	3.55

Table 1: Incomplete simultaneous equation model. Number of components and corresponding value of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ achieved by Algorithm 1.

We also investigated the usefulness of the initial approximation (2)-(3) by comparing the values of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ obtained by Algorithm 1 without the transformation; see Table 1. It is evident that the initial transformation is useful since, for example, the values obtained for ESS for 3 components and sample size $I = 1000$ are, through (10), 571 and 220 with and without the transformation respectively.

Evidently, the posterior results are quite close to Hoogerheide et al. (2007) but the effective sample size expressed through (10) and the relative numerical efficiency indicate the advantage of importance sampling over MCMC. Algorithm 2 produced $\tilde{\mathcal{W}}(\boldsymbol{\alpha}') = 0.80$. This is very similar to the best values obtained with 3 components via Algorithm 1. The final parameter estimates are reported in the Appendix.

3.2 Mixture GARCH model

The mixture GARCH model of Ausín and Galeano (2007) is formulated as

$$y_t = \mu + h_t^{1/2}\varepsilon_t$$
$$h_t = \omega + \alpha(y_{t-1} - \mu)^2 + \beta h_{t-1}$$

where h_t denotes the instantaneous volatility at time t, $t = 1, \ldots, T$, $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$ with probability ρ and $\varepsilon_t \sim \mathcal{N}(0, \sigma^2\lambda^{-1})$ with probability $1-\rho$, $0 < \lambda < 1$, $\sigma^2 = (\rho + (1-\rho)/\lambda)^{-1}$ and the parameter vector to be estimated is $\theta = (\mu, \omega, \alpha, \beta, \rho, \lambda)$. To impose covariance stationarity, we restrict $\omega > 0$ and $\alpha, \beta \geq 0$ with $\alpha + \beta < 1$.

The initial value h_0 is treated as a known constant set as the sample variance y_t. Following Bastürk et al. (2017), we use the S&P 500 index percentage log-returns (100 times the change of the logarithm of the closing

	π_1^*	π_2^*	β
location parameters, Student-t	-0.023, 0.017, 0.033	-0.024, 0.017, 0.024	0.71, 0.25, 0.32
scale parameters, Student-t	1.82, 2.36, 0.25	0.31, 0.44, 0.78	0.67, 2.57, 4.43
d.f. parameters, Student-t	1.34, 5.72, 9.44	1.77 3.81 6.13	2.40, 7.17, 15.32
mixing probabilities, Student-t	0.24, 0.32, 0.44	0.27, 0.33, 0.40	0.14, 0.45, 0.41
beta-Liouville, α_j	(0.44, 0.81, 0.92), (0.12, 0.24, 0.71), (0.25, 0.32, 0.61)		
beta-Liouville, (a, b)	(3.81, 7.44), (2.52, 6.33), (4.41, 9.32)		
mixing probabilities, beta-Liouville	0.334, 0.541, 0.125		

Table 2: Incomplete simultaneous equation model. Final estimates of Algorithm 1 based on 3 components.

	π_1^*	π_2^*	β
posterior mean	0.0197	0.0158	0.6355
posterior mean, HKD	0.0200	0.0158	0.6357
posterior standard error $\times 20$	1.2×10^{-4}	1.4×10^{-4}	0.0071
posterior standard error $\times 20$, HKD	1.2×10^{-4}	1.4×10^{-4}	0.0070
relative numerical efficiency	0.9715	0.9822	0.9533
relative numerical efficiency, HKD	0.3622	0.3586	0.2211
Posterior standard deviation	0.0944	0.0934	3.0742
Posterior standard deviation, HKD	0.0945	0.0934	3.0745
$\mathcal{W}(\boldsymbol{\alpha}')$	0.75		
$\mathcal{W}(\boldsymbol{\alpha}')$, HKD	1.47		

Table 3: Incomplete simultaneous equation model. Summary of results: HKD refers to the results of Hoogerheide et al. (2007).

price) from January 2, 1998 to December 26, 2002. For the dimensional parameter vector θ we place a uniform prior on $[-1,1] \times (0,1]^3 \times (\frac{1}{2},1]$. The likelihood function, hence the posterior density under an uninformative prior, may have non-elliptical shapes (Zivot, 2009). We illustrate our importance sampling strategy by comparing it with the results of Bastürk et al. (2017). Moreover, we also included in our comparison the Hamiltonian Monte Carlo algorithm of Girolami and Calderhead (2011) started at the first-stage GMM estimators and run for 50,000 iterations with 10,000 iterations as burn-in.

The results are based on 10^4 draws of the final importance function of Algorithm 1 based on two mixture components. Tables 4 and 5 present values of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ and comparisons with other competing methods with respect to CPU times. Note that the CPU times reported here for AdMit and MitISEM are smaller than the ones reported by Bastürk et al. (2017) because are based on the same mainframe computer as Algorithm 1 for fair comparison. The CPU time of Algorithm 1 refers to the time required for both the adaptation and the sampling effort. The results indicate that we do better than the other importance sampling methods with respect to ESS and as well as the Hamiltonian Monte Carlo, which, of course, requires more draws because of the Markovian dependency of the sampler and more effort because of the necessity to derive second derivatives of the likelihood function. However, our sampling method is slower with respect to computing time. Notice the trade-off between efficiency and precision between Algorithms 1 and 2: it seems that here Algorithm 2 is preferable, it has only slightly larger $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ while it uses 70% of the CPU time used by Algorithm 1. We also report in Table 4 the improvement of the initial parameter transformation (2)-(3) and in Table 6 the final estimates of Algorithm 1 based on two mixture components. Here it seems that the initial parameter transformation does not offer a great improvement in the efficiency of the algorithm. Finally, Algorithm 2 resulted in $\tilde{\mathcal{W}}(\boldsymbol{\alpha}') = 0.92$ which is very satisfactory; see Appendix for the corresponding parameter estimates.

	number of components	$\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$	$\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ without the transformation	CPU time in seconds
AdMit	5	1.99	-	17.57
MitISEM	3	0.99	-	6.12
Algorithm 1	1	2.35	3.42	-
Algorithm 1	2	0.88	0.97	215.3
Algorithm 2	5	0.92	-	151.45

Table 4: Mixture GARCH model. Values of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ and CPU times. AdMit is the Adaptive mixture of Student t-distributions approach and MitISEM is the mixture of t-distributions importance sampling using the EM algorithm for crafting the mixture approximation. Transformation refers to the initial transformation (2)-(3). The results of $\tilde{\mathcal{W}}(\boldsymbol{\alpha}')$ for AdMit and MitISEM are taken from Bastürk et al. (2017). All CPU times are based on the same mainframe computer. CPU time for AdMit and Algorithm 1 includes adaptation time.

	Posterior mean				numerical standard error ×100			
	AdMit	MitISEM	HMC	Algorithm 1	AdMit	MitISEM	HMC	Algorithm 1
ω	0.08	0.08	0.080	0.081	0.07	0.05	0.041	0.04
λ	0.37	0.37	0.369	0.372	0.17	0.17	0.132	0.13
β	0.86	0.86	0.863	0.862	0.07	0.05	0.032	0.03
α	0.10	0.10	0.103	0.102	0.06	0.03	0.025	0.02
ρ	0.79	0.79	0.792	0.792	0.30	0.23	0.151	0.15
μ	0.03	0.03	0.031	0.032	0.07	0.05	0.020	0.02

Table 5: Summary of results for the mixture GARCH model. The results of AdMit and MitISEM are taken from Bastürk et al. (2017). AdMit is the Adaptive mixture of t-distributions approach and MitISEM is the mixture of Student t-distributions importance sampling using the EM algorithm for crafting the mixture approximation. HMC is the Hamiltonian Monte Carlo of Girolami and Calderhead (2011).

	μ	ω	α	β	ρ	λ
location parameters, Student-t	0.17, 0.28	0.10, 0.32	0.15, 0.41	0.28, 0.53	0.22, 0.35	0.21, 0.35
scale parameters, Student-t	0.11, 0.18	0.04, 0.09	0.08, 0.15	0.03, 0.11	0.04, 0.09	0.02, 0.04
d.f parameters, Student-t	3.23, 8.33	5.12, 11.3	1.82, 9.85	3.15, 7.12	4.12, 9.81	3.18, 8.16
mixing probabilities, Student-t	0.31, 0.69	0.35, 0.65	0.21, 0.79	0.61, 0.39	0.66, 0.34	0.77, 0.23
beta-Liouville, α_j	(0.12, 0.19), (0.05, 0.36), (0.10, 0.61), (0.14, 0.64), (0.15, 0.33), (0.22, 0.53)					
beta-Liouville, (a, b)	(1.82, 6.14), (1.30, 4.71)					
mixing probabilities, beta-Liouville	0.713, 0.287					

Table 6: Mixture GARCH model. Final estimates of Algorithm 1 based on two components.

3.3 Marginal likelihood calculation: The EGARCH model

We consider the parameter rich EGARCH-type model of Durham and Geweke (2014) which allows for more than one volatility factor and a finite mixture of normals structure for the disturbance term. The model is as follows:

$$\begin{aligned} y_t &= \mu_Y + \sigma_Y \exp\left(\sum_{k=1}^{K} v_{kt}/2\right)\varepsilon_t,\ t = 1, \ldots, T \\ v_{kt} &= \alpha_k v_{k,t-1} + \beta_k\left(|\varepsilon_{t-1}| - \sqrt{2/\pi}\right) + \gamma_k \varepsilon_{t-1},\ k = 1, \ldots, K \end{aligned}$$

where y_t represents the returns of an asset, v_{kt} are volatility factors and $\mu_Y, \sigma_Y, \alpha_k, \beta_k, \gamma_k$ parameters that are restricted with the usual covariance stationarity restrictions. The disturbance density $p(\varepsilon_t)$ is modelled as a finite mixture of normal densities

$$p(\varepsilon_t) = \sum_{i=1}^{L} p_i \phi(\varepsilon_t; \mu_i, \sigma_i^2).$$

where $\phi(\varepsilon; \mu, \sigma^2)$ represents the normal density with mean μ and variance σ^2. This specification is completed with the zero mean and unit variance conditions

$$\sum_{i=1}^{L} p_i \mu_i = 0,\ \sum_{i=1}^{L} p_i\left(\mu_i^2 + \sigma_i^2\right) = 1.$$

The models are indexed by K, the number of volatility factors, and L, the number of components in the return disturbance normal mixture. The original EGARCH model due to Nelson (1991) has $K = L = 1$. Durham and Geweke (2014) craft carefully a novel sequential Monte Carlo sampler to perform Bayesian analysis of this model.

We compare our results directly with that of Durham and Geweke (2014) so we use price log-differences as returns of the S&P 500 beginning January 3, 1990 ($t = 1$) and ending March 31, 2010 ($T = 5,100$).

Algorithm 1 converged to 3 components with values of $\tilde{\mathcal{W}}(\alpha')$ being 2.35, 1.46 and 0.33 for 1,2 and 3 components respectively, whereas Algorithm 2 resulted in $\tilde{\mathcal{W}}(\alpha') = 0.42$. The corresponding final estimates are presented in the Appendix.

Our comparison with the results in Durham and Geweke (2014) is provided in Table 7. They used a Bayesian sequential Monte Carlo sampling strategy with 2^{16} particles organised in a computationally efficient way in 2^6 groups of 2^{10} particles each. Their estimates of log marginal likelihood are taken from their Table 1, p. 24 corresponding to their 'Hybrid Step 2' which has lower numerical standard error compared to 'Hybrid Step 1'.

EGARCH (K, I)	Durham and Geweke (2014) LML	Durham and Geweke (2014) NSE	Importance sampling LML	Importance sampling NSE	ESS without transformation	ESS of Algorithm 1	ESS with Gaussian copula
(1,1)	16,641.69	0.0541	16,642.11	0.0542	82.42%	95.42%	5.46%
(1,2)	16,713.60	0.0799	16.713.40	0.0781	81.51%	98.51%	4.90%
(2,1)	16,669.39	0.0929	16,668.76	0.0932	75.32%	94.32%	4.48%
(2,2)	16,736.89	0.0864	16,736.73	0.0872	74.44%	97.45%	3.77%
(2,3)	**16,750.83**	0.0869	16,750.25	0.0903	63.24%	98.23%	3.52%
(3,2)	16,734.94	0.0735	16,735.65	0.0633	61.12%	98.12%	2.99%
(3,3)	16,748.75	0.0646	16,748.55	0.0645	72.25%	96.25%	2.62%
(3,4)	16,748.64	0.0716	16,748.33	0.0711	75.14%	98.15%	2.37%
(4,3)	16,745.61	0.0725	16,745.20	0.0722	64.12%	97.13%	1.84%
(4,4)	16,745.54	0.0643	16,745.39	0.0642	72.21%	98.26%	1.72%

Table 7: Log marginal likelihood estimation, EGARCH model. LML = log marginal likelihood; NSE = numerical standard error; ESS= Effective sample size as as % of number of draws. Transformation refers to the initial transformation (2)-(3).

Compared to Durham and Geweke (2014) our method does not always deliver lower numerical standard errors although the differences are not very large and estimates of the log marginal likelihood are comparable. Notably, both approaches agree that $K = 2$ and $L = 3$ works best for this data in terms of marginal likelihood.

Finally, Table 7 includes results from an experimental exercise that illustrates the need to adopt the initial transformation (2)-(3) and a mixture of beta-Liouville densities rather than, for example, a simple Gaussian copula. The transformation offers only a small improvement but clearly the Gaussian copula produces a very inefficient importance sampler with very low ESS.

In terms of computing time, the algorithm of Durham and Geweke (2014) clearly outperforms ours in terms of CPU time. For example, our importance sampling took 178 and 14, 890 seconds for EGARCH(1,1) and EGARCH(4,4) models respectively, whereas the corresponding reported values in Durham and Geweke (2014) are 65 and 2685 seconds respectively. Note that the algorithms of Durham and Geweke (2014) have used parallel computing environment with full GPU implementation.

4 Latent Gaussian models

We focus on a very large family of latent Gaussian models that have a wide range of applications in all areas of econometrics. In these models, the density of the response variable y_t, $t = 1, \ldots, T$, is assumed to belong to an exponential family and is written as $p(Y|\theta, H)$, where $H = (H_1, H_2, \ldots, H_T)$ denotes a vector of T latent Gaussian variables with mean zero and a precision matrix which specifies the prior structure imposed to H. Temporal dependence is introduced by treating the latent process as a structured time series model. The Bayesian treatment of these models requires to treat H as an extra set of parameters and obtain a sample from $p(\theta, H|Y)$. When integrating out the latent variables H is not possible, the dimension of the posterior densities increases with the number of observations, so the required approximation we propose in this paper requires special treatment which

is presented in this Section. Clearly, this family of models represents a challenging task for our sampling strategy.

4.1 Full conditional approximations

Our approximation to the posterior marginals can be enriched by results in Bayesian inference for latent Gaussian models through nested Laplace approximations as developed by Rue et al. (2009). The Gaussianity of the latent paths allows estimation of posterior marginal densities and such approximations may be used as initialisations of the optimisations required in our construction of mixtures of Student-t densities. We propose here a faster approximation which is based on approximating the posterior full conditional densities. Although the resulting approximation may not be as accurate as that of the Laplace approximation, it is adequate for approximating an importance sampling density and it involves only lower-dimensional optimisations.

First, consider the simple, but often met, case in which the modes of the full conditionals $p(H_t|H_{-t}\theta, Y)$, $\hat{H}_t$, are available with low computational cost, and the corresponding second derivatives at the mode, D_t, are analytically available. Then, one can just reduce the dimension of $\boldsymbol{\alpha}_{pj}$ by setting $g = 1$ and replacing the vector of σ^2_{j1} with $-\tau D_t^{-1}$. Thus, only one parameter (τ) is maximized for all variances of marginal densities. As an example, consider the univariate stochastic volatility model:

$$\begin{aligned} y_t &= h_t^{1/2}\varepsilon_t, \\ \log h_t &= \alpha + \rho \log h_{t-1} + v_t, \end{aligned} \tag{12}$$

where $\varepsilon_t \sim \mathcal{N}(0,1)$, $v_t \sim \mathcal{N}(0, \sigma_v^2)$, $t = 1, \ldots, T$. The parameters are $\theta = (\alpha, \rho, \sigma_v, H)$, where $H = (\log h_1, \ldots, \log h_T)$. By setting $\mu_t = [\alpha(1-\rho) + \rho(\log h_{t+1} + \log h_{t-1})](1+\rho^2)^{-1}$ and $\sigma^2 = \sigma_v^2(1+\rho^2)^{-1}$, the log-posterior full conditional of H_t is given by

$$\log p(H_t|H_{t-1}, H_{t+1}, \theta, Y) = -\tfrac{1}{2}H_t - \frac{y_t^2}{2}\exp(-H_t) - \frac{(H_t - \mu_t)^2}{2\sigma^2}$$

and it is concave. Its mode satisfies the equation

$$-\frac{1}{2} + \frac{y_t^2}{2}\exp(-\hat{H}_t) - \frac{\hat{H}_t - \mu_t}{\sigma^2} = 0 \tag{13}$$

and the second derivative at the mode is $D_t = -y_t^2\exp(-\hat{H}_t)/2 - \sigma^{-2}$. Since the mode and the second derivative can be computed at, practically, no cost, we can use as an approximation to the posterior marginals $p(H_t|Y)$ a Student-t density centred at a location which is obtained via (13) with numerical optimisation and variance equal to $-\tau D_t^{-1}$, where τ is optimised in Algorithms 1 or 2 and it is the same across all t.

4.2 Nested beta-Liouville approximations

We present here an adaptive strategy for the beta-Liouville density used for the approximation of the copula function in Gaussian latent models. Specifically, ignoring θ and focusing only on H, approximation of $\tilde{c}(u_1, \ldots, u_T)$ can be obtained by sequentially optimising with respect to $\boldsymbol{\alpha_c}$ for different subsets of the data based on ℓ windows of $\xi = T/\ell$ observations. Note that each data point y_t provides information, through the likelihood function, to the parameter H_t. We denote by $y_{1:t}$ the data vector $y_1, y_2, \ldots, y_t$ and we construct a series of nested approximations $\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \ldots, \boldsymbol{\alpha_\ell}$ based on data $y_{1:\xi}, y_{1:2\xi}, \ldots, y_{1:T}$ respectively. Thus, the final approximation of $\boldsymbol{\alpha_c} = \boldsymbol{\alpha_\ell}$ is achieved through a path of consecutive approximations based on nested data vectors. More precisely, first we construct the ξ-dimensional beta-Liouville approximation $\tilde{c}_{\boldsymbol{\alpha_1}}$ based on data $y_{1:\xi}$. For the window $y_{1:2\xi}$ the 2ξ-dimensional density $\tilde{c}_{\boldsymbol{\alpha_2}}$ is constructed by retaining the optimum values of α_1 from the initial window and optimising only with respect to the remaining ξ parameters. Initial values for this optimisation are the optimum values of $\boldsymbol{\alpha_1}$, which are usually very good approximations due to the dependency imposed by the Gaussian prior. The approach is performed ℓ

times and only ξ-dimensional optimisations are required each time. Although this procedure does not guarantee that the mixture of beta-Liouville densities $\tilde{c}(u_1, \ldots, u_T)$, is based on the best values of $\boldsymbol{\alpha_c}$, we have found that by nesting the approximations the optimisations are much quicker than maximising in the T-dimensional space and the resulting importance function is adequate for our sampling strategy.

4.3 A stochastic volatility example

We apply the full conditional and the nested beta-Liouville approximations to the stochastic volatility model (12). We report a direct comparison with the data used in Kim et al. (1998) based on a sample of 946 observations for the U.K. Sterling / U.S. dollar exchange rate from 1/10/81 to 28/6/85. We adopted the same parameterisation and priors as in Kim et al. (1998), we set $\ell = 20$ and applied the initial transformations based on (2) and (3). Algorithm 1 produced $\tilde{\mathcal{W}}(\boldsymbol{\alpha}') = 0.30$ with one mixture component whereas Algorithm 2 resulted in $\tilde{\mathcal{W}}(\boldsymbol{\alpha}') = 0.47$. The results of Algorithm 1 are shown in Table 8 and the results of Algorithm 2 in the Appendix. Our comparisons based on the simulation inefficiency factor of the sampler as explained in (Kim et al., 1998, p. 368-369) are shown in Table 9. We sampled 200,000 draws from the resulting importance sampling function of Algorithm 1. Evidently, our approximation delivers much the same results with comparable Monte Carlo errors (not reported here) and lower computational inefficiency factors. To illustrate the importance of the initial parameter transformation, the last column of Table 9 presents how the inefficiency factors are increased when the transformation is not applied.

	ρ	σ_η	β
location parameters, Student-t	0.981	0.59	0.66
scale parameters, Student-t	0.11	0.034	0.13
d.f. parameters, Student-t	3.25	12.10	7.33
copula, (a, b)	(2.33,5.85)		

Table 8: Stochastic volatility model. Final estimates of Algorithm 1 based on one mixture component.

	mean Kim et al. (1998)	mean Algorithm 1	Inefficiency Kim et al. (1998)	Inefficiency Algorithm 1	Inefficiency, Algorithm 1 without transformation
ρ	0.97779	0.97744	29.776	15.420	27.44
σ_η	0.15850	0.15831	155.42	32.61	55.12
β	0.64733	0.64729	4.3264	2.311	7.19

Table 9: A Stochastic volatility model. Comparison of Importance sampling based on full conditional and nested beta-Liouville approximations with Kim et al. (1998). IS denotes our proposed importance sampling strategy.

The performance of the nested beta-Liouville approximations with respect to different window sizes can be investigated by inspecting the corresponding inefficiency factors; see Table 10. Overall, as the window size increases the inefficiency factor decreases but the posterior means remain almost the same. The nested approximation can be automated as it depends only on window width, ℓ, and involves only minimal additional computation time due to re-optimizations. The initial parameter transformation provides some improvement in the efficiency of the importance sampling.

4.4 A time-varying parameter vector autoregressive model

In this Section we present a high-dimensional empirical application based on a large time-varying parameter vector autoregressive (TVP-VAR) model. It is well known that we can write a VAR model in the form

$$y_t = X_t \beta_t + \varepsilon_t,$$

	$\ell = 5$		$\ell = 10$		$\ell = 20$	
	mean	inefficiency	mean	inefficiency	mean	inefficiency
ρ	0.97744	11.444	0.97730	9.352	0.97738	7.348
σ_η	0.15822	45.88	0.15815	22.17	0.15821	15.220
β	0.64707	1.2415	0.64712	1.1925	0.64709	1.1172
Results without the transformation (2)-(3)						
ρ	0.97512	14.25	0.97526	14.546	0.97536	12.313
σ_η	0.15417	49.31	0.15424	31.817	0.15220	21.215
β	0.62715	1.715	0.64255	1.832	0.64114	2.212

Table 10: A Stochastic volatility model. Performance of nested beta-Liouville approximations based on different window sizes, with or without the initial parameter transformation (2)-(3).

and allow for parameter variation using a random walk specification

$$\beta_{t+1} = \beta_t + u_t,$$

where y_t is an $M \times 1$ time series vector, X_t is $M \times k$ matrix of lagged values of y_t where, typically, $k = LM$ and L denotes the number of lags, and the error terms are i.i.d. $\varepsilon_t \sim \mathcal{N}(0, \Sigma_t)$, $u_t \sim \mathcal{N}(0, \Omega_t)$ and independent of one another. Each equation contains a $k \times 1$ vector of regressors, say x_t, and $X_t = I \otimes x_t'$. From standard Kalman filter results one can readily deduce that $\beta_t | y^{t-1} \sim \mathcal{N}(\beta_{t|t-1}, P_{t|t-1})$, where $\beta_{t-1}|y^{t-1} \sim \mathcal{N}(\beta_{t-1|t-1}, P_{t-1|t-1})$, the expressions for $\beta_{t-1|t-1}, P_{t-1|t-1}$ are widely available and $P_{t|t-1} = P_{t-1|t-1} + \Omega_t$. Koop and Korobilis (2013) have proposed replacing Σ_t and Ω_t by estimates based on forgetting factors, and exploit analytical expressions for the posterior in order to deal with the insurmountable problems in large VAR models. They also propose to use a similar approximation to remove the need for a posterior simulation algorithm for multivariate stochastic volatility in the measurement equation. Specifically, they use

$$P_{t|t-1} = \zeta P_{t-1|t-1}, \tag{14}$$

$$\hat{\Sigma}_t = \kappa \hat{\Sigma}_{t-1} + (1-\kappa)\hat{\varepsilon}_t \hat{\varepsilon}_t', \tag{15}$$

where $\hat{\varepsilon}_t = y_t - X_t \beta_{t|t-1}$ is the one-step ahead prediction error produced by the Kalman filter and κ is a certain constant, which they choose to be $\kappa = 0.96$. Here, $0 < \zeta \leq 1$ is a forgetting factor which Koop and Korobilis (2013) propose to estimate from the data. In (14) one avoids the need to update elements of Ω_t and in (15) the multivariate stochastic volatility is filtered out. Although (15) is a simple form of multivariate stochastic volatility, this specification works well in terms of predictive accuracy as shown by Koop and Korobilis (2013). The model is clearly a restricted version of a multivariate GARCH model of the form $\Sigma_t = A\Sigma_{t-1} + B\varepsilon_t(\beta_t)\varepsilon_t(\beta_t)'$, where $\varepsilon_t(\beta_t) := y_t - X_t\beta_t$, $A = \kappa I$ and $B = (1-\kappa)I$.

In this example we test the performance of our importance sampling strategy in expectations expressed as predictive densities. For comparison purposes we use the same data as in Koop and Korobilis (2013) . The data set comprises 25 major quarterly US macroeconomic variables from 1959:Q1 to 2010:Q2, so the TVP-VAR model contains 25 equations, for details see Section 3.4 in Koop and Korobilis (2013). Following, for example, Stock and Watson (2009) and recommendations in Carriero et al. (2015) we transform all variables to stationarity.

We used $\kappa = 0.96$ as in Koop and Korobilis (2013) so the parameter vector consists of ζ and the vector of β's. For the constant coefficients we adopt a standard Minnesota prior as in Koop and Korobilis (2013). Algorithm 1 enriched with full conditional and nested beta-Liouville approximations based on $\ell = 5$ converged in one mixture with $\tilde{\mathcal{W}}(\boldsymbol{\alpha}') = 0.61$ and Algorithm 2 resulted to $\tilde{\mathcal{W}}(\boldsymbol{\alpha}') = 0.67$.

The predictive densities based on the importance function derived from Algorithm 1 resulted to the mean

	GDP, forecast horizon							
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
Koop & Korobilis	1.02	1.05	1.03	1.06	1.06	1.08	1.07	1.09
`TVP-VAR`	1.03	1.08	1.03	1.05	1.04	1.07	1.07	1.09
	Inflation, forecast horizon							
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
Koop & Korobilis	1.01	1.03	1.03	0.95	1.01	1.04	0.97	1.02
`TVP-VAR`	1.01	1.02	1.02	0.95	1.01	1.02	0.96	1.02
	Interest rate equation, forecast horizon							
	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$	$h=7$	$h=8$
Koop & Korobilis	1.05	0.94	1.05	0.97	0.98	1.00	0.92	0.91
`TVP-VAR`	1.05	0.92	0.95	0.97	0.99	0.99	0.98	0.97

Table 11: Large TVP-VAR model. Mean squared forecast errors. The results of Koop and Korobilis (2013) correspond to their Table 1, row corresponding to Large TVP-VAR, $\zeta = 0.99$ and β following a random walk.

squared forecast errors reported in Table 11. We compare our results for $h = 1, 2, \ldots, 8$ forecast horizons with the corresponding errors of Koop and Korobilis (2013) TVP-VAR-DMS model. The row that corresponds to the results of Koop and Korobilis (2013) paper were taken from their Table 1, row corresponding to Large TVP-VAR, $\zeta = 0.99$ and β_{T+h} following a random walk model as this row seems to give the most favourable results. Our estimate of ζ was 0.98.

It turns out that our model provides similar results relative to Koop and Korobilis (2013) who also use estimated Σ and Ω matrices through their simplified versions of the Kalman filter updates.

5 Concluding remarks

We have contributed to the long tradition of importance sampling methodology for the implementation of the Bayesian paradigm by introducing a new flexible class of importance functions. Our new methodological ingredient is based on the observation that easy to sample multivariate densities can be constructed as a product of univariate mixture densities and a mixture density on a hypercube, resulting to a copula-based approximation.

The key to construct our proposed importance function was to choose a mixture of beta-Liouville densities on hypercube as a copula density. An important advantage of these densities is that they can capture a rich association structure and at the same time sampling from them is straightforward. Together with the flexible mixture of Student-t densities chosen for the posterior marginal densities, we have provided a copula-based approximation that can be adequately approximate high-dimensional posterior densities.

We have tested our method to a range of examples ranging from a very challenging posterior shape produced by an incomplete simultaneous equation model to high-dimensional latent Gaussian models. Our adaptive strategy performed well compared with other competitive methods. For comparison purposes we also demonstrated results for a non-adaptive strategy based on fixed number of mixture components that does not require sequential adaptation of the importance function parameters. The adaptive scheme (Algorithm 1) requires more computational effort but results in better importance sampling functions.

Our proposed importance sampling strategy may seem very inefficient when the computational effort of the adaptation is added to the estimation samples. For example, the stochastic volatility model requires $I + \ell I_2$ samples to be estimated. Even if the adaptation sample size is much lower than $I_2 = 100,000$ that we used in all our examples, this computational effort is not comparable with efficient purposed-built MCMC algorithms for stochastic volatility models. However, the power of the importance sampling algorithms is that they have a broad applicability to many posterior densities. Moreover, when good initial values are available the adaptation phase can be much more efficient. For example, in the stochastic volatility example we could utilise the full conditional

approximations presented in Section 4.1.

There is a plethora of methodological tools that have been proposed to adaptively fit an importance sampling density for Bayesian computation. One may consider adopting tools such as EM algorithm and machine leaning and investigate how the adaptation may be improved. We believe that in the area of sequential Monte Carlo the importance functions we propose will turn out to be very valuable when multivariate filtering is required, since copula decompositions have not been exploited at all in this rich area of research.

We did not exploit our samplers to construct efficient proposals for Metropolis-Hastings samplers. The experience of Hoogerheide et al. (2012) shows that there is a duality when a good adaptive function is constructed, in the sense that it can be successfully used to both independent and dependent sampling strategies. Although independent Metropolis-Hastings algorithms can be applied immediately using our proposal densities, the challenge here is to construct efficient random walk Metropolis algorithms which can be used in more complex Bayesian models.

We have presented two approximation methods to optimise the parameters of the importance function in latent Gaussian models. There is wide scope to investigate other related optimisation strategies borrowing ideas from quick, but less accurate Bayesian implementation strategies such as variational approximations.

Acknowledgements: The authors wish to thank Gary Koop and Herman K. van Dijk for constructive comments on an earlier version of the article.

References

Arakelian, V. and D. Karlis (2014). Clustering dependencies via mixtures of copulas. *Communications in Statistics-Simulation and Computation 43*(7), 1644–1661.

Ardia, D., L. F. Hoogerheide, and H. K. Van Dijk (2009). Admit: adaptive mixtures of student-t distributions. *The R Journal 1*(1), 25–30.

Ausín, M. C. and P. Galeano (2007). Bayesian estimation of the Gaussian mixture GARCH model. *Computational Statistics & Data Analysis 51*(5), 2636–2652.

Bastürk, N., S. Grassi, L. Hoogerheide, A. Opschoor, and H. van Dijk (2017). The R package MitiSEM: Efficient and robust simulation procedures for Bayesian inference. *Journal of Statistical Software 79*(1), 1–40.

Bauwens, L., C. S. Bos, H. K. Van Dijk, and R. D. Van Oest (2004). Adaptive radial-based direction sampling: some flexible and robust Monte Carlo integration methods. *Journal of Econometrics 123*(2), 201–225.

Belzile, L. R. and J. G. Nešlehová (2017). Extremal attractors of Liouville copulas. *Journal of Multivariate Analysis 43*, 68–92.

Burda, M. and A. Prokhorov (2014). Copula based factorization in Bayesian multivariate infinite mixture models. *Journal of Multivariate Analysis 127*, 200–213.

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*(4), 447–459.

Carriero, A., T. E. Clark, and M. Marcellino (2015). Bayesian VARs: specification choices and forecast accuracy. *Journal of Applied Econometrics 30*(1), 46–73.

Cornebise, J., É. Moulines, and J. Olsson (2008). Adaptive methods for sequential importance sampling with application to state space models. *Statistics and Computing 18*(4), 461–480.

Cornebise, J., E. Moulines, and J. Olsson (2014). Adaptive sequential Monte Carlo by means of mixture of experts. *Statistics and Computing 24*(3), 317–337.

Drèze, J. H. (1976). Bayesian limited information analysis of the simultaneous equations model. *Econometrica 44*(5), 1045–75.

Drèze, J. H. (1977). Bayesian regression analysis using poly-t densities. *Journal of Econometrics 6*(3), 329–354.

Durham, G. and J. Geweke (2014). Adaptive sequential posterior simulators for massively parallel computing environments. In I. Jeliazkov and D. J. Poirier (Eds.), *Bayesian Model Comparison (Advances in Econometrics, Volume 34)*, pp. 1–44. Emerald Group Publishing Limited.

Evans, M. (1991). Adaptive importance sampling and chaining. In N. Flournoy and R. Tsutakawa (Eds.), *Statistical Numerical Integration, Contemporary Mathematics*, Volume 115, pp. 137–143. Amer. Math. Soc., Providence, RI.

Fang, K., S. Kotz, and K. W. Ng (1990). *Symmetric multivariate and related distributions*. Chapman & Hall, London.

Geweke, J. (1988). Antithetic acceleration of Monte Carlo integration in Bayesian inference. *Journal of Econometrics 38*(1-2), 73–89.

Geweke, J. et al. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica 57*(6), 1317–39.

Girolami, M. and B. Calderhead (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(2), 123–214.

Givens, G. H. and A. E. Raftery (1996). Local adaptive importance sampling for multivariate densities with strong nonlinear relationships. *Journal of the American Statistical Association 91*(433), 132–141.

Hoogerheide, L., A. Opschoor, and H. K. Van Dijk (2012). A class of adaptive importance sampling weighted em algorithms for efficient and robust posterior and predictive simulation. *Journal of Econometrics 171*(2), 101–120.

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.

Hua, L. (2016). A note on upper tail behavior of Liouville copulas. *Risks 4*(4), 40.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with arch models. *The Review of Economic Studies 65*(3), 361–393.

Kleibergen, F. and H. K. Van Dijk (1994). On the shape of the likelihood/posterior in cointegration models. *Econometric Theory 10*(3-4), 514–551.

Kleibergen, F. and H. K. Van Dijk (1998). Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory 14*(06), 701–743.

Kloek, T. and H. van Dijk (1978). Bayesian estimates of equation system parameters: An application of integration by monte carlo. *Econometrica 46*(1), 1–19.

Kong, A., J. S. Liu, and W. H. Wong (1994). Sequential imputations and Bayesian missing data problems. *Journal of the American statistical association 89*(425), 278–288.

Koop, G. and D. Korobilis (2013). Large time-varying parameter vars. *Journal of Econometrics 177*(2), 185–198.

Liu, J. S. (2008). *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, USA.

Marshall, A. W. and I. Olkin (1988). Families of multivariate distributions. *Journal of the American statistical association 83*(403), 834–841.

McDonald, J. B. and Y. J. Xu (1995). A generalization of the beta distribution with applications. *Journal of Econometrics 66*(1), 133–152.

McNeil, A. J. (2008). Sampling nested archimedean copulas. *Journal of Statistical Computation and Simulation 78*(6), 567–581.

McNeil, A. J. and J. Nešlehová (2010). From Archimedean to Liouville copulas. *Journal of Multivariate Analysis 101*(8), 1772–1790.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica 59*(2), 347–370.

Oh, D. H. and A. J. Patton (2017). Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics 35*(1), 139–154.

Oh, M.-S. and J. O. Berger (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association 88*(422), 450–456.

Richard, J.-F. and W. Zhang (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics 141*(2), 1385–1411.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology) 59*(4), 731–792.

Rubinstein, R. Y. and D. P. Kroese (2013). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, New York.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society, Series B 71*(2), 319–392.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris 8*, 229–231.

Smith, M. S. and W. Maneesoonthorn (2016). Inversion copulas from nonlinear state space models. *arXiv preprint arXiv:1606.05022*.

Stock, J. H. and M. Watson (2009). Forecasting in dynamic factor models subject to structural instability. In N. Shephard and J. Castle (Eds.), *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry*, pp. 1–57. Oxford University Press, Oxford.

Van Dijk, H. K. and T. Kloek (1980). Further experience in Bayesian analysis using Monte Carlo integration. *Journal of Econometrics 14*(3), 307–328.

West, M. (1993). Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society, Series B 55*, 409–422.

Zellner, A., L. Bauwens, and H. K. Van Dijk (1988). Bayesian specification analysis and estimation of simultaneous equation models using Monte Carlo methods. *Journal of Econometrics 38*(1), 39–72.

Zivot, E. (2009). Practical issues in the analysis of univariate GARCH models. In T. Mikosch, J.-P. Kreiß, R. A. Davis, and T. G. Andersen (Eds.), *Handbook of financial time series*, pp. 113–155. Springer Berlin Heidelberg.

APPENDIX

Incomplete simulaneous equation model: Parameter estimates for Algorithm 2

	π_1^*	π_2^*	β
location parameters, Student-t	-0.32 0.28 0.44 0.02 0.01	0.31 0.57 1.81 -0.03 0.03	0.71 0.25 0.32 0.01 0.04
scale parameters, Student-t	1.82 2.36 0.25 0.02 0.05	0.31 0.44 0.78 0.05 0.08	0.81 0.25 1.54 0.01 0.04
d.f. parameters, Student-t	1.34 5.72 9.44 4,32, 6.81	1.81 3.67 5.82 4.81 7.81	2.40 7.17 15.32 5.43 9.44
mixing probabilities, Student-t	0.21 0.30 0.42 0.03 0.04	0.25 0.30 0.40 0.02 0.03	0.10 0.43 0.37 0.04 0.06
mixing probabilities, beta-Liouville	0.11, 0.80, 0.04		
copula, α_j	(0.11, 0.80, 0.04), (0.10, 0.20, 0.80), (0.01, 0.10, 0.89), (0.03,0.64,0.33),(0.02,0.07, 0.91)		
copula, (a, b)	(3.81, 7.44), (2.52, 6.33), (4.41, 9.32),(0.12,0.40),(0.07,0.18)		

GARCH mixture model: Parameter estimates for Algorithm 2

	μ	ω	α	β	ρ	λ
location parameters, Student-t	(0.12, 0.33,0.07, -0.02, -0.01)	(0.13, 0.25, 0.07, 0.03, 0.01)	(0.17, 0.35, 0.05, 0.03, 0.01)	(0.33, 0.45, 0.01, 0.03, 0.04)	(0.15, 0.23, 0.01, 0.02, 0.0.03)	(0.14, 0.27, 0.12, 0.05, 0.03)
scale parameters, Student-t	(0.07, 0.12, 0.03, 0.02, 0.05)	(0.03, 0.06, 0.01, 0.02, 0.02)	(0.06, 0.10, 0.01, 0.02, 0.03)	(0.042, 0.091, 0.02, 0.02, 0.02)	(0.023, 0.071,0.004, 0.02, 0.03)	(0.012, 0.033, 0.005, 0.007, 0.007)
d.f. parameters, Student-t	(1.16, 2.33, 4.12, 5.15, 7.13)	(3.31, 5.52, 6.45, 8.32, 9.44)	(2.32, 3.51, 4.84, 6.71, 9.44)	(2.33, 4.41, 5.32, 7.18, 8.43)	(1.51, 3.49, 5.32, 6.17, 12.21)	(2.10, 4.14, 5.32, 6.41, 8.14)
mixing probabilities, Student-t	(0.30, 0.60, 0.02, 0.03, 0.05)	(0.40, 0.50, 0.03, 0.06, 0.01)	(0.13, 0.85, 0.005, 0.005, 0.01)	(0.43, 0.52, 0.01, 0.01, 0.03)	(0.30, 0.65, 0.02, 0.02, 0.01)	(0.87, 0.10, 0.01, 0.01, 0.01)
mixing probabilities, beta-Liouville	0.051, 0.075, 0.144, 0.189, 0.322, 0.219					
copula, α_j	(0.32, 0.51, 1.44, 1.71, 2.33), (0.87, 0.44, 1.71, 2.32, 3.12), (1.13, 0.65, 2.32, 3.44, 5.12), (0.45, 0.73, 1.82, 2.44, 3.17), (0.88, 1.7, 2.33, 3.17, 4.10)					
copula, (a, b)	(2.89, 4.83), (2.13. 2.24), (2.37, 5.61), (3.35, 7.81), (2.14, 4.17)					

EGARCH model: Parameter estimates for Algorithm 1: Student-t mixtures, 3 components

	μ_Y	σ_Y	α_{k1}	α_{k2}	β_{k1}	β_{k2}	γ_{k1}
location	-0.17, 0.07, 0.12	0.025, 0.041, 0.062	0.17, 0.32, 0.55	0.19, 0.44, 0.60	0.39, 0.41, 0.48	0.61, 0.35, 0.39	0.33, 0.12, 0.35
scale	0.04, 0.055, 0.081	0.015, 0.021, 0.037	0.012, 0.024, 0.044	0.021, 0.034, 0.061	0.021, 0.033, 0.051	0.022, 0.018, 0.035	0.017, 0.021, 0.044
d.f	4.12, 5.65, 8.21	4.12, 6.14, 9.1	3.17, 8.12, 11.15	4.12, 7.49, 9.31	3.14, 6.12, 8.04	3.21, 5.67, 7.71	2.21, 3.15, 4.44
mixing	0.30, 0.60, 0.10	0.25, 0.17, 0.58	0.13, 0.22, 0.75	0.21, 0.45, 0.34	0.32, 0.35, 0.33	0.17, 0.23, 0.60	0.07, 0.22, 0.69
	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	γ_{k2}
location	-0.17, 0.032, 0.061	-0.047, 0.042, 0.063	-0.012, 0.017, 0.003	0.0062, 0.012, 0.0031	0.0017, 0.0032, 0.0081	0.0011, 0.041, 0.016	0.21, 0.23, 0.35
scale	0.028, 0.035, 0.052	0.012, 0.024, 0.033	0.013, 0.024, 0.045	0.013 0.0014, 0.021	0.0021, 0.0033, 0.032	0.013, 0.021, 0.037	0.007, 0.012, 0.044
d.f	2.12, 3.13, 8.15	2.61, 2.89, 6.15	1.88, 3.25, 9.63	2.37, 4.25, 8.12	1.67, 2.54, 7.32	1.82, 2.44, 6.71	2.89, 4.14, 6.81
mixing	0.21, 0.42, 0.37	0.25, 0.47, 0.28	0.24, 0.43, 0.33	0.14, 0.32, 0.54	0.19, 0.24, 0.57	0.21, 0.48, 0.31	0.09, 0.23, 0.58

Notes: The parameters a and b for beta-Liouville densities were $(2.13, 4.89)$, $(3.12, 8.13)$ and $(12.30, 18.15)$ whereas the mixing parameters were $(0.72, 0.23, 0.05)$. The values of α for each component of the beta-Liouville density were $(0.012, 0.015, 0.222, 0.251, 0.367, 0.554, 0.717, 1.081, 1.337, 1.330)$, $(0.015, 0.167, 0.442, 0.556, 0.771, 1.115, 1.321, 1.446, 1.603, 1.701)$ and $(0.033, 0.045, 0.061, 0.082, 0.093, 0.098, 1.125, 1.233, 1.335, 1.515)$.

EGARCH model: Parameter estimates for Algorithm 2

	μ_Y	σ_Y	α_{k1}	α_{k2}	β_{k1}	β_{k2}	γ_{k1}	γ_{k2}
location parameters, Student-t	(-0.11, 0.044, 0.052, -0.01, -0.02)	(0.018, 0.021, 0.044, 0.003, 0.009)	(0.12 0.28, 0.19, 0.02, 0.04)	(0.14, 0.35, 0.43, 0.01, 0.01)	(0.35, 0.43, 0.33, 0.01, 0.02)	(0.60, 0.21, 0.30, 0.02, 0.03)	(0.27, 0.09, 0.26, 0.01, 0.03)	(0.12, 0.15, 0.17, 0.03, 0.02)
scale parameters, Student-t	(0.028, 0.044, 0.060, 0.02, 0.04)	(0.010, 0.015, 0.028, 0.03, 0.03)	(0.032, 0.057, 0.068, 0.02, 0.04)	(0.012, 0.018, 0.025, 0.01, 0.03)	(0.017, 0.021, 0.032, 0.01, 0.01)	(0.015, 0.021, 0.044, 0.02, 0.02)	(0.015, 0.020, 0.023, 0.03, 0.03)	(0.003, 0.015, 0.043, 0.02, 0.03)
d.f. parameters, Student-t	(5.44, 7.32, 12.20, 21.32, 33.40)	(8.12, 10.25, 13.55, 27.8, 41.4)	(5.13, 7.15, 9.44, 38.4, 45.2)	(2.28, 7.15, 12.20, 25.5, 37.7)	(5.25, 7.19, 11.33, 19.49, 39.5)	(3.12, 5.15, 8.22, 35.6, 40.1)	(2.12, 5.67, 8.13., 33.16, 52.1)	(3.15, 7.12, 12.20, 43.2, 56.4)
mixing probabilities, Student-t	(0.30, 0.60, 0.05, 0.02, 0.03)	(0.10, 0.21, 0.60, 0.03, 0.06)	(0.18, 0.70, 0.04, 0.04, 0.04)	(0.41, 0.55, 0.03, 0.005, 0.005)	(0.30, 0.42, 0.21, 0.06, 0.04)	(0.15, 0.17, 0.65, 0.01, 0.02)	(0.03, 0.20, 0.65, 0.05, 0.07)	(0.11, 0.34, 0.52, 0.02, 0.01)
	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3		
location parameters, Student-t	(-0.020, 0.048, -0.015, 0.012, 0.016)	(-0.050, 0.044, -0.023, 0.030, 0.007)	(-0.015, 0.012, -0.010, 0.017, 0.003)	(0.0085, 0.012, 0.001, 0.012, 0.001)	(0.0044 0.0052, 0.001, 0.020, 0.002)	(0.0051, 0.020, 0.019, 0.012, 0.014)		
scale parameters, Student-t	(0.028, 0.032, 0.007, 0.009, 0.012)	(0.016, 0.019, 0.005, 0.007, 0.009)	(0.015, 0.013, 0.003, 0.003, 0.005)	(0.013, 0.007, 0.002, 0.003, 0.003)	(0.007, 0.012, 0.006, 0.007, 0.009)	(0.004, 0.0055, 0.004, 0.004, 0.007)		
d.f. parameters, Student-t	(3.44, 8.15, 14.51, 25.32, 35.12)	(1.15, 6.77, 15.5, 25.7, 33.8)	(5.88, 12.10, 25.1, 35.1, 44.4)	(4.13, 9.32, 13.10, 28.12, 35.22)	(5.24, 9.71, 22.5, 28.9, 39.3)	(3.12, 13.15, 25.5, 30.4, 42.8)		
mixing probabilities, Student-t	(0.20, 0.70, 0.03, 0.03, 0.04)	(0.03, 0.92, 0.01, 0.03, 0.01)	(0.10, 0.84, 0.01, 0.03, 0.02	(0.11, 0.81, 0.04, 0.03, 0.01)	(0.06, 0.90, 0.02, 0.01, 0.01)	(0.12, 0.82, 0.03, 0.01, 0.02)		
copula, (a, b)	(2.44, 5.75), (4.17, 12.51), (0.51, 0.72), (0.313, 0.32), (0.10, 0.25)							

Notes: The parameter values α_j for the beta-Liouville mixture are available on request.

Stochastic volatility model: Parameter estimates for Algorithm 2

	ρ	σ_η	β
location parameters, Student-t	(0.512, 0.773, 0.830, 0.969, 0.99)	(0.050, 0.01, 0.01, 0.02, 0.01)	(0.51, 0.01, 0.01, 0.02, 0.03)
scale parameters, Student-t	(0.12,0.002, 0.002, 0.002, 0.003)	(0.033, 0.0032, 0.0035, 0.0041, 0.0045)	(0.12, 0.003, 0.004, 0.004, 0.006)
d.f. parameters, Student-t	(3.12, 12.55, 15.43, 17.21, 25.4)	(12.10, 22.50, 28.12, 37.14, 45.51)	(6.87, 33.5, 37.21, 41.5, 50.2)
mixing probabilities, Student-t	(0.03, 0.17, 0.22, 0.35, 0.23)	(0.03, 0.17, 0.33, 0.25, 0.32)	(0.09, 0.13, 0.21, 0.22, 0.45)

Notes: We set ℓ=5. The parameter values for the Student-t mixtures of the latent volatility path and the beta-Liouville mixture are available on request.