TI 2010-059/4
Tinbergen Institute Discussion Paper

A Comparative Study of Monte Carlo Methods for Efficient Evaluation of Marginal Likelihood

David Ardia[1]
Nalan Baştürk[2]
Lennart F. Hoogerheide[2,3]
Herman K. van Dijk[2,3]

[1] *University of Fribourg, and aeris CAPITAL AG,Switzerland;*
[2] *Erasmus University Rotterdam, Tinbergen Institute;*
[3] *Econometric Institute.*

Tinbergen Institute

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

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

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

A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihood

David Ardia* Nalan Baştürk† Lennart F. Hoogerheide‡ Herman K. van Dijk‡

June 2010

Abstract

Strategic choices for efficient and accurate evaluation of marginal likelihoods by means of Monte Carlo simulation methods are studied for the case of highly non-elliptical posterior distributions. A comparative analysis is presented of possible advantages and limitations of different simulation techniques; of possible choices of candidate distributions and choices of target or warped target distributions; and finally of numerical standard errors. The importance of a robust and flexible estimation strategy is demonstrated where the complete posterior distribution is explored. Given an appropriately yet quickly tuned adaptive candidate, straightforward importance sampling provides a computationally efficient estimator of the marginal likelihood (and a reliable and easily computed corresponding numerical standard error) in the cases investigated in this paper, which include a non-linear regression model and a mixture GARCH model. Warping the posterior density can lead to a further gain in efficiency, but it is more important that the posterior kernel is appropriately wrapped by the candidate distribution than that is warped.

Keywords: marginal likelihood; Bayes factor; importance sampling; bridge sampling; adaptive mixture of Student-t distributions.
JEL codes: C11, C15, C52

1 Introduction

This paper provides a comparative study on the efficiency of some commonly used Monte Carlo estimators of marginal likelihood in the context of highly non-elliptical posterior distributions. As the key ingredient in Bayes factors, the marginal likelihood lies at the heart of model selection and model discrimination in Bayesian statistics, see e.g., Kass and Raftery (1995). In several cases of scientific analysis, e.g., in non-linear regression models, mixture

*Department of Quantitative Economics, University of Fribourg, Switzerland, and aeris CAPITAL AG, Switzerland.

†Tinbergen Institute and Econometric Institute, Erasmus University Rotterdam, The Netherlands, corresponding author.

‡Tinbergen Institute and Econometric Institute, Erasmus University Rotterdam, The Netherlands

models, and instrumental variables models, one deals with target distributions that may have very non-elliptical contours and that are not members of a known class of distributions.

The focus is on the situation in which one uses either Importance Sampling (IS; due to Hammersley and Handscomb (1964), introduced in econometrics and statistics by Kloek and Van Dijk (1978)), or the independence chain Metropolis-Hastings algorithm (MH; Metropolis et al. (1953), Hastings (1970)) for posterior simulation. That is, our analysis is especially relevant for those cases where the model structure implies that Gibbs sampling (Geman and Geman (1984)) is not feasible; e.g., non-linear models like the example model of Ritter and Tanner (1992) that we will consider in Section 4. Obviously, the Griddy-Gibbs sampler of Ritter and Tanner (1992) is still feasible in such cases, but we discard this approach due to the computational efforts that it requires. For the Griddy-Gibbs sampler the computing time required for obtaining results with a high precision is typically enormously larger than for the IS and MH approaches.

The marginal likelihood is given by

$$p(y) = \int_{\theta \in \Theta} k(\theta \mid y) d\theta = \int_{\theta \in \Theta} p(y \mid \theta) p(\theta) d\theta, \tag{1}$$

where θ denotes the set of parameters of interest, typically a scalar, a vector, a matrix, or a set of these mathematical objects; Θ is the parameter space; $k(\theta \mid y) = p(y \mid \theta)p(\theta)$ is the kernel function of the joint posterior $p(\theta \mid y)$; $p(y \mid \theta)$ is the likelihood function of θ for the vector of observations $y = (y_1 \cdots y_T)'$; $p(\theta)$ is the exact prior density of θ, i.e., not merely a prior kernel. This marginal likelihood (sometimes also referred to as model likelihood; see e.g., Frühwirth-Schnatter (2001)) is equal to the normalizing constant of the joint posterior density. The estimation of $p(y)$ can be a difficult task in practice, especially for complex statistical models.

The aim of this paper is to investigate the effect that strategic choices may have on the results when estimating a marginal likelihood. We argue that these choices are important for the following issues:

(i) the sensitivity to the choice of the particular sampling procedure (either IS or MH);

(ii) the sensitivity to the choice of the candidate distribution (e.g., a Student-t distribution or a mixture of Student-t distributions);

(iii) the impact of aiming at the posterior density kernel or aiming at a 'warped' version of it;

(iv) the reliability of different types of numerical standard errors (NSE's) as signals for the uncertainty on the respective estimators.

The analysis of the robustness and efficiency of these estimators in the context of non-elliptical posteriors has not been much investigated so far. Frühwirth-Schnatter (2004) provides an

excellent survey but it is restricted to the special case of mixture models. Our results demonstrate the importance of a robust and flexible estimation strategy which explores the full joint posterior. A poor choice of the importance density may lead to a huge loss of efficiency, where the numerical standard error may be highly unreliable. On the other hand, given an appropriately chosen candidate density, the straightforward IS approach provides the most efficient marginal likelihood estimator (with a reliable numerical standard error). The approach of Hoogerheide et al. (2007) that constructs an adaptive mixture of Student-t distributions (AdMit) is particularly useful for automatically obtaining an appropriate candidate density.

This article proceeds as follows. Section 2 provides a summary of some commonly used Monte Carlo estimators of the marginal likelihood. Section 3 gives a brief overview of the AdMit approach. In Section 4 we investigate the robustness and efficiency of these estimators in the case of a three-dimensional highly non-elliptical example distribution, a posterior distribution in a non-linear regression model. In Section 5 we consider the reliability of numerical standard errors. In Section 6 we analyze the performance in a mixture GARCH model. Section 7 concludes.

2 Some Monte Carlo methods for marginal likelihood estimation

We first summarize some of the most commonly used Monte Carlo estimators of marginal likelihood. For more details, see Ardia, Hoogerheide and Van Dijk (2009). We extend the overview of Frühwirth-Schnatter (2004) on Monte Carlo estimators of marginal likelihoods by including the approach of Chib and Jeliazkov (2001), and addressing some more details on implementation, advantages and drawbacks of alternative methods. We especially pay attention to the case of the one-block independence chain MH approach. Further review papers that deal with a comparative review of marginal likelihood estimation methods are Han and Carlin (2001) and Miazhynskaia and Dorffner (2006).

Importance sampling (IS) The IS estimator estimator (Hammersley and Handscomb (1964), Kloek and Van Dijk (1978), Van Dijk and Kloek (1980), Geweke (1989)) is given by

$$\hat{p}_{\mathrm{IS}}(y) = \frac{1}{L} \sum_{l=1}^{L} \frac{k(\theta^{[l]} \mid y)}{q(\theta^{[l]})}, \tag{2}$$

where $\{\theta^{[l]}\}_{l=1}^{L}$ are i.i.d. draws from the exact importance density q which should approximate the joint posterior density $p(\theta \mid y)$. The IS approach of marginal likelihood estimation is a *globally oriented* method that aims at directly evaluating the integral $\int_{\theta \in \Theta} k(\theta \mid y) d\theta$ over the whole parameter space Θ. An importance density which *globally* matches the joint posterior closely will lead to efficient estimation. For this purpose, the tails of q should not be thinner

than the tails of the posterior. That is, q should 'wrap' the posterior density in the sense that all areas of the parameter space Θ that contain substantial posterior probability mass must be 'wrapped' with a reasonable amount of candidate probability mass.

Reciprocal importance sampling (RIS) The RIS estimator (Gelfand and Dey (1994)) is given by

$$\hat{p}_{\mathrm{RIS}}(y) = \left[\frac{1}{M} \sum_{m=1}^{M} \frac{q_{\mathrm{aux}}(\theta^{[m]})}{k(\theta^{[m]} \,|\, y)} \right]^{-1}, \tag{3}$$

where $\{\theta^{[m]}\}_{m=1}^{M}$ are (correlated) posterior draws from an MCMC sampler. q_{aux} is an exact auxiliary density from which we do not require draws. That is, even if the MCMC draws $\{\theta^{[m]}\}_{m=1}^{M}$ are simulated using a candidate density, then this candidate density should generally not be q_{aux}. The RIS approach makes use of the fact that *for each* $\theta \in \Theta$ there holds $p(y) = k(\theta \,|\, y)/p(\theta \,|\, y)$. High efficiency is most likely to result if q_{aux} roughly matches the posterior density. However, the RIS estimator is still consistent if q_{aux} only covers a small part of the parameter space Θ. For stability of the estimator, the tails of $q_{\mathrm{aux}}(\theta)$ should not be fatter than those of the posterior in order to keep the ratio $q_{\mathrm{aux}}(\theta)/k(\theta \,|\, y)$ bounded. Van Dijk and Kloek (1980), Hop and Van Dijk (1992) and Gelfand and Dey (1994) propose a multivariate Gaussian or Student-t density whose mean vector and covariance matrix are estimated from the joint posterior sample. Geweke (1999) proposes the use of a multivariate Gaussian density, truncated to a subspace of Θ.

An advantage of the RIS estimator is that the auxiliary density q_{aux} does not have to cover the whole posterior. Still, we do require that the MCMC draws $\{\theta^{[m]}\}_{m=1}^{M}$ are representative of the whole posterior distribution: otherwise the RIS estimator is no longer consistent.

A special case of (3) is the harmonic mean estimator by Newton and Raftery (1994), in which the prior $p(\theta)$ is used as the auxiliary density. However, it is well-known that this estimator is unstable. Therefore, we do not investigate the version of the harmonic mean.

(Optimal) bridge sampling (BS) The BS estimator (Meng and Wong (1996)) is obtained as the limit of the sequence

$$\hat{p}_{\mathrm{BS}}^{(t)}(y) = \hat{p}_{\mathrm{BS}}^{(t-1)}(y) \cdot \frac{\frac{1}{L} \sum_{l=1}^{L} \frac{\hat{p}(\theta^{[l]} \,|\, y)}{Lq(\theta^{[l]}) + M\hat{p}(\theta^{[l]} \,|\, y)}}{\frac{1}{M} \sum_{m=1}^{M} \frac{q(\theta^{[m]})}{Lq(\theta^{[m]}) + M\hat{p}(\theta^{[m]} \,|\, y)}}, \tag{4}$$

where $\hat{p}(\theta \,|\, y) = k(\theta \,|\, y)/\hat{p}_{\mathrm{BS}}^{(t-1)}(y)$ and the initial value $p_{\mathrm{BS}}^{(0)}(y)$ is set to (2), for instance. The $\{\theta^{[m]}\}_{m=1}^{M}$ are (correlated) posterior draws from an MCMC sampler and $\{\theta^{[l]}\}_{l=1}^{L}$ are i.i.d. draws from the importance density q. Usually, we set $M = L$. Convergence of the bridge sampling technique requires few steps in practice (i.e., typically less than ten iterations). Moreover, these steps do not require many additional computational efforts: no extra draws or evaluations of candidate or target densities are needed. The BS estimator provides

(asymptotically) the optimal combination of draws $\{\theta^{[m]}\}_{m=1}^{M}$ and $\{\theta^{[l]}\}_{l=1}^{L}$ for the estimation of a (ratio of) normalizing constant(s). That is, the BS estimator gives the optimal *bridge* between the posterior kernel and the candidate density q. The original BS estimator in (4) is optimal if the draws $\{\theta^{[m]}\}_{m=1}^{M}$ are i.i.d. We refer to this estimator as the BS1 estimator. A simple correction for correlated draws is proposed by Meng and Schilling (2002). This correction means that one substitutes M by an 'effective number of draws' $\tilde{M}$, defined as $\tilde{M} = M(1-\rho_1)/(1+\rho_1)$ with ρ_1 the first order serial correlation of the likelihood evaluations of the $\{\theta^{[m]}\}_{m=1}^{M}$. We refer to this estimator as the BS2 estimator.

In general, an advantage of the BS estimator is that its variance depends on a ratio that is bounded regardless of the tail behavior of the importance density q, which renders the estimator robust. A disadvantage is that we require both a set of draws from the posterior and a set of independent candidate draws. Further, it requires some implementation cost. It has been investigated by Frühwirth-Schnatter (2004) in the context of mixture models, where it has shown a good performance.

The optimal bridge sampling estimator is a special case of the general bridge sampling (GBS) estimator

$$\hat{p}_{\mathrm{GBS}}(y) = \frac{\frac{1}{L}\sum_{l=1}^{L} \alpha(\theta^{[l]}) k(\theta^{[l]} \,|\, y)}{\frac{1}{M}\sum_{m=1}^{M} \alpha(\theta^{[m]}) q(\theta^{[m]})}\,. \tag{5}$$

The IS and RIS estimators are also members of this class of GBS estimators: these correspond to the choices of $\alpha_{\mathrm{IS}}(\theta) = 1/q(\theta)$ and $\alpha_{\mathrm{RIS}}(\theta) = 1/k(\theta \,|\, y)$, respectively. The BS1 estimator corresponds to the choice

$$\alpha_{\mathrm{BS1}}(\theta) \propto \frac{1}{L\, q(\theta) + M\, p(\theta \,|\, y)},$$

that asymptotically minimizes the relative error of the GBS estimator $\hat{p}_{\mathrm{GBS}}(y)$ if the posterior draws $\{\theta^{[m]}\}_{m=1}^{M}$ are independent.

Chib and Jeliazkov (2001) (CJ) The CJ estimator for marginal likelihood estimation on the basis of MH draws is given by

$$\hat{p}_{\mathrm{CJ}}(y) = \frac{k(\theta^* \,|\, y)}{\hat{p}(\theta^* \,|\, y)}\,, \tag{6}$$

where θ^* is a certain point in the parameter space Θ with $p(\theta^* \,|\, y) > 0$. In the case of the independence chain MH algorithm, the estimated density $\hat{p}(\theta^* \,|\, y)$ of the CJ estimator is given by

$$\hat{p}(\theta^* \,|\, y) = q(\theta^*) \frac{\frac{1}{M}\sum_{m=1}^{M} \alpha_{\mathrm{MH}}(\theta^{[m]}, \theta^*)}{\frac{1}{L}\sum_{l=1}^{L} \alpha_{\mathrm{MH}}(\theta^*, \theta^{[l]})}\,, \tag{7}$$

with $\alpha_{\mathrm{MH}}(\theta, \theta')$ the probability that a transition from θ to θ' is accepted in the MH algorithm,

$$\alpha_{\mathrm{MH}}(\theta, \theta') = \min\left\{1, \frac{k(\theta' \,|\, y)}{k(\theta \,|\, y)} \frac{q(\theta)}{q(\theta')}\right\}.$$

The CJ approach can be applied for each $\theta^* \in \Theta$ with $p(\theta^* \mid y) > 0$. However, for efficiency, the point θ^* must be taken to be a high-density point in Θ, typically the posterior mode. In the case of a highly non-elliptical posterior distribution it may be a bad strategy to use the (estimated) posterior mean, as this may have a low (or even zero) posterior density value.

The CJ estimator is another member of the class of GBS estimators, corresponding to the choice of

$$\alpha_{\mathrm{CJ},\theta^*}(\theta) = \min\left\{\frac{q(\theta^*)}{q(\theta)}, \frac{k(\theta^* \mid y)}{k(\theta \mid y)}\right\}.$$

See Meng and Schilling (2002) and Mira and Nicholls (2004) who show that also other variations proposed by Chib and Jeliazkov (2001) are individual cases of bridge sampling. This suggests that the CJ approach should always be dominated by the optimal BS method. However, BS1 is only optimal: (i) asymptotically; and (ii) if the posterior draws were i.i.d.. For the BS2 estimator, the optimality is also asymptotical and the 'effective number of draws' may provide a crude correction. Therefore, it still makes sense to compare the performance of the CJ and BS methods.

Of the approaches that we consider, the CJ method is the *most local* method: we only estimate the posterior density in one point θ^*. This is in sharp contrast with the IS approach where the whole posterior is 'wrapped' by a fat-tailed candidate. In between we have the RIS method, where (possibly a subspace of) the parameter space is covered by a thin-tailed auxiliary density. A graphical overview of these methods is given by Figure 1.

The Gibbs sampler is a special case of the MH approach, so that the method of Chib (1995) that estimates the marginal likelihood from Gibbs draws, is a special case of the CJ method. In the case of IS we can in principle use the prior as the importance density. However, we do not consider this option in this paper, as this approach is typically very inefficient; see Van Dijk (1999). In general, the prior has much higher variance than the posterior, so that the IS estimate would then be based on only a few IS weights (i.e., likelihood evaluations), with most likelihood values being close to zero.

Warping The methods above can be used in combination with another technique: warping the target posterior (see Meng and Schilling (2002)). If we assume that the parameter space of θ is $\Theta = \mathbb{R}^d$, then

$$p(y) = \int_{\theta\in\Theta} k(\theta \mid y) d\theta = \int_{\theta\in\Theta} \frac{1}{2}\left[k(\theta \mid y) + k(-\theta + 2\theta_0 \mid y)\right] d\theta\,. \tag{8}$$

This implies that application of the aforementioned methods to the *warped* posterior kernel

$$\tilde{k}(\theta \mid y) = \frac{1}{2}\left[k(\theta \mid y) + k(-\theta + 2\theta_0 \mid y)\right]\,, \tag{9}$$

rather than to the posterior kernel $k(\theta \mid y)$, also yields an estimator of the marginal likelihood. The *warped* posterior kernel $\tilde{k}(\theta \mid y)$ is point symmetric around θ_0, where one typically chooses θ_0 as the (estimated) posterior mean. This gain in symmetry may substantially improve the approximation of the target density by the candidate density, typically a symmetric

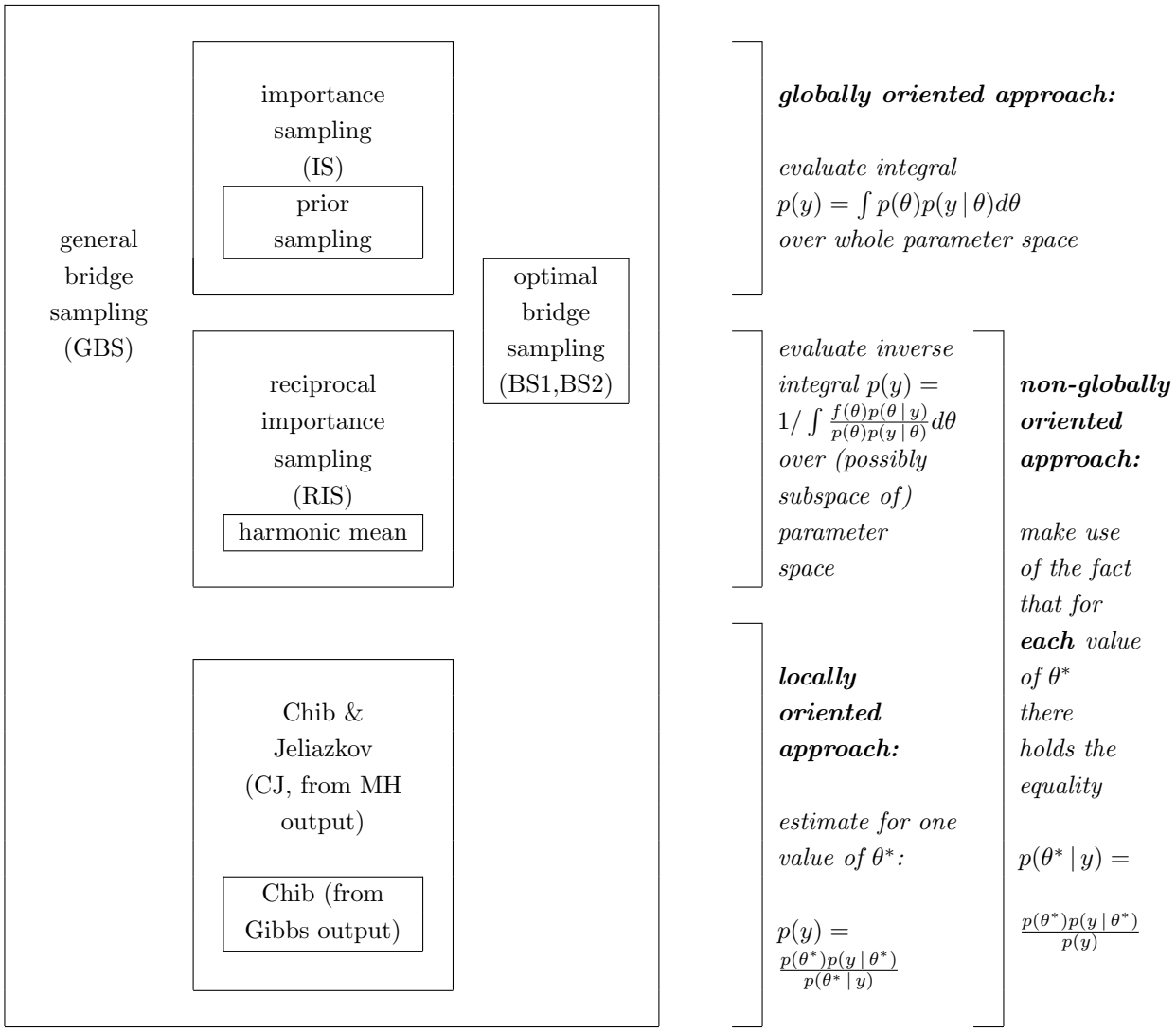


Figure 1: Classification of some well-known methods for estimating marginal likelihoods. All estimators are members of the class of general bridge sampling estimators.

density (e.g., Gaussian and Student-t). This may yield a substantial increase in efficiency. However, a disadvantage is that for each candidate draw we now require two evaluations of the posterior density kernel instead of one. We refer to the transformation in (9) as the Warp1 transformation.

In the two terms of the Warp 1 transformation in (9) we either take the original parameter vector θ or the 'mirror image' of all elements. A further gain in symmetry is obtained by taking an average over all 2^d combinations where individual elements of θ may be 'mirrored'. Obviously, a disadvantage is that for increasing values of the dimension d, the number of posterior kernel evaluations per candidate draw increases exponentially. We refer to this transformation as the Warp2 transformation.

Meng and Schilling (2002) use the name Warp-III for both these Warp1 and Warp2 transformations: Warp-I and Warp-II correspond to adapting the location and variance of the target density to the candidate. We always use candidate distributions of which the location and variance are adapted to the target, so that we only explicitly make use of the Warp-III type transformation that eliminates asymmetries via mixtures of the target.

Table 1 provides an overview of the number of candidate draws and function evaluations that are required by different methods. The candidate distributions that we will consider are Student-t distributions and mixtures of Student-t distributions. The auxiliary densities (of RIS) will be truncated Gaussian. Evaluations of these densities and the simulation of pseudo-random draws from these distributions is done easily and quickly. Therefore, the computational efforts mainly depend on the number of posterior kernel evaluations. For a *fair* comparison between methods, we apply these in such a way that the numbers of posterior kernel evaluations are equal. The IS and RIS estimators are members of the general bridge sampling (GBS) class of which the BS2 estimator is (approximately, asymptotically) optimal. However, this result holds for L and M taken equal in IS, RIS and BS. In this paper the equal numbers of posterior kernel evaluations imply that we take L_{IS} and M_{RIS} twice as large as $L_{\text{BS}} = M_{\text{BS}}$, so that IS and RIS could very well outperform BS.

We focus on the cases of IS and the independence chain MH algorithm. So, we compare the following strategies:

(IS) use all candidate draws in the IS estimator (2);

(RIS, CJ) transform all candidate draws to a sequence of MH draws (plus a burn-in) and use these in the RIS estimator (3) or the CJ estimator (6);

(BS) transform 50% of the candidate draws to a sequence of MH draws (plus a burn-in) and combine these with the other 50% of the candidate draws in the BS1 estimator (4) – with M substituted by the 'effective number of draws' $\tilde{M}$ for the BS2 estimator.

In Sections 4, 5 and 6 the methods will be applied to several target distributions. In the next section we briefly review the method of Hoogerheide et al. (2007) that uses an adaptive mixture of Student-t distributions (AdMit) as the importance or candidate distribution.

Table 1: Computations required by different marginal likelihood estimation approaches, in case we make use of IS or the independence chain MH algorithm. L is the number of candidate draws that are not used in the MH algorithm. M is the number of independence chain MH draws from the posterior. Warp1 and Warp2 refer to the Warp transformations of Meng and Schilling (2002) where one aims at a mixture of 2 or 2^d 'mirror images' of the posterior density that is typically more symmetric than the posterior itself. Further explanations are given in Section 2.

	number of posterior kernel evaluations	number of candidate draws	number of candidate density evaluations	number of auxiliary density evaluations
IS	L	L	L	-
RIS	M	M	M	M
BS	$L+M$	$L+M$	$L+M$	-
CJ	$L+M$	$L+M$	$L+M$	-
Warp1 IS	$2L$	L	L	-
Warp1 BS	$2(L+M)$	$L+M$	$L+M$	-
Warp2 IS	2^dL	L	L	-
Warp2 BS	$2^d(L+M)$	$L+M$	$L+M$	-

3 The Adaptive Mixture of Student-t method

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

of freedom, the risk is small that the tails of the candidate are thinner than those of the target distribution. Finally, Zeevi and Meir (1997) showed that under certain conditions any density function may be approximated to arbitrary accuracy by a convex combination of basis densities; the mixture of Student-t distributions falls within their framework.

The AdMit approach determines the number of mixture components, the mixing probabilities, the modes and scale matrices of the components in such a way that the mixture density approximates the target density $p(\theta \mid y)$ of which we only know a kernel function $k(\theta \mid y)$. The AdMit strategy consists of the following steps:

(0) Initialization: computation of the mode and scale matrix of the first component (typically the posterior mode and minus the inverse Hessian of the log-posterior evaluated at the mode), and drawing a sample from this Student-t distribution.

(1) Iterate on the number of components: add a new component that covers a part of the space of θ where the previous mixture density was relatively small, as compared to $k(\theta \mid y)$. The new component is based on the ratio of the target density kernel $k(\theta \mid y)$ and the previous mixture of Student-t densities. It is located where this ratio is *relatively* high, which does not depend on the normalizing constant of the target density.

(2) Optimization of the mixing probabilities: the mixing probabilities are chosen such that the coefficient of variation, i.e., the standard deviation divided by the mean, of the IS weights is minimized. This coefficient of variation does not depend on the normalizing constant of the target density.

(3) Drawing a sample from the new mixture.

(4) Evaluation of IS weights: if the coefficient of variation of the IS weights has converged, then stop. Otherwise, go to step (1).

For more details on the AdMit procedure we refer to Hoogerheide et al. (2007), Ardia et al. (2009a) and Ardia et al. (2009b). The package `AdMit` (Ardia et al. (2008)), an R implementation (R Development Core Team 2008), is available from the Comprehensive R Archive Network (CRAN) at `http://cran.r-project.org/package=AdMit`.

The AdMit approach has been successfully applied to the simulation of posterior draws from non-elliptical posterior distributions, where the reason for non-elliptical shapes is typically *local non-identification* of certain parameters. Examples are the IV model with weak instruments, or mixture models where one component has weight close to zero. This paper provides the first analysis of the AdMit method's performance in the case of marginal likelihood estimation.

4 Application 1: non-linear regression model

In this section we apply our methods in order to estimate the marginal likelihood in a non-linear regression model. We consider the biochemical oxygen demand (BOD) data from Marske (1967) that are analyzed by Bates and Watts (1988) and Ritter and Tanner (1992).

We consider the non-linear model of Bates and Watts (1988)

$$y_i = \theta_1(1 - \exp(-\theta_2\, x_i)) + \varepsilon_i\,, \tag{10}$$

with independent errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, where y_i is the BOD at time x_i $(i = 1, \ldots, 6)$.

Following Ritter and Tanner (1992), we specify a flat prior. However, we use a flat prior on a bounded domain: $(\theta_1, \theta_2, \sigma) \in [-20, 50] \times [-2, 6] \times [0, 20]$. Ritter and Tanner (1992) do not restrict the interval of σ; for the identification of a marginal likelihood we make this choice in order to have a proper prior. Obviously, the marginal likelihood will crucially depend on the prior specification. We consider the model and data from Ritter and Tanner (1992) in order to compare the efficiency of alternative estimation methods and illustrate the results in the case of a well-known, three-dimensional highly non-elliptical posterior distribution for a very small data set. In Section 6 we will consider a marginal likelihood and posterior distribution for a large data set.

The top-left panel of Figure 2 gives an illustration of the shapes of this posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$; it shows a Highest Posterior Density (HPD) credible set. Note the bimodality and the curved shapes of the larger mode. The sets $\{\theta : \theta_1 > 0, \theta_2 > 0\}$ and $\{\theta : \theta_1 < 0, \theta_2 < 0\}$ correspond to concave and convex increasing functions (through the origin) in (10), respectively. The smaller mode reflects the small posterior probability of a convex function.

For the IS and independence chain MH algorithms we consider three candidate distributions:

1. the mixture of Student-t distributions resulting from the AdMit procedure;
2. an 'adaptive' Student-t distribution where the mode and scale have been iteratively updated by several IS steps (starting with the posterior mode and iteratively using the estimated posterior mean and covariance as the mode and scale in the next iteration);
3. a so-called 'naive' Student-t distribution around the posterior mode.

In order to minimize the risk that the candidate 'misses' parts of the posterior, we specify very fat-tailed candidates: we choose one degree of freedom (i.e., Cauchy tails). Figure 2 shows the shapes of the three candidate distributions. Notice that the AdMit candidate nicely 'wraps' the relevant areas of the parameter space with candidate probability mass. Figure 3 illustrates how the AdMit approach has constructed this 'wrapping' distribution. Starting with the naive Student-t distribution around the mode, it finds that a Student-t distribution parallel with the θ_2 axis must be added, yielding a cross shape. After that, a

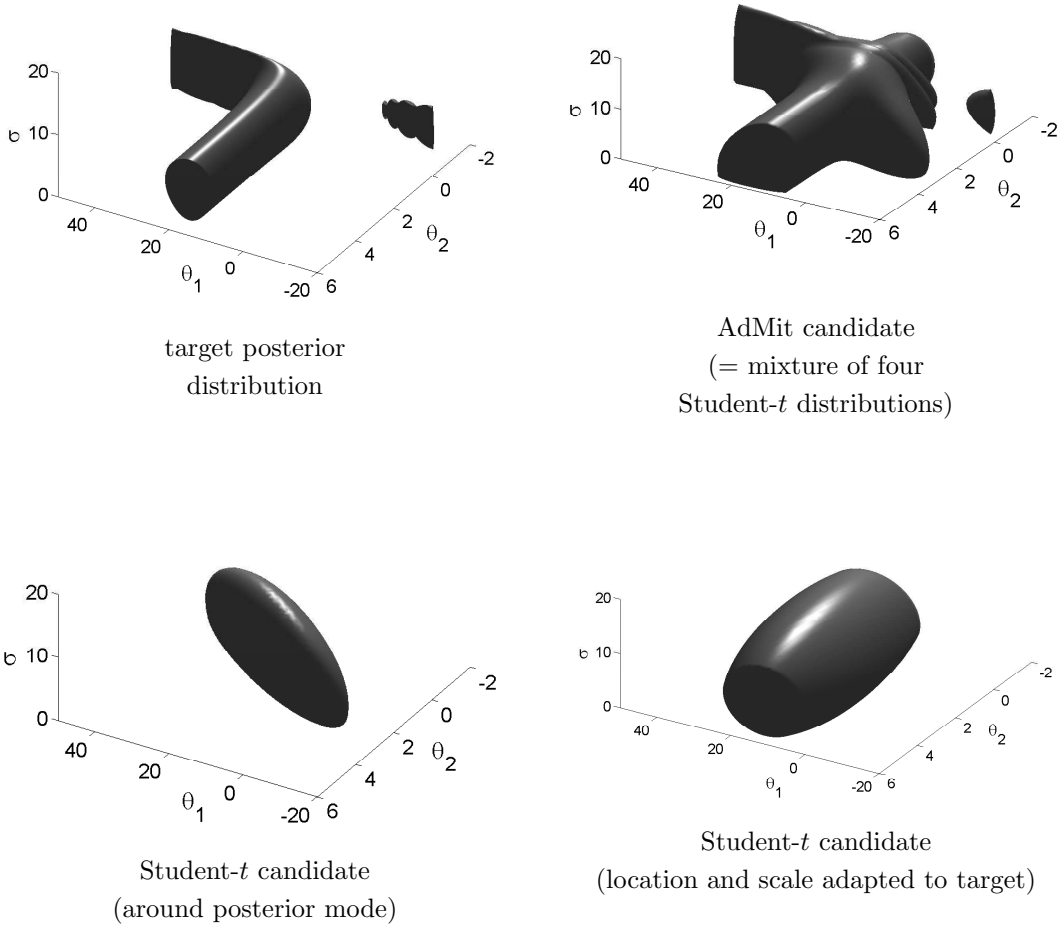


Figure 2: Non-linear regression model (10): Highest Posterior Density credible region of $\theta = (\theta_1, \theta_2, \sigma)'$ (top left) and 'Highest Candidate Density regions' for mixture of Student-t (AdMit, top right), 'naive' Student-t (bottom left) and adaptive Student-t (bottom right) candidate distributions.

third Student-t distribution parallel with the θ_1 axis is added, leading to a wrapping of the whole larger posterior mode. Finally, the fourth Student-t distribution in the mixture wraps the smaller posterior mode, so that the resulting mixture of four Student-t distributions covers the whole posterior distribution. This whole procedure took merely 11 seconds on a 2006 Intel (R) Centrino Duo Core processor.

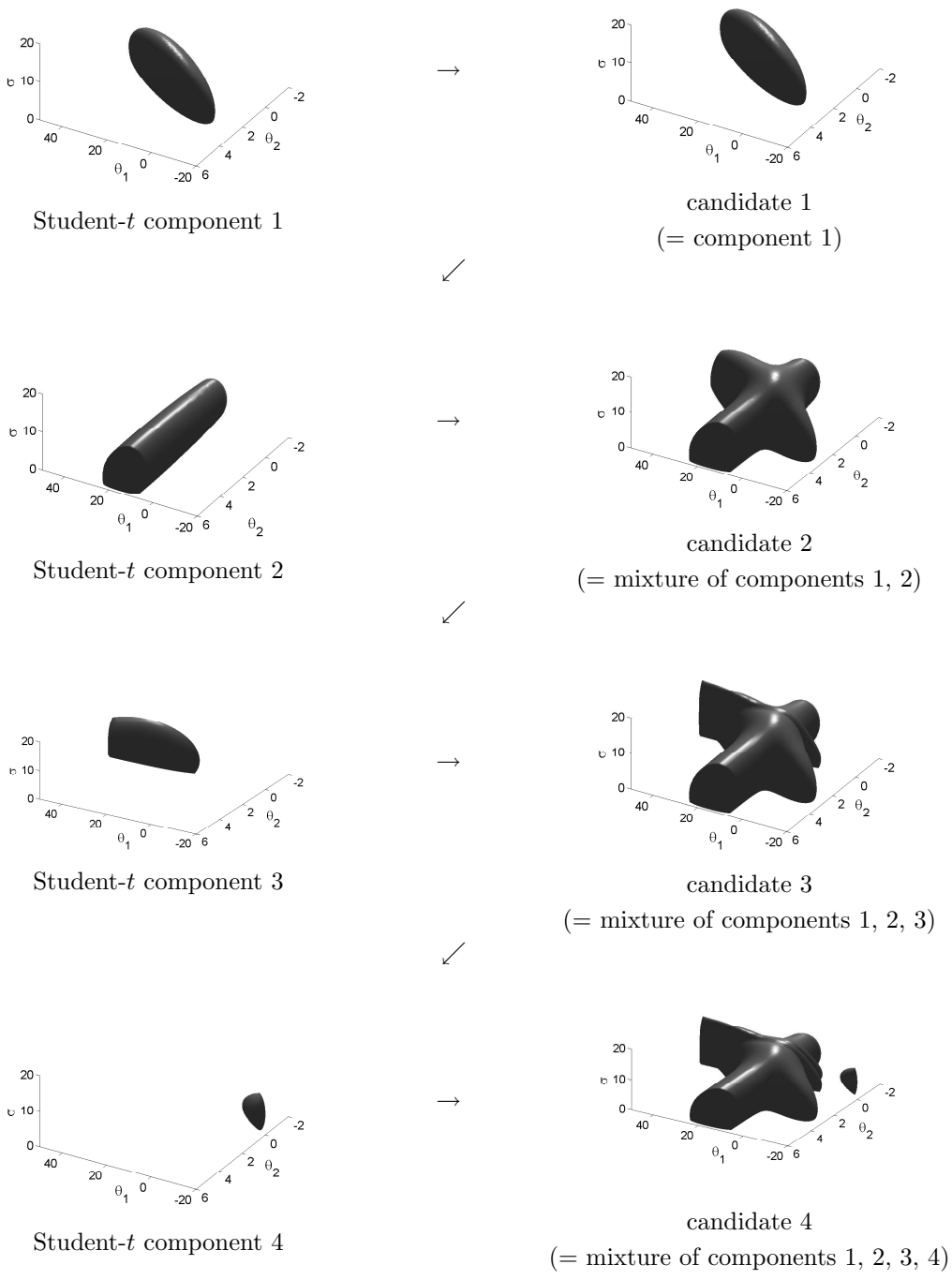


Figure 3: Non-linear regression model (10): The AdMit algorithm (automatically and) iteratively approximates the non-elliptical posterior shapes of $\theta = (\theta_1, \theta_2, \sigma)'$ by a mixture of Student-t distributions.

We will now use these three candidate distributions in combination with the marginal likelihood estimators of Section 2. For the IS estimator we generate $L = 100000$ candidate draws. For the RIS and CJ estimators we take $M = 100000$ independence chain MH draws; we use a burn-in of 1000 draws, so that we actually generate 101000 draws. The reason for not including the burn-in in the 100000 draws is that a burn-in of fewer than 1000 draws may suffice. For the BS estimators we use $L = 50000$ candidate draws and $M = 50000$ MH draws, again not counting a burn-in of 1000 draws.

For the RIS estimator we use a truncated Gaussian auxiliary density around the posterior mode. For the CJ estimator we choose θ^* as the posterior mode.

For each estimator, we repeat the simulation 500 times. Simulation results are reported in Table 2. Boxplots of the 500 marginal likelihood estimates are given in Figures 4. The real value of the marginal likelihood is (rounded to two digits) $12.79 \cdot 10^{-10}$. This real value is computed by deterministic integration which is still feasible (but quite time-consuming) in this three-dimensional example.

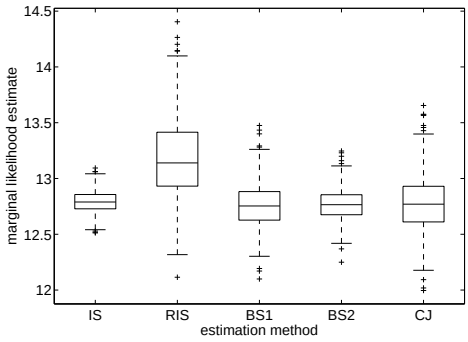
Table 2: Non-linear regression model (10): Estimation of the marginal likelihood (ML) based on 100000 draws from AdMit mixture of four Student-t distributions, adaptive Student-t or naive Student-t distribution. Mean and standard deviation of 500 estimates of $10^{10}\cdot$ ML from 500 simulation runs. True value is ML$= 12.79 \cdot 10^{-10}$.

$10^{10}\cdot$ ML	AdMit		adaptive		naive	
	mean	st.dev.	mean	st.dev.	mean	st.dev.
IS	12.7906	0.0962	12.7899	0.1791	12.7317	1.0945
RIS	13.1803	0.3435	12.8792	0.9456	12.8846	2.5144
BS1	12.7621	0.1984	12.8348	0.4238	13.0995	4.3776
BS2	12.7636	0.1405	12.7890	0.2739	13.0877	4.2780
CJ	12.7816	0.2568	12.7814	0.2841	13.1030	4.4004

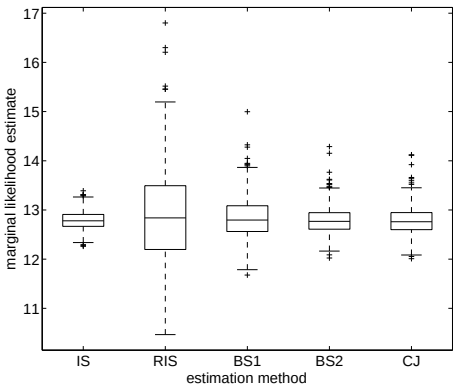
First, notice the very inefficient estimators that make use of the naive Student-t candidate distribution. Even though this naive Student-t distribution is chosen very fat-tailed (one degree of freedom), the resulting estimators have much higher variance than the estimators based on the AdMit and adaptive candidates. The boxplots show that the naive Student-t candidate may result in extreme outliers for all marginal likelihood estimators. This stresses the importance of wisely specifying an appropriate candidate distribution.

Second, the AdMit candidate clearly outperforms the adaptive Student-t candidate: iteratively adding Student-t distributions to the mixture candidate distribution leads to far more precise estimators than merely iteratively adapting the location and scale of the Student-t candidate.

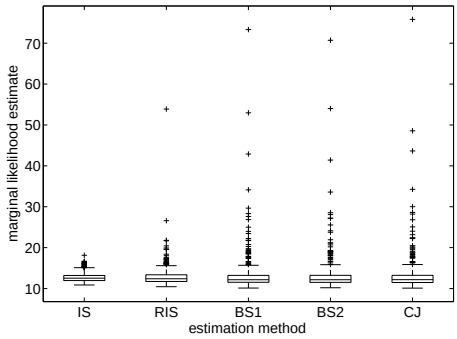
Third, the IS estimator is the best, whereas the RIS estimator is clearly the worst. The BS2, BS1 and CJ are typically ranked second to fourth, although in case of the adaptive candidate the CJ estimator outperforms the BS1 estimator. In that case, the difference between the 'i.i.d. optimal' BS1 estimator and the 'serial correlation corrected' BS2 estimator



candidate draws from AdMit distribution



candidate draws from adaptive Student-t distribution



candidate draws from naive Student-t distribution

Figure 4: Non-linear regression model (10): Estimates of $10^{10}\cdot$ marginal likelihood based on 100000 draws from AdMit mixture of four Student-t distributions, adaptive Student-t or naive Student-t distribution (500 simulation runs).

is substantial, reflecting the high serial correlation in the MH chain.

In this example, the winner is clearly the AdMit-IS estimator, the IS estimator based on the AdMit candidate. It outperforms the alternative estimators (including the BS estimators) that make use of the same number of candidate draws and function evaluations.

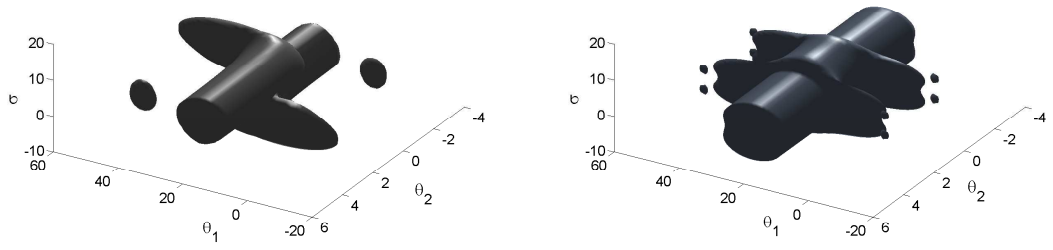
Simulating draws from a mixture of Student-t distributions takes hardly more time than generating draws from a Student-t distribution. The AdMit approach does require the evaluation of multiple Student-t densities, in our case four, instead of one; but the little extra computing time required for this is typically very small compared to the time required for evaluation of the posterior density kernel. Further, the 'victory' of the IS estimator over alternative estimators is actually slightly larger than represented by the tables: the burn-in of the MCMC draws is neglected and the implementation of the IS estimation of the marginal likelihood and its numerical standard error are relatively straightforward.

In this example, one comparison is still to be made: the comparison with methods aimed at the 'warped' target density. Figure 5 shows the shapes of the warped posterior kernels. These are more symmetric than the posterior kernel itself; especially the Warp2 distribution looks 'closer to' a Student-t distribution than the original posterior distribution. This illustrates the elimination of asymmetries by using mixtures of the posterior distribution. Table 3 shows the results of IS, BS1 and BS2 (the three best performing algorithms) for Warp1 and Warp2 transformations in combination with an adaptive Student-t candidate. The rows with 100000 posterior kernel evaluations correspond to IS with 50000 and 12500 draws (BS with 25000+25000 and 6250+6250 draws) for Warp1 and Warp2, respectively. The Warp1-IS results are comparable to the regular IS results with an adaptive Student-t candidate. The Warp1-BS estimators are somewhat better than the 'unwarped' BS estimators. The Warp2 results are worse than their 'unwarped' counterparts; the obvious reason is that the number of candidate draws is now much smaller in order to keep the number of posterior kernel evaluations equal to 100000.

Table 3: Non-linear regression model (10): Marginal likelihood estimation making use of Warp1 or Warp2 transformations in combination with an adaptive Student-t candidate distribution. Standard deviation of 500 estimates of $10^{10}\cdot$ ML from 500 simulation runs.

st.dev. $10^{10}\cdot$ ML	IS	BS1	BS2
Warp1 (100000 posterior kernel evaluations)	0.1750	0.3535	0.2250
Warp2 (100000 posterior kernel evaluations)	0.3097	0.5813	0.4054
Warp1 (100000 candidate draws)	0.1250	0.2575	0.1623
Warp2 (100000 candidate draws)	0.1182	0.2131	0.1522

Even if we use the same number of *candidate draws*, thereby requiring two or eight times more posterior kernel evaluations in the Warp1 and Warp2 approach, the resulting estimators do not outperform the AdMit-IS estimator. This confirms that the AdMit-IS estimator is clearly the winner. In this example, warping may provide a slight improvement, but



Warp1 (mixture of 2 posterior transformations)

Warp2 (mixture of $8 = 2^3$ posterior transformations)

Figure 5: Non-linear regression model (10): Warping of posterior density kernel. A mixture of the posterior density and its 'mirror images' (that naturally have the same normalizing constant) can have shapes that are much closer to an elliptical distribution than the original posterior.

appropriately *wrapping* the posterior yields a much larger gain in computational efficiency than *warping* it!

We now briefly pay attention to the implications that an unreliable marginal likelihood estimator may have. Suppose we face the choice between the non-linear regression model (10) and the linear regression model

$$y_i = \beta_1 + \beta_2 \, x_i + \varepsilon_i \,, \tag{11}$$

with independent errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The linear model ignores that for $x = 0$ we should have $y = 0$: the purpose of considering these two models is purely illustrative. Suppose we specify a conjugate prior that is approximately as 'non-informative' as the prior we used for the non-linear regression model (10), the Normal-Gamma prior

$$\beta \,|\sigma^{-2} \sim \mathcal{N}\left(\underline{\beta}, \, \sigma^2 \, \underline{V}\right) \qquad \sigma^{-2} \sim \mathcal{G}\left(\underline{s}^{-2}, \underline{\nu}\right) \,,$$

with

$$\underline{\beta} = \begin{pmatrix} 8 \\ 4 \end{pmatrix} \qquad \underline{V} = \frac{1}{100} \begin{pmatrix} 16 & 0 \\ 0 & 4 \end{pmatrix} \qquad \underline{s}^2 = 100 \qquad \underline{\nu} = 3 \,.$$

Under the Normal-Gamma prior the marginal likelihood can be analytically computed, see e.g., Koop (2003); here it equals $12.40 \cdot 10^{-10}$. The Bayes factor in favor of the non-linear model is 1.0315, so that under equal prior probabilities the posterior model probabilities for the non-linear and linear models are 0.5078 and 0.4922, respectively. Figure 4 shows that only for the AdMit-IS estimator all 500 repetitions of the simulation yield marginal likelihood estimates above $12.40 \cdot 10^{-10}$, leading (under equal prior probabilities) to a 'correct' model choice. Here we use the term 'correct' to denote that the model choice is optimal given our data and prior assumptions, and not determined by simulation 'noise'. For all other approaches, estimates smaller than $12.40 \cdot 10^{-10}$ are observed, resulting in an 'incorrect' model choice. Arguably,

in this situation one should consider Bayesian model averaging (BMA) rather than model choice. Under equal prior probabilities, appropriate model weights are 0.5078 and 0.4922. The extreme overestimation of the non-linear model's marginal likelihood that may occur for estimators using the naive candidate distribution, would result in highly 'incorrect' model weights. This simple example illustrates that an appropriate marginal likelihood estimator (using a suitable candidate distribution) is important, both for model selection and for model combination.

Until now we have considered the standard deviations of the estimators, when the simulation process is repeated 500 times. In practice, we usually do not compute standard deviations in such a time consuming way. Instead, we estimate the standard deviation by a numerical standard error based on a single simulation run. In the next section we consider the reliability of numerical standard errors.

5 Numerical standard errors

For the IS estimator, the computation of a numerical standard error (NSE) is particularly straightforward. One simply divides the standard deviation of the terms $k(\theta^{[l]} \mid y) \,/\, q(\theta^{[l]})$ $(l = 1, \ldots, L)$ by $\sqrt{L}$. However, for the RIS, BS1, BS2 and CJ estimators we make use of the usual *delta rule*. Moreover, the latter four estimators make use of correlated MCMC draws where we need to take into account serial correlation. In this section we will consider three methods for computing the standard error of a sample mean of such correlated series; that is an estimate of the standard deviation of

$$\hat{g} = \frac{1}{M} \sum_{m=1}^{M} g(\theta^{[m]}) \,, \tag{12}$$

where $\{\theta^{[m]}\}_{m=1}^{M}$ is a series of MCMC draws.

The first estimate of the variance $\mathrm{var}(\hat{g})$ that we consider, is the estimate of Newey and West (1987)

$$\widehat{\mathrm{var}}_{\mathrm{NW}}(\hat{g}) = \frac{1}{M} \left[\hat{\gamma}_0 + 2 \sum_{i=1}^{b} \left(1 - \frac{i}{b+1} \right) \hat{\gamma}_i \right] \,, \tag{13}$$

where b is a constant that should represent the lag at which the autocorrelation tapers off, $\hat{\gamma}_0$ is the sample variance of the series $\{g(\theta^{[m]})\}_{m=1}^{M}$, and $\hat{\gamma}_i$ is its i-th order sample autocovariance. This Newey-West (NW) estimate is used by Chib (1995) and Chib and Jeliazkov (2001), who set b equal to 10 and 40, respectively. We choose a bandwidth of $b = 40$.

The second and third estimate we consider are from Geyer (1992): the initial positive sequence estimator and the initial monotone sequence estimator. These are specialized for reversible Markov chains such as the series of MH draws. Theorem 3.1 of Geyer (1992) states the following. For a stationary, irreducible, reversible Markov chain with autocovariance γ_i

let $\Gamma_t = \gamma_{2t} + \gamma_{2t+1}$ be the sums of adjacent pairs of autocovariances. Then Γ_t is a strictly positive, strictly decreasing, strictly convex function of t.

The initial positive sequence estimator (IPSE) estimator is now given by

$$\widehat{\text{var}}_{\text{IPSE}}(\hat{g}) = \frac{1}{M}\left[\hat{\gamma}_0 + 2\sum_{t=0}^{2h+1}\hat{\gamma}_t\right] = \frac{1}{M}\left[-\hat{\gamma}_0 + 2\sum_{t=0}^{h}\hat{\Gamma}_t\right], \tag{14}$$

where $\hat{\Gamma}_t = \hat{\gamma}_{2t} + \hat{\gamma}_{2t+1}$ and where h is chosen to be the largest integer such that $\hat{\Gamma}_t > 0$ for $t = 1, \ldots, h$.

In the initial monotone sequence estimator (IMSE) the value of h is chosen to be the largest integer such that $\hat{\Gamma}_{t-1} > \hat{\Gamma}_t$ *and* such that $\hat{\Gamma}_t > 0$ for $t = 1, \ldots, h$. Therefore, the resulting estimates satisfy: $\widehat{\text{var}}_{\text{IMSE}}(\hat{g}) \leq \widehat{\text{var}}_{\text{IPSE}}(\hat{g})$. For derivations of NSE's for normalizing constants we refer to Chen, Shao, and Ibrahim (2000).

We now inspect the NSE in the example from the previous section. Figure 6 shows boxplots, comparing the numerical standard errors to the standard deviations for the three candidate distributions. For the naive Student-t candidate distribution the NSE is often unreliable: huge underestimation of the uncertainty in the marginal likelihood estimator often occurs. For the adaptive Student-t candidate distribution the NSE is more reliable than in the naive case. However, for all estimators a substantial underestimation of the uncertainty may still occur. The NSE based on the IPSE should be preferred over the NSE from the IMSE and NW formula. For the AdMit candidate distribution the NSE is more reliable than for the other candidates. Especially for the AdMit-IS estimator, the 'winner' of Section 4, the NSE seems reliable. For the BS1, BS2 and CJ estimators, the NSE from the IPSE should again be preferred over the NSE from the IMSE or NW approach. Only for the RIS estimator, which anyway performs poorly in this example, the IMSE provides a NSE that yields a huge overestimation of the uncertainty.

Another way of assessing the performance of the numerical standard errors is to inspect the coverage rate of estimated 90% intervals

$$\left(\hat{p}(y) - 1.645 \cdot \text{NSE}_{\hat{p}(y)},\ \hat{p}(y) + 1.645 \cdot \text{NSE}_{\hat{p}(y)}\right).$$

Table 4 gives these coverage rates. In (approximately) 90% of the simulations, the interval should include the true value $p(y)$, whereas the situations with too low or too high intervals should both occur in (about) 5% of the simulations. For the naive candidate distribution, significant deviations from the correct rates can be found for the intervals of all estimators. For the adaptive Student-t candidate, the coverage rates are incorrect for all but the IS estimator. This confirms the unreliable character of the NSE for the naive or adaptive candidate distributions. For the AdMit-IS estimator the coverage rates are correct, whereas for the BS1, BS2 and CJ estimators using AdMit draws only the IPSE and IMSE provide (approximately) correct rates.

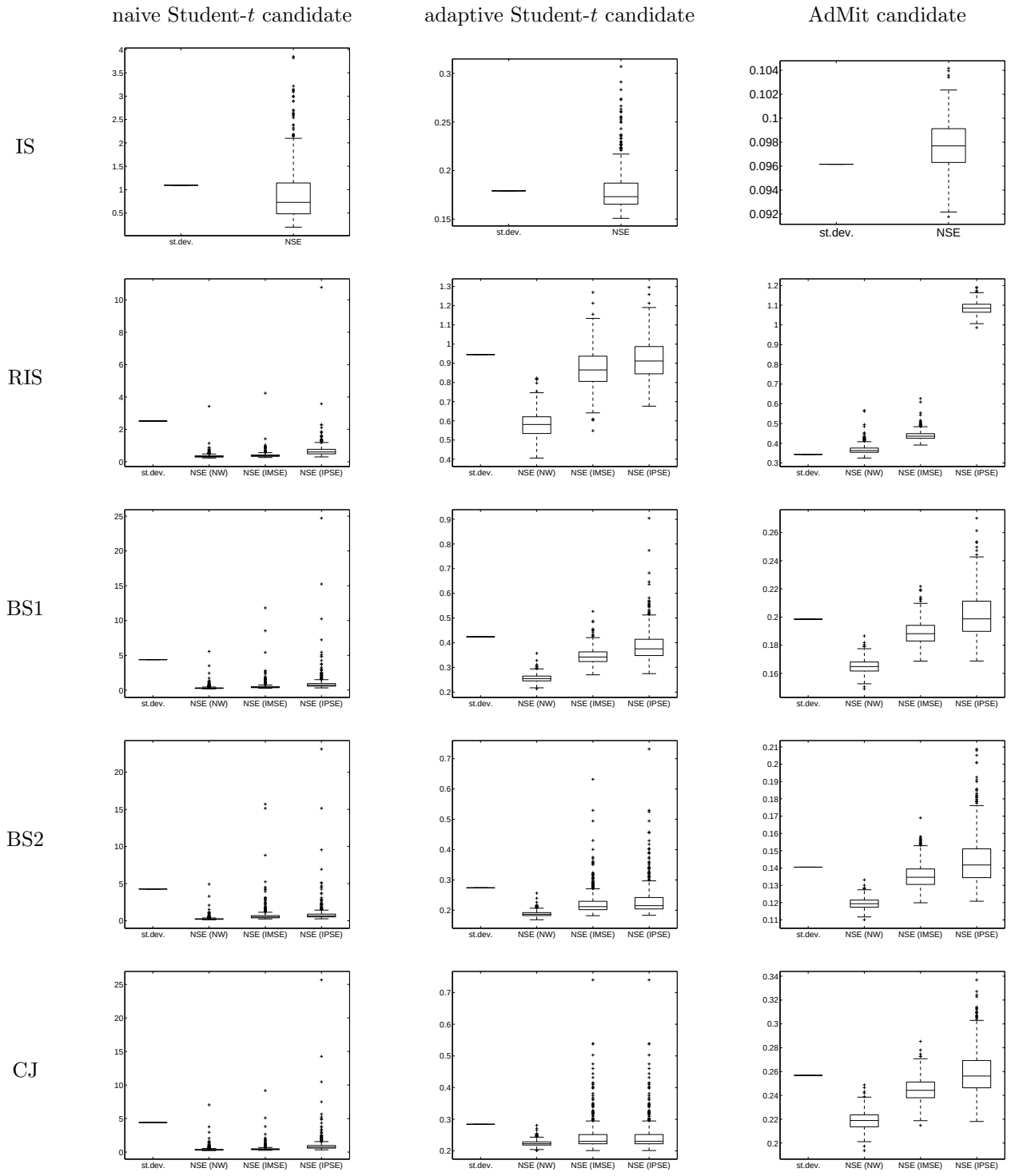


Figure 6: Non-linear regression model (10): Boxplots of 500 numerical standard errors (NSE's) for estimates of $10^{10}\cdot$ marginal likelihood based on 100000 candidate draws from the *'naive' Student-t, adaptive Student-t and AdMit (mixture of four Student-t) candidate distribution.* The standard deviation of the 500 marginal likelihood estimates is shown as the horizontal line in the first column. NSE's are computed using the delta rule, where NW, IMSE, IPSE refer to the approach of Newey and West (1987), the initial monotone sequence estimator and the initial positive sequence estimator (Geyer (1992)) for taking into account the serial correlation in the MH draws.

Table 4: Non-linear regression model (10): Coverage rate of estimated 90% interval for $p(y)$ based on different NSE's (in 500 repetitions). For the IS estimators there is no serial correlation in the series of draws, so that only one (straightforward) NSE formula is used.

	90% interval from Newey-West NSE			90% interval from IMSE NSE			90% interval from IPSE NSE		
	too low	ok	too high	too low	ok	too high	too low	ok	too high
AdMit candidate									
IS	0.056	0.902	0.042	0.056	0.902	0.042	0.056	0.902	0.042
RIS	0.002	0.730	0.268	0.002	0.836	0.162	0.000	1.000	0.000
BS1	0.106	0.824	0.070	0.068	0.886	0.046	0.052	0.912	0.036
BS2	0.102	0.844	0.054	0.082	0.884	0.034	0.072	0.900	0.028
CJ	0.092	0.834	0.074	0.058	0.880	0.062	0.038	0.908	0.054
adaptive Student-t candidate									
IS	0.052	0.902	0.046	0.052	0.902	0.046	0.052	0.902	0.046
RIS	0.440	0.312	0.248	0.412	0.360	0.228	0.338	0.532	0.130
BS1	0.128	0.728	0.144	0.080	0.846	0.074	0.068	0.872	0.060
BS2	0.118	0.772	0.110	0.082	0.864	0.054	0.080	0.874	0.046
CJ	0.092	0.834	0.074	0.086	0.866	0.048	0.086	0.866	0.048
naive Student-t candidate									
IS	0.258	0.740	0.002	0.258	0.740	0.002	0.258	0.740	0.002
RIS	0.440	0.312	0.248	0.412	0.360	0.228	0.338	0.532	0.130
BS1	0.548	0.220	0.232	0.490	0.316	0.194	0.354	0.546	0.100
BS2	0.578	0.172	0.250	0.450	0.416	0.134	0.368	0.536	0.096
CJ	0.518	0.266	0.216	0.484	0.314	0.202	0.342	0.564	0.094

We conclude that also in terms of the reliability of the NSE and confidence intervals the AdMit-IS approach performs best. For other AdMit estimators (BS1, BS2 and CJ) the initial monotone sequence estimator of Geyer (1992) provides a useful NSE. For the adaptive (and naive) candidate we find that all three types of NSEs may be (highly) unreliable. The reason for the failure of the NSE based on the Newey-West formula is partly that the 'bandwidth' $b = 40$ is simply a too small value. Still, also the IPSE and IMSE that automatically adapt the 'bandwidth' to the autocorrelation in the given series of MCMC draws (slightly) fail in case of the naive (and adaptive) candidate distribution. Therefore, the fixed value of $b = 40$ is arguably not always the only reason for its failure.

6 Application 2: mixture GARCH model

In this section we apply our methods in order to estimate the marginal likelihood in a two-component Gaussian mixture GARCH(1,1) model, a model with six parameters. Ausín and Galeano (2007) propose a Griddy-Gibbs sampler for Bayesian estimation of this model, and note that MH algorithms could improve the efficiency of this method despite the drawback of the effort required to calibrate the candidate distribution in the latter case. We provide an additional estimation method for the model and show that given an appropriately tuned candidate density, straightforward IS provides an efficient method for parameter estimation. We extend the study of Ausín and Galeano (2007) by providing an efficient estimation method for the marginal likelihood. The data consist of 1859 daily closing prices of the SMI, for the period 1/Jul/1991 - 14/Aug/1998. Daily nominal log-returns are expressed in percentages.

A two-component Gaussian mixture GARCH$(1,1)$ model for the series y_t is defined by

$$\begin{aligned} y_t &= \mu + \sqrt{h_t}\,\varepsilon_t, \\ h_t &= \omega + \alpha\,(y_{t-1} - \mu)^2 + \beta h_{t-1}, \\ \varepsilon_t &\sim \begin{cases} N(0, \sigma^2) & \text{with probability } \rho, \\ N(0, \sigma^2/\lambda) & \text{with probability } 1-\rho, \end{cases} \end{aligned} \tag{15}$$

where h_t is the conditional variance of y_t given previous information $I_{t-1} = \{y_{t-1}, y_{t-2}, \dots\}$, $0 < \lambda < 1$ and $\sigma^2 = 1/\left(\rho + (1-\rho)/\lambda\right)$, so that $\mathrm{var}\,(\varepsilon_t) = 1$. Similar to Ausín and Galeano (2007), we assume that the initial variance h_0 is a known constant, $\varepsilon_t \sim$ mixture Gaussian (λ, ρ), and the following parameter restrictions hold: $\omega > 0$, $\alpha \geq 0$, $0.5 \leq \rho < 1$, $\beta \geq 0$ and $\alpha + \beta < 1$. Notice that these parameter restrictions ensure positivity of h_t, and that there is a higher prior probability that an observation falls into the state with smaller variance. Following Ausín and Galeano (2007), we specify a flat prior. However, we truncate the domain for μ and ω to finite (wide) intervals to have a proper (non-informative) prior: $(\rho, \lambda, \mu, \alpha, \beta, \omega) \in [0.5, 1] \times [0, 1] \times [-1, 1] \times [0, 1] \times [0, 1] \times (0, 1]$ with $\alpha + \beta < 1$.

For the IS and independence chain MH algorithms, we consider three candidate distributions based on Student-t densities with Cauchy tails: the mixture of Student-t distributions

resulting from the AdMit procedure, an 'adaptive' Student-t distribution where the mode and scale have been iteratively updated by several IS steps and a 'naive' Student-t distribution around the posterior mode.

Figure 7 shows the shapes of the three candidate distributions, together with the conditional posterior density of (λ, ρ); parameters $(\omega, \beta, \alpha, \mu)$ are fixed at their posterior mean values. Figure 7 illustrates that the AdMit candidate outperforms both adaptive and naive Student-t candidates: the relevant subdomain of the posterior density is wrapped by the AdMit candidate. In particular, the naive Student-t candidate is quite inadequate for wrapping the posterior.

In order to illustrate the local non-identification in the model and the corresponding irregularity in the posterior density, we consider the posterior density of $(1/\lambda, \rho)$. The posterior density for $(1/\lambda, \rho)$ is shown in Figure 8, where the other parameters are fixed at posterior means.

For $\rho \to 1$, $1/\lambda$ becomes unidentified since the corresponding large variance regime disappears from the model. For $\lambda \to 1$, the conditional variances in both states are identical, hence the mixing probability ρ cannot be identified. This explains why the shapes of the posterior density are far from elliptical, and wisely choosing a candidate that can approximate this non-elliptical shape can provide considerable efficiency gains.

We will now use these three candidate distributions, using 100000 draws for IS and independence chain MH, where we use a burn-in of 1000 draws for the latter. Parameter estimates and NSE's for all cases, together with Ausín and Galeano (2007) estimates are reported in Table 5. Notice that Ausín and Galeano (2007) consider log-returns instead of log-returns in percentages, hence the parameter estimates for ω and μ are scaled differently in our estimation. Estimates under the adaptive and AdMit approaches are similar to Ausín and Galeano (2007). Further, we find two main results. First, the naive Student-t density has the worst performance among the candidate densities we compare. Both IS and MH estimates under the naive Student-t candidate are biased for 100000 draws. This shows that both IS and independence chain MH fail to provide accurate results using a poor candidate density. Hence, in the rest of our analysis, we compare the performances of only the adaptive Student-t and AdMit candidates. Second, the AdMit candidate clearly outperforms the adaptive Student-t candidate: NSE's obtained using the AdMit candidate are much smaller than those obtained using the adaptive Student-t candidate. According to NSE's, AdMit-IS has the best performance.

The next step is to estimate the marginal likelihood for the mixture GARCH model. We analyze the sensitivity of the marginal likelihood estimators in Section 2 to the choices of the candidate distribution. For marginal likelihood estimation, we consider the IS, BS1, BS2 and CJ estimators. We do not report results for RIS since this estimator already gave particularly bad results in Section 4. Table 6 reports marginal likelihood estimates and the respective NSE's. For the IS, BS1 and CJ estimators, NSE's are calculated by IPSE, as it was shown to be the most accurate estimator in Section 5. First, the AdMit candidate yields more precise

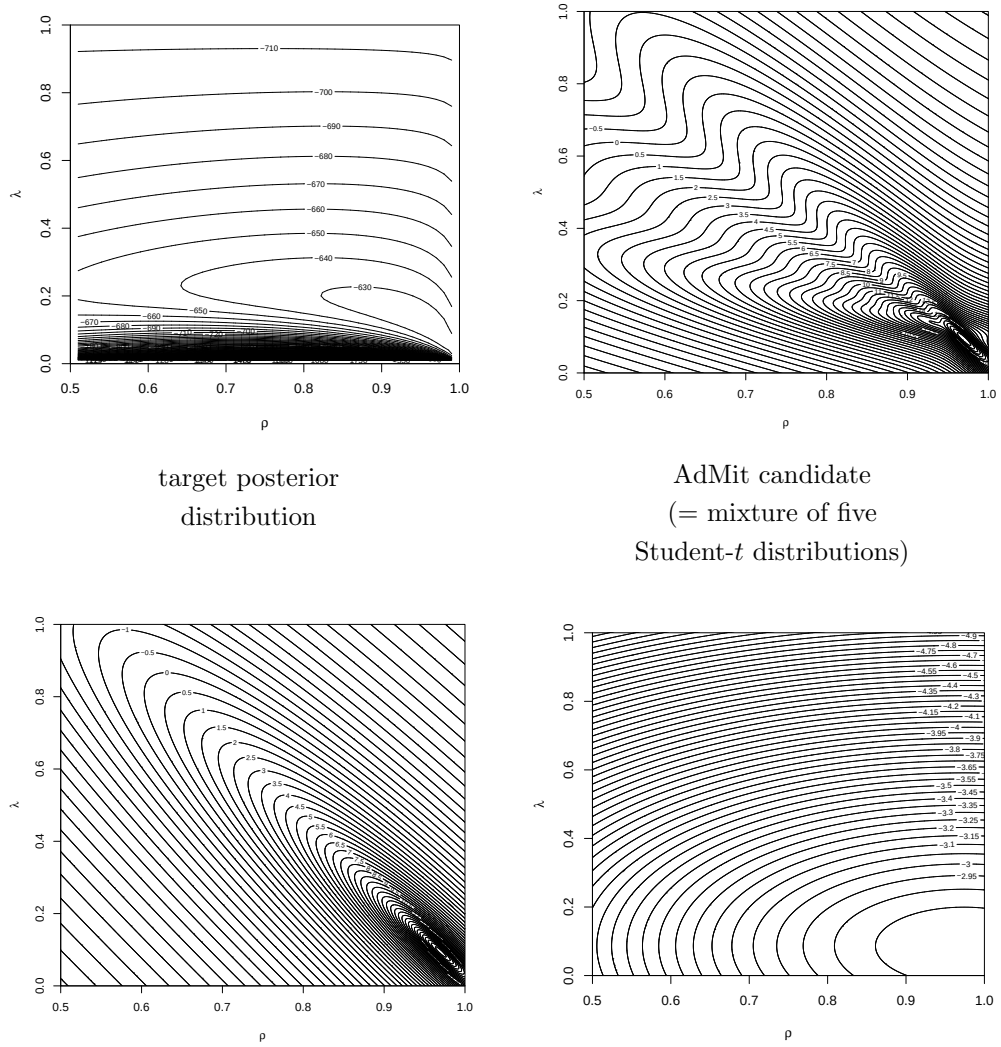


Figure 7: Mixture GARCH(1,1) model (15): Contour plots for (the logarithm of) the conditional posterior density of (ρ, λ) given $(\mu, \omega, \alpha, \beta)$ equal to the posterior mean. Conditional posterior density (top left) and candidate density contours for mixture of Student-t (AdMit, top right), adaptive Student-t (bottom left), 'naive' Student-t around the posterior mode (bottom right).

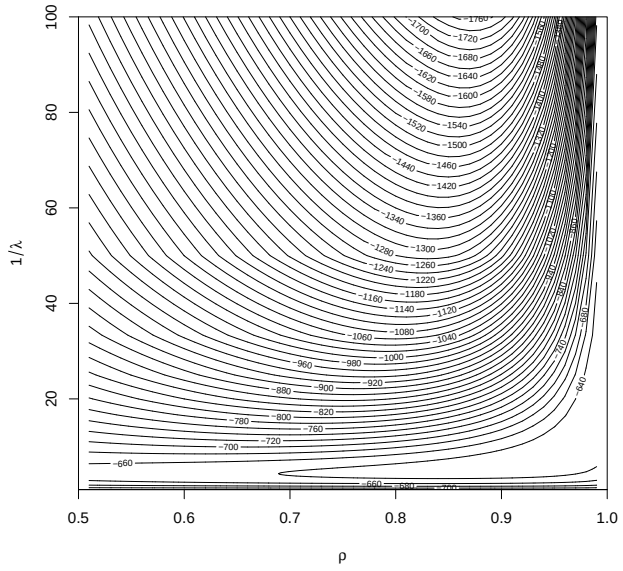


Figure 8: Mixture GARCH(1,1) model (15): Contour plot for (the logarithm of) the conditional posterior density of $(\rho, 1/\lambda)$ given $(\mu, \omega, \alpha, \beta)$ equal to the posterior mean.

Table 5: Mixture GARCH(1,1) model (15): Estimated posterior means and corresponding NSE's, obtained by IS and independence chain MH algorithms under different candidate densities. NSE's for MH algorithm are calculated by Newey-West method, bandwidth 40.

	IS estimates						
	Ausín & Galeano (2007)	AdMit		adaptive		naive	
	mean	mean	NSE ·100	mean	NSE ·100	mean	NSE ·100
ω	$1.130 \cdot 10^{-3}$	0.08	0.02	0.07	0.08	0.33	2.20
λ	0.13	0.12	0.03	0.12	0.15	0.27	2.20
β	0.74	0.80	0.03	0.80	0.12	0.45	2.05
α	0.15	0.13	0.02	0.13	0.06	0.20	1.68
ρ	0.92	0.95	0.02	0.95	0.16	0.56	0.13
μ	1.13	0.11	0.01	0.11	0.03	0.08	0.45
	independence chain MH estimates						
	Ausín & Galeano (2007)	AdMit		adaptive		naive	
	mean	mean	NSE ·100	mean	NSE ·100	mean	NSE ·100
ω	$1.130 \cdot 10^{-3}$	0.08	8.00	0.07	10.24	0.15	32.34
λ	0.16	0.12	13.29	0.12	15.81	0.14	22.73
β	0.74	0.80	14.37	0.80	18.26	0.72	50.74
α	0.15	0.13	8.81	0.13	10.50	0.13	26.27
ρ	0.92	0.95	10.53	0.95	12.52	0.90	31.86
μ	1.13	0.11	5.59	0.11	6.37	0.12	12.96

estimates than the adaptive Student-t candidate. This points out the importance of wisely specifying an appropriate candidate both for IS and independence chain MH algorithms. In particular for the IS estimator, this gain in efficiency is quite significant: NSE's for IS estimates decrease at least 50% when the posterior density is wrapped nicely by the AdMit candidate. Second, the smallest NSE is achieved by the IS estimator using AdMit Student-t candidate. Given a suitable candidate, IS performs better than the independence chain MH algorithm.

Table 6: Mixture GARCH(1,1) model (15): NSE's from IS and independence chain MH based on adaptive Student-t and AdMit candidates in combination with the Warp1 method. Warp1* refers to 10^5 posterior kernel evaluations (i.e., $0.5 \cdot 10^5$ candidate draws), whereas Warp1† refers to 10^5 candidate draws (i.e., $2 \cdot 10^5$ posterior kernel evaluations).

adaptive					
		IS	BS1	BS2	CJ
$10^{280}\cdot$ estimate	'unwarped'	3.96	3.95	3.90	3.96
$10^{282}\cdot$ NSE	'unwarped'	11.12	11.88	12.07	10.95
	Warp1*	4.97	12.19	12.28	14.11
	Warp1†	4.20	8.77	8.52	10.11
AdMit					
		IS	BS1	BS2	CJ
$10^{280}\cdot$ estimate	'unwarped'	4.06	4.04	4.02	4.06
$10^{282}\cdot$ NSE	'unwarped'	2.95	4.49	3.70	5.49
	Warp1*	2.48	7.03	5.90	8.62
	Warp1†	2.46	4.92	4.13	6.05

We now consider marginal likelihood estimates resulting from aiming at a 'warped' version of the posterior kernel using the Warp1 method. We do not consider the Warp2 method because of its computational cost: for this six-dimensional posterior each draw would require $2^6 = 64$ posterior kernel evaluations. For the adaptive Student-t candidate, the NSE's are reported in the top panel of Table 6. For BS and CJ estimators, NSE's are calculated by IPSE. For the IS estimators, straightforward NSE calculation is used, as there is no serial correlation between the draws. Considering BS1, BS2 and CJ estimators, warping the posterior density leads to smaller NSE's only when the number of candidate draws are constant. Hence this gain in efficiency is related both to warping the posterior and the increased number of posterior kernel evaluations. For the IS estimator on the other hand, Warp1 transformation does provide efficiency gains, even when the number of kernel evaluations is the same as for the 'unwarped' counterpart. Notice that IS estimator provides the smallest NSE's compared to BS1, BS2 and CJ methods. Finally, IS estimates using the adaptive Student-t density in combination with the Warp1 method are less precise than the 'unwarped' IS estimate using the AdMit candidate, which is shown in the bottom panel of Table 6. Hence the gain from

'warping' the posterior is smaller than that of 'wrapping' the posterior.

Until now, we have considered 'warping' and 'wrapping' the posterior kernel separately. A natural extension would be to combine these methods. Hence we analyze the changes in NSE's when the posterior kernel is both 'warped' and 'wrapped'. The bottom panel in Table 6 shows the NSE's estimated by IS, BS1, BS2 and CJ algorithms making use of the AdMit candidate and the Warp1 method, together with the 'unwarped' counterparts. For the BS1, BS2 and CJ, reported NSE's are achieved by IPSE. Warp1 transformation does not lead to efficiency gains in BS1, BS2 and CJ estimators according to IPSE. The Warp1 transformation leads to a small decrease in NSE for the IS algorithm, given the same number of posterior kernel evaluations. Therefore we conclude that 'warping' and 'wrapping' the posterior at the same time increases the efficiency of the IS algorithm, but most of this efficiency gain stems from constructing an appropriate candidate density.

We make a final comparison for the NSE's achieved by the independence chain MH algorithms according to Newey-West estimator, IPSE and IMSE of Section 5. Table 7 reports NW, IPSE and IMSE standard errors for the independence chain MH sampler under adaptive and AdMit candidates, without or with a Warp1 transformation (10^5 candidate draws). For the NW estimator, we choose a bandwidth of 40. NSE's using all estimators are still larger than those of IS using the AdMit candidate. Hence the victory of AdMit-IS is not related to the choice of the NSE estimators. Furthermore, NW estimates are quite different from IMSE and IPSE values. Notice that this result is in line with Section 5 where we show that the NW estimator is less reliable than the IPSE and IMSE.

Table 7: Mixture GARCH(1,1) model (15): numerical standard errors based on the Newey-West, IPSE and IMSE methods for independence chain MH sampler under adaptive and AdMit candidates, without or with Warp1 transformation (10^5 candidate draws).

adaptive				adaptive, Warp1 combination			
$10^{282}\cdot$ NSE				$10^{282}\cdot$ NSE			
	BS1	BS2	CJ		BS1	BS2	CJ
NW	7.20	6.07	7.30	NW	4.39	3.65	4.23
IPSE	11.88	12.07	10.95	IPSE	8.77	8.52	10.11
IMSE	11.88	12.07	10.95	IMSE	8.77	8.52	10.11
AdMit				AdMit, Warp1 combination			
$10^{282}\cdot$ NSE				$10^{282}\cdot$ NSE			
	BS1	BS2	CJ		BS1	BS2	CJ
NW	5.08	3.83	5.43	NW	4.13	3.88	4.23
IPSE	4.49	3.70	5.49	IPSE	4.92	4.13	6.05
IMSE	4.32	3.70	5.49	IMSE	3.65	4.13	4.38

7 Concluding remarks

We have considered two very different model structures (for data sets with different sample sizes), a non-linear regression model (for a very small data set) and a mixture GARCH model (for a large data set), with clearly different non-elliptical posterior shapes. Still, we obtain roughly the same findings. Given a suitable candidate distribution, which can be obtained by the AdMit method, the IS algorithm delivers a computationally efficient marginal likelihood estimator (and a reliable, easily computed numerical standard error), which outperforms the RIS, BS1, BS2 and CJ estimators. Warping the posterior density can lead to a further gain in efficiency, but it is more important that the posterior kernel is appropriately wrapped by the (AdMit) candidate distribution than that is warped. Moreover, warping requires evaluations of the warped posterior density kernel which are only used for marginal likelihood estimation. For the straightforward IS estimator of the marginal likelihood only computations are required that are typically already performed for parameter estimation or forecasting; usually no *extra* computations are required for marginal likelihood estimation.

If one uses a marginal likelihood estimator on the basis of serially correlated MCMC draws, the IPSE of Geyer (1992) performs best among the considered methods for computing numerical standard errors.

In further research, we intend to consider different empirical applications. We will further compare the performance of different types of bridge sampling estimators with the approach of Chib (1995) in cases of non-elliptical posteriors where the Gibbs sampler is applicable. We will also consider the quality of the estimators when these are applied in combination with the radial-based transformation of Bauwens et al. (2004). Another possibility is to consider the path sampling method of Gelman and Meng (1998), which extends the bridge sampling approach.

Acknowledgements

The authors thank Siddhartha Chib, participants of seminars, the editor and two anonymous reviewers for several useful comments on an earlier version, which have led to a substantial revision. We also thank Maria Concepcion Ausín Olivera for providing the SMI data used in Section 6. We emphasize that only the authors remain responsible for any errors. The first author is grateful to the Swiss National Science Foundation (under grant #FN PB FR1-121441) for financial support.

References

[1] Ardia D, Hoogerheide LF and Van Dijk HK 2009 To Bridge, to Warp or to Wrap? A Comparative Study of Monte Carlo Methods for Efficient Evaluation of Marginal Likelihoods. Tinbergen Institute discussion paper TI 2009-17/4.

[2] Ardia D, Hoogerheide LF and Van Dijk HK 2008 AdMit: Adaptive mixture of Student-t distribution in R.
URL `http://CRAN.R-project.org/package=AdMit`.

[3] Ardia D, Hoogerheide LF and Van Dijk HK 2009 Adaptive mixture of Student-t distributions as a flexible candidate distribution for efficient simulation: The R package **AdMit**. *Journal of Statistical Software* **29**(3), 1-32,
URL `http://www.jstatsoft.org/v29/i03`.

[4] Ardia D, Hoogerheide LF and Van Dijk HK 2009 AdMit: Adaptive mixture of Student-t distributions. *The R Journal* **1**(1), 25-30.

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

[6] Bates DM and Watts DG 1988 *Nonlinear Regresion Analysis and Its Applications*. Wiley, New York.

[7] Bauwens L, Bos CS, Van Dijk HK and Van Oest RD 2004 Adaptive radial-based direction sampling: Some flexible and robust Monte Carlo integration methods. *Journal of Econometrics* **123**, 201–225.

[8] Chen M-H, Shao Q-M and Ibrahim JG 2000 Monte Carlo Methods in Bayesian Computation. Springer Series in Statistics. New York: Springer.

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

[10] Chib S and Jeliazkov I 2001 Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **96**(453), 270–281.

[11] Frühwirth-Schnatter S 2001 Markov chain Monte Carlo estimation of classical and dynamic switching models. *Journal of the American Statistical Association* **96**, 194–209.

[12] Frühwirth-Schnatter S 2004 Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques. *Econometrics Journal* **7**(1), 143–167.

[13] Gelfand AE and Dey DK 1994 Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society B* **56**, 501–514.

[14] Gelman A and Meng X-L 1991 A note on bivariate distributions that are conditionally normal. *The American Statistician* **45**(2), 125–126.

[15] Gelman A and Meng X-L 1998 Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13**(2), 163–185.

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

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

[18] Geweke J 1999 Using simulation methods for Bayesian econometric models: Inference, development, and communication. *Econometric Reviews* **18**, 1–73.

[19] Geyer CJ 1992 Practical Markov chain Monte Carlo. *Statistical Science* **7**(4), 473-511.

[20] Hammersley JM and Handscomb DC 1964 *Monte Carlo Methods*. Methuen, London.

[21] Han C and Carlin BP 2001 Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association* **96**, 1122–1132.

[22] Hastings WK 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

[23] Hoogerheide LF, Kaashoek JF and Van Dijk HK 2007 On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: An application of flexible sampling methods using neural networks. *Journal of Econometrics* **139**(1), 154–180.

[24] Hop JP and Van Dijk HK 1992 SISAM and MIXIN: Two algorithms for the evaluation of posterior moments and densities using Monte Carlo integration. *Computer Science in Economics and Management*, (now *Computational Economics*) **5**, 183–220; reprinted in *Bulletin of the International Statistical Institute*, Cairo, vol. LIV, book 3, 29 pages.

[25] Kass RE and Raftery AE 1995 Bayes factors. *Journal of the American Statistical Association* **90**(430), 773–795.

[26] Kloek T and Van Dijk HK 1978 Bayesian estimates of equation system parameters: An application of integration by Monte Carlo. *Econometrica* **46**, 1–20.

[27] Koop G 2003 *Bayesian Econometrics*. Wiley.

[28] Marske 1967 Biomedical Oxygen Demand Data Interpretation Using Sums of Squares Surface, unpublished master's thesis, University of Wisconsin.

[29] Meng X-L and Schilling S 2002 Warp bridge sampling. *Journal of Computational & Graphical Statistics* **11**(3), 552–586.

[30] Meng X-L and Wong WH 1996 Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* 6, 831–860.

[31] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E 1953 Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* **21**, 1087–1092.

[32] Mira A and Nicholls G 2004 Bridge estimation of the probability density at a point. *Statistica Sinica* 14, 603-612.

[33] Miazhynskaia T and Dorffner G 2006 A comparison of Bayesian model selection based on MCMC with an application to GARCH-type models. *Statistical Papers* **47**, 525-549.

[34] Newey WK and West KD 1987 A simple positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* **55**, 703-708.

[35] Newton MA and Raftery AE 1994 Approximate Bayesian inference by the weighted likelihood bootstrap, *Journal of the Royal Statistical Society B* **56**, 3–48.

[36] R Development Core Team 2008 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria,
URL `http://www.R-project.org`.

[37] Ritter C and Tanner MA 1992 Facilitating the Gibbs sampler: the Gibbs Stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association* **87**, 861–868.

[38] Van Dijk HK 1999 Some Remarks on the simulation revolution in Bayesian econometric inference. *Econometric Reviews* **18**(1), 105–112.

[39] Van Dijk HK and Kloek T 1980 Further experience in Bayesian analysis using Monte Carlo integration. *Journal of Econometrics* **14**, 307–328.

[40] Zeevi AJ and Meir R 1997 Density estimation through convex combinations of densities: Approximation and estimation bounds. *Neural Networks* **10**, 99–106.