TI 2009-017/4
Tinbergen Institute Discussion Paper

To Bridge, to Warp or to Wrap? A Comparative Study of Monte Carlo Methods for Efficient Evaluation of Marginal Likelihoods

David Ardia[1]
Lennart Hoogerheide[2]
Herman K. van Dijk[2]

[1] *University of Fribourg, and Aeris Capital AG, Switzerland;*
[2] *Erasmus University Rotterdam, Econometric and Tinbergen Institutes, The Netherlands.*

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

To Bridge, to Warp or to Wrap?
A comparative study of Monte Carlo methods for efficient evaluation of marginal likelihoods

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

February 2009

Tinbergen Institute report 09-017/4

Abstract

Important 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. We focus on the situation where one makes use of importance sampling or the independence chain Metropolis-Hastings algorithm for posterior analysis. 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. In this respect, the adaptive mixture of Student-t distributions of Hoogerheide et al. (2007) works particularly well. Given an appropriately

*Department of Quantitative Economics, University of Fribourg, Switzerland, and Aeris CAPITAL AG, Pfäffikon SZ, Switzerland.

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

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

yet quickly tuned candidate, straightforward importance sampling provides the most efficient estimator of the marginal likelihood in the cases investigated in this paper, which include a non-linear regression model of Ritter and Tanner (1992) and a conditional normal distribution of Gelman and Meng (1991). A poor choice of candidate density may lead to a huge loss of efficiency where the numerical standard error may be highly unreliable.

Keywords: marginal likelihood; Bayes factor; importance sampling; Markov chain Monte Carlo; bridge sampling; adaptive mixture of Student-t distributions.

1 Introduction

In this article we provide a comparative study 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 or instrumental variables models, one deals with a target distribution that has very non-elliptical contours and that is not a member of a known class of distributions. It is therefore of interest to investigate the performance of some widely used estimators for such cases.

In this paper we restrict our focus to 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.

In Bayesian econometrics, a joint posterior density is given by:

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

where θ denotes the set of parameters of interest, typically a scalar, a vector, a matrix, or a set of these mathematical objects; $p(y|\theta)$ 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. In (1) we define $k(\theta|y) = p(y|\theta)p(\theta)$ as the kernel function of the joint posterior and

$$p(y) = \int_{\theta\in\Theta} k(\theta|y)d\theta \tag{2}$$

as the marginal likelihood. It is clear that the 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 article is to investigate which choices have to be made when estimating a marginal likelihood. We argue that these choices are important. We consider the following issues:

(i) the sensitivity to the choice of the particular estimation procedure (e.g., making use of 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) has considered the special case of mixture models. This article demonstrates 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).

This article proceeds as follows. Section 2 provides a review of some commonly used Monte Carlo estimators of the marginal likelihood. These methods are all members of the class of general bridge sampling estimators. Section 3 gives a brief overview of the approach of Hoogerheide et al. (2007) that uses an adaptive mixture of Student-t distributions (AdMit) as the candidate or importance distribution. In Section 4 we investigate the robustness and efficiency of these estimators in the case of a highly non-elliptical example distribution, a posterior distribution in a non-linear regression model discussed by Ritter and Tanner (1992). In Section 5 we consider the reliability of numerical standard errors. In Section 6 we analyze the performance in conditionally normal distributions of Gelman and Meng (1991). Section 7 concludes.

2 A review of some Monte Carlo methods for marginal likelihood estimation

We first review some of the most commonly used Monte Carlo estimators of marginal likelihood. Their performance will be analyzed later. We extend the overview by Frühwirth-Schnatter (2004), by including the approach of Chib and Jeliazkov (2001), and addressing some more details on implementation, advantages and drawbacks of the methods. Moreover, we especially pay attention to the case of the one-block independence chain MH approach.

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

$$\hat{p}_{\text{IS}}(y) = \frac{1}{L}\sum_{l=1}^{L}\frac{k(\theta^{[l]}|y)}{q(\theta^{[l]})}\,, \tag{3}$$

where $\{\theta^{[l]}\}_{l=1}^{L}$ are i.i.d. draws from the exact importance density $q(\cdot)$ which should approximate the joint posterior density $p(\theta|y)$. The IS estimator in (3) stems from

$$p(y) = \int_{\theta\in\Theta} k(\theta|y)d\theta = \int_{\theta\in\Theta}\frac{k(\theta|y)}{q(\theta)}q(\theta)d\theta = E_q\left[\frac{k(\theta|y)}{q(\theta)}\right].$$

where $E_q[.]$ denotes the expectation over the importance density $q(\cdot)$. The IS approach of marginal likelihood estimation is a *globally oriented* method that aims at directly evaluating the integral $\int_{\theta\in\Theta} k(\theta|y)d\theta$ 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 $q(\cdot)$ must also be fatter than the tails of the posterior. That is, $q(\cdot)$ should 'wrap' the posterior density. An advantage of the IS estimator is that its derivation and implementation are straightforward. A possible disadvantage is that for efficiency we require a suitable importance density that covers the whole posterior density: all areas of the parameter space Θ that contain substantial posterior probability mass must be 'wrapped' with a reasonable amount of candidate probability mass. Finding an appropriate importance or candidate density can be troublesome, especially if the posterior density is asymmetric or multimodal. However, we focus on the case in which we make use of IS or the independence chain MH algorithm, where we anyway require an appropriate candidate distribution to efficiently generate our candidate draws, so that this requirement of the IS marginal likelihood estimator does not really pose an extra problem.

The **Reciprocal Importance Sampling (RIS)** estimator (Gelfand and Dey (1994)) is given by:

$$\hat{p}_{\text{RIS}}(y) = \left[\frac{1}{M}\sum_{m=1}^{M}\frac{q_{\text{aux}}(\theta^{[m]})}{k(\theta^{[m]}|y)}\right]^{-1}, \tag{4}$$

where $\{\theta^{[m]}\}_{m=1}^{M}$ are (correlated) posterior draws from an MCMC sampler. $q_{\text{aux}}(\cdot)$ 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 $q_{\text{aux}}(\cdot)$. The RIS estimator (4) stems from

$$\frac{1}{p(y)} = \int_{\theta\in\Theta} \frac{q_{\text{aux}}(\theta)}{p(y)} d\theta = \int_{\theta\in\Theta} \frac{q_{\text{aux}}(\theta)}{k(\theta|y)} p(\theta|y) d\theta = E_p\left[\frac{q_{\text{aux}}(\theta)}{k(\theta|y)}\right] \tag{5}$$

where $E_p[.]$ denotes the expectation over the posterior density $p(\theta|y)$. The second equality stems from

$$p(y) = \frac{k(\theta|y)}{p(\theta|y)}. \tag{6}$$

The RIS approach is a *locally oriented* approach: it makes use of the fact that *for each* $\theta \in \Theta$ there holds (6). High efficiency is most likely to result if $q_{\text{aux}}(\cdot)$ roughly matches the posterior density. However, the RIS estimator is still consistent if $q_{\text{aux}}(\cdot)$ only covers a small part of the parameter space Θ, since under mild conditions (5) holds for each density $q_{\text{aux}}(\theta)$ on the parameter space Θ. For stability of the estimator, the tails of $q_{\text{aux}}(\theta)$ must be thinner than those of the posterior in order to keep the ratio $\frac{q_{\text{aux}}(\theta)}{k(\theta|y)}$ bounded.

Gelfand and Dey (1994) propose a multivariate normal 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 normal density, truncated to a subspace $\hat{\Theta}$ of Θ

$$q_{\text{aux}}(\theta) = \frac{1}{(1-p)\,(2\pi)^{d/2}} |\hat{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta})\right] 1(\theta \in \hat{\Theta})$$

where $\hat{\theta}$ and $\hat{\Sigma}$ can be chosen as estimates of the posterior mean and covariance matrix, $1(\cdot)$ is the indicator function, d is the dimension of θ; the parameter subspace $\hat{\Theta}$ is defined as

$$\hat{\Theta} = \left\{\theta : (\theta-\hat{\theta})'\hat{\Sigma}^{-1}(\theta-\hat{\theta}) \leq \chi^2_{1-c}(d)\right\}$$

where $\chi^2_{1-c}(d)$ is the $(1-c)$th quantile of the Chi-squared distribution with d degrees of freedom. The value of c can be chosen to minimize the numerical standard error of the resulting marginal likelihood estimator. The additional cost of trying several different values for c is very low, as this requires no extra draws or evaluations of candidate or target densities. In the case of (almost) elliptical posterior distributions, one would

expect small values of c (e.g., $c = 0.01$) to work best, since then more draws will be included when estimating the marginal likelihood. In the case of a (highly) non-elliptical posterior, one should choose $\hat{\theta}$ as the posterior mode rather than the posterior mean, as the posterior density kernel may be low (or even 0) around the posterior mean. Further, the optimal value of c can then be much lower (e.g., $c = 0.40$), since in certain directions the posterior density kernel may quickly drop.

An advantage of the RIS estimator is that the *local* character of the approach implies that the auxiliary density $q_{\text{aux}}(\cdot)$ 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 (4) is the harmonic mean estimator by Newton and Raftery (1994), in which the prior $p(\theta)$ is used as the importance density. However, it is well-known that this estimator is unstable because we have

$$\frac{q_{\text{aux}}(\theta^{[m]})}{k(\theta^{[m]}|y)} = \frac{p(\theta)}{p(\theta)p(y|\theta)} = \frac{1}{p(y|\theta)},$$

where the inverse likelihood function typically does not have a finite variance. The reason is that some of the likelihood terms in the sum are near zero, leading to extreme values of $\frac{1}{p(y|\theta)}$. Therefore, we do not investigate the version of the harmonic mean.

The **(optimal) Bridge Sampling (BS)** estimator (Meng and Wong (1996)) is obtained as the limit of the sequence

$$\hat{p}_{\text{BS}}^{(t)}(y) = \hat{p}_{\text{BS}}^{(t-1)}(y) \times \frac{\frac{1}{L}\sum_{l=1}^{L} \frac{\hat{p}(\theta^{[l]} \,|\, y)}{Lq(\theta^{[l]}) + M\hat{p}(\theta^{[l]} \,|\, y)}}{\frac{1}{M}\sum_{m=1}^{M} \frac{q(\theta^{[m]})}{Lq(\theta^{[m]}) + M\hat{p}(\theta^{[m]} \,|\, y)}}, \tag{7}$$

where $\hat{p}(\theta \,|\, y) = k(\theta \,|\, y)/\hat{p}_{\text{BS}}^{(t-1)}(y)$ and the initial value $p_{\text{BS}}^{(0)}(y)$ is set to (3), 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 $q(\cdot)$. 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 $q(\cdot)$. The original BS estimator in (7) is optimal if the draws $\{\theta^{[m]}\}_{m=1}^{M}$ are i.i.d. We will 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 size' $\tilde{M}$, defined as $\tilde{M} = M(1-\rho)/(1+\rho)$ with ρ the first order serial correlation of the likelihood evaluations of the $\{\theta^{[m]}\}_{m=1}^{M}$. We will 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 $q(\cdot)$, 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 BS estimator stems from the following results. Let $\alpha(\cdot)$ be an arbitrary function such that

$$\int_{\theta\in\Theta} \alpha(\theta)p(\theta|y)q(\theta)d\theta > 0.$$

Then we have

$$1 = \frac{\int_{\theta\in\Theta} \alpha(\theta)p(\theta|y)q(\theta)d\theta}{\int_{\theta\in\Theta} \alpha(\theta)q(\theta)p(\theta|y)d\theta} = \frac{E_q[\alpha(\theta)p(\theta|y)]}{E_p[\alpha(\theta)q(\theta)]}.$$

Multiplying both sides by $p(y)$ yields:

$$p(y) = \frac{E_q[\alpha(\theta)k(\theta|y)]}{E_p[\alpha(\theta)q(\theta)]}.$$

Substituting sample averages for these expectations results in the general bridge-sampling (GBS) estimator:

$$\hat{p}_{\mathrm{GBS}}(y) = \frac{\frac{1}{L}\sum_{l=1}^{L} \alpha(\theta^{[l]})k(\theta^{[l]}|y)}{\frac{1}{M}\sum_{m=1}^{M} \alpha(\theta^{[m]})q(\theta^{[m]})} \tag{8}$$

The IS and RIS estimators are 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.

The estimator of **Chib and Jeliazkov (2001)** for marginal likelihood estimation on the basis of Metropolis-Hastings draws is given by:

$$\hat{p}_{\mathrm{CJ}}(y) = \frac{k(\theta^*|y)}{\hat{p}(\theta^*|y)} \tag{9}$$

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 Chib-Jeliazkov (CJ) estimator is given by:

$$\hat{p}(\theta^*|y) = q(\theta^*) \frac{\frac{1}{M}\sum_{m=1}^{M} \alpha_{\mathrm{MH}}(\theta^{[m]}, \theta^*)}{\frac{1}{L}\sum_{l=1}^{L} \alpha_{\mathrm{MH}}(\theta^*, \theta^{[l]})} \tag{10}$$

with $\alpha(\theta, \theta')$ 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 estimator stems from the fact that for each $\theta^* \in \Theta$ we have $p(y) = \frac{k(\theta^*|y)}{p(\theta^*|y)}$. The idea behind equation (10) is that we have:

$$p(\theta^*|y) = q(\theta^*) \frac{E_p[\alpha_{\mathrm{MH}}(\theta, \theta^*)]}{E_q[\alpha_{\mathrm{MH}}(\theta^*, \theta)]} = \frac{\int_{\theta\in\Theta} q(\theta^*)\alpha_{\mathrm{MH}}(\theta, \theta^*)p(\theta|y)d\theta}{\int_{\theta\in\Theta} \alpha_{\mathrm{MH}}(\theta^*, \theta)q(\theta)d\theta},$$

which follows from the MH chain's key property, the reversibility condition:

$$p(\theta|y)q(\theta^*)\alpha_{\mathrm{MH}}(\theta, \theta^*) = p(\theta^*|y)q(\theta)\alpha_{\mathrm{MH}}(\theta^*, \theta). \tag{11}$$

That is, in the Markov chain of MH draws, moves from θ to θ^* (left-hand side of (11)) are observed as often as moves from θ^* to θ (right-hand side of (11)).

The CJ approach can be applied for each $\theta^* \in \Theta$ with $p(\theta^*|y) > 0$. 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 0) posterior density value.

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

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

since substituting this choice of $\alpha_{\mathrm{CJ},\theta^*}(\theta)$ into (8) gives:

$$\begin{aligned}
\hat{p}(y) &= \frac{\frac{1}{L}\sum_{l=1}^{L}\alpha_{\mathrm{CJ},\theta^*}(\theta^{[l]})k(\theta^{[l]}|y)}{\frac{1}{M}\sum_{m=1}^{M}\alpha_{\mathrm{CJ},\theta^*}(\theta^{[m]})q(\theta^{[m]})} \\
&= \frac{\frac{1}{L}\sum_{l=1}^{L}\min\left\{\frac{q(\theta^*)}{q(\theta^{[l]})}, \frac{k(\theta^*|y)}{k(\theta^{[l]}|y)}\right\}k(\theta^{[l]}|y)}{\frac{1}{M}\sum_{m=1}^{M}\min\left\{\frac{q(\theta^*)}{q(\theta^{[m]})}, \frac{k(\theta^*|y)}{k(\theta^{[m]}|y)}\right\}q(\theta^{[m]})} \\
&= \frac{\frac{1}{L}\sum_{l=1}^{L}\min\left\{\frac{k(\theta^{[l]}|y)}{k(\theta^*|y)}\frac{q(\theta^*)}{q(\theta^{[l]})}, 1\right\}k(\theta^*|y)}{\frac{1}{M}\sum_{m=1}^{M}\min\left\{1, \frac{k(\theta^*|y)}{k(\theta^{[m]}|y)}\frac{q(\theta^{[m]})}{q(\theta^*)}\right\}q(\theta^*)} \\
&= k(\theta^*|y)\Bigg/\left[q(\theta^*)\frac{\frac{1}{M}\sum_{m=1}^{M}\alpha_{\mathrm{MH}}(\theta^{[m]},\theta^*)}{\frac{1}{L}\sum_{l=1}^{L}\alpha_{\mathrm{MH}}(\theta^*,\theta^{[l]})}\right].
\end{aligned}$$

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 size' of the sample 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 a subspace 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 Importance Sampling 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. 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 (=likelihood evaluations), with most likelihood evaluations being close to 0.

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 the integral

$$p(y) = \int_{\theta\in\Theta} k(\theta|y)d\theta = \int_{\theta\in\Theta} k((\theta - \theta_0) + \theta_0|y)d\theta \tag{12}$$

remains equal if we take the 'mirror image' around a certain point $\theta_0 \in \Theta$:

$$p(y) = \int_{\theta\in\Theta} k(-(\theta - \theta_0) + \theta_0|y)d\theta = \int_{\theta\in\Theta} k(-\theta + 2\theta_0|y)d\theta. \tag{13}$$

Combining (12) and (13) yields:

$$p(y) = \int_{\theta\in\Theta} k(\theta|y)d\theta = \int_{\theta\in\Theta} \frac{1}{2}\left[k(\theta|y) + k(-\theta + 2\theta_0|y)\right] d\theta. \tag{14}$$

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

$$\tilde{k}(\theta|y) = \frac{1}{2}\left[k(\theta|y) + k(-\theta + 2\theta_0|y)\right] \tag{15}$$

rather than to the posterior kernel $k(\theta|y)$, also yields an estimator of the marginal likelihood. The *warped* posterior kernel $\tilde{k}(\theta|y)$ is point symmetric around θ_0, where we choose 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 distribution (e.g., Student-t or Gaussian). 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 will refer to the transformation in (15) as the Warp1 transformation.

In the two terms of the Warp 1 transformation in (15) 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'. For example, in the two-dimensional case this yields:

$$k^*(\theta|y) = \frac{1}{4}\left[k(\theta_1, \theta_2|y) + k(\theta_1, \theta_2'|y) + k(\theta_1', \theta_2|y) + k(\theta_1', \theta_2'|y)\right] \tag{16}$$

where $\theta' = -\theta + 2\theta_0$. Obviously, a disadvantage is that for increasing values of the dimension d, the number of posterior kernel evaluations per candidate draw increases exponentially. We will refer to this transformation, of which the two-dimensional version is given by (16), 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 normal. 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 we shall take L_{IS} and M_{RIS} twice as large as $L_{\mathrm{BS}} = M_{\mathrm{BS}}$, so that IS and RIS could very well outperform BS.

We focus on the cases of importance sampling and the independence chain Metropolis-Hastings algorithm. So, we compare the following strategies:

(IS) use all candidate draws in the IS estimator (3);

(RIS, CJ) transform all candidate draws to a sequence of MH draws (plus a burn-in) and use these in the RIS estimator (4) or the CJ estimator (9);

(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 (7) – with M substituted by the effective size $\tilde{M}$ 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 candidate or importance distribution.

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

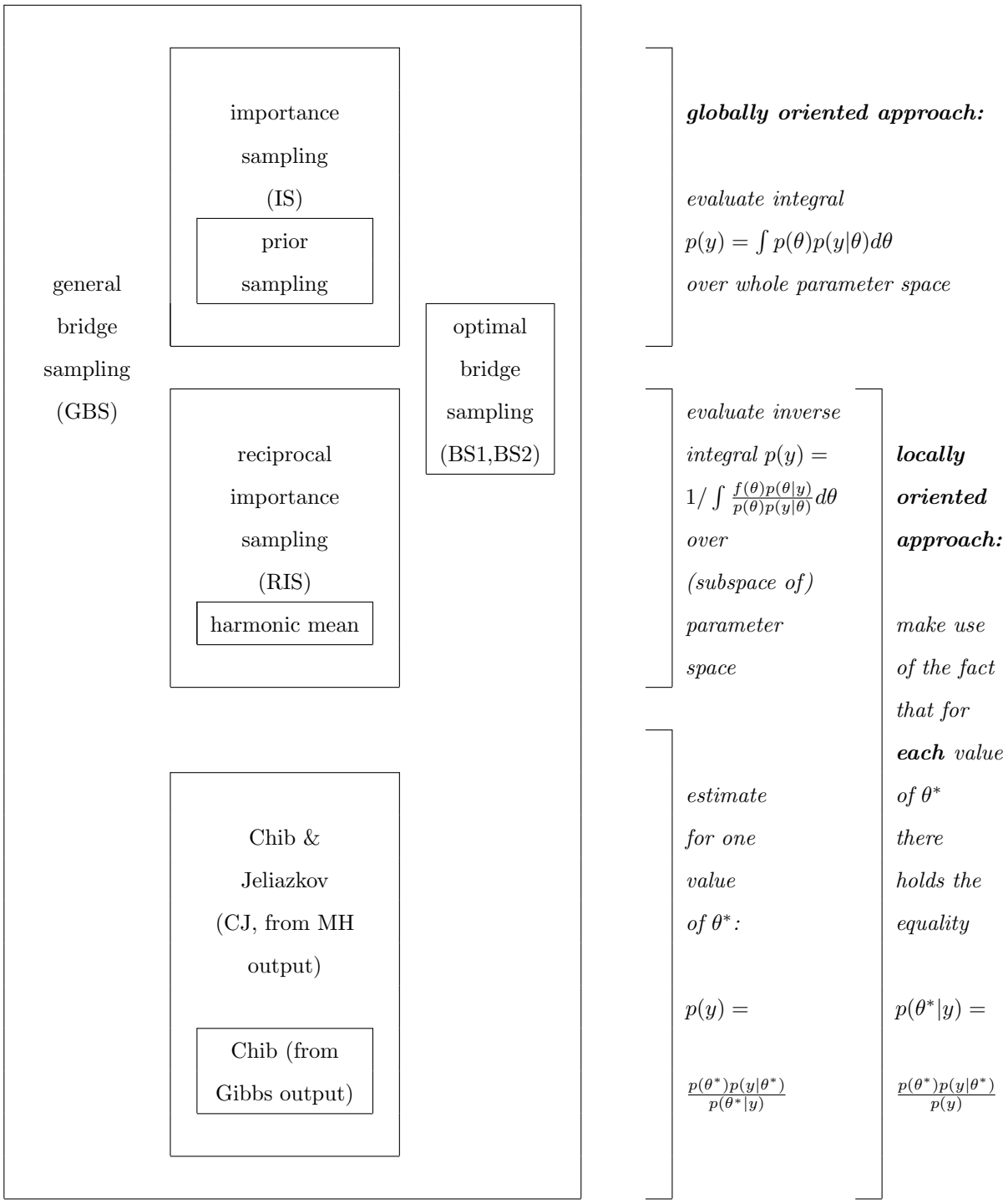


Table 1: Computations required by different marginal likelihood estimation approaches, in case we make use of IS or the independence chain MH algorithm. L is the number of candidate draws that are not used in the MH algorithm. M is the number of independence chain MH draws from the posterior. Warp1 and 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 t (AdMit) 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 importance sampling (or as a candidate density in the independence chain Metropolis-Hastings 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 importance sampling the candidate density is unimodal. If the target distribution is multimodal then some draws may have huge importance weights or some modes may even be completely missed. Thus, an important problem is the choice of the importance density, especially when little is known 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 normal 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 $p(\theta|y)$ of which we only know a kernel function $k(\theta|y)$ with $\theta \in \mathbb{R}^d$. 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 $k(\theta|y)$;

(2) Optimization of the mixing probabilities;

(3) Drawing a sample from the new mixture;

(4) Evaluation of importance sampling weights: if the coefficient of variation, the standard deviation divided by the mean, of the IS weights has converged, then stop. Otherwise, go to step (1).

For more details we refer to Hoogerheide et al. (2007).
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`. Its use is discussed and illustrated by Ardia et al. (2009).

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 (and the first application of AdMit to a non-linear regression model).

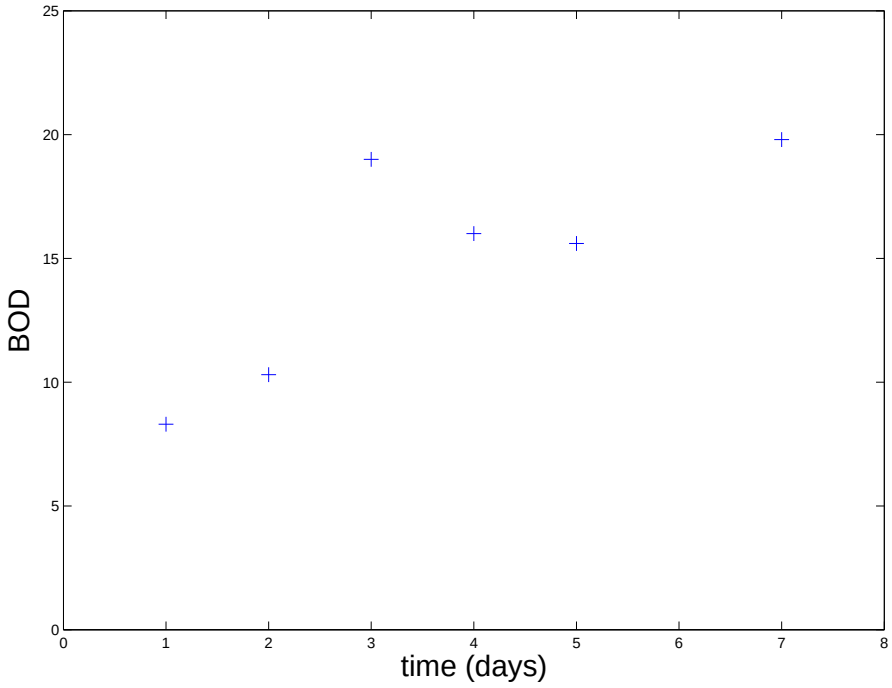


Figure 2: Data from Marske (1967): Biochemical Oxygen Demand (BOD) versus time.

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) (see Figure 2).

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

$$y_i = \theta_1(1 - \exp(-\theta_2\, x_i)) + \varepsilon_i \tag{17}$$

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

Following Ritter and Tanner (1992), we specify a flat prior on a bounded interval: $(\theta_1, \theta_2, \sigma) \in [-20, 50] \times [-2, 6] \times [0, 20]$. (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.)

The top-left panel of Figure 3 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 (17), respectively. The smaller mode reflects the small posterior probability of a convex function.

For the importance sampling and independence chain Metropolis-Hastings algorithms we consider three candidate distributions:

(1) the mixture of Student-t distributions resulting from the AdMit procedure of Hoogerheide et al. (2007);

(2) an 'adaptive' Student-t distribution where the mode and scale have been iteratively updated by several importance sampling 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 3 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 4 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 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.)

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 normal auxiliary density around the posterior mode where the optimal value of c appeared to be (approximately) $c = 0.40$. This result differs from the low value of c, e.g., $c = 0.01$, that is typically optimal in case of (nearly) elliptical posteriors. 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. Since one often works with the (natural) logarithm of the marginal likelihood, we display results for both the marginal likelihood and its logarithm. Boxplots of the 500 marginal likelihood estimates are given in Figures 5. The real value of the marginal likelihood is (rounded to two digits) 12.79×10^{-10} (with logarithm -20.48). This real value is computed by deterministic integration which is still feasible (but quite time-consuming) in this three-dimensional example.

First of all, 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 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 6 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.

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 here it is better to *wrap* the posterior than to *warp* 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 (17) and the linear regression model

$$y_i = \beta_1 + \beta_2 \, x_i + \varepsilon_i \tag{18}$$

with independent errors $\varepsilon_i \sim N(0, \sigma^2)$ $(i = 1, \ldots, 6)$. 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 (17), the Normal-Gamma prior

$$\beta \left| \frac{1}{\sigma^2} \sim N\left(\underline{\beta},\, \sigma^2 \, \underline{V}\right) \right. \qquad \frac{1}{\sigma^2} \sim \text{Gamma}\left(\underline{s}^{-2}, \underline{\nu}\right),$$

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

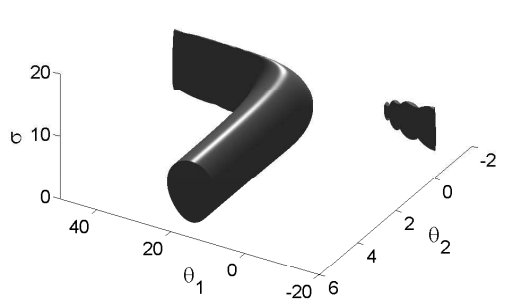
non-linear model is 1.0315, so that under equal prior probabilities the posterior model probabilities for the non-linear and linear models are 0.5078 and 0.4922, respectively. Figure 5 shows that only for the AdMit-IS estimator all 500 repetitions of the simulation yield marginal likelihood estimates 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. We conclude 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 such standard deviations. 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.

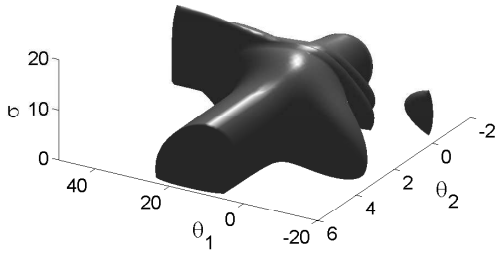
Table 2: Posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$ in non-linear regression model (17): estimation of the marginal likelihood (ML) based on 100000 draws from AdMit mixture of four Student-t distributions, adaptive Student-t or naive Student-t distribution (500 repetitions). True values are ML= 12.79×10^{-10} and log(ML)= -20.48.

$10^{10}\cdot$ 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

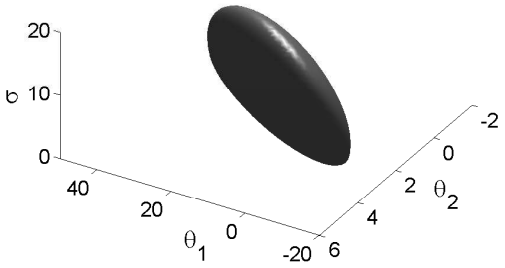
log(ML)	AdMit		Adapt		Naive	
	mean	st.dev.	mean	st.dev.	mean	st.dev.
IS	-20.4772	0.0075	-20.4773	0.0140	-20.4853	0.0824
RIS	-20.4475	0.0260	-20.4729	0.0729	-20.4810	0.1354
BS1	-20.4795	0.0155	-20.4743	0.0328	-20.4797	0.1976
BS2	-20.4794	0.0110	-20.4776	0.0212	-20.4798	0.1949
CJ	-20.4781	0.0200	-20.4782	0.0221	-20.4796	0.1986



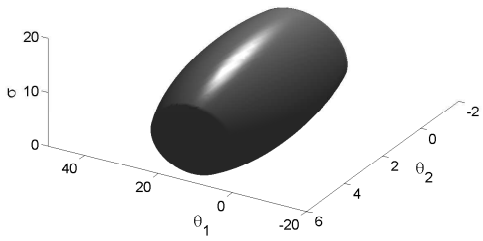
target posterior distribution



AdMit candidate (= mixture of four Student-t distributions)

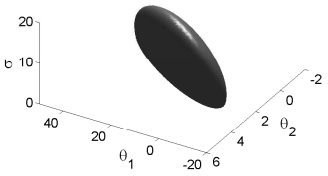


Student-t candidate (around posterior mode)



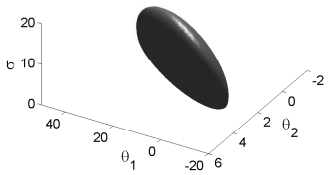
Student-t candidate (location and scale adapted to target)

Figure 3: Posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$ in non-linear regression model (17): Highest Posterior Density credible region (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.



Student-t component 1

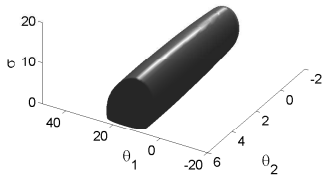
→



candidate 1

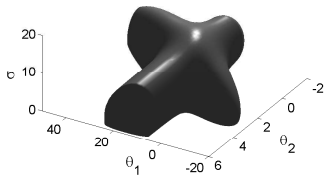
(= component 1)

↙



Student-t component 2

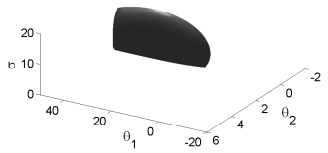
→



candidate 2

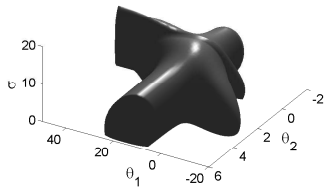
(= mixture of components 1, 2)

↙



Student-t component 3

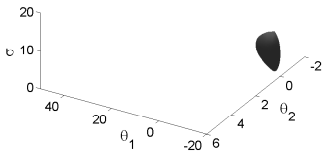
→



candidate 3

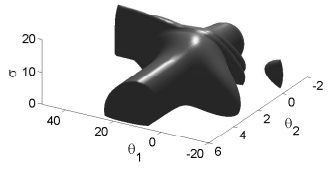
(= mixture of components 1, 2, 3)

↙



Student-t component 4

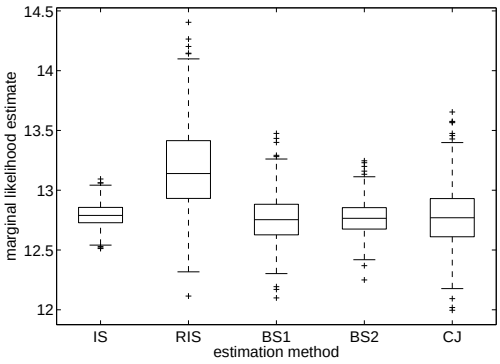
→



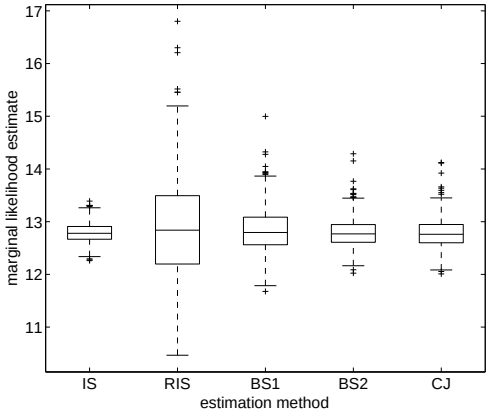
candidate 4

(= mixture of components 1, 2, 3, 4)

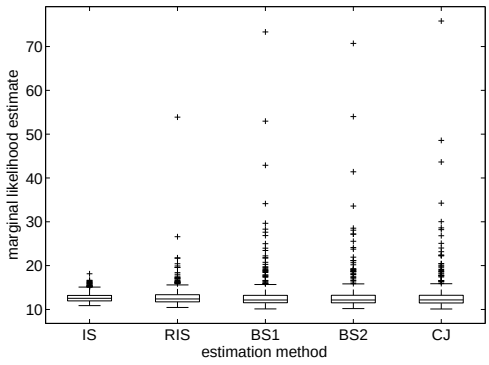
Figure 4: Posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$ in non-linear regression model (17): the AdMit algorithm (automatically and) iteratively approximates the non-elliptical posterior shapes by a mixture of Student-t distributions.



candidate draws from AdMit distribution

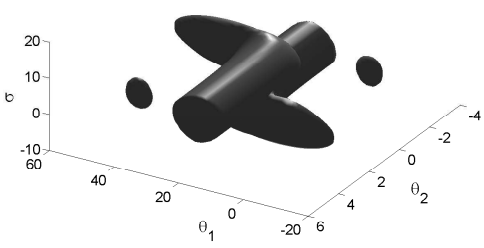


candidate draws from adaptive Student-t distribution

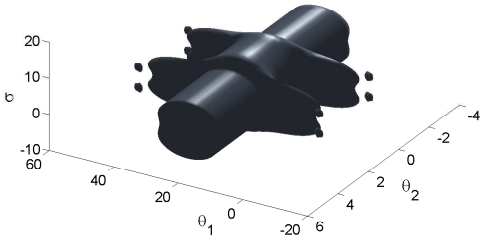


candidate draws from naive Student-t distribution

Figure 5: Posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$ in non-linear regression model (17): estimates of $10^{10}\times$ marginal likelihood based on 100000 draws from AdMit mixture of four Student-t distributions, adaptive Student-t or naive Student-t distribution (500 repetitions).



Warp1 (mixture of 2 posterior transformations)



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

Figure 6: Posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$ in non-linear regression model (17): 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.

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 $\frac{k(\theta^{[l]}|y)}{q(\theta^{[l]})}$ $(l = 1, \ldots, L)$ 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 stdev$(\hat{g})$ of

$$\hat{g} = \frac{1}{M} \sum_{m=1}^{M} g(\theta^{[m]}) \tag{19}$$

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

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

$$\widehat{\text{var}}_{\text{NW}}(\hat{g}) = \hat{\gamma}_0 + 2 \sum_{i=1}^{b} \left(1 - \frac{i}{b+1}\right) \hat{\gamma}_i, \tag{20}$$

Table 3: Posterior distribution of $\theta = (\theta_1, \theta_2, \sigma)'$ in non-linear regression model (17): marginal likelihood estimation making use of Warp1 or Warp2 transformations in combination with an adaptive Student-t candidate distribution (500 repetitions).

$10^{10}\cdot$ ML		IS	BS1	BS2
		st.dev.	st.dev.	st.dev.
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

log(ML)		IS	BS1	BS2
		st.dev.	st.dev.	st.dev.
Warp1	(100000 posterior kernel evaluations)	0.0137	0.0276	0.0176
Warp2	(100000 posterior kernel evaluations)	0.0242	0.0454	0.0316
Warp1	(100000 candidate draws)	0.0098	0.0201	0.0127
Warp2	(100000 candidate draws)	0.0092	0.0167	0.0119

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 Metropolis-Hastings 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:

$$\widehat{\text{var}}_{\text{IPSE}}(\hat{g}) = \hat{\gamma}_0 + 2\sum_{t=0}^{2h+1}\hat{\gamma}_t = -\hat{\gamma}_0 + 2\sum_{t=0}^{h}\hat{\Gamma}_t \tag{21}$$

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: $\widehat{\text{var}}_{\text{IMSE}}(\hat{g}) \leq \widehat{\text{var}}_{\text{IPSE}}(\hat{g})$.

We now inspect the NSE in the example from the previous section. Figures 7, 8 and 9 show boxplots, comparing the numerical standard errors to the standard deviations for the three candidate distributions.

Figure 7 shows that for the naive Student-t candidate distribution the NSE is often unreliable: huge underestimation of the uncertainty in the marginal likelihood estimator often occurs. Figure 8 depicts that for an 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. Figure

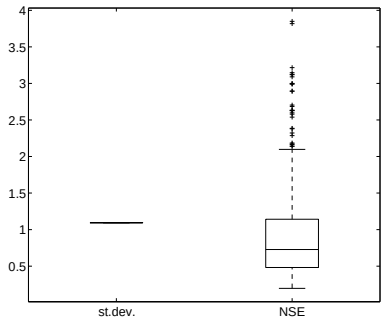
9 shows that for the AdMit candidate distribution the NSE is more reliable than for the other candidates. Especially for the AdMit-IS estimator, the 'winner' of Section 4, the NSE seems reliable. For the BS1, BS2 and CJ estimators, the NSE from the IPSE should again be preferred over the NSE from the 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

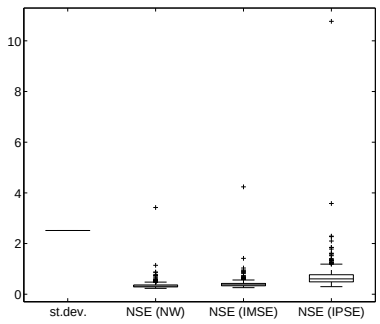
$$(\widehat{p}(y) - 1.645 \times \mathrm{NSE}_{\hat{p}(y)},\ \widehat{p}(y) + 1.645 \times \mathrm{NSE}_{\hat{p}(y)}).$$

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.

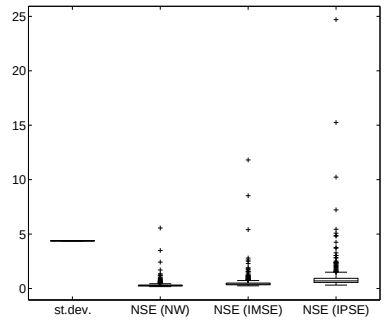
We conclude that also in terms of the reliability of the NSE and confidence intervals the AdMit-IS approach performs best. For other AdMit estimators (BS1, BS2 and CJ) the initial monotone sequence estimator of Geyer (1992) provides a useful NSE. For the adaptive (and naive) candidate we find that all 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.



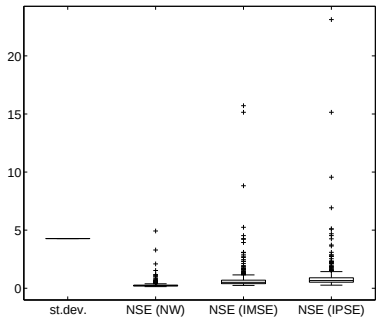
IS



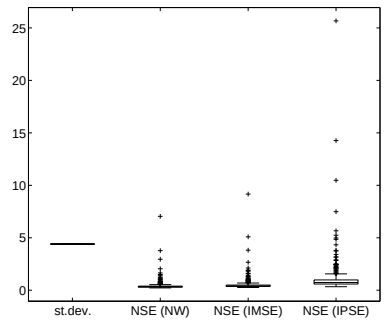
RIS



BS1

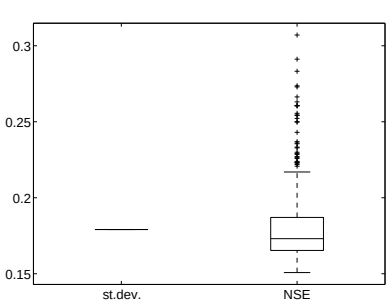


BS2

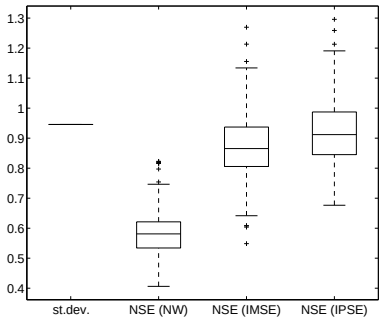


CJ

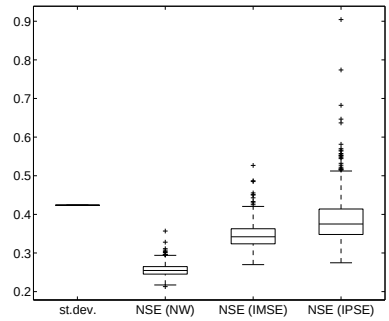
Figure 7: 500 estimates of $10^{10}\times$ marginal likelihood in the non-linear regression model (17), based on 100000 candidate draws from the *'naive' Student-t candidate distribution*: standard deviation (horizontal line in first column) versus 500 numerical standard errors (boxplots in other columns). 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.



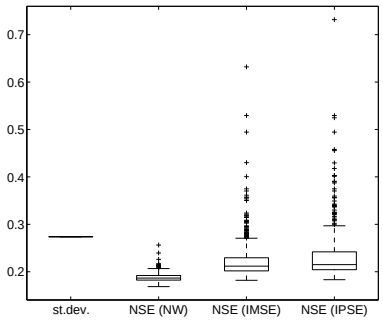
IS



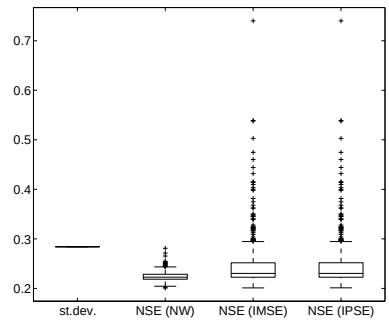
RIS



BS1

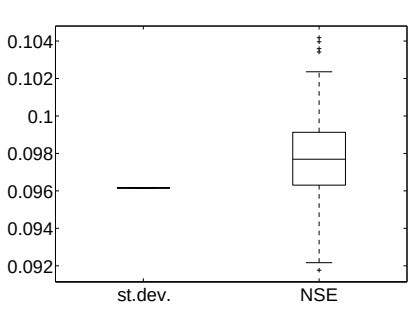


BS2

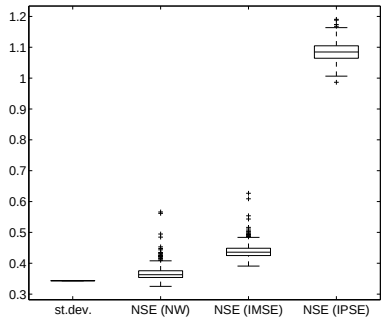


CJ

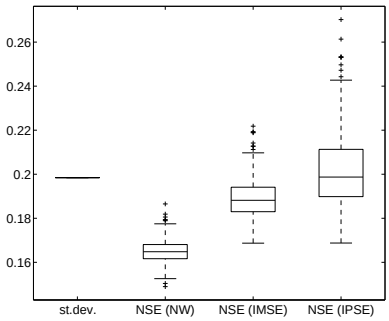
Figure 8: 500 estimates of $10^{10}\times$ marginal likelihood in the non-linear regression model (17), based on 100000 candidate draws from the *'adaptive' Student-t candidate distribution*: standard deviation (horizontal line in first column) versus 500 numerical standard errors (boxplots in other columns). 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.



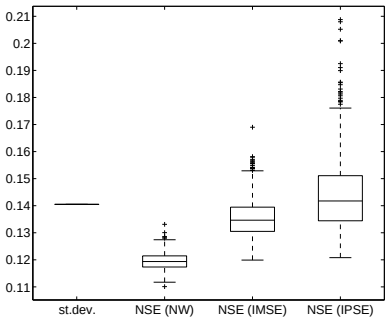
IS



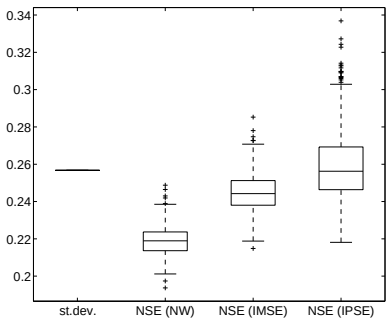
RIS



BS1



BS2



CJ

Figure 9: 500 estimates of $10^{10}\times$ marginal likelihood in the non-linear regression model (17), based on 100000 candidate draws from the *AdMit (mixture of four Student-t) candidate distribution*: standard deviation (horizontal line in first column) versus 500 numerical standard errors (boxplots in other columns). 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: Estimation of marginal likelihood $p(y)$ in non-linear regression model (17): coverage rate of 90% interval for $p(y)$ based on different NSE's (in 500 repetitions).

	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

* For the IS estimators there is no serial correlation in the series of draws, so that only one (straightforward) NSE formula is used.

6 Application 2: conditionally normal distributions of Gelman and Meng (1991)

Gelman and Meng (1991) discuss the class of conditionally normal distributions. Suppose we consider a conditionally normal distribution for $\theta \in \mathbb{R}^d$. Then after location and scale transformations in each variable, the joint density kernel of θ is given by:

$$k(\theta) \propto \exp\left(-\frac{1}{2}\sum_j A_j\, \theta_1^{c_{1j}} \cdots \theta_d^{c_{dj}}\right) \tag{22}$$

where the c_{ij} attain the values 0, 1 or 2, and where the summation is possibly over 3^d terms. The 3^d coefficients A_j are allowed to take on any real values for which the joint density kernel (22) has a finite integral.

In this section we consider the estimation of the normalizing constant (NC) of a joint density kernel (22) with dimension $d = 10$. Since in this case the target density kernel does not correspond to a *posterior* distribution, this normalizing constant does not have the interpretation of a marginal likelihood. Nevertheless, for the evaluation of the quality of our estimation methods this is not an essential difference. The advantage of the class of conditionally normal distributions is that we can simply choose the dimension and easily 'tune' the shapes of the target distribution.

We analyze a highly non-elliptical example distribution where the shapes of the marginal distribution of (θ_1, θ_2) are depicted by Figure 10. This example can roughly be interpreted as a ten-dimensional extension of the posterior in the non-linear regression model of Section 4.

We again apply our methods with 100000 candidate draws (and 100000 target density evaluations); for each estimator we repeat the simulation 50 times. The results are given by Table 5. To a large extent, conclusions are similar to those of Section 4. The AdMit-IS estimator performs best, with reliable numerical standard errors. The construction of the AdMit candidate distribution, again a mixture of four Student-t distributions, took 26 seconds on a 2006 Intel (R) Centrino Duo Core processor.

In this application, the 'victory' over the naive and adaptive Student-t candidate distributions is larger. All estimators based on the naive candidate are downwards

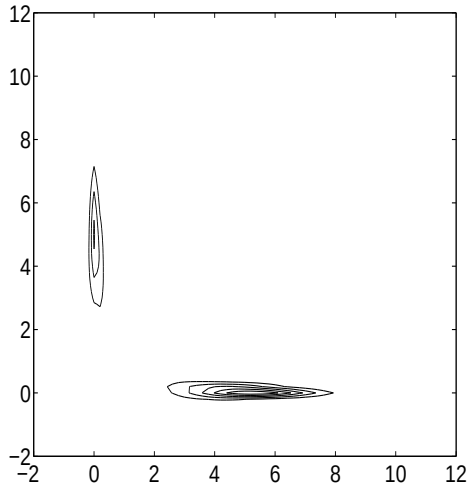


Figure 10: Contour plot: marginal density of (θ_1, θ_2) in a ten-dimensional conditionally normal distribution (Gelman and Meng (1991)).

biased, since the second mode (with $(\theta_1, \theta_2) \approx (0, 5)$) is completely 'missed'. The standard deviations (and all standard errors) are deceptively low: these obviously do not signal that part of the target distribution is not 'covered'. In this ten-dimensional space, the smaller mode is simply not 'found'. For this reason, the 'adaptive' Student-t distribution is not simply iteratively obtained by starting with the distribution around the posterior mode and iteratively using the estimated posterior mean and covariance as the mode and scale in the next iteration, as this approach would also yield a candidate that 'misses' the second mode. A robust optimization of the IS weight function is required. Still, after this optimization the resulting estimators are less precise than their AdMit based counterparts, also if we apply the Warp1 transformation to the target distribution (even if we use twice as many evaluations of the target density kernel). We did not compute the Warp2 estimators, since these would require $2^{10} = 1024$ target density evaluations per candidate draw. The AdMit-RIS estimator has a remarkably small standard deviation, but seems upwards biased. The NSE's based on the IMSE and IPSE are again often more reliable than the NSE's based on the Newey-West approach (with bandwidth 40).

Table 5: Simulation results for estimation of the logarithm of the normalizing constant (NC) of a 10-dimensional highly non-elliptical, conditionally normal distribution of Gelman and Meng (1991), based on 100000 candidate draws from an AdMit, adaptive Student-t or naive Student-t candidate distribution (50 repetitions). 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 MH draws. For the IS estimators there is no serial correlation in the series of draws, so that only one (straightforward) NSE formula is used.

		log(NC) estimate		NSE (NW)		NSE (IMSE)		NSE (IPSE)	
		mean	st.dev.	mean	st.dev.	mean	st.dev.	mean	st.dev.
Without Warping:									
AdMit	IS	20.9231	0.0070	0.0071	0.0007	0.0071	0.0007	0.0071	0.0007
(mixture	RIS	20.9773	0.0097	0.0147	0.0001	0.0194	0.0001	0.0616	0.0003
of 4	BS1	20.9324	0.0150	0.0108	0.0003	0.0127	0.0008	0.0192	0.0035
Student-t)	BS2	20.9331	0.0124	0.0091	0.0003	0.0117	0.0010	0.0194	0.0027
candidate	CJ	20.9273	0.0633	0.0493	0.0189	0.0508	0.0191	0.0510	0.0192
Adaptive	IS	20.9235	0.0217	0.0217	0.0005	0.0217	0.0005	0.0217	0.0005
Student-t	RIS	20.9143	0.0384	0.0201	0.0007	0.0362	0.0020	0.0378	0.0028
candidate	BS1	20.9272	0.0574	0.0420	0.0023	0.0607	0.0047	0.0643	0.0063
(location & scale	BS2	20.9159	0.0264	0.0253	0.0003	0.0302	0.0015	0.0312	0.0023
adapted to target)	CJ	20.9439	0.1262	0.0939	0.0148	0.1186	0.0207	0.1222	0.0216
Naive	IS	20.6580	0.0060	0.0053	0.0015	0.0053	0.0015	0.0053	0.0015
Student-t	RIS	20.6582	0.0083	0.0068	0.0002	0.0072	0.0004	0.0076	0.0020
candidate	BS1	20.6582	0.0073	0.0074	0.0002	0.0078	0.0003	0.0081	0.0014
(round	BS2	20.6577	0.0062	0.0064	0.0002	0.0067	0.0004	0.0072	0.0015
mode)	CJ	20.6594	0.0616	0.0454	0.0254	0.0462	0.0259	0.0462	0.0259
Warp 1 transformation:									
	100000 target density kernel evaluations:								
Adaptive	IS	20.9236	0.0218	0.0214	0.0006	0.0214	0.0006	0.0214	0.0006
Student-t	BS1	20.9119	0.0530	0.0413	0.0021	0.0520	0.0056	0.0561	0.0035
candidate	BS2	20.9226	0.0270	0.0259	0.0004	0.0299	0.0017	0.0314	0.0032
(location and	100000 candidate draws (= 200000 target density kernel evaluations):								
scale adapted	IS	20.9220	0.0133	0.0152	0.0003	0.0152	0.0003	0.0152	0.0003
to target)	BS1	20.9306	0.0371	0.0292	0.0010	0.0372	0.0019	0.0390	0.0026
	BS2	20.9249	0.0214	0.0184	0.0002	0.0215	0.0012	0.0224	0.0021

7 Concluding remarks

In the title we posed the question: to bridge, to warp or to wrap? In our examples of non-elliptical distributions where we use an importance sampling or independence chain Metropolis-Hastings approach for posterior simulation, we have the following findings on different marginal likelihood estimators. Given a wisely specified candidate or importance density that appropriately 'wraps' the posterior, the straightforward importance sampling estimator outperforms the bridge sampling estimators and also the importance or bridge sampling estimators that are aimed at a warped target density. So, in our applications the answer is: to wrap!

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 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 2008 AdMit: Adaptive mixture of Student-t distribution in R. R package version 1.01-01,
URL `http://CRAN.R-project.org/package=AdMit`

[2] 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`

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

[4] 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.

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

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

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

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

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

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

[11] 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.

[12] 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.

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

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

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

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

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

[18] 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.

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

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

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

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

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

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

[25] 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.

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

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

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

[29] 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`.

[30] 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.

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

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