BAYESIAN ANALYSIS OF DSGE MODELS - SOME COMMENTS

MALIN ADOLFSON, JESPER LINDÉ, AND MATTIAS VILLANI

Sungbae An and Frank Schorfheide have provided an excellent review of the main elements of Bayesian inference in Dynamic Stochastic General Equilibrium (DSGE) models. Bayesian methods have, for reasons clearly outlined in the paper, a very natural role to play in DSGE analysis, and the appeal of the Bayesian paradigm is indeed strongly evidenced by the flood of empirical applications in the area over the last couple of years. We expect their paper to be the natural starting point for applied economists interested in learning about Bayesian techniques for analyzing DSGE models, and as such the paper is likely to have a strong influence on what will be considered best practise for estimating DSGE models.

The authors have, for good reasons, chosen a stylized six-equation model to present the methodology. We shall here use the large-scale model in Adolfson, Laséen, Lindé and Villani (2005), henceforth ALLV, to illustrate a few econometric problems which we have found to be especially important as the size of the model increases. The model in ALLV is an open economy extension of the closed economy model in Christiano, Eichenbaum and Evans (2005). It consists of 25 log-linearized equations, which can be written as a state space representation with 60 state variables, many of them unobserved. 15 observed unfiltered time series are used to estimate 51 structural parameters. An additional complication compared to the model in An and Schorfheide's paper is that some of the coefficients in the measurement equation are non-linear functions of the structural parameters. The model is currently the main vehicle for policy analysis at Sveriges Riksbank (Central bank of Sweden) and similar models are being developed in many other policy institutions, which testifies to the model's practical relevance. The version considered here is estimated on Euro area data over the period 1980Q1-2002Q4. We refer to ALLV for details.

Posterior Sampling Efficiency

The one-block Random Walk Metropolis (RWM) algorithm is by far the most widely used posterior sampling algorithm for DSGE models. It is a robust algorithm, but it can be inefficient, especially in high-dimensional parameter spaces, even when the scaling factor (c^2 in An and Schorfheide's notation) in the proposal density is chosen optimally. A commonly used measure of numerical efficiency for MCMC samplers is the inefficiency factor $1 + 2\sum_{k=1}^{K} \rho_k$, where ρ_k is the autocorrelation at the kth lag in the MCMC chain for a given parameter and K is an upper limit of the lag length such that $\rho_k \approx 0$ for all $k > K$. The inefficiency factor approximates the ratio of the numerical variance of the posterior mean from the MCMC chain

Adolfson: *Sveriges Riksbank, malin.adolfson@riksbank.se.* Lindé: *Sveriges Riksbank and CEPR, jesper.linde@riksbank.se.* Villani: *Sveriges Riksbank and Stockholm University, mattias.villani@riksbank.se.* This paper has been invited as a comment on Sungbae An and Frank Schorfheide's review paper 'Bayesian Analysis of DSGE Models', which is forthcoming in *Econometric Reviews.* The authors thank Paolo Giordani and Ivan Jeliazkov for insightful discussions. The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Executive Board of Sveriges Riksbank. Villani was partially financially supported by a grant from the Swedish Research Council (Vetenskapsrådet, grant no. 412-2002-1007).

to that from hypothetical iid draws. As an example, the average inefficiency factors across the parameters in the model in ALLV is 217.13, when a scaling factor of $c^2 = 0.1$ is used. The mean acceptance rate is roughly 25%, which is in the optimal range. A doubling of the scaling factor decreases the mean acceptance rate to 10%, but leaves the inefficiency factors largely unaltered; the gain from taking larger proposal steps is offset by the decreased probability of a move.

We conduct the following experiment to find out to if these inefficiencies can be mostly attributed to the non-linearity of the DSGE model, with a possibly complicated (*e.g.* containing bimodality or ridges) posterior density, or if they simply reflect the high dimensionality of the parameter space. In the experiment we assume that the true posterior density is proportional to the kernel of a 51-dimensional multivariate t density with a mode and scale matrix equal to the posterior mode and (negative) inverse Hessian at the mode, respectively, from the DSGE model in ALLV. This setup tries to mimic the posterior density in the DSGE model, but excludes any potential pathological behavior (which may result from potential weak identification). The degrees of freedom in the multivariate t is set to 50. The RWM algorithm is used to generate $1,000,000$ draws from the posterior. The scale factor is set to 0.2, resulting in an acceptance rate of roughly 27%. No problems with convergence was detected and the true posterior was well approximated by the posterior draws. The average inefficiency factor in this case is 155.23, which is not dramatically lower than in the DSGE model. To investigate if the inefficiencies result from the correlation structure in the posterior (two pairs of parameters have posterior correlations below -0.8), which may not be adequately captured by the inverse Hessian, we repeated the same experiment but this time with a mode and scale matrix of the target t distribution equal to the zero vector and identity matrix, respectively. The average inefficiency factor in this case was nearly identical to the correlated case. We thus conclude that the inefficiencies in the 51-dimensional DSGE model are large, but not excessively so, and that most of the inefficiency in the RWM algorithm comes from the dimensionality of the parameter space.

There is clearly room for big improvements in simulation efficiency, however. One possibility that immediately comes to mind is to split the parameter vector in blocks and to draw each block conditional on all parameters in the other blocks. There are two main reasons for why this may not be a good idea in DSGE models. First, none of the full conditional posteriors will be easily simulated distributions. Second, any blocking of parameters which requires the computationally expensive likelihood to be re-evaluated for more than one of the blocks is likely to be time-consuming. We shall restrict attention to the single-block case in the following. Still, future research in large-dimensional DSGEs may find it worthwhile to explore blocking. One possibility is then to use the correlation structure in the inverse Hessian at the posterior mode to form blocks so that strongly correlated parameters are placed in the same block.

Another venue for improving posterior sampling efficiency is re-parametrization. An and Schorfheide have chosen to work with the original bounded parameter space in the posterior sampling stage. It is not uncommon to have a least a few parameters with marginal posteriors which are markedly non-normal in DSGE applications. For example, the persistence coefficients in some of the exogenous processes tend to be rather close to their upper bound of unity, which makes these posteriors skewed to the left. Applying log-transformations to strictly positive parameters and logit-transformations on parameters restricted to the unit interval (we shall refer to these transformations as the natural re-parametrization) is typically beneficial for improving the normal approximation of the posterior, and we have indeed found this to be the case in our applications. Figure 1 shows the effect of applying the natural parameter transformation in the model in ALLV. The figure displays the log posterior sliced through the posterior mode in the direction of the coordinate axis of three of the model parameters. The chosen parameters are: λ_f, the average mark-up in the domestic prices, ρ_{tech} and σ_{tech}, the persistence and innovation standard deviation of a stationary technology shock. The second row shows that the posterior densities in transformed parameter space are substantially closer to normal than the corresponding posterior densities in the original parametrization (first row).[1] Improving the approximation to normality is expected to make the posterior sampling more efficient. This is confirmed here, the average inefficiency factor from applying RWM in the original parameter space is 339.25, which should be compared to 217.13 when sampling is performed in the transformed parameter space. Another disadvantage of the original bounded parametrization is that it requires an adjustment of the modified harmonic estimator of the marginal data density if the truncating ellipsoid (see equation 49 in An and Schorfheide's paper) is not fully contained in the parameter space (Geweke, 1999a).

Other commonly used Metropolis-Hastings algorithms, like the Independence Metropolis-Hastings (IMH, Tierney, 1994) or the Accept-Reject Metropolis-Hastings (ARMH) algorithm (Tierney, 1994 and Chib and Greenberg, 1995), are rarely discussed or used in the DSGE context. One reason for this neglect is surely that the RWM algorithm is the only algorithm available in the user friendly software package DYNARE, but there are more substantial motives. Single-block independence samplers can get stuck for long spells when the parameter space is high-dimensional, and this is indeed also the case in DSGE models. As an example, we generated posterior samples of $1,000,000$ draws from the posterior of the DSGE model in ALLV using both the IMH and the ARMH algorithm[2]. The estimated percentiles of the marginal posteriors from the three different algorithms are very similar. Plots of the draws for the domestic sticky price parameter for all three algorithms are displayed in Figure 2, where it is clearly seen that the two independence samplers occasionally get stuck for long spells. This happens when the posterior-to-proposal ratio, which is what essentially determines the probability of a move, is large in some regions of the parameter space. One possible explanation for excessive posterior-to-proposal ratios may be that the posterior has prolongued tails or ridges (weak identification is some dimensions) which are not adequately captured in the proposal density[3]. The non-movement of the independence chains will of course inflate the inefficiency factors, but these are still large when the longest spells are excluded. For example, the average inefficiency factor based on the IMH subsample with iterations $550,000$-$650,000$ is 199.16. Contrary to the RWM case, the large inefficiency factors in the IMH case cannot be explained by the dimensionality of the parameter space. The inefficiency factors are roughly 7

[1] The superimposed normal approximation based on the diagonal elements of the Hessian matrix indicates that the posterior standard deviations based on the inverse Hessian are slightly too large also for the transformed parameters. Although this is an admittedly limited one-dimensional view of the posterior, it does suggest that there is some room for improvement by modifying the scale component of the Hessian matrix to better fit the sliced posterior. One simple way of modifying the scale component of the Hessian is to fit a quadratic function to the univariate log posterior slices, extract the second order term to form a diagonal matrix of correction factors which is subsequently used to pre- and postmultiply the inverse Hessian. The quadratic fit to the sliced posterior are the dotted curves in Figure 1.

[2] We use a tailored multivariate-t density with $v = 10$ degrees of freedom as proposal. The mean and scaling matrix is set equal to the posterior mode and the inverse Hessian at the mode, respectively. The parametrization of the t density is such that the covariance of the proposal density is $v/(v-2)$ times the scaling matrix. We also experimented with additional scaling factors and other choices of degrees of freedom in the proposal density. The performance of the samplers varied, but all of them got stuck for a substantial number of iterations.

[3] The independence samplers got stuck at parameter draws which did not have an unusally high posterior density or a remarkably low proposal density, but where the ratio of the two was very large in comparison to the other proposed draws.

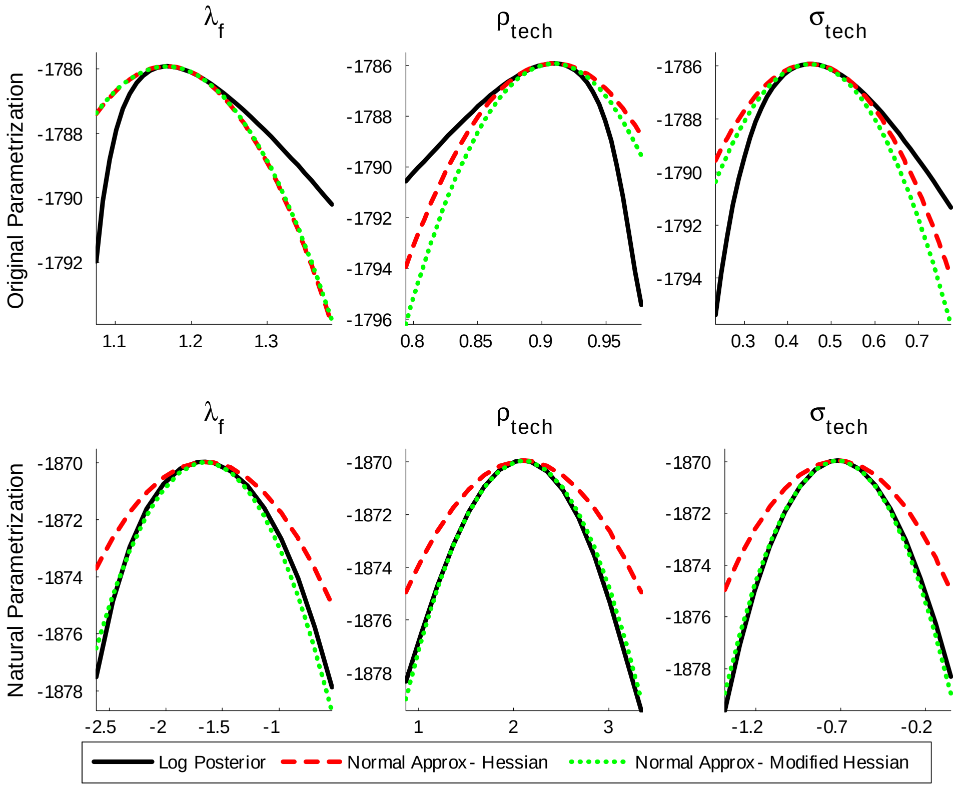

FIGURE 1. The effect of parameter transformations on the log posterior density. The first row displays slices of the log posterior (solid lines) for three different parameters in the original parametrization. The normal approximation based on the diagonal elements of the inverse Hessian are superimposed (dashed). The graphs also displays a normal approximation based on a local quadratic fit to the sliced log posterior (dotted). The second row gives the same information as the first row, but for the transformed parameters.

in the above reported experiment where the posterior density is a well-behaved 51-dimensional DSGE-mimicing multivariate t density.

Computing time is a major concern when Bayesian methods are employed to analyze large-scale DSGE models. As an example, a full posterior analysis of the log-linearized model in ALLV (numerical optimization of the posterior followed by 500,000 iterations with the one-block Random Walk Metropolis (RWM) algorithm) takes roughly 20 hours on a Pentium Xeon 3.4 GHz processor using our Matlab code. The time-consuming part of the posterior sampling, at least when used on large-scale linearized DSGE models, is iterating the Kalman filter. Specifically, the algorithm spends most of the time updating the covariance matrix of the state variables, the dominant operation being simple matrix-to-matrix multiplication. It is therefore somewhat puzzling that the authors find that Gauss outperforms Matlab, at least to the extent reported in their paper.

Marginal Likelihood Estimation

The marginal data density, or the marginal likelihood, can sometimes be difficult to estimate, especially in highly non-linear models with high-dimensional parameter spaces, like the ones at the frontier of DSGE analysis. An and Schorfheide compare the numerical stability of

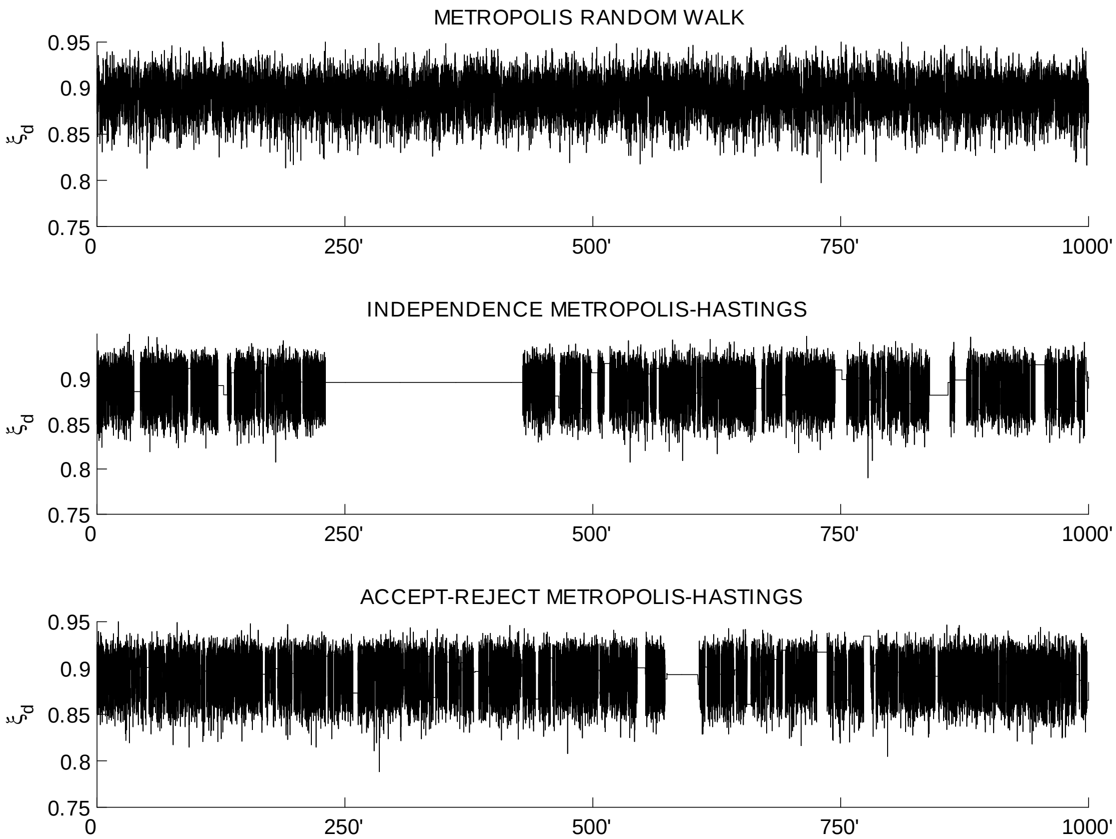

FIGURE 2. Sampling paths for the domestic sticky price parameter for three different posterior sampling algorithms.

two different marginal likelihood estimators (modified harmonic and Chib-Jeliazkov) in their stylized model. We shall here take the opportunity to complement their findings from the perspective of large-scale models. We consider the three above mentioned methods for posterior sampling: RWM, IMH and ARMH. For each of these posterior sampling algorithms we investigate the performance of three marginal likelihood estimators: modified harmonic mean (MHM), Chib-Jeliazkov (CJ) and the importance weights (IW) estimator. The MHM and CJ estimators are discussed by An and Schorfheide. The IW estimator is simply the mean of the importance weights (the ratio of the posterior and proposal densities for the proposed parameter draws). With independent proposal draws the IW estimator is equivalent to the usual importance sampling estimator of the marginal likelihood first used in Geweke (1989b). The IW estimator has been generalized to the Metropolis-Hastings environment by Geweke (1999a, Theorem 4.1.2). The normalizing constant of the ARMH proposal is unknown and the IW estimator is therefore not applicable for that algorithm. The original CJ estimator was developed for the Metropolis-Hastings case, but a similar estimator has been proposed in Chib and Jeliazkov (2005) for the ARMH case. We illustrate the behavior of the estimators in the model in ALLV. In the IMH and the ARMH case we used the same tailored multivariate-t density as in the previous section. We use a scaling factor of 0.1 in the RWM case.

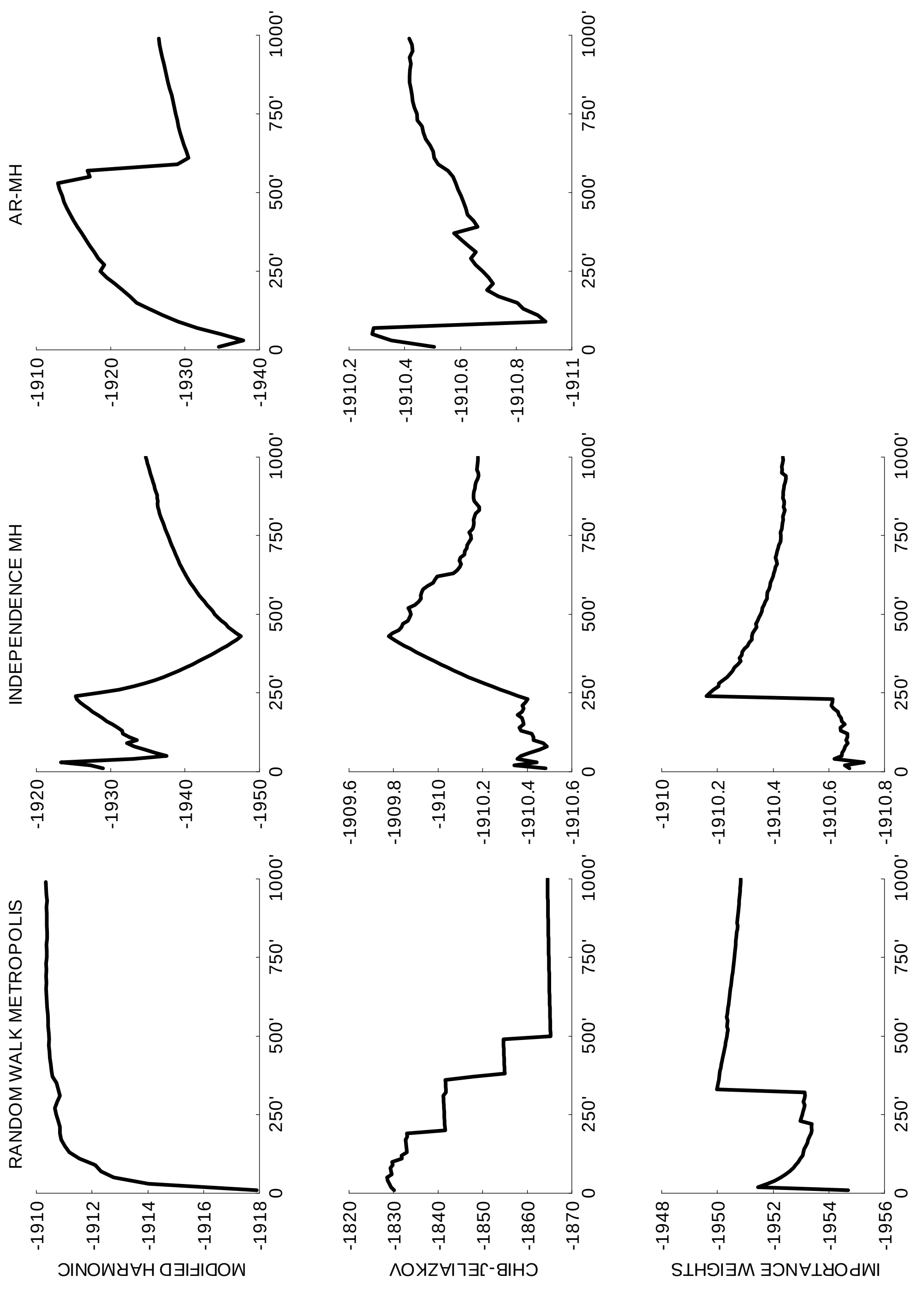

FIGURE 3. Sequential marginal likelihood estimates from the large-scale DSGE model for different combinations of posterior sampling algorithms and marginal likelihood estimators.

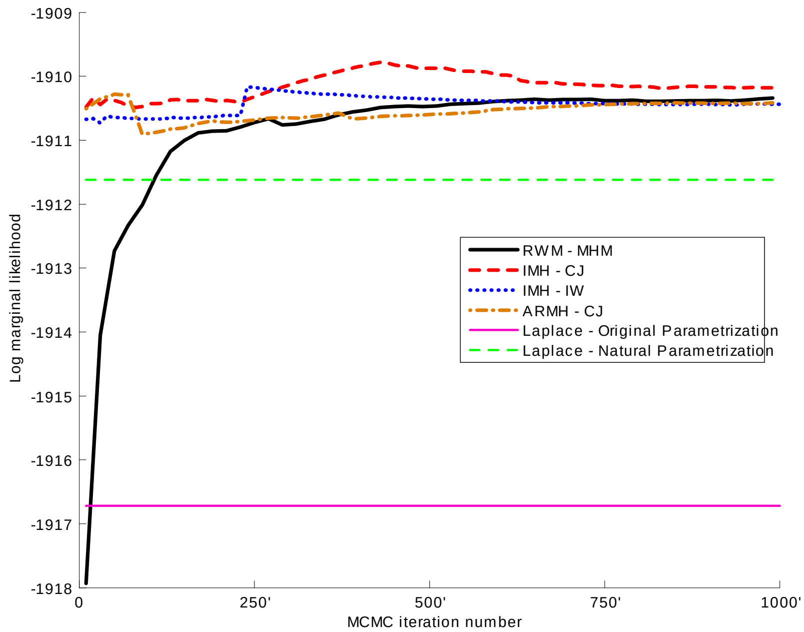

FIGURE 4. Sequential marginal likelihood estimates from the large-scale DSGE model for a subset of marginal likelihood estimators.

Figure 3 reports the sequentially estimated marginal likelihoods for each combination of posterior sampler and marginal likelihood estimator using different scales in each subgraph, while Figure 4 displays a subset of the results from Figure 3 on the same scale (Figure 4 also displays Laplace approximations of the marginal likelihood, which we discuss later on). The results are striking. In the RWM case, only the modified harmonic estimator seems to work well. The odd behavior of the CJ estimator in the RWM case is explained in detail below. Turning to the IMH case, the performance of the three estimators is reversed: here the MHM estimator converges very slowly and is adversely affected by the long spells where the IMH sampler does not move (compare with Figure 2). The periods of non-movement also have an effect on the CJ and IW estimators, but to a much lesser extent, and the performance of both these estimators is still excellent: they are already accurate after $10,000$ post burn-in draws, see Figure 4. The same is true for the CJ estimator in the ARMH, and also here is the MHM estimator thrown off track when the ARMH sampler does not move.

The main result in Figure 3 is that the MHM estimator performs well for the RWM case, but is poor for the two independence samplers. The opposite is true for the CJ and IW estimators. To see if this finding is peculiar to the specific DSGE application, we conduct the following experiment. We assume that the true posterior density is proportional to the kernel of a 10-dimensional multivariate t density with a zero mean, identity scale matrix and 50 degrees of freedom. We generate 5,500 draws using the RWM algorithm and subsequently compute the three marginal likelihood estimators on the 5,000 last draws in the sample. This procedure is repeated a thousand times for several different scaling factors in the tailored proposal density. Figure 5 displays the median marginal likelihood estimate over the thousand replications and the associated 90% numerical error bands. The true marginal likelihood, which is given by the inverse of the normalization constant of the multivariate t, is indicated by the horizontal dashed line. The mean acceptance rate decreases monotonically from 65% down to 24% as the scaling factor increases (not shown). When the scaling factor is small, the inefficiencies are

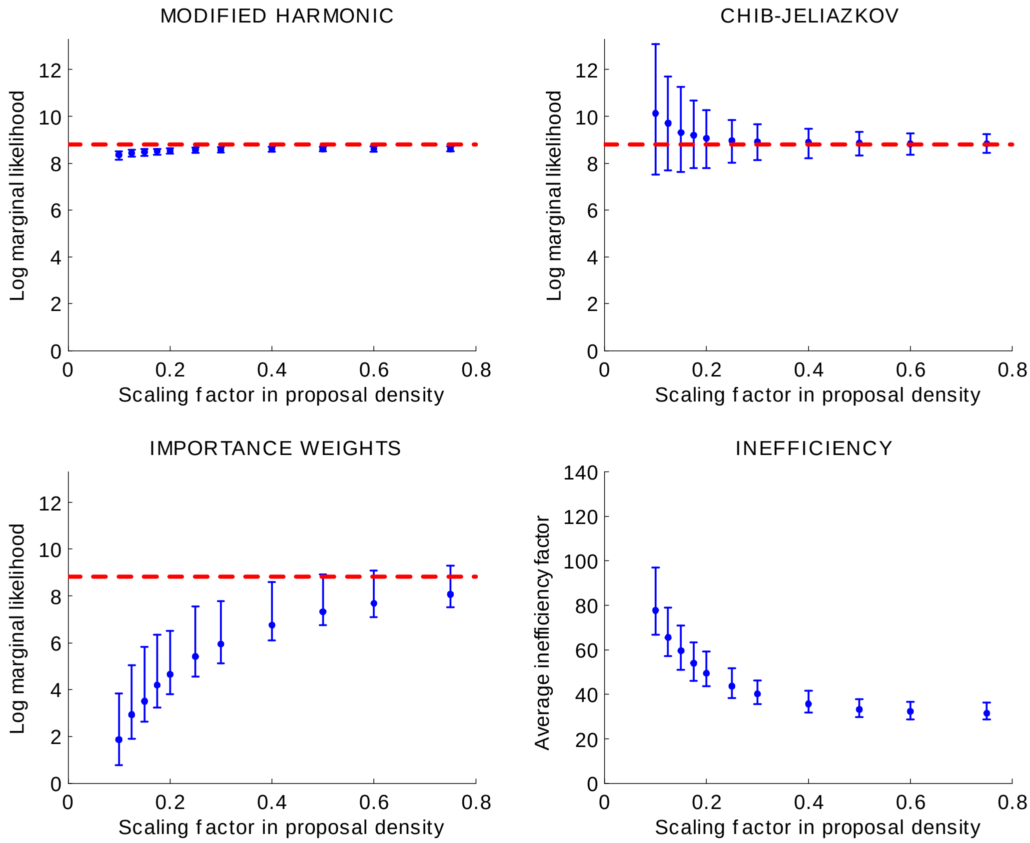

FIGURE 5. Performance of marginal likelihood estimators as a function of the scaling factor c^2 in the RWM algorithm. The target posterior is the kernel of a 10-dimensional multivariate t density with 50 degrees of freedom, zero mean and identity scale matrix. The true log marginal likelihood (the dashed line) is 8.812.

large and the MHM clearly outperforms the CJ and IW estimators. Chib and Jeliazkov (2001) made the point that the variability of CJ increases with the inefficiency factors. As the scaling factor increases, the CJ estimator becomes more accurate, but still has a larger numerical variability than the MHM estimator. However, the bias in the CJ estimator disappears more quickly that that of the MHM estimator as the efficiency increases. The IW estimator is not competitive for any scaling factor. These results confirm the results from the DSGE model where the MHM estimator was the best estimator in the RWM case.

Turning to the same experiment for the IMH case in Figure 6, we see that the roles are here reversed. The CJ estimator, and in particular the IW estimator, outperforms the MHM estimator uniformly over the scaling factor. The MHM estimator is particularly poor when the inefficiencies are large[4]. These results are again consistent with the findings in the DSGE model.

The different behavior of the MHM estimator in the RWM, on the one hand, and the IMH and ARMH on the other, are to a large extent explained by the long spells of non-movement

[4]The mean acceptance rate is 8% for the smallest scaling factor. It then increases with the scaling factor up $c^2 = 1$, reaching a maximal value of 81%, and then decreases to 6% for the largest scaling factor.

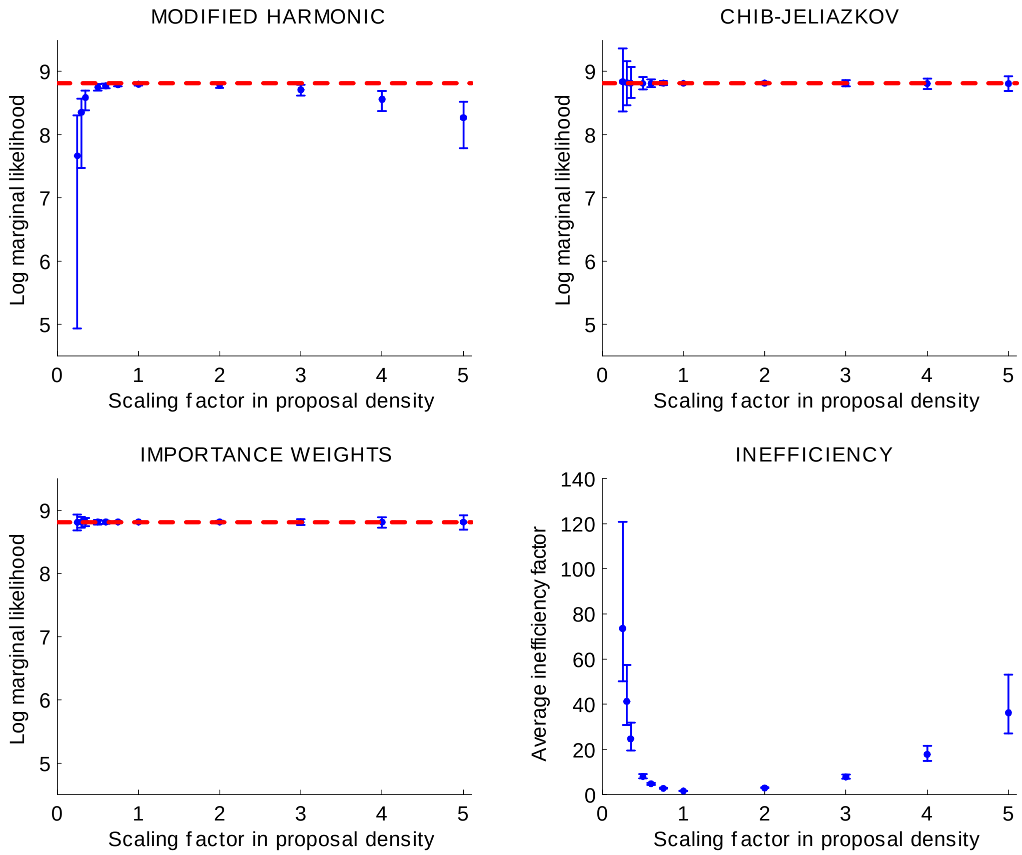

FIGURE 6. Performance marginal likelihood estimators as a function of the scaling factor c^2 in the tailored multivariate t proposal density with 10 degrees of freedom in the IMH sampler. The target posterior is the kernel of a 10-dimensional multivariate t density with 50 degrees of freedom, zero mean and identity scale matrix. The true log marginal likelihood (the dashed line) is 8.812.

in the independence samplers. A dissemination of the components of the CJ estimator in the RWM case and the IMH cases gives insight into the radically different behavior of the CJ estimator in these two cases.[5] Setting the evaluation point in the CJ estimator to the posterior mode $\tilde{\theta}$ implies in the RWM case that $\alpha(\theta^{(m)}, \tilde{\theta}|Y) = 1$ for all $\theta^{(m)}$, and the posterior ordinate estimate in equation (53) in An and Schorfheide's paper simplifies to

$$\frac{n_{sim}^{-1} \sum_{s=1}^{n_{sim}} q(\theta^{(s)}, \tilde{\theta}|Y)}{J^{-1} \sum_{j=1}^{J} \alpha(\tilde{\theta}, \theta^{(j)}|Y)}, \tag{1}$$

where the average in the numerator is based on the draws from the posterior and the average in the denominator is based on the draws from the proposal density $q(\tilde{\theta}, \theta^{(j)}|Y)$, with $\tilde{\theta}$ kept fixed. In the case of the IMH sampler we have $q(\theta^{(s)}, \tilde{\theta}|Y) = q(\tilde{\theta})$ and the posterior ordinate

[5] We have verified that our code is correct by replicating the calculations in Example 3.1 in Chib and Jeliazkov (2001). We are grateful to Ivan Jeliazkov for providing the data.

estimate can be written

$$q(\tilde{\theta})\frac{n_{sim}^{-1}\sum_{s=1}^{n_{sim}}\alpha(\theta^{(s)},\tilde{\theta}|Y)}{J^{-1}\sum_{j=1}^{J}\alpha(\tilde{\theta},\theta^{(j)}|Y)}. \tag{2}$$

A major difference between the RWM in (1) and IMH case in (2) is the presence of the average

$$n_{sim}^{-1}\sum_{s=1}^{n_{sim}} q(\theta^{(s)},\tilde{\theta}|Y) \propto n_{sim}^{-1}\sum_{s=1}^{n_{sim}} \exp[-\frac{1}{2}(\theta^{(s)}-\tilde{\theta})'(c^2\tilde{\Sigma})^{-1}(\theta^{(s)}-\tilde{\theta})] \tag{3}$$

in the RWM case. This average is also lacking in the ARMH-CJ estimator, which explains its similarity to IMH-CJ. To see the effect of this term, let us assume that the posterior distribution is normal and that the inverse Hessian $\tilde{\Sigma}$ is a good approximation of the posterior covariance matrix. It then follows that $z_\theta = (\theta-\tilde{\theta})'(c^2\tilde{\Sigma})^{-1}(\theta-\tilde{\theta}) \sim Gamma(p/2, 2/c^2)$, such that $\mathrm{E}(z_\theta) = p/c^2$ and $\mathrm{Var}(z_\theta) = 2p/c^4$, where p is the dimension of θ. z_θ will thus have a large mean and variance when the dimensionality of θ is large and the scaling constant in the RWM algorithm is small. Thus, for a high-dimensional θ, the CJ estimate in the RWM case is likely to be dominated by the few posterior draws closest to $\tilde{\theta}$[6]. This is exactly what happens in Figure 3, where a handful of influential draws cause the sequential estimates to make abrupt jumps. The behavior of the CJ estimator in Figure 10 in An and Schorfheide's paper displays a little bit of the same jumpy behavior, but, because of the much simpler model with a lower-dimensional parameter space, it is not nearly of the same magnitude as the jumps here. Note also that it is generally not possible to increase c to reduce the variance of z_θ since that will quickly drive the RWM acceptance rate to zero in high-dimensional parameter spaces. It should also be clear that it is extremely important to discard a burn-in sample before applying the CJ estimator in the RWM case if the MCMC chain is initialized in the neighborhood of the high density point $\tilde{\theta}$. This is because once the MCMC chain moves away from $\tilde{\theta}$, it is very unlikely to return to a point *sufficiently* near $\tilde{\theta}$ (with respect to the z_θ-metric), even though $\tilde{\theta}$ is by definition located in a high-density region, and the first few draws are likely to dominate the marginal likelihood estimate.

It should also be noted that the CJ estimator requires draws both from the posterior and the proposal distribution, conditional on the posterior mode. The additional draws from the proposal distribution, which are different from the ones generated in the course of the posterior sampling, are not time-consuming *per se*, but in order to compute $\alpha(\tilde{\theta},\theta^{(j)}|Y)$ in (1) the likelihood function needs to be evaluated for each such draw, which is very computationally expensive in DSGE models. With $n_{sim} = J$, this means twice as many likelihood evaluations, and therefore nearly twice the computing time. This is not as bad as it may seem at first sight, though. First, when independence samplers are used one may directly re-use the proposal draws and the corresponding likelihood evaluations obtained throughout the posterior sampling. Thus, in this situation, the CJ estimate is available directly after the completion of the posterior sampling. Second, in the RWM case it possible to transform the already generated proposal draws to draws from $q(\tilde{\theta},\theta^{(j)}|Y)$ by reweighting[7]. This would allow us to

[6]One way to see this is by observing that $z_{\tilde{\theta}} = 0$ and that

$$\Pr(z_\theta \le h) = 1 - e^{-hc^2/2}\sum_{y=0}^{p/2-1}\frac{(hc^2/2)^y}{y!}.$$

When p is large and c is small $\Pr(z_\theta \le h)$ is essentially zero for all h which are sufficiently small to have an impact on (3). The expression for $\Pr(z_\theta \le h)$ holds when p is even, and can be derived from a well-known relationship between the Gamma and Poisson distributions.

[7]We are grateful to Ivan Jeliazkov for this suggestion.

estimate the denominator of (1) without re-evaluating the likelihood function for each draw from $q(\tilde{\theta}, \theta^{(j)}|Y)$. Finally, our experience is that the denominator in (1) converges much faster than the numerator in the RWM case, which therefore suggest setting $J \ll n_{sim}$.

To sum up, using the RWM algorithm in high-dimensional spaces typically requires a small scaling factor in the proposal density in order to get a decent acceptance rate. The combination of a high-dimensional parameter space and a small tuning factor tends to make the CJ estimator inefficient. The MHM appears to be more reliable for this situation, whereas the IW estimator does really poorly. When independence samplers are used in high-dimensional spaces they tend to get stuck for long spells. This has a severe adverse effect on the MHM estimator, but only a minor effect on the CJ and IW estimators, which are both rapidly converging to the true value (assuming this to be the point, within a decimal, where half of the estimators in Figure 3 converge to). The latter fact suggests that even though independence samplers tend to get stuck for long spells, they are useful, in combination with CJ or IW, to get a quick preliminary estimate of the marginal likelihood. This is important in the model building process where such an estimate can be used as a screening device to detect bad models without generating a full posterior sample. It does not seem unreasonable to assume that additional fine tuning of the ARMH algorithm can make it competitive to the RWM algorithm, and then the ARMH-CJ estimator would be a natural choice for marginal likelihood estimation.

An alternative preliminary estimate of the marginal likelihood is given by the Laplace approximation method. This estimate of the marginal likelihood uses only the posterior mode and Hessian matrix, and is therefore available directly after the completion of the optimization stage. The Laplace approximation is accurate when the posterior density is close to normal, so it is expected that re-parametrization will matter here. This is indeed the case in the ALLV model. Figure 4 displays the Laplace estimates of the marginal likelihood in both the original and the natural parametrization, where it is clear that only in the natural parametrization is the Laplace estimate accurate enough to be useful as a screening device.

Additional references

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm, *American Statistican* **49**, 327–35.

Chib, S. and Jeliazkov, I. (2005). The accept-reject Metropolis-Hastings algorithm and marginal likelihood estimation. *Statistica Neerlandica,* **13**, 339-368.

Geweke, J. (1989b). Exact predictive densities in linear models with ARCH disturbances, Journal of Econometrics **40**, 63-86.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *Annals of Statistics* **22**, 1701–1762.