Discussion on "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations" by H. Rue, S. Martino, and N. Chopin,

Christian P. Robert, CEREMADE, Université Paris Dauphine and CREST, INSEE

Rue, Martino and Chopin are to be congratulated on their impressive and wide-ranging attempt at overcoming the difficulties in handling latent Gaussian structures. In time series as well as spatial problems, the explosion in the dimension of the latent variable is indeed a stumbling block for MCMC implementations and convergence, and recent solutions all resort to approximations of sorts whose impact was not always completely assessed (see, e.g., Polson et al., 2008). The double Laplace approximation proposed by the authors is therefore an interesting and competing alternative in these areas.

Nonetheless, as much as I respect the other major contributions to (Bayesian) Statistics, Mathematics, and other branches of Science, of my fellow Norman, Pierre Simon de Laplace, and despite earlier uses in Bayesian Statistics (Tierney et al., 1989), I have always (Robert, 1992, 1994) been wary of Laplace's approximation because

- it is not parameterisation invariant,
- it requires some analytic computation or/and some black-box numerical differentiation, while being based on a standard second-order Taylor approximation to the log-density, and
- it misses a clean evaluation of the associated error.

In the present paper by Rue *et al.*, the amount of recourse to this approximation is particularly intense since both $\pi(\mathbf{x}|\theta,\mathbf{y})$ and $\pi(x_i|\theta,\mathbf{x}_{-i},\mathbf{y})$ are approximated by multi-level (nested) Laplace approximations. I however contend that a less radical interpretation of the approximation (3) as a proposal could lead to an manageable MCMC implementation, at least in some settings.

My first reservation is that the calibration of those Laplace approximations seems to require a high level of expertise that conflicts with the down-the-shelf features advertised in the final section of the paper. Designing the approximation then represents an overwhelming portion of the analysis time, while computation indeed becomes close to instantaneous, unless one abandons the analytical derivation for a numerical version that is difficult to entirely trust. After reading the paper, I certainly feel less than confident in implementing this type of approximation, although I did not try to use the generic Open Source software carefully developed by Rue and Martino. Attempting at applying this approach to the standard stochastic volatility model using our own programming thus took us quite a while, even though it produces decent approximations to the marginal $\pi(\theta|\mathbf{y})$ (Casarin and Robert, 2008). I am however wondering whether or not the approximation (3) is a proper density for any and every model, since it writes as $|\mathbf{Q}(\theta)|^{1/2}|\mathbf{Q}(\theta)+\text{diag}(c(\theta))|^{-1/2}\,\pi(\mathbf{x}^\star(\theta),\theta,\mathbf{y})$, where the dependence of both $\mathbf{x}^\star$ and c on θ is quite obscure.

My second reservation is that, while the pseudo-marginal (3) seems to be an acceptably manageable version of the marginal posterior of θ, the additional Laplace approximations in the derivation of the marginals of the x_i's do not appear as crucial or necessary. Indeed, once (3) is available as a (numerical) approximation to $\pi(\theta|\mathbf{y})$, given that θ has a limited number of components, as hinted at on several occurrences of the paper, a regular MCMC strategy targeted at this distribution is likely to work. This would result in a much smaller cost than the discretisation underlying Figure 1 (which cannot resist the curse of dimensionality, unless cruder approximations as in Section 6.5 are used, but at a cost in accuracy). Once simulations of the θ's are available, simulations of

the $\mathbf{x}$'s can also be produced using the Gaussian approximation as a proposal and the true target as a Metropolis-Hastings correction. (It is thus possible to envision the whole simulation of the pair $(\theta, \mathbf{x})$.) Indeed, the derivation of $\tilde{\pi}(x_i|\theta, y)$ is awfully complex and thus difficult to assess. In particular, the construction is partly numeric and must be repeated for a large number of θ's, even though I understand this does not induce high computing costs. Given that the $\tilde{\pi}(x_i|\theta, y)$'s are averaged over several values of θ, it somehow is delicate to get convinced that this complex construction is worthwhile, when compared with a original Gaussian approximation coupled with an MCMC simulation. Simulating a single x_i or the whole vector $\mathbf{x}$ from the Gaussian approximation has the same precision (in x_i), therefore the dimension of $\mathbf{x}$ cannot be advanced as having a negative impact on the convergence of the MCMC algorithm. Furthermore, a single run of the chain produces approximations for all x_i's.

My last reservation is that the error resulting from this approximation is not, despite the authors' efforts, properly assessed. We can indeed check on the simpler examples that the error resulting from one of the schemes is indeed minimal but the $\mathcal{O}(n_d^{-1})$ error rates do little for my reassurance as (a) they involve the sample size, even though we are dealing with a fixed sample, and not a computational effort that seems to be restricted to the size of the grid in the θ space and (b) they do not calibrate the error resulting from the Bayesian inference based on this approximation. Using and comparing different approximations all based on the same principle (Laplace's!) does not provide a clear indicator of the performances of those approximations. Furthermore, resorting to Spiegelhalter et al. (2002) measure of effective dimension does not bring an additional validation to the approximation.

A minor side remark is that, from a Bayesian point of view, I find rather annoying that latent variables and parameters or hyperparameters with Gaussian priors are mixed together in $\mathbf{x}$ (as in the stochastic volatility example) and that θ coalesces all leftovers without paying any attention to the model hierarchy (as in Section 1.3 with θ_1 versus θ_2). Of course, this does not impact the sampling performances of the method, but it still feels awkward. In addition, this may push towards a preference for Gaussian priors, since the more (hyper-) parameters with Gaussian priors, the smaller m and the less costly the numerical integration step.

Given the potential advances resulting from [as well as the challenges posed by] this paper, both in terms of modelling and of numerical approximation, I am unreservedly glad to second the vote of thanks.

References

Casarin, R. and Robert, C. (2008). Discussion of "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations" by H. Rue, S. Martino, and N. Chopin. *J. Royal Statist. Society Series B*. (To appear.).

Polson, N., Stroud, J., and Müller, P. (2008). Practical filtering with sequential parameter learning. *J. Royal Statist. Society Series B*, 70(2):413–428.

Robert, C. (1992). *L'Analyse Statistique Bayésienne*. Economica, Paris.

Robert, C. (1994). *The Bayesian Choice*. Springer-Verlag, New York.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Royal Statist. Society Series B*, 64(2):583–639.

Tierney, L., Kass, R., and Kadane, J. (1989). Fully exponential Laplace approximations to expectations and variances of non-positive functions. *J. American Statist. Assoc.*, 84:710–716.