Christian P. Robert *(CREST and CEREMADE, Université Dauphine)*

This paper aims at developing general strategies for improving jumps between models in reversible jump Markov chain Monte Carlo (RJMCMC) algorithms, which is quite an important and timely goal. Indeed, in practical implementations of the method, one usually finds that the choice of proposals is paramount: in many cases, the "natural choice" leads to a zero acceptance probability and the construction of well-tuned moves is often quite costly. Given that the RJMCMC method is an essential part of the Bayesian toolbox, at least in Bayesian exploratory analysis, a debate is called for more global strategies on the choice of proposals.

A first appealing feature, at the core of the paper, is that image parameters that give a Metropolis–Hastings probability of *one* should be identified as *pivotal quantities,* just like the current value is a pivot for the random walk Metropolis–Hastings move. The authors then propose "higher order" methods where some derivatives of the probability are set to zero, but I find this less appealing, because it considerably adds to the complexity of the algorithm.

Obviously, the authors mostly focus on *scale,* rather than *location-scale* tuning. The choice of the centering function $c_{i,j}(\theta_i)$ is not so much discussed further in the paper, even in the case of nested models. For instance, the resolution of

$$L_i(\mathrm{data}|\theta_i) = L_j\{\mathrm{data}|f_{i,j}[\theta_i, b_{i,j}(\theta_i)]\},$$

suggested in §2,2, usually is intractable. Moreover, by considering only the likelihood, it may provide values with very small prior probabilities. In addition, the moves between *non-nested models* are not necessarily natural. Take for instance a set of generalized linear models M_{ij} including some distributions f_i and link functions, ℓ_j, such that $y|x \sim f_i(y|\ell_j(x^T\beta_{ij}))$; the parameter β_{ij} depends on the choice of both f_i and ℓ_j and the move from model M_{ij} to model M_{uv} should not be centered in β_{ij}. At the very least, one must exhibit an invariant over the models like some moments of y or the likelihood, but the resolution of the corresponding equations is likely to be costly. Similarly, weak non-identifiability does not work well in this non-nested example.

This is an opportunity to stress a general *statistical* problem with reversible jump methods, which is the systematic recourse to the *same* parameters when jumping between nested models (§2.1.1). Most authors on model choice state that, on the contrary, the interpretation of parameters *should* change between models. For instance, when considering an AR(p) and an

AR($p+1$) model, the meaning of a_1 is not the same in both models. In this setup, under stationarity, the correct parameterization is either through *partial autocorrelations,* which are rather unrelated for orders p and $p+1$, or through the [inverse] roots λ^p of the lag-polynomial

$$\prod_{i=1}^{p} (1-\lambda_i B)\ X_t = \epsilon_t\,, \tag{1}$$

as in Huerta and West (1999), which, again, do not remain similar between models and thus make the "natural" centering

$$c_{p,p+1}(\lambda^p) = (\lambda^p, 0) \tag{2}$$

questionable. Under uniform priors for the real and complex roots λ_j,

$$\frac{1}{\lfloor k/2 \rfloor + 1} \prod_{\lambda_i \in \mathbb{R}} \frac{1}{2} \mathbb{I}_{|\lambda_i|<1} \prod_{\lambda_i \notin \mathbb{R}} \frac{1}{\pi} \mathbb{I}_{|\lambda_i|<1}$$

and proposals based on the priors, a standard reversible jump algorithm does not encounter convergence problems (Figure 1). Nonetheless, if the uniform prior is replaced with a Beta type proposal,

$$\varphi(\lambda_i) \propto |\lambda_i|^\beta\,, \quad \beta > 0\,,$$

where β is to be optimised, the centering equation (8) [in the paper] is not defined when (2) above is used. (This points out a usual measure-theoretic difficulty with criteria which depend on the value of a density at a specific point and suggests a more *integrated* alternative for the centering equation.)

Another very appealing feature of the paper is to consider in parallel a kind of saturated model, called the *dual space approach*, where the parameter to be simulated is of constant dimension $n_{\max}$ by incorporation of the remaining auxiliary variables, a feature reminiscent of Carlin and Chib' (1995) and Godsill (2001). I find this representation of reversible techniques very natural. At a deeper level, it seems to me that this could lead to more general (incl. non-reversible) MCMC algorithms in variable dimension models, since moves between models are then Gibbs steps.

The idea of using the $u_{i,r}$'s as auxiliary variables is particularly exciting, in that it eliminates the apparent difficulty of moving between spaces of different dimensions. It also clarifies what happens to the $u_{i,j}$'s, when the move is not

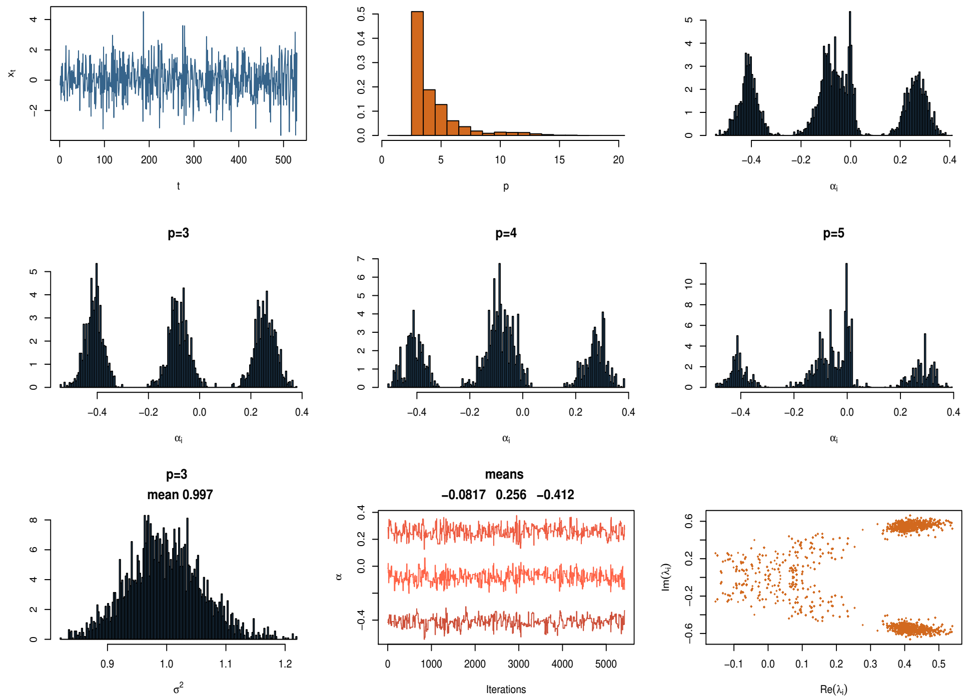

Figure 1: Graphical representation of the performances of a reversible jump algorithm for the root parameterisation of the $AR(p)$ model, based on a simulated dataset of 530 points *(upper left)* with true parameters α_i $(-0.1, 0.3, -0.4)$ and $\sigma = 1$. The first histogram gives the posterior distribution of p, the following histograms the distribution of the natural parameters α_i, for different values of p, and of σ^2. The final graph is a scatterplot of the complex roots of the lag-polynomial and the one before shows the evolution of $\alpha_1, \alpha_2, \alpha_3$.

between M_i and M_j. It however presupposes some degree of homogeneity between models in that the $u_{i,r}$'s remain (almost) the same between moves, that is, some generic feature in the $u_{i,r}$'s, like uniformity, and thus a more elaborate construction of the transforms $f_{i,j}$. This is particularly true for (dimension) correlated proposals. For instance, while the correlated auxiliary variable methods are performing better than the "vanilla" algorithm in the mixture example, it is quite difficult to fathom why this is so and this seems to indicate that the proposed methods require harder tuning than alluded in the paper. (The comparison between the three methods in §9 definitely is unfair, in that the truncation on $k \leq 5$ seems to invalidate the "vanilla" algorithm, while, in Richardson and Green (1997), mixing was quite satisfactory. As shown in Figure 2, using a C program of my own, the truncation of k at $k \leq 5$ does not appear to hamper convergence in such a dramatic way.)

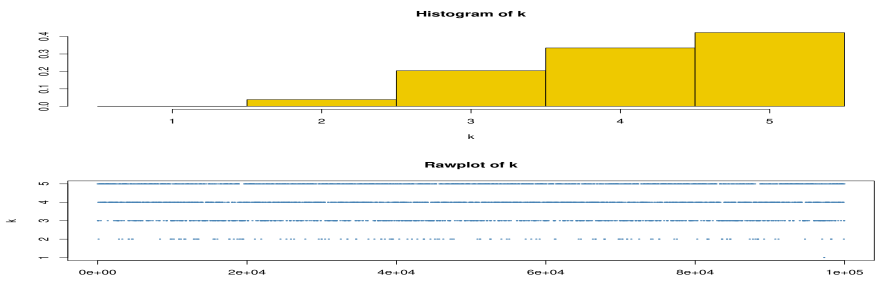

Figure 2: Histogram and rawplot of a sequence of $100,000$ k produced by Richardson and Green's (1997) algorithm for the enzyme dataset, under the imposed constraint $k \leq 5$.

This paper offers a well-thought reflection on the reversible jump technology, in particular by rephrasing the moves in terms of a center in the arrival space, contains path-breaking innovations in designing reversible jump proposals, opens new and broad avenues for designing reversible jump algorithms, while calling for further research in producing generic proposals. **I thus have great pleasure in proposing the vote of thanks!**

References

Huerta, G. and West, M. (2000) Bayesian inference on periodicities and component spectral structure in time series. *J. Time Series Analysis* (to appear).