BAYESIAN STATISTICS 7, pp. 307–326
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Non-Centered Parameterisations for Hierarchical Models and Data Augmentation

OMIROS PAPASPILIOPOULOS GARETH O. ROBERTS
Lancaster University, UK
o.papaspiliopoulos@lancaster.ac.uk g.o.roberts@lancaster.ac.uk

MARTIN SKÖLD
Australian National University, Australia
martin.skold@anu.edu.au

SUMMARY

In this paper, we will compare centered and non-centered parameterisations for classes of hierarchical models. Our examples will include variance component models, random effect models, hidden Markov process models, and partially observed diffusion models. We will investigate the construction of non-centered methods by the use of state space expansion techniques, and will introduce methods for devising partially non-centered parameterisations, many of which are data-dependent.

Keywords: PARAMETERIZATION OF HIERARCHICAL MODELS; MISSING DATA PROBLEMS; CENTRED AND NON-CENTRED PARAMETRIZATIONS.

1. INTRODUCTION

For at least the last 15 year or so, hierarchical models in various guises have revolutionised Bayesian statistical methodology. Their key advantages have been their flexibility, interpretability and the ease of inference using MCMC techniques, see for example Smith and Roberts (1993), Gilks *et al.* (1994). The simplest possible hierarchical model can be described by the directed graphical model in Figure 1a, where θ is a generic collection of hyperparameters, Y represents the observed data, and X can take the role of population parameters, missing data, hidden Markov process, or various other possibilities. Therefore even this very simple setup encompasses a huge diversity of model types and statistical contexts. Most examples we consider here can be loosely described by Figure 1a, though there will often be other parameters present in the model not explicitly represented there. Convergence of MCMC methods, particularly when using the Gibbs sampler or related techniques, depends crucially on the parameterisation used for the unknown quantities. From a modeling and interpretation perspective, the natural *centered parameterisation* (denoted in this paper by CP) for Figure 1a is to use just $\theta,\ X$. This utilises the conditional independence inherent within the model which makes updating θ computationally less challenging in general. This is particularly true where conditional conjugacy is present, as is often the case by design. Thus an algorithm for sampling from the joint distribution of θ, X, which we will call parameters and missing data respectively, given the observed data Y might proceed by alternating between:

1. update θ from a Markov chain with stationary distribution $\theta \mid X$
2. update X from a Markov chain with stationary distribution $X \mid \theta, Y$ (1)

Algorithms like (1) will be called *Hastings-within-Gibbs* algorithms, the assumption being that each update is carried out using an appropriate Metropolis-Hastings update which preserves the relevant conditional distribution.

However, according to Figure 1a, θ and X exhibit *a priori* dependence and in many contexts this dependence is very strong. The presence of data tends to diminish the magnitude of that dependence, but the efficiency of the above successive substitution scheme will depend crucially on the extent to which this is the case, as we shall see in Section 2.

On the other hand, we might be able to find an alternative parameterisation, $(X, \theta) \rightarrow (\tilde{X}, \theta)$, of the model in Figure 1a, i.e where the new missing data $\tilde{X}$ is some function of the previous missing data X and the parameters θ, such that $\tilde{X}$ is *a priori* independent of θ. This type of reparameterisation of the hierarchical model in Figure 1a, is called *non-centered parameterisation* (denoted by NCP) and the corresponding graphical model is given in Figure 1b. The MCMC algorithm corresponding to (1) for simulating from the posterior distribution of $(\tilde{X}, \theta)$ iterates

1. update θ from a Markov chain with stationary distribution $\theta \mid \tilde{X}, Y$
2. update $\tilde{X}$ from a Markov chain with stationary distribution $\tilde{X} \mid \theta, Y$ (2)

and the motivation behind the NCP is that in many contexts the convergence properties of (2) might be better than those of (1). Notice that where the conditional distributions in the above algorithms can be sampled directly, we shall refer to them as the *Gibbs sampler*.

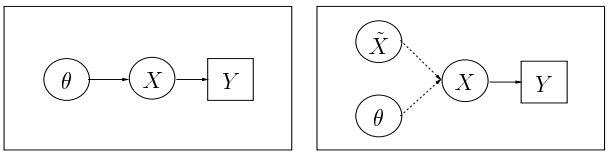

Figure 1. *(a): Graphical model of the centered hierarchical parameterisation (CP), (b): Graphical model of the non-centered hierarchical parameterisation (NCP).*

1.1. *To Center or not to Center?*

NCPs have been used in many contexts dating back to the introduction of data-augmentation in Tanner and Wong (1987) and probably well before. On the other hand, Gelfand *et al.* (1996) give strong arguments for the adoption of CPs particularly for simple classes of hierarchical structures such as random effects models. Moreover, the conditional independence of θ and Y given X often means that Gibbs sampling can be implemented for the CP but not for the NCP, leading to a significant computational edge in favour of the CP. So is there any need to consider non-centered approaches at all?

In this paper, we shall argue that there is an important role for the NCP in many contexts. It provides a general reparameterisation strategy to improve convergence of MCMC in cases where the latent process is relatively weakly identified by the data, since the prior independence between θ and $\tilde{X}$ will then ensure weak posterior dependence. We shall consider the construction of NCPs, and also classes of parameterisations which lie on continua between the CP and the NCP, so called partially non-centered parameterisations (PNCP). The first examples of the use of PNCP were introduced (originally in the context of the EM algorithm but subsequently for MCMC too) by Meng and van Dyk (1997, 1999).

Our work here should be taken as complementary to the marginal augmentation techniques introduced in Meng and van Dyk (1999), Liu and Wu (1999). The marginal augmentation technique can be superior to both the CP and NCP in context where it can be applied, although it relies on implementation strictly by Gibbs sampling whereas our methodology is designed to be implemented in conjunction with Hastings-within-Gibbs strategies.

One problem with the NCP is the requirement of orthogonality between $\tilde{X}$ and θ. In many models this can be hard to achieve in practice, and this is therefore a major limitation on the use of non-centered methods. We will introduce state-space expansion techniques in this paper that allow easy implementation of the NCP in a wide variety of stochastic models.

Our three major examples will particularly emphasise the use of the CP, the NCP and the PNCP in the context where X is a hidden stochastic process. In our examples in Section 5, X either represents a Markov chain (or process) or in the geostatistical example of Section 5.1, X is a Gaussian random field.

1.2. *Example: Hierarchical Linear Models*

A toy example that serves to illustrate the main ideas in this paper and which is totally understood theoretically, is the following Normal hierarchical model (NHM), written as

$$\begin{aligned} Y_{ij} &= X_i + \sigma_y \epsilon_{ij},\ j = 1, \ldots, n \\ X_i &= \theta + \sigma_x z_i,\ i = 1, \ldots, m \end{aligned} \tag{3}$$

This model has also been used for pedagogical purposes in Liu and Wu (1999). Here ϵ_{ij} and z_i are independent standard normal random variables, θ is assigned a uniform improper prior and the variances are considered known for the time being at least. Due to the sufficiency of $\sum_j Y_{ij}/n$, which is again Gaussian, there is no loss of generality in assuming a single observation per random effect X_i and therefore from now on we will take $n = 1$ and drop the j subscript. The parameterisation $(\theta, X),\ X = (X_1, \ldots, X_m)$ is known as the *centered parameterisation* (CP), see Gelfand *et al.* (1995), and depicted graphically in Figure 1a illustrating the independence between θ and $Y = (Y_1, \ldots, Y_m)$ conditional on X.

The name *non-centered parameterisation* was originally used for the NHM, see Gelfand *et al.* (1995). In this context the NCP writes the model as

$$\begin{aligned} Y_i &= \tilde{X}_i + \theta + \sigma_y \epsilon_{ij} \\ \tilde{X}_i &= \sigma_x z_i,\ i = 1, \ldots, m. \end{aligned} \tag{4}$$

Notice that $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_m)$ and θ are *a priori* independent (see Figure 1b) but conditional on the data, they are dependent.

Gibbs sampling can be applied very easily in this context using either the CP or the NCP and it is of interest to know which gives the most rapidly convergent sampler. Fortunately, in this simple context, this question has a definitive answer as we shall see in Section 2.1. However, the result depends explicitly on the relative sizes of σ_x^2 and σ_y^2. We shall also investigate whether there are alternative parameterisations which are capable of improving on both the CP and NCP. Again in this simple context (but also for a very general form of (3)), we can give a definitive and positive answer to this - there always exists a *partially non-centered parameterisation* (PNCP) which gives more rapid convergence than either of these two alternatives.

1.3. *Summary of Paper*

In this paper we will introduce, analyse and apply general forms of NCPs. In Section 2, we shall describe existing theory on rates of convergence for the Gibbs sampler which allows us

to compare different parameterisation schemes, at least in the Gaussian context, and apply this theory to the linear Gaussian model example introduced in Section 1.3. Section 3 introduces NCPs in more generality, including methods which involve the expansion of the state space. Section 4 considers PNCPs, while Section 5 considers three examples of the effective use of NCPs: one involving a non-linear geostatistical model; another involving the problem of inference for partially observed diffusions; while the third considers a class of stochastic volatility models currently popular in finance.

2. RATES OF CONVERGENCE OF THE GIBBS SAMPLER

Let $Z = (Z_1, Z_2)$ denote a random variable with density π, partitioned into two components, Z_1, Z_2, of arbitrary dimension. We term a 2-component Gibbs sampler on π, under the parameterisation (Z_1, Z_2), the algorithm that iterates the following procedure.

1. sample Z_1 from the conditional distribution of $Z_1 \mid Z_2$
2. sample Z_2 from the conditional distribution of $Z_2 \mid Z_1$ (5)

It is well beyond the scope of this article to discuss rates of convergence of algorithms in any detail, though see Roberts and Tweedie (2002), Jones and Hobert (2001) for recent summaries. However where the 2-component Gibbs sampler can be implemented, it admits a very complete (if not always practically useful) theory that we very briefly describe here. Let $P^n(x, \cdot)$ denote the distribution of the 2-component Gibbs sampler after n iterations, where x denotes an arbitrary starting value for the (Z_1, Z_2) pair. Amit (1991) observed that the $\mathcal{L}_2$ distance from stationarity decays as $A(x)b(n)\rho^n$ for some function $b(n)$ which varies slower than an exponential function. The constant $\rho \leq 1$ is defined as

$$\rho^{1/2} = \sup \text{Corr}(f(Z_1), g(Z_2)) \tag{6}$$

and the supremum above is taken with respect to all real-valued non-constant functions f and g which admit finite variances under π. It turns out that other common norms (such as total variation distance) can also be shown to have this rate, at least for a large class of plausible target distributions (see Roberts and Tweedie (2001)). The covariance structure of the 2-component Gibbs sampler has also been studied in detail by Liu *et al.* (1994).

It has long been recognised that the correlation structure of the target distribution determines the convergence behaviour of the corresponding Gibbs sampler, see Hills and Smith (1992), Gelfand *et al.* (1995). In the two component case, this result makes the connection precise. The characterisation (6) is of little practical use, since in general it will be impossible to evaluate the supremum in (6), although one important exception is for Gaussian π, when suprema of the kind appearing in (6) are achieved exclusively by linear functions. However, (6) has been used to compare different augmentation schemes (see for example Meng and van Dyk (1999), Liu and Wu (1999)).

For Gibbs samplers with larger numbers of components, it is not possible to find an explicit statement analogous to (6) which relates the rate of convergence of the algorithm to the target distribution correlation structure, though Amit (1991) does give some related inequalities. In the Gaussian target distribution case however, explicit formulae are available for rates of convergence of the sampler in terms of the target distribution correlation matrix, see Roberts and Sahu (1997). We shall use these results to compare parameterisation exactly for the NHM.

2.1. *Rates of Convergence for CP and NCP of the NHM*

The following results are taken directly from Roberts and Sahu (1997). We wish to sample from the joint posterior distribution of X and θ of the model (3) using the Gibbs sampler as

in (1). Since this is a multivariate Gaussian distribution we can explicitly calculate the rate of convergence, denoted by ρ_c, using the results of Roberts and Sahu (1997),

$$\rho_c = 1 - \kappa, \text{ where} \tag{7}$$

$$\kappa = \frac{(\sigma_x^2 + \sigma_y^2)^{-1}}{(\sigma_x^2)^{-1}} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2}. \tag{8}$$

The first expression for κ in (8) writes κ as the ratio of observed by augmented information for θ under the CP, defined as $1/\text{Var}(\theta \mid Y)$ and $1/\text{Var}(\theta \mid X, Y) = 1/\text{Var}(\theta \mid X)$ respectively. In this context, $1 - \kappa$ is the *Bayesian fraction of missing information* in the sense defined by Rubin (1987). The relationship between observed and augmented information and rates of convergence of algorithms was noted first in a very general framework for the EM algorithm (see Meng and van Dyk (1997)), but can be translated to the data augmentation methodology in this specialised linear model context (see Sahu and Roberts (1999)).

Therefore, the CP will perform well when $\kappa \to 1$, i.e when the data are relatively very informative in the sense that the observed data contain almost as much information about the parameter as the augmented. This result can also be expressed in terms of (6) since

$$\text{Corr}(\overline{X}, \theta \mid Y) = \sqrt{1 - \kappa}, \qquad \overline{X} = \sum_{i=1}^{m} X_i/m.$$

We return now to the NCP (4). In this case, we use the Gibbs sampler algorithm (2) and the rate of convergence, ρ_{nc}, turns out to be

$$\rho_{nc} = \kappa. \tag{9}$$

When the one parameterisation produces very slow mixing for the Gibbs sampler the other will be performing extremely well. For this model with the flat priors specified, the relative performance of the CP and NCP is explicit since $\rho_{nc} = 1 - \rho_c$. Note however that this relation does not hold when proper priors are used for θ, since in this case both algorithms have faster convergence than the rates given in (7) and (9).

2.2. *Example: State-space models*

Another class of Gaussian models that we can analyse is that of linear state space models defined in the following way.

$$\begin{aligned} Y_i &= X_i + \sigma_y \epsilon_i \\ X_i &= \phi X_{i-1} + \theta(1 - \phi) + \sigma_x (1 - \phi^2)^{1/2} z_i, \; i = 1, \ldots, m \end{aligned} \tag{10}$$

where σ_x, σ_y and $\phi \in [0, 1]$ are assumed to be known. Here the stationary moments of $X = (X_1, \ldots, X_m)$ are $\text{E}(X_i) = \theta$, $\text{Var}(X_i) = \sigma_x^2$, therefore (10) extends (3) to allow dependence among the X_is. The two-component Gibbs sampler alternates by updating θ given X, and X given θ and $Y = (Y_1, \ldots, Y_m)$, using for instance forward filtering-backward smoothing techniques as described in Carter and Kohn (1994). The NCP is derived by setting $\tilde{X}_i = X_i - \theta$.

For this example, explicit rates of convergence are easily calculable using the results of Roberts and Sahu (1997). It can be shown that as $m \to \infty$:

$$\frac{1 - \rho_{nc}}{1 - \rho_c} \approx \frac{\sigma_y^2}{\tau \sigma_x^2} \tag{11}$$

where τ is the integrated correlation time of X. A similar expression can be found in Pitt and Shephard (1999) together with considerably more detailed convergence rate analysis. Thus highly correlated hidden Markov models, with large marginal variance, favour the use of the CP over the NCP. Similar empirical results are obtained for a geostatistical model in Section 5.1.

In fact for both the CP and NCP, when X is updated as a block rather than by single-site updating, the corresponding rate of convergence does not converge to 1 for large m and thus suggests the use of forward-backward filtering as part of MCMC routines for general classes of hidden Markov processes (see Carter and Kohn (1994)) (without taking computational cost into account).

2.3. *Example: Linear non-Gaussian models*

The following toy example is very simple, but its results are quite striking. Suppose we modify the NHM (3) such that ϵ_{ij} has a standard Cauchy distribution, while z_i remains standard Gaussian. In this context, the heavy tailed nature of the Cauchy distribution makes the observation equation relatively un-informative for extremal values of X. Following the intuition gained from studying the NHM, we might expect the CP to perform poorly in some way in the tail regions in relation to the NCP. Figure 2 shows output from the Gibbs sampler for both CP and NCP (where $n = m = 1$) for different starting values for θ.

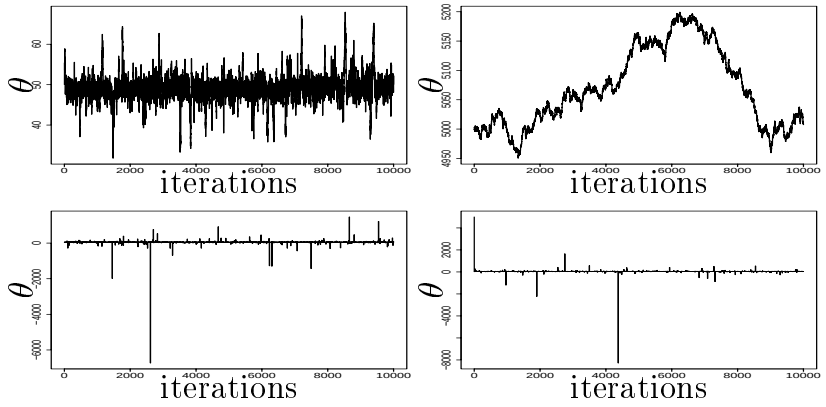

Figure 2. *Gibbs sampler output for θ in the Normal-Cauchy model described in Section 2.3 with $Y = 49.06$. Top: centered parameterisation started from $\theta_0 = 50$ (left) and $\theta_0 = 5000$ (right). Bottom: non-centered parameterisation for the same starting values. All chains were run for 10^4 iterations. Notice the different scales in the plots*

The CP exhibits unstable heavy-tailed excursions characteristic of algorithms which fail to be geometrically ergodic (see Roberts (2001)), while the NCP appears to return to the distribution mode very rapidly. In fact the following result (which we state without proof) holds. For the linear non-Gaussian model (3) with Cauchy observation equation and Gaussian hidden equation, the CP converges slower than geometrically, while the NCP converges uniformly quickly from all starting values (uniformly ergodic). This can be linked to the results obtained by Dawid (1973), where it is shown that, under the above setting, as $\theta \to \infty$, $\tilde{X}$ is independent of θ and Y and its distribution converges to the prior. Therefore, the NCP will have no difficulty making excursions far into the target distribution tails and swftly come back to the modal area. For Cauchy hidden equation with Gaussian observation error (as used for example in Wakefield *et al.* (1994)), the opposite result holds with the NCP failing to be geometrically ergodic while the CP is uniformly ergodic. Considerably more general results relating tail behaviour to qualitative convergence properties, will be found in Roberts (2002).

3. GENERAL NON-CENTERED PARAMETERISATIONS

The important feature of the NCP for the NHM that can be extracted to a much more general context, is the orthogonality of the prior structure. Specifically, we find $\tilde{X}$ which is *a priori* independent of θ and from which X can be constructed via a deterministic function

$$X = h(\tilde{X}, \theta). \tag{12}$$

In the Gaussian context under NCP (4), it is easy to identify h as $h(\tilde{x}, \theta) = \theta + \tilde{x}$. For the general hierarchical model in Figure 1a, such a function h always exists, but is not unique. However, it can be difficult to identify such a function h which is analytically sufficiently tractable to be of practical use.

The effect of this reparameterisation is described by Figure 1b. The NCP does not exhibit conditional independence between θ and the data Y conditional on $\tilde{X}$. Once an NCP has been identified, it is therefore unlikely that Gibbs sampling in its pure form can be used to update the two components θ and $\tilde{X}$.

From experience in the NHM context, we would expect an NCP to be more effective than its CP rival when X is poorly identified by the data and remains highly correlated with θ. For this reason, it is particularly important in contexts where the dimensionality of X increases as the data set becomes larger, and this in turn explains its relevance to latent structure models such as hidden Markov models.

A technique that allows us to construct NCPs for a wide range of hierarchical models (for example most of the models considered in Lee and Nelder (1996)) is the expansion of the state-space. In an NCP we allow the function $h(\cdot, \theta)$ to be non-invertible so that for example $\tilde{X}$ can exist on a higher dimensional space than X. As an example, suppose that X is distributed according to a Gamma with shape parameter α and scale parameter 1. We can take $\tilde{X}$ to be a standard Gamma process in $[0, \infty)$ and set $X = \tilde{X}(\alpha)$. The link between infinite divisible distributions and Lévy processes can be used for similar parameterisations. Due to lack of space, we will not give the details of this construction here, but the interested reader is referred to Papaspiliopoulos *et al.* (2002a). The implementation of a state-space expanded NCP is not more difficult than the corresponding CP, although it is more computationally intensive. We can also use these ideas to orthogonalise a stochastic process from its parameters and an example is given in Section 5.2.

4. PARTIALLY NON-CENTERED ALGORITHMS

We have already defined and analysed the convergence properties of the CP and NCP for the normal hierarchical model in Section 1.2. Consider the following parameterisation for the same model,

$$\begin{aligned} Y_i &= w\theta + \tilde{X}_i^w + \sigma_y \epsilon_i \\ \tilde{X}_i^w &= (1-w)\theta + \sigma_x z_i, \ i = 1, \ldots, m. \end{aligned} \tag{13}$$

where w is a fixed number in $[0, 1]$. It can easily be seen that

$$\tilde{X}_i^w = (1-w)X_i + w\tilde{X}_i$$

where X_i and $\tilde{X}_i$ as defined in (3) and (4) respectively, and therefore (13) defines a continuum of parameterisation strategies with the CP at one extreme, for $w = 0$ and the NCP at the other, for $w = 1$. The motivation behind this construction is the following: from the results of Section 2.1 we know that the CP is the optimal algorithm when the relative observation error σ_y/σ_x

tends to zero, i.e when the data are most informative, while the NCP is optimal when this error tends to infinity, i.e in absence of any data. We would therefore like to construct an algorithm that will adapt to the quantity of information present in the observed data, in order to provide samplers with superior convergence properties to both the CP and NCP. This can be achieved by (1), which we will call *partially-non-centered* (PNCP), while it is the optimal Gibbs sampling algorithm for a specific choice of w.

The joint posterior distribution of $\tilde{X}^w = (\tilde{X}_1^w, \ldots, \tilde{X}_m^w)$ and θ is still Gaussian and therefore we can calculate the rate of convergence of the Gibbs sampler (1) under this parameterisation which is given and plotted against w in Figure 3.

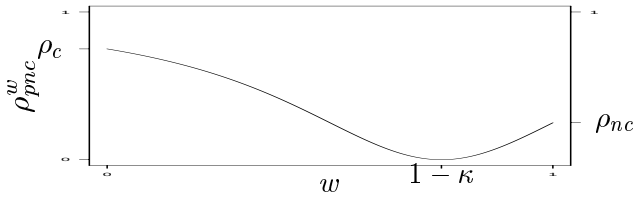

Figure 3. *Rate of convergence* $\rho_{pnc}^w = (w-(1-\kappa))^2/(w^2\kappa + (1-w)^2(1-\kappa))$ *for the PNCP on the NHM with* $\sigma_x = 1,\ \sigma_y = 3$.

We can easily derive that $\rho_{pnc}^w \leq \max(\rho_c, \rho_{nc}),\ \forall w \in (0,1)$ and $\rho_{pnc}^w = 0$ for $w = 1-\kappa$, which suggests not only that the PNCP algorithm (13) can outperform both CP and NCP, but also it can be tuned appropriately to produce IID samples, by setting $w = 1-\kappa$.

A similar result can be obtained for the general NHM (see Gelfand *et al.* (1995))

$$\begin{aligned} Y_i &= C_{1i}X_i + (\sigma_y^2 I_{n_i})^{1/2}\epsilon_i \\ X_i &= C_2\theta + D^{1/2}z_i,\ i = 1, \ldots, m \end{aligned} \tag{15}$$

where Y_i is an $n_i \times 1$ vector, X_i a $p \times 1$, θ a $q \times 1$, D is the prior $p \times p$ variance matrix, and $C_{1i},\ n_i \times p$ and $C_2,\ p \times q$ are design matrices, $C_2^T C_2$ assumed to be invertible. Then the NCP and PNCP are defined as

$$\begin{aligned} \tilde{X}_i &= X_i - C_2\theta \\ \tilde{X}_i^w &= X_i - W_i C_2\theta \end{aligned} \tag{16}$$

and IID samples can be produced by setting

$$W_i = B_i D^{-1},\ B_i^{-1} = \sigma_y^{-2} C_{1i}^T C_{1i} + D^{-1}. \tag{17}$$

The rate of convergence of the CP for (15) is given by the maximal modulus eigenvalue of $m^{-1}\sum_{i=1}^m W_i$. The importance of this matrix in assessing the convergence properties of the CP was observed by Gelfand *et al.* (1995).

Notice that in (16) the proportion of θ subtracted from each X_i varies with i unlike (13), reflecting the varying informativity of each Y_i about the underlying X_i present in (16).

4.1. *Partial Non-centering Outside the Gaussian Context*

Partial non-centering can be used for many models outside the Gaussian context. In general there is no unique way of defining a continuum of partial non-centering strategies. However often there will be a natural one suggested strongly by the model structure. Outside the Gaussian context, it is rare that pure Gibbs sampling can be used in conjunction with a PNCP, so that as

with the NCP and often the CP, appropriate Hastings-within-Gibbs strategies will be necessary. Thus optimal PNCP will not produce IID observations from the target distribution.

Sensitivity of algorithm performance to data is very common in many classes of hierarchical models, since it is often the case that the information about X_i contained in Y_i will depend on Y_i. Therefore, to extend the PNCP to other models in an efficient way, we will need to allow w to vary across $i = 1, \ldots, m$, as in (16), and possibly be a function of the corresponding data Y_i.

To make this more concrete, consider the random effects model:

$$\begin{aligned} Y_i &\sim f(\cdot|X_i) \\ X_i &= \theta + \sigma_x z_i,\ i = 1, \ldots, m \end{aligned} \tag{18}$$

for some class of densities $f(\cdot \mid \cdot)$.

A quadratic expansion of the log-likelihood, $\ell = \log f$ gives a rough indication into the information content present in Y_i about X_i. We set $I(y) = -\partial^2 \ell(y|x)/\partial x^2$ evaluated at the MLE $\hat{x}$ (ignoring the latent structure). Other approximations of information may be more appropriate in certain cases. In the NHM, $I(y) = (\sigma_y^2)^{-1}$, but more generally I will depend on y. Data-dependent non-centering then sets

$$\tilde{X}_i^w = X_i - w_i(y)\theta \tag{19}$$

where $w_i(y) = (1 + I(Y_i)/\sigma_x^2)^{-1}$.

A multivariate generalisations of the technique outlined above is described in the geostatistical example of Section 5.1. A further spatial application of the PNC appears in Higdon (1998).

5. EXAMPLES

5.1. *Spatial GLMM*

In this section we will consider a special case of a GLMM model introduced in Breslow and Clayton (1993) proposed in the spatial context by Diggle *et al.* (1998). Similar modeling approaches have received much attention recently but strong posterior correlation between parameters and latent variables makes a fully Bayesian approach difficult without the use of complex MCMC algorithms and careful reparameterisations. We will here consider a spatial Poisson log-Normal model also studied in Diggle *et al.* (1998) for modeling radioactive counts and in Christensen and Waagerpetersen (2002) for modeling counts of weed. For a more detailed description of the model refer to Diggle *et al.* (1998).

The data consists of recorded observations $Y = (Y_1, \ldots, Y_m)$ with

$$\begin{aligned} Y_i &\sim \text{Po}(\exp(X_i)) \\ X &\sim \text{N}(\theta \mathbf{1}, \sigma^2 R) \end{aligned} \tag{20}$$

Here $X = (X_1, \ldots, X_m) = (X(t_1), \ldots, X(t_m))$ are (unobserved) values from a stationary isotropic Gaussian random field $\boldsymbol{X} = \{X(t), t \in \boldsymbol{R}^2\}$ with mean θ, standard deviation σ and correlation function $r(u) = \text{Corr}(X(s_1), X(s_2)) = \exp(-u/\alpha),\ u = \|s_2 - s_1\|$ (Euclidean distance), and $\mathbf{1}$ is a $m \times 1$ vector of 1s.

Thus the unknown components in this model are the parameters θ, α and σ, together with the underlying field X. It is beyond the scope of this article to fully describe partially non-centered methods which can be applied effectively to this problem. We shall instead concentrate on part of the algorithm, in which the partial non-centering is applied to θ while the remaining parameters σ and α remain fixed. Similar non-centering strategies can be applied to σ and α

and also directly to X in order to break down the posterior correlation structure present in the field X. This will be reported in Christensen *et al.* (2002).

Both Diggle *et al.* (1998) and Christensen and Waagerpetersen (2002) use the NCP, *i.e.*, they alternate between updating $\tilde{X} = X - \mathbf{1}\theta$ and θ. Diggle *et al.* (1998) use single site Gibbs updating of $\tilde{X}$ and Christensen and Waagerpetersen (2002) use the Metropolis adjusted Langevin algorithm (MALA, see for example Roberts and Tweedie (1996)) which can give considerable convergence advantages for large m. Here we shall extend the data-dependent partial non-centering ideas of Section 4.1 to the present case of spatially varying X.

In the absence of covariates, the partial centering parameters for the NHM with $m = 1$ are given by $W = BD^{-1} = (\text{diag}(\sigma_x^{-2}) + D^{-1})^{-1}D^{-1}$, where D is the $n \times n$ prior variance matrix and $\text{diag}(\sigma_x^{-2})$ is the $n \times n$ matrix with σ_x^{-2} on the diagonal and zeros elsewhere. The latter matrix being constant along the diagonal reflects the fact that the variance of the error distribution $Y|X$ is independent of X and thus the loss of information about θ is equal along the components of Y. This is a feature not shared by model (20) since here the error-distribution depends on X in a nonlinear way through the logarithmic link-function and large values in Y tend to be more informative about the mean than small ones.

Outside the Gaussian context, there is no direct analogue of the B_i matrices in (17), but a quadratic expansion of the likelihood as in Section 4.1 suggests we set

$$\hat{B}^{-1} := -\frac{d^2}{dX^2}\log \pi(X|Y,\theta)|_{X=\hat{X}} = -(\text{diag}(\exp(\hat{X})) + \sigma^{-2}R^{-1}) \approx -(\text{diag}(Y) + \sigma^{-2}R^{-1})$$

(where $\hat{X}$ is the MLE from the observation equation alone) leading to the partial centering for θ

$$\tilde{X}^w = X - W\theta.$$

with $W = \hat{B}\sigma^{-2}R^{-1}$. Where $\alpha = 0$ this reduces to the random effects case described by (19).

A simulation study involving 100 observations equally spaced on the unit square was carried out. It involved using two different combinations of θ and σ each under two levels of dependence. Although updates of $\tilde{X}^w$ need to be carried out by a suitable Hastings algorithm in this context for each of our parameterisations, our comparison is based on pure Gibbs updates (approximated by running multiple MALA steps for $\tilde{X}^w$ and θ). This eliminates the possibility that any difference between simulation performance could be due to varying efficiency of the MALA updates. ACFs summarising our results are given in Figure 4.

The CP performs better in relation to the NCP as the dependence in $\boldsymbol{X}$ becomes stronger (as measured by α), which is in agreement with (11). Moreover, while the performance of the CP and NCP vary considerably as the parameters change, the data-dependent PNCP performs extremely well in all cases.

5.2. *Non-Gaussian Ornstein-Uhlenbeck Stochastic volatility models*

Non-centering was recently applied in Dellaportas *et al.* (2001) to construct efficient algorithms for Bayesian inference for the class of non-Gaussian stochastic volatility models introduced in Barndorff-Nielsen and Shephard (2001). The parameterisation structures that we present for this model, are more complicated than the CP and NCP shown in Figure 1, however the general methodology of non-centering developed in the previous sections can also be extended in a natural way.

A stochastic volatility model describes the continuous time movement of the logarithm of an asset price, $Y(t)$, through the stochastic differential equation (SDE)

$$dY(t) = v(t)^{1/2}dB(t),\ t \in [0,T] \tag{21}$$

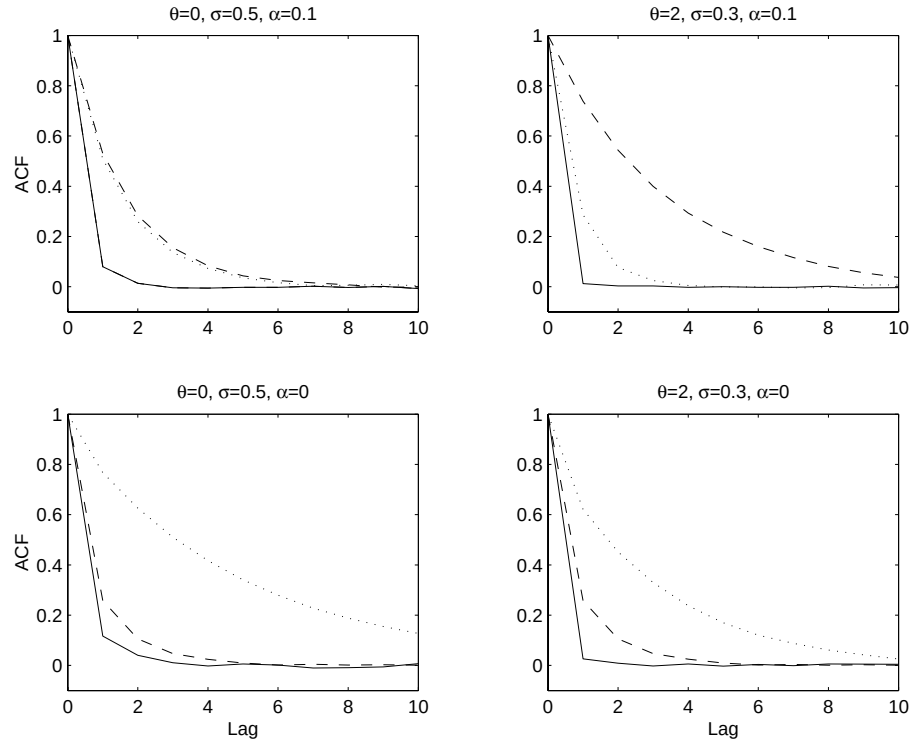

Figure 4. *The spatial GLMM of Section 5.1 and its special case, the random effects model of Section 4.1 (showed in the bottom two plots). ACF for θ using CP (dotted), NCP (dashed) and data-dependent PNCP (solid) for various parameter values.*

where $B(\cdot)$ is a standard Brownian motion. The volatility $v(t)$ is assumed to be unobserved, stochastic and independent of $B(t)$. Defining $v^*(t_{i-1}, t_i)$ to equal volatility integrated over $[t_{i-1}, t_i]$, (20) implies that the conditional distribution of the increments of the data, $Y(t_i) - Y(t_{i-1})$, given v^* is $\mathrm{N}(0, v^*(t_{i-1}, t_i))$ and with conditional independence between disjoint time intervals. In Barndorff-Nielsen and Shephard (2001) the volatility process is modeled as a stationary non-Gaussian Ornstein-Uhlenbeck process described through an SDE with solution

$$v(t) = e^{-\mu t} v(0) + \sum_{j=1}^{N(t)} e^{-\mu(t-c_j)} \epsilon_j. \tag{22}$$

We will focus on the case where $N(t)$ is a Poisson process with rate λ and arrival times $c_1 < \ldots < c_{N(t)}$, and $\epsilon_j \sim \mathrm{Ex}(\phi)$ independently of the Poisson process so that (22) defines a process with stationary distribution $\mathrm{Ga}(\nu = \lambda/\mu, \phi)$. The data consists of discrete observations $Y = \{Y(t_1), \ldots Y(t_n)\}$. Although it is straightforward to write down the conditional density of the data given the corresponding integrated volatilities $v^*(t_{i-1}, t_i),\ i = 1, \ldots, n$, the likelihood function $f(Y \mid \phi, \lambda, \mu)$ is unavailable since it involves integrating out the volatility process which is practically impossible (for details see Section 5.4 of Barndorff-Nielsen and Shephard (2001)). On the other hand a data augmentation method that augments the $v^*(t_{i-1}, t_i)$ is very inefficient, as shown in Barndorff-Nielsen and Shephard (2001). Instead, we explicitly augment $v(0)$, the jump times $\{c_j\}$ and the jump sizes $\{\epsilon_j\}$ and denote by Ψ the marked point process containing these pairs, since they uniquely determine v^*. Therefore, the missing data are $(\Psi, v(0))$, and conditionally on this $\lambda,\ \phi$ are independent of Y. Implementational details of the algorithm constructed for this parameterisation appear in Dellaportas *et al.* (2001).

The algorithm described above is a CP for λ and ϕ. Therefore, from the intuition acquired

from the simpler examples of Section 2.1 and Section 3, we would expect this method to exhibit poor convergence properties when the information about λ, ϕ contained in Ψ and $v(0)$ strongly dominates the marginal information about these parameters.

An NCP for this model, that makes Ψ and the parameters *a priori* independent, can naturally be produced using the state-space expansion technique mentioned in Section 3. Ψ is *a priori* a Poisson process on $[0,T] \times (0,\infty)$ with points $\{(c_j, \epsilon_j)\}$ and mean measure $\lambda\phi e^{-\phi\epsilon} dc\, d\epsilon$. Let $\tilde{\Psi}$ be a Poisson process on the higher dimensional space $[0,T] \times (0,\infty) \times (0,\infty)$ with points denoted by $\{(\tilde{c}_j, m_j, \tilde{\epsilon}_j)\}$ and mean measure $e^{-\tilde{\epsilon}} d\tilde{c}\, dm\, d\tilde{\epsilon}$. Clearly $\tilde{\Psi}$ is *a priori* independent of the parameters. Consider the following transformation illustrated in Figure 5. Choose all points of $\tilde{\Psi}$ with $m_j < \lambda$ and project them to $[0,T] \times (0,\infty)$. Denote those points by $(c_j, \tilde{\epsilon}_j)$. Then take $\epsilon_j = \tilde{\epsilon}_j/\phi$. It can easily be shown (see Dellaportas *et al.* (2001)) that the set of the resulting points $\{(c_j, \epsilon_j)\}$ has the same distribution as Ψ. Hence the NCP augments $\tilde{\Psi}$ and the transformation corresponding to (12) is described above.

Extensive simulation study was carried out in Dellaportas *et al.* (2001) to study the properties of CP and NCP for this model, revealing robustness of the NCP to a variety of different parameter values and time-series length. Unfortunately, due to lack of space, we can't reproduce these results here. It was also observed that the CP has similar behaviour as in the linear model with Cauchy observation equation, presented in Section 2.3.

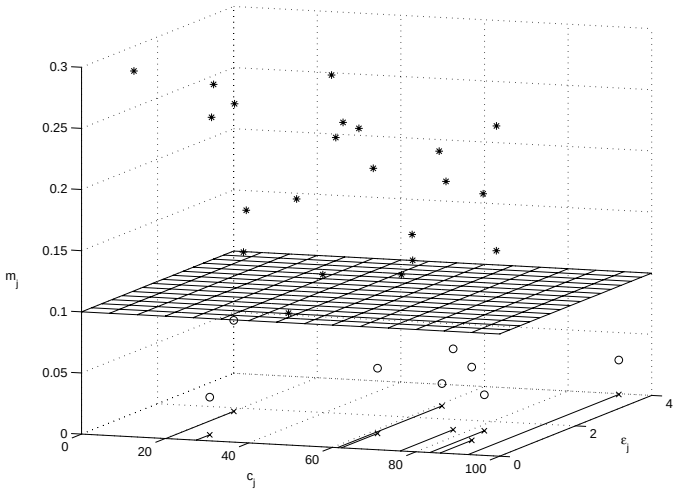

Figure 5. *Construction of the NCP of Section 5.2. Transformation of $\tilde{\Psi}$ to Ψ; using the current value of $\lambda = 0.1$, choose all points of $\tilde{\Psi}$ with $m_j \leq \lambda$ (denoted by circles, where points with $m_j > \lambda$ by asterisks), project them to $[0,T] \times (0,\infty)$, where $T = 100$; denote them as $\{c_j, \tilde{\epsilon}_j\}$. Set $\epsilon_j = \tilde{\epsilon}_j/\theta$ using the current value of $\theta = 1$.*

5.3. *Partially observed diffusion models*

Here we briefly review the techniques introduced in Roberts and Stramer (2001). We shall consider the simplest case of these models where the volatility is an unknown constant

$$dX(t) = \sigma dB(t) + \xi(\theta, X(t))\, dt\ . \tag{23}$$

We wish to make inference about σ and θ from observations $\{Y(t_1), \ldots, Y(t_n)\}$. The data is assumed to be insufficiently fine to allow useful approximations of the continuous time reality, so it is natural to treat this as a missing data problem.

Here we shall consider the theoretical algorithm which imputes the entire missing data, $X = \{X(s), t_i < s < t_{i+1},\ 1 \leq i \leq n-1\}$. In practice, observations at a fine partition of time points need to be imputed, but here we shall assume that continuous imputation is feasible.

The CP which alternates updating the parameters and X is actually *reducible*, since σ never changes as a result of the quadratic variation identity

$$\lim_{m\to\infty}\sum_{i=2}^{m}(X(i(t_n-t_1)/m)-X((i-1)(t_n-t_1)/m))^2=\sigma^2(t_n-t_1)\ .$$

We call such an identity linking X and the parameters an *ergodicity constraint*.

A natural non-centering of the scale parameter σ sets

$$\tilde{X}(t)=X(t)/\sigma$$

(though this is not strictly a non-centering of σ without an associated transformation of the diffusion drift term). However, the corresponding NCP is again reducible. This time the problem is caused by the continuity of sample paths of (23) at the observed points $Y(t_j)$, which is violated if σ is updated.

Here there is essentially a unique piecewise linear transformation of $\tilde{X}$, giving a PNCP which leads to an irreducible algorithm. For $t_i<t<t_{i+1}$ it is described by

$$\tilde{\tilde{X}}(t)=\tilde{X}(t)-\frac{(t_{i+1}-t)Y(t_i)/\sigma+(t-t_i)Y(t_{i+1})/\sigma}{t_{i+1}-t_i}\ .$$

Much more detail and intuition into this construction can be obtained from Roberts and Stramer (2001). The problems of the CP are caused by the fact that the augmented data contains infinite information about σ while the observed finite data set can hardly do this. More complex SDE models can be treated by adaptations of the techniques outlined above.

6. DISCUSSION

In this paper we have considered many strategies for constructing non-centered and partially non-centered parameterisations across wide-ranging classes of models. We have tried to provide a balance of theory, methodology, statistical insight and examples. Our motivation has been the search for effective parameterisations for the use in MCMC algorithms.

However, it has not been the purpose of this paper to promote unreservedly non-centered methods. In fact there are many situations where there is little reason to look further than the CP for the construction of effective algorithms. On the other hand, as we have seen in a number of examples, there are many situations where the posterior dependence between θ and X is prohibitively strong so that non-centered methods are needed. We have also shown how the tail behaviour of the observation and hidden equations can affect the behaviour of the CP and the NCP. Thus we have tried to present a balanced view of the available parameterisations. Moreover, we have discussed PNCP methods that offer a solution to finding algorithms with increased robustness to data.

Other work to which the ideas in this paper appear well-suited involves inference using stochastic epidemic models, and this is subject to ongoing work with Peter Neal (Neal *et al.* (2002)). Non-centering methods ought also to be well-suited for problems in Bayesian non-parametrics and other situations where hidden structure is allowed to vary flexibly and we are currently working on this (see Papaspiliopoulos *et al.* (2002b)). We are also working on non-centering for multivariate θ, where there is the option of non-centering the parameters individually or simultaneously, and on hierarchical models with more than two stages.

ACKNOWLEDGEMENTS

All authors would like to acknowledge the support of TMR network FMRX-CT960095 on Spatial and Computational Statistics. The first author would like to thank the Onasis foundation and Lancaster University for support and similarly the third author thanks the Helmut Hertz and Wennergren foundations. We thank Petros Dellaportas for motivating discussions, and Ole Christensen and Paul Fearnhead for constructive comments on earlier drafts.

REFERENCES

Amit Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Analysis* **38**, 82–99.

Barndorff-Nielsen O. and Shephard N. (2001). Non-Gaussian Ornstein-Ulhenbeck-based models and some of their uses in financial economics. *J. Roy. Statist. Soc. B* **63**, 167–241.

Breslow N. E. and Clayton D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9–25.

Carter C. K. and Kohn R. (1994). On Gibbs sampling for state space models. *Biometrika* **81**, 541–553.

Christensen O. F., Roberts G. O. and Sköld M. (2002). Bayesian analysis of spatial GLMM using partially non-centered MCMC methods. *In preparation.*

Christensen O. F. and Waagepetersen R. P. (2002). Bayesian prediction of spatial count data using generalised linear mixed models. *Biometrics* **58**, 280–286.

Dawid. A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664–667.

Dellaportas P., Papaspiliopoulos O. and Roberts G. O. (2001). Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility processes. *Submitted for publication.*

Diggle P. J., Tawn J.A. and Moyeed R. A. (1998). Model-based geostatistics. *J. Roy. Statist. Soc. C* **47**, 299–350, (with discussion).

Gelfand A. E., Sahu S. K. and Carlin B. P. (1995). Efficient parametrization for normal linear mixed models. *Biometrika* **82**, 479–488.

Gelfand A. E., Sahu S. K. and Carlin B. P. (1996). Efficient parameterizations for generalised linear models. *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 479–488.

Gilks W. R., Thomas A., and Spiegelhalter D. J. (1994). A language and program for complex bayesian modelling. *The Statistician* **43**, 169–178.

Higdon. D. M. (1998). Auxiliary variable methods for Markov Chain Monte Carlo with applications. *J. Amer. Statist. Assoc.* **93**, 585–596.

Hills S. E. and A. F. M. Smith. (1992). Parameterization issues in Bayesian inference. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 227–246.

Jones G. L. and Hobert J. P. (2001). Honest exploration of intractible probability distributions via Markov chain Monte Carlo. *Statist. Sci.* **16**, 312–334.

Knuth. D. E. (1979). *TEX and METAFONT, New Directions in Typesetting*. New York: Digital Press.

Lee Y. and Nelder J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. B* **58**, 619–656, (with discussion).

Liu J. S. and Wu Y. (1999). Parameter expansion for data augmentation. *J. Amer. Statist. Assoc.* **94**, 1264–1274.

Liu J. S., Wong W. H. and Kong A. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81**, 27–40.

Meng X. and van Dyk D. (1997). The EM algorithm — an old folk song sung to a fast new tune. *J. Roy. Statist. Soc. B* **59**, 511–567, (with discussion).

Meng X. and van Dyk D. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**, 301–320.

Neal P., Roberts G. O. and Viallefont V. (2002). Robust MCMC algorithms for inference for stochastic epidemic models. *In preparation.*

Papaspiliopoulos O., Roberts G. O. and Sköld M. (2002a). State-space expansion parameterisations for MCMC. *In preparation*.

Papaspiliopoulos O., Roberts G. O. and Sköld M. (2002b). Non-centred MCMC methods for Bayesian non-parametrics. *In preparation*.

Pitt M. K. and Shephard N. (1999). Analytic convergence rates and parameterisation issues for the Gibbs sampler applied to state space models. *J. Time Ser. Anal.* **20**, 63–85.

Roberts G. O. and Sahu S. K. (1997). Updating schemes, Correlation Structure, Blocking and Parameterisation for the Gibbs Sampler. *J. Roy. Statist. Soc. B* **59**, 291–397.

Roberts G. O. (2001). Linking theory and practice of MCMC. *Highly Structured Stochastic Systems*.

Roberts G. O. (2002). Convergence of MCMC for hierarchical models with heavy tailed-links. *In preparation*.

Roberts G. O. and Stramer O. (2001). Bayesian inference for incomplete observations of diffusion processes. *Biometrika* **88**, 203–221.

Roberts G. O. and Tweedie R. L. (1996). Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli* **2**, 341–364.

Roberts G. O. and Tweedie R. L. (2001). Geometric $L2$ and $L1$ convergence are equivalent for reversible Markov chains. *J. Appl. Probability* **38A**, 37–41.

Roberts G. O. and Tweedie R. L. (2002). *Understanding MCMC*. Berlin: Springer.

Rubin D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Sahu S. K. and Roberts G. O. (1999). On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing* **9**, 55–64.

Smith A. F. M. and Roberts G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. B* **55**, 3–24.

Tanner M. A. and Wong W. H. (1987). The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.* **82**, 528–540.

Wakefield J. C., Smith A. F. M., Racine-Poon A. and Gelfand A. E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Appl. Statist.* **43**, 201–221.

DISCUSSION

ALAN E. GELFAND *(Duke University, USA)*

Continuing the tradition of previous work by Roberts and his collaborators, this paper is clever, provocative and potentially quite useful. Anyone who has invested time in implementing challenging MCMC model fitting will have appreciated the importance of choice of parametrization. This contribution provides some strategic, though generally terse, guidance for many hierarchical modeling settings.

My comments will be confined to spatial and nonparametric modeling contexts. For the former, consider first a Gaussian spatial process model of the form $Y(\boldsymbol{s}_i) = \boldsymbol{X}^T(\boldsymbol{s}_i)\boldsymbol{\beta} + \alpha(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i)$ where $\alpha(\boldsymbol{s}_i)$ is a 0 mean Gaussian process with variance σ^2 and arbitrary correlation function ρ while $\epsilon(\boldsymbol{s}_i)$ is a white noise process with variance τ^2. If σ^2, τ^2 and ρ are known and $\boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\beta} = \mu$ then $(\mu, \boldsymbol{\alpha})$ is the noncentred parametrization, $(\mu, \boldsymbol{\eta})$ where $\eta_i = \mu + \alpha_i$ is the centred one and $(\mu, \boldsymbol{\eta}_{cent})$ where $\boldsymbol{\eta}_{cent} = \boldsymbol{\eta} - \mu W 1$ is partially noncentred (PNCP) . It is straightforward to compute that, provided $var(\mu|\boldsymbol{Y})$ exists,

$$cov(\boldsymbol{\eta}_{cent}, \mu|\boldsymbol{Y}) = \left[I - W - \left(I + \frac{\tau^2}{\sigma^2} R^{-1}\right)\right]^{-1} var(\mu|\boldsymbol{Y})$$

where R is such that $R_{ij} = corr(\alpha(\boldsymbol{s}_i), \alpha(\boldsymbol{s}_j))$. Hence $\boldsymbol{\eta}_{cent}$ and μ are uncorrelated *a posteriori* if $W = (I + \frac{\sigma^2}{\tau^2} R)^{-1}$. This choice of W is exactly the one obtained by the authors in section 5.1 to optimize the convergence rate, ρ_{NCP} derived under a normal hierarchical model (NHM).

Naturally arising questions are the following. Is the focus on finding a parametrization to achieve prior independence (as the introduction suggests) or on posterior uncorrelatedness? The foregoing calculation is done as a cross-correlation under the posterior while the rate of

convergence is obtained as an L_2 distance between the distribution at the current iteration and the posterior. Is the former a legitimate way to obtain the PNCP under a NHM? The foregoing calculation produces the same PNCP for any prior on μ as long as $var(\mu|\boldsymbol{Y})$ exists. Is there a similar robustness to prior in rate of convergence? If we return to a general $\boldsymbol{X}^T(\boldsymbol{s}_i)\boldsymbol{\beta}$ with intercept, what sort of centering, partial centering, is sensible? If σ^2, τ^2 and ρ are not known would an estimated W using parameter estimates be appropriate to use?

Suppose in the above that we have a non Gaussian first stage with canonical link. That is, again taking $\boldsymbol{X}^T(\boldsymbol{s}_i)\boldsymbol{\beta} = \mu$, $f(Y(\boldsymbol{s}_i)|\mu, \alpha(\boldsymbol{s}_i)) = c\exp(Y(\boldsymbol{s}_i)\theta(\boldsymbol{s}_i) - b(\theta(\boldsymbol{s}_i)))$ where $\theta(\boldsymbol{s}_i) = \mu + \alpha(\boldsymbol{s}_i)$ with $\alpha(\boldsymbol{s}_i)$ as above. If we expand $b(\theta(\boldsymbol{s}_i))$ about μ to second order, we can again compute W to obtain $cov(\boldsymbol{\eta}_{cent}, \mu) \approx 0$. This W only involves $b''(\mu)$ rather than $\{b''(\mu+\alpha(\boldsymbol{s}_i))\}$ suggesting that in estimating W we would use only say $\overline{Y}$ rather than $\{Y(\boldsymbol{s}_i)\}$ as in the authors' expression in Section 5.1. Is it clear that the latter is a better data-based choice?

Next we consider two nonparametric illustrations. First, consider the quantal bioassay problem using Dirichlet processes as in, e.g., Gelfand and Kuo (1991). Here $Y_i|p_i$ are independent $Bi(n_i, p_i)$, $i = 1, \cdots, k$ with $p_i = F(d_i)$ where d_i are increasing dosage levels and F is a random c.d.f. from a Dirichlet process, i.e., $F \sim DP(\alpha F_0)$ with F_0 a parametric family of c.d.f.'s indexed by say location μ and scale σ. In this situation there is misalignment between the likelihood which is of the form $\prod_i p_i^{y_i}(1-p_i)^{n_i-y_i}$ and the prior which is of the form $p_1^{\gamma_1-1}(p_2-p_1)^{\gamma_2-1}\cdots(1-p_k)^{\gamma_{k+1}-1}$. The customary fitting approach introduces latent multinomial variables to break the misaligment but a slowly converging chain with high autocorrelation typically results.

Can a state space expansion approach work here? The centred parametrization is given by $((\alpha, \mu, \sigma), \{p_i\})$. If, analogous to the Gamma process example in section 3.1, we replace p_i with $F_i(.)$ we will be unable to ensure the hard order restrictions on the p_i. If we introduce a single $F(.)$ then given $\{d_i\}$ we can define $p_i = h_i(F(.), \mu, \sigma) = F((d_i - \mu)/\sigma)$ and convert to a noncentred parametrization $(F(.), \alpha, \mu, \sigma)$. (We can set $\sigma = 1$ for convenience of illustration). Would the size of the n_i determine which parametrization is preferred? Would the choice really matter since we still have the misalignment problem?

Lastly, consider a Dirichlet process mixing setting. Here $Y_i \sim f(Y_i|\theta_i, \phi)$ are conditionally independent and θ_i are conditionally independent from F. Finally, $F \sim DP(\alpha F_0)$ where, again F_0 is indexed by say μ and σ. See, e.g., Gelfand and Kottas (2002) for details and computational discussion. Here, the centered parametrization is $((\alpha, \mu, \sigma), F, \{\theta_i\}, \phi)$. Customary marginalization over F still yields a centred parametrization, which, under MCMC, exhibits high auto and cross-correlation.

Instead, we could consider marginalizing over $\{\theta_i\}$. Since F is a.s. discrete we could introduce a partial sum approximation $\tilde{F}$ to carry out the marginalization yielding the parametrization $((\alpha, \mu, \sigma), \tilde{F}, \phi)$. Is this parametrization partially noncentred? If Y_i does not inform well about θ_i, do we expect marginalizing over $\{\theta_i\}$ to be "better" than marginalizing over F? Are there other promising possibilities? Perhaps my real question is what suggestions the authors might have with regard to choice of parametrization in such nonparametric models.

OLE F. CHRISTENSEN *(Lancaster University, U.K.)*

I would like to congratulate the authors on providing a very useful characterisation of different parameterisations whereby one can get an intuition on how they behave in practice. Basically, the hierarchical parameterisation illustrated by Figure 1a works well when the information in data is relatively strong, and the a priori independence parameterisation illustrated by Figure 1b works well when the information in the data is relatively weak. They also propose

several strategies for constructing a robust parametrisation that works well in both cases. The authors use the terminology "non-centered" and "partial non-centered" for the later two types of parameterisation, which suggest that they find them un-natural. I think these two types of parameterisation deserve better names.

My particular interest is the spatial example in Section 5.1 where I can confirm that the robust parameterisation suggested by the authors performs well for all the data sets I have considered. Updating the covariance parameters σ^2 and α is more challenging since α enters the target density in a complicated way through the correlation matrix.

My last comment is about the possibility of marginalising out parameters when using conjugate priors or limits of conjugate priors. Consider the simple example where the latent variable X is Gaussian with mean $D\beta$, covariance matrix Σ, and $\pi(\beta) \propto 1$. The marginal posterior distribution of X is

$$f(x \mid y) \propto f(y \mid x) \exp(-x^T(\Sigma^{-1} - \Sigma^{-1}D(D^T\Sigma^{-1}D)^{-1}D^T\Sigma^{-1})x/2),$$

and MCMC sampling can be done without updating β. In practice marginalisation is an advantage, since one avoids having to tune the proposal variance for β. Could the authors comment on how well marginalisation would work compared to their approach of updating both the latent variables and the parameters ?

DARREN J. WILKINSON *(University of Newcastle, UK)*

The authors are to be congratulated on a most interesting paper. They consider the common problem of inference for "latent process" models, with parameters θ latent process X and data Y such that $\theta \perp\!\!\!\perp Y \mid X$.

The classic data augmentation scheme alternately samples $\theta \mid X, Y$ and $X \mid \theta, Y$ and often works well in the situation where the two steps can both be carried out exactly as pure Gibbs steps. Note in particular that the step $X \mid \theta, Y$ is often carried out as a sequence of smaller Gibbs moves but in many cases, such as the case of systems that are linear Gaussian conditional on θ, this is not necessary and block-updating schemes generally perform much better despite some increase in computational overhead (Rue 2001, Wilkinson and Yeung 2002a). Two-block updating schemes perform particularly well when there is high dependence within X and weak (posterior) dependence between θ and X. However, as noted in the paper, in the case of high posterior dependence between θ and X the classic data augmentation scheme can break down. The authors suggest that non-centred or partially non-centred parameterisations may provide a solution. It should be noted that there are often other possible strategies which can work even better and are often easier to implement. Like NCPs, such techniques involve sacrificing a Gibbs move for a Metropolis-Hastings scheme, but constructed in such a way as to get around the strong dependence between θ and X. The simplest way is to integrate X out of the problem completely, and simply construct a Metropolis-Hastings scheme for $\theta \mid Y$. If a new $\theta^\star$ is proposed from some kernel $q(\theta^\star \mid \theta)$, it is accepted with probability $\min\{1, A\}$ where

$$A = [\pi(\theta^\star)L(\theta^\star; Y)q(\theta \mid \theta^\star)]/[\pi(\theta)L(\theta; Y)q(\theta^\star \mid \theta)].$$

Note that this depends on the marginal likelihood of the data $L(\theta; Y)$, but this is usually tractable whenever sampling $X \mid \theta, Y$ in a single block is possible (Rue 2001, Wilkinson and Yeung 2002b). Even in cases where this is not tractable, effective updating schemes can be created by constructing "single-block" updating schemes where a new $(\theta^\star, X^\star)$ is sampled in two stages: first $\theta^\star$ is sampled from $q(\theta^\star \mid \theta)$, and then a new $X^\star$ is sampled from a tractable approximation to $X \mid \theta^\star, Y$ (Knorr-Held and Rue 2002). The acceptance probability is primarily related to the

closeness of the approximation. By simulating $X^\star$ to be consistent with $\theta^\star$, the dependence between them is overcome without the need to resort to NCPs.

My second observation concerns the application of NCPs to discretely observed diffusions. This at first appears to be a "killer application" for NCPs, as CPs are well-known to exhibit pathologically poor mixing. However, despite a statement to the contrary in the discussion of Roberts and Stramer (2001), the techniques do not generalise easily from the univariate to the multivariate case. Even for the simple bivariate log-Gaussian stochastic volatility model

$$dX_t = \mu X_t\, dt + \exp\{Y_t/2\} X_t\, dW_t$$

$$dY_t = -\lambda(Y_t - \nu)\, dt + \tau\, dW_t'$$

it seems to be impossible (at least in practice) to find a transformation to constant volatility, a required step in the construction of an effective NCP.

REPLY TO THE DISCUSSION

We would like to thank the discussants for their insightful contributions. Many important points have been raised, and we'll try to address most of these.

Alan Gelfand asks about what we should really be looking for in a suitable parameterisation, and suggests that in many cases, it is sufficient to consider posterior correlation structure.

In this paper, our approach is different. We are primarily interested in constructing and assessing the performance of NCPs. Of course there are many situations where neither of these methods is adequate, and when a CP and an NCP exist, we try to find ways to construct intermediate parameterisations that adapt according to the information in the data, so that the user doesn't have to choose *apriori* between the two extreme parameterisations. Actually , the PNCP in the NHM can also be obtained using a posterior uncorrelatedness argument, as Alan suggests, because in the Gaussian context the maximal correlation described in Amit (1991) is attained by linear functions. The same is true in the geostatistical example that mimics the construction for the NHM. However, outside the Gaussian context it is possible for posterior correlations to be very unreliable for the purposes of predicting convergence properties (see *e.g.*, Roberts, (1992)). Moreover, for generalised hierarchical mixed models (such as those considered in Section 3 of Gelfand *et al.* (1996)), we have found examples where there exist parameterisations (essentially based on the weighted average form of the posterior expectation of X given θ, see formula 3.3 of Gelfand *et al.* (1996)) under which the missing data and the parameters are uncorrelated, but which possess convergence properties inferior to the CP (due to higher order dependence between the random effects and the parameters). Finally, it is unclear how to generalise the construction based on minimising posterior correlation to models where the missing data live on non-Euclidean spaces (as in the example of Section 5.2).

Alan also asks about the effect of the prior on θ. It turns out that although the prior on θ affects the rates of convergence of all the competing algorithms, the optimal PNCP and the relative sizes of ρ_c and ρ_{nc} are both unaffected by the prior.

An important question raised in Alan's discussion asks how to proceed when multiple parameters are to be updated as part of the MCMC cycle. Using notation of the geostatistical example in Section 5.1, we can easily construct a multivariate NCP by setting $\tilde{X} = h(X, \theta, \sigma, \alpha) := (X - \theta 1)R(\alpha)^{-1/2}\sigma^{-1}$, making all parameters *a priori* independent. This construction also makes the components X_i, X_j of the latent field *a priori* independent, and potentially eases updating of X when this has to be performed with a Hastings step. The approach adopted in Christensen *et al.* (2002) proceeds by constructing a PNCP of the form $\tilde{X} = (X - \theta 1 W)W^{-1/2}$ where $W = W(\sigma, \alpha)$ is chosen dynamically as a function of the current values of (σ, α). This

works well in practise, although the extra computational expense involved means that for really large problems some shortcuts, involving less $W(\sigma, \alpha)$ evaluations, are necessary.

The idea behind the method sketched in Section 4.1 is to make a Gaussian approximation of the likelihood and then apply the PNCP given by (17). There are of course many competing strategies for partial non-centering, and even if we decide to determine the extent of non-centering on the basis of a Taylor series expansion, we still need to decide where to expand around. The most robust estimates of information content will be achieved by expanding about the posterior mode, but of course that information is unavailable *apriori*. While we carry out the Taylor expansion about $\hat{\boldsymbol{x}}$, Alan suggests performing the expansion about $\bar{Y}$, which in turn leads to a PNCP which non-centers each datum equally. Whilst this approach has the advantage of simplicity, it seems reasonable that the success of this approach depends on the quality of the Gaussian approximation and our experience has been that in situations like this where information contained in a particular datum depends strongly on the observed data value there is much to be gained by a heterogeneous degree of non-centering. However it might turn out that expanding about a weighted average of $\hat{\boldsymbol{x}}$ and $\bar{Y}$ (perhaps using some simple kriging estimate) would perform even better than $\hat{\boldsymbol{x}}$.

Ole Christensen asks to what extent marginalising parameters will be beneficial to MCMC mixing. It turns out that this generally improves convergence when a pure Gibbs sampler can be applied (see for example Liu (1994)). In practice however, marginalisation will sometimes complicate the likelihood surface of the latent variables X, and if X has to be updated with a Hastings step, this step can become very problematic after marginalisation.

Bayesian non-parametric and semi-parametric analysis is ideally suited to the use of NCPs. We have been working on general non-centering constructions to Lévy processes which lead to natural applications in Bayesian non-parametrics. In particular, there are very natural ways of constructing non-centering of Dirichlet processes using their representation in terms of the Gamma process. Similarly we can provide NCPs of Polya trees (and similar models) by working directly with the Gamma random variables used in their sequential construction. Thus we intend to implement NCP in the Dirichlet mixture application suggested by Alan (see Papaspiliopoulos *et al.* (2002b)).

We also believe that the bioassay problem suggested by Alan is a natural one for our methods, and we look forward to implementing NCPs in this context. We agree strongly that the size of the n_i parameters here will be crucial in determining the effectiveness of the methods.

Darren Wilkinson is indeed correct that the generalisation of the Roberts-Stramer methodology to multiple dimensional problems in full generality is difficult. To see this consider a d-dimensional diffusion satisfying

$$d\boldsymbol{X}_t = \Sigma(\boldsymbol{X}_t)^{1/2} d\boldsymbol{B}_t + b(\boldsymbol{X}_t)dt$$

where Σ and b are functions of $\boldsymbol{X}_t$, t and unknown parameters. In this context, the search for a non-centered parameterisation decoupling the parameters governing Σ from $\boldsymbol{X}$ involves finding an invertible function $\boldsymbol{h} : \boldsymbol{R}^d \to \boldsymbol{R}^d$ which satisfies

$$\nabla \boldsymbol{h} (\nabla \boldsymbol{h})' = \Sigma^{-1} .$$

In full generality, this matrix differential equation is impossible to solve in practice, although there are important special cases which do admit solution very easily (for instance where Σ is diagonal and for all i, Σ_{ii} is not a function of X_j for any $j \neq i$). Unfortunately, this special case does not cover the stochastic volatility example cited by Darren.

ADDITIONAL REFERENCES IN THE DISCUSSION

Gelfand, A. E. and Kottas, A. (2002). A computational approach for full nonparametric Bayesian inference in single and multiple sampling problems. *J. Comp. Graph. Statist.* **11**, 289–305.

Gelfand, A. E. and Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrics* **78**, 657–666.

Knorr-Held, L. and H. Rue (2002). On block updating in Markov random fields for disease mapping. *Scandinavian J. Statist.* (to appear).

Liu J. S. (1994). The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem *J. Amer. Statist. Assoc.* **89**, 958–966.

Roberts G. O. (1992). Discussion on Parameterization issues in Bayesian inference. *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 227–246.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *J. Roy. Statist. Soc. B* **63**, 325–338.

Wilkinson, D. J. and S. K. H. Yeung (2002a). Conditional simulation from highly structured Gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models. *Statist. Computing* **12**, 287–300.

Wilkinson, D. J. and S. K. H. Yeung (2002b). A sparse matrix approach to Bayesian computation in large linear models. *Comput. Statist. and Data Analysis* (to appear).