Biometrika (1999), **86**, 2, *pp.* 301–320
© 1999 Biometrika Trust
Printed in Great Britain

Seeking efficient data augmentation schemes via conditional and marginal augmentation

By XIAO-LI MENG

Department of Statistics, The University of Chicago, Chicago, Illinois 60637, U.S.A.
meng@galton.uchicago.edu

AND DAVID A. VAN DYK

Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
vandyk@hustat.harvard.edu

Summary

Data augmentation, sometimes known as the method of auxiliary variables, is a powerful tool for constructing optimisation and simulation algorithms. In the context of optimisation, Meng & van Dyk (1997, 1998) reported several successes of the 'working parameter' approach for constructing efficient data-augmentation schemes for fast and simple EM-type algorithms. This paper investigates the use of working parameters in the context of Markov chain Monte Carlo, in particular in the context of Tanner & Wong's (1987) data augmentation algorithm, via a theoretical study of two working-parameter approaches, the conditional augmentation approach and the marginal augmentation approach. Posterior sampling under the univariate t model is used as a running example, which particularly illustrates how the marginal augmentation approach obtains a fast-mixing positive recurrent Markov chain by first constructing a nonpositive recurrent Markov chain in a larger space.

Some key words: Auxiliary variable; EM algorithm; Incomplete data; Markov chain Monte Carlo; PXEM algorithm; Rate of convergence; Working parameter.

1. Introduction and overview

Research in data augmentation, which is a generic term for methods that require the construction of unobserved 'data', as a computational tool has produced a multitude of powerful algorithms both for mode finding and for distributional sampling. Perhaps the best-known example is the EM algorithm (Dempster, Laird & Rubin, 1977). One of the major contributions of Dempster et al. (1977) is the recognition that, by purposely constructing 'missing data', one can apply the EM algorithm to many models that have no missing data in the traditional sense. The idea of augmenting the observed data to a larger dataset for which model fitting is easier was extended to posterior sampling by Tanner & Wong (1987), who developed the data augmentation algorithm which can be viewed as a two-step Gibbs sampler.

Although these algorithms are flexible and often simple to implement, their slow convergence is a common complaint. For the EM algorithm, many proposals have been made for speeding the convergence; see Meng & van Dyk (1997) and Meng (1997) for a brief review.

Among them, we prefer methods that aim at increasing the speed without unduly sacrificing simplicity or stability in the resulting algorithms. An effective method for achieving this goal is to search for efficient data-augmentation schemes, where 'efficient' refers to both easy implementation and fast convergence of the resulting algorithms. Such algorithms are developed in Meng & van Dyk (1997, 1998), Liu, Rubin & Wu (1998) and van Dyk (2000) via a 'working parameter' which is introduced solely for the purpose of efficient augmentation, and thus is not a part of the original observed-data model. Currently, there are two methods of making use of such working parameters.

The first approach, which we call conditional augmentation, seeks to condition on a fixed value of the working parameter that results in selecting an optimal or nearly optimal algorithm from a class of candidate algorithms indexed by the working parameter. For EM, the optimality refers to the minimisation of the expected observed Fisher information from the augmented-data likelihood/posterior as a function of the working parameter, since the so-called 'fraction of missing information' (Dempster et al., 1977) determines the rate of convergence of EM; the less we augment, in terms of the Fisher information, the greater the theoretical speed of convergence of the algorithm. The key constraint for this minimisation is that the resulting algorithm should be easy to implement. Meng & van Dyk (1997, 1998) give three applications where this balance is achieved.

The second approach, which we call marginal augmentation, also aims at reducing the augmented information but, instead of conditioning on the value of the working parameter that minimises the augmented information, the methodology aims to marginalise over, i.e. integrate out, the working parameter. Not introducing a working parameter is, in fact, implicitly conditioning on a special value of an invisible working parameter. By actively avoiding this conditioning, we can increase the variability in the augmented data and thus reduce the augmented information. The idea of marginal augmentation was stimulated by our study of the expanded-parameter EM algorithm of Liu et al. (1998), which maximises the augmented-data loglikelihood as a function of the working parameter within each EM iteration. This contrasts the conditional augmentation method adopted in Meng & van Dyk (1997), where one seeks the optimal value of the working parameter before EM iterations.

In general Markov chain Monte Carlo, marginal augmentation refines the standard auxiliary variable method, e.g. Besag & Green (1993) and Green (1997), and has connections with several other approaches for improving mixing rates, as discussed in § 4. It also suggests that one can often obtain a fast mixing positive recurrent Markov chain by constructing a larger nonpositive recurrent Markov chain; it is nonpositive recurrent because of the non-identifiability, and thus improper posterior, of the working parameter. This finding may add an interesting dimension to the recent debate over the usefulness of nonpositive recurrent Markov chains in Markov chain Monte Carlo, e.g. Casella (1996), Berger (1996) and George (1996), and calls for further investigation of Markov chain Monte Carlo theory with improper invariant distributions.

The focus of this paper is to introduce conditional augmentation and marginal augmentation in the context of constructing sampling algorithms and to apply these approaches to the univariate t model. Applications to other common models, such as mixed-effect models and probit regression, will be reported elsewhere. For clarity and succinctness, we will avoid inconsequential mathematical details, and we assume background knowledge of the EM algorithm, compare Dempster et al. (1977), Meng & van Dyk (1997) and McLachlan & Krishnan (1997), and of Markov chain Monte Carlo, in particular the Gibbs sampler; compare Casella & George (1992), Tierney (1994), Gilks, Richardson & Spiegelhalter (1995) and Gelfand (1997).

2. Conditional augmentation

2·1. *Theoretical background*

Suppose Y_{obs} are the observed data and $p(\theta | Y_{\text{obs}}) \propto p(Y_{\text{obs}} | \theta)p(\theta)$ is the posterior density of θ upon which inference will be based. Here $p(Y_{\text{obs}} | \theta)$ is a probability density/function with respect to a measure $\mu(.)$, and $p(\theta)$ is a proper or improper density on $\Theta \in R^d$. Typically, we want to find the modes of $p(\theta | Y_{\text{obs}})$ and/or to sample from it. These tasks, however, are often nontrivial, in which case the idea of data augmentation may be useful. We start by relating the observed data, Y_{obs}, to the so-called augmented data, Y_{aug}, through a many-to-one mapping $Y_{\text{obs}} = \mathscr{M}(Y_{\text{aug}})$. The partially unobserved Y_{aug} is given a model $p(Y_{\text{aug}} | \theta)$ that preserves the marginal model of interest, $p(Y_{\text{obs}} | \theta)$, namely

$$\int_{\mathscr{M}(Y_{\text{aug}}) = Y_{\text{obs}}} p(Y_{\text{aug}} | \theta)\mu(dY_{\text{aug}}) = p(Y_{\text{obs}} | \theta). \tag{2·1}$$

With appropriate choices of $p(Y_{\text{aug}} | \theta)$, we can achieve two objectives simultaneously: (i) it is relatively straightforward to sample from or maximise $p(\theta | Y_{\text{aug}})$, and (ii) it is relatively easy to sample from or perform analytical calculation with respect to $p(Y_{\text{aug}} | Y_{\text{obs}}, \theta)$. For maximising $p(\theta | Y_{\text{obs}})$, with such choices, we can construct an easily implemented EM algorithm (Dempster et al., 1977) that iterates between the E-step, which computes $Q(\theta | \theta^{(t)}) = E\{\log p(\theta | Y_{\text{aug}}) | Y_{\text{obs}}, \theta^{(t)}\}$, and the M-step, which maximises $Q(\theta | \theta^{(t)})$ to determine the next iterate $\theta^{(t+1)}$. To sample from $p(\theta | Y_{\text{obs}})$, we can implement Tanner & Wong's (1987) data augmentation algorithm, which iterates between sampling $Y_{\text{aug}}^{(t+1)}$ from $p(Y_{\text{aug}} | Y_{\text{obs}}, \theta^{(t)})$ and sampling $\theta^{(t+1)}$ from $p(\theta | Y_{\text{aug}}^{(t+1)})$.

To speed up EM-type algorithms, the 'working parameter' approach (Meng & van Dyk, 1997) introduces a hidden parameter, α, into (2·1) that is only identifiable given Y_{aug}:

$$\int_{\mathscr{M}(Y_{\text{aug}}) = Y_{\text{obs}}} p(Y_{\text{aug}} | \theta, \alpha)\mu(Y_{\text{aug}}) = p(Y_{\text{obs}} | \theta). \tag{2·2}$$

Thus, introducing α does not alter the model we are fitting. The following univariate t example, adopted from Meng & van Dyk (1997), illustrates the construction of such an augmented-data model.

Suppose y follows a location-scale t distribution with v degrees of freedom. Then we can write

$$y = \mu + \frac{\sigma z}{\sqrt{\tilde{q}}}, \quad z \sim N(0, 1), \quad \tilde{q} \sim \chi_v^2/v, \quad z \perp \tilde{q}, \tag{2·3}$$

where $\perp$ denotes independence. Consequently, an obvious data-augmentation scheme for fitting (2·3) to $Y_{\text{obs}} = \{y_1, \ldots, y_n\}$, where y_i is assumed to be the ith independent realisation of y, is $Y_{\text{aug}} = \{(y_i, \tilde{q}_i), i = 1, \ldots, n\}$. More generally, we can view (2·3) as a special case of the following model with $\alpha = 0$:

$$y = \mu + \frac{\sigma^{1-\alpha} z}{\sqrt{q}}, \quad z \sim N(0, 1), \quad q | \alpha \sim \sigma^{-2\alpha}\chi_v^2/v, \quad z \perp q. \tag{2·4}$$

Model (2·4) suggests that the standard augmentation using (2·3) is in fact conditioning on a specific value of the working parameter, that is $\alpha = 0$, and there may be other values of α that result in better algorithms. Indeed, Meng & van Dyk (1997) show that using $\alpha = 1/(1+\nu)$ yields the optimal EM algorithm in the sense of maximising the theoretical speed of convergence. Although the optimal EM algorithm differs trivially from the standard EM algorithm corresponding to $\alpha = 0$, its convergence is always faster and often much faster.

The basic idea underlying the t example is in fact general. Once $p(Y_{\text{aug}} \mid \theta, \alpha)$ is constructed, it yields a class of EM implementations indexed by α. We can then search for the optimal implementation by minimising the theoretical matrix rate of convergence $DM^{\text{EM}}(\alpha) = I - I_{\text{obs}} I_{\text{aug}}^{-1}(\alpha)$ over α in a suitable class $\mathscr{A}_0$. Here

$$I_{\text{obs}} = -\left.\frac{\partial^2 \log p(\theta \mid Y_{\text{obs}})}{\partial\theta\,\partial\theta}\right|_{\theta=\theta^*}, \quad I_{\text{aug}}(\alpha) = E\left\{-\left.\frac{\partial^2 \log p(\theta \mid Y_{\text{aug}}, \alpha)}{\partial\theta\,\partial\theta}\right| Y_{\text{obs}}, \theta, \alpha\right\}\Bigg|_{\theta=\theta^*},$$

where θ^* is the limit of $\{\theta^{(t)}, t \geqslant 0\}$. Since I_{obs} does not depend on α, it is sufficient to minimise $I_{\text{aug}}(\alpha)$, in the sense of a semipositive definite order, over α. Note that $I - I_{\text{obs}} I_{\text{aug}}^{-1}$ is the fraction of missing information mentioned in § 1. We call this minimisation approach conditional augmentation because it seeks a fixed value of the working parameter to be conditioned upon while constructing an algorithm.

A general method of introducing α can be formalised as follows. Typically, we start with a standard augmentation scheme $\tilde{Y}_{\text{aug}} = \{\tilde{Y}_{\text{mis}}, Y_{\text{obs}}\}$ with density/distribution $\tilde{p}(\tilde{Y}_{\text{mis}} \mid Y_{\text{obs}}, \theta) p(Y_{\text{obs}} \mid \theta)$. We then define a more general data augmentation, $Y_{\text{aug}} = \{Y_{\text{mis}}, Y_{\text{obs}}\}$, via

$$Y_{\text{mis}} = \mathscr{D}_{\alpha,\theta}(\tilde{Y}_{\text{mis}}), \tag{2·5}$$

where the mapping $\mathscr{D}_{\alpha,\theta}(.)$ is one-to-one for any given θ and α; for continuous $\tilde{Y}_{\text{mis}}$ we also assume this mapping is differentiable. For instance, with the t example, $\tilde{Y}_{\text{mis}} = \{\tilde{q}_1, \ldots, \tilde{q}_n\}$ and $Y_{\text{mis}} \equiv \mathscr{D}_{\alpha,\theta}(\tilde{Y}_{\text{mis}}) = \{\sigma^{-2\alpha}\tilde{q}_1, \ldots, \sigma^{-2\alpha}\tilde{q}_n\}$. The distribution of Y_{aug} is then given by

$$p(Y_{\text{aug}} \mid \theta, \alpha) = p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta, \alpha) p(Y_{\text{obs}} \mid \theta) = \tilde{p}(\mathscr{D}_{\alpha,\theta}^{-1}(Y_{\text{mis}}) \mid Y_{\text{obs}}, \theta) |J(Y_{\text{mis}} \mid \theta, \alpha)| p(Y_{\text{obs}} \mid \theta), \tag{2·6}$$

where $J(Y_{\text{mis}} \mid \theta, \alpha)$ is the Jacobian for the inverse transformation, $\mathscr{D}_{\alpha,\theta}^{-1}(Y_{\text{mis}})$, or 1 if Y_{mis} is discrete. It is easy to see that

$$\int p(Y_{\text{aug}} \mid \theta, \alpha)\, dY_{\text{mis}} = \int \tilde{p}(\tilde{Y}_{\text{mis}} \mid Y_{\text{obs}}, \theta) p(Y_{\text{obs}} \mid \theta)\, d\tilde{Y}_{\text{mis}} = p(Y_{\text{obs}} \mid \theta);$$

that is, (2·6) is a legitimate augmentation for any α such that $\mathscr{D}_{\alpha,\theta}(.)$ is a one-to-one mapping for any $\theta \in \Theta$. We denote the set of all such α's by $\mathscr{A}$; note that $\mathscr{A}_0$ may be a proper subset of $\mathscr{A}$ and α need not be a scalar.

2·2. *Finding efficient data-augmentation schemes: From* EM *to the Gibbs sampler*

To apply our augmentation approaches to the Gibbs sampler, we might consider choosing α on the basis of the so-called geometric rate of convergence of the Gibbs sampler. For the data augmentation algorithm, Amit (1991) has shown that the geometric rate of convergence is the square of the maximal correlation between θ and Y_{aug} under the joint stationary density $p(\theta, Y_{\text{aug}} | Y_{\text{obs}}, \alpha)$, namely

$$\begin{aligned}\lambda^{\text{DA}}(\alpha) &= \sup_{h:\text{var}\{h(\theta)|Y_{\text{obs}}\}=1} \text{var}[E\{h(\theta)|Y_{\text{aug}}, \alpha\}|Y_{\text{obs}}, \alpha] \\ &= 1 - \inf_{h:\text{var}\{h(\theta)|Y_{\text{obs}}\}=1} E[\text{var}\{h(\theta)|Y_{\text{aug}}, \alpha\}|Y_{\text{obs}}, \alpha]. \end{aligned} \qquad (2{\cdot}7)$$

The right-most term in (2·7) relates $\lambda^{\text{DA}}(\alpha)$ to the maximal fraction of missing information (Liu, 1994a). Throughout the rest of the paper we will assume all expectation calculations are well defined, as in (2·7).

However, it is clear from (2·7) that $\lambda^{\text{DA}}(\alpha)$ is not a practical criterion for choosing α except in special cases, e.g. with Gaussian models. A more manageable criterion is the lag-1 autocorrelation, a common measure for studying the mixing rate of a Markov chain. If the chain from a data augmentation algorithm has reached equilibrium, Liu (1994a) establishes that, for any non-constant scalar-valued function $h(\theta)$,

$$\text{corr}\{h(\theta^{(t)}), h(\theta^{(t+1)})\} = \frac{\text{var}[E\{h(\theta)|Y_{\text{aug}}, \alpha\}|Y_{\text{obs}}, \alpha]}{\text{var}\{h(\theta)|Y_{\text{obs}}\}}.$$

Consequently, the maximum autocorrelation over linear combinations $h(\theta) = x^{\text{T}}\theta$, for $x \neq 0$, is given by

$$\sup_{x\neq 0} \text{corr}(x^{\text{T}}\theta^{(t)}, x^{\text{T}}\theta^{(t+1)}) = \sup_{x \neq 0} \frac{x^{\text{T}} \text{var}\{E(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\}x}{x^{\text{T}} \text{var}(\theta|Y_{\text{obs}})x} = \rho(\mathscr{F}_B(\alpha)), \qquad (2{\cdot}8)$$

where $\mathscr{F}_B(\alpha)$ is the Bayesian fraction of missing information for θ under $p(Y_{\text{aug}}|\theta, \alpha)$,

$$\begin{aligned}\mathscr{F}_B(\alpha) &= \{\text{var}(\theta|Y_{\text{obs}})\}^{-1} \text{var}\{E(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\} \\ &= I - \{\text{var}(\theta|Y_{\text{obs}})\}^{-1} E\{\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\},\end{aligned}$$

that is the fraction of the posterior variance of θ explained by the unobserved part of Y_{aug} (Rubin, 1987, p. 86), and $\rho(A)$ denotes the spectral radius of A. Thus, in order to reduce autocorrelation, we would like to maximise $E\{\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\}$ over α using the semi-positive definite ordering.

The autocorrelation criterion typically is still not practical because it requires the calculation of $E\{\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\}$ which itself may require simulation. Fortunately, the $I_{\text{aug}}(\alpha)$ criterion discussed in § 2·1 for choosing the optimal EM algorithm turns out to be a rather useful approximation here too, at least in the applications we have encountered. This approximation becomes exact when $p(\theta, Y_{\text{aug}}|Y_{\text{obs}}, \alpha)$ is normal, in which case

$$I_{\text{aug}}^{-1}(\alpha) = E\{\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\}$$

and $\lambda^{\text{DA}}(\alpha) = \rho(\mathscr{F}_B(\alpha)) = \rho(DM^{\text{EM}}(\alpha))$; e.g. Roberts & Sahu (1997). Even in cases where $I_{\text{aug}}^{-1}(\alpha)$ does not approximate $E\{\text{var}(\theta|Y_{\text{aug}}, \alpha)|Y_{\text{obs}}, \alpha\}$ well, its maximiser can still be a very good approximation to the maximiser of the latter. The t example discussed in § 2·1 illustrates this point.

Suppose we want to use a data augmentation algorithm to sample from the posterior $p(\mu, \sigma^2|Y_{\text{obs}})$, under the t-model given in (2·3) using the standard non-informative prior $p(\mu, \log(\sigma^2)) \propto 1$. Under the new augmentation scheme $Y_{\text{aug}} = \{(y_i, q_i), i = 1, \ldots, n\}$, defined by (2·4), we have

$$q_i \,\Big|\, \mu, \sigma^2, Y_{\text{obs}}, \alpha \sim \frac{\sigma^{-2\alpha}}{(y_i - \mu)^2/\sigma^2 + \nu} \chi^2_{\nu+1}, \qquad (2{\cdot}9)$$

independently for $i = 1, \ldots, n$,

$$\mu \,\Big|\, \sigma^2, Y_{\rm aug}, \alpha \sim N\left(\hat{\mu}, \frac{\sigma^{2(1-\alpha)}}{\sum_{i=1}^n q_i}\right), \qquad (2{\cdot}10)$$

where $\hat{\mu} = \sum_{i=1}^n q_i y_i / \sum_{i=1}^n q_i$, and

$$p(\sigma^2 \,|\, Y_{\rm aug}, \alpha) \propto \sigma^{n\{\alpha(v+1)-1\}-(1+\alpha)} \exp\left[-\frac{\sigma^{2\alpha}}{2} \sum_{i=1}^n q_i \left\{\frac{(y_i - \hat{\mu})^2}{\sigma^2} + v\right\}\right]. \qquad (2{\cdot}11)$$

Thus, the two steps of each iteration consist of first drawing $q = (q_1, \ldots, q_n)$ from (2·9) and then drawing $\theta = (\mu, \sigma^2)$ jointly using (2·10) and (2·11). The density in (2·11) is not easily sampled unless $\alpha = 0$ or $\alpha = 1$. Gilks (1997) uses the adaptive rejection Metropolis sampling to implement this step and his empirical result, with $n = 100$, shows that the closer α is to the optimal value for implementing EM, $\alpha^{\rm EM}_{\rm opt} = 1/(1+v)$, the smaller the autocorrelation, $\mathrm{corr}(\tau^{(t)}, \tau^{(t+1)})$, where $\tau = \sigma^{-2}$.

Gilks's (1997) empirical validation of the conjecture made in Meng & van Dyk (1997) that $\alpha^{\rm EM}_{\rm opt}$ should also work well for the Gibbs implementation would not be of much interest if $p(\theta, q \,|\, Y_{\rm obs}, \alpha)$ could be well approximated by a multivariate normal, which is not the case; a plot of $p(\tau \,|\, Y_{\rm obs}, \alpha)$ will show obvious nonnormal character. Furthermore, Gilks's (1997) implementation was actually a three-step Gibbs sampler which drew from $p(\tau \,|\, \mu, Y_{\rm aug}, \alpha)$ in place of (2·11), and thus was not the data augmentation algorithm on which (2·7) and (2·8) are based. This is rather encouraging; the usefulness of the $\alpha^{\rm EM}_{\rm opt}$ approximation is apparent because it is typically much more difficult to compute $E\{\mathrm{var}(\theta \,|\, Y_{\rm aug}, \alpha) \,|\, Y_{\rm obs}, \alpha\}$ than to compute $I_{\rm aug}(\alpha)$. For example, calculating $E\{\mathrm{var}(\tau \,|\, Y_{\rm aug}, \alpha) \,|\, Y_{\rm obs}, \alpha\}$ requires first finding $\mathrm{var}(\sigma^{-2} \,|\, Y_{\rm aug}, \alpha)$ with respect to (2·11) and then averaging it over $p(q \,|\, Y_{\rm obs}, \alpha)$. Neither of the two steps is analytically tractable; in computation, we used numerical integration in the first step and Monte Carlo integration in the second step, using the data augmentation algorithm to simulate from $p(q \,|\, Y_{\rm obs}, \alpha)$.

Table 1 provides the results with $n = 100$ and the same values of v and α, with the addition of $\alpha = 0{\cdot}9$ and $\alpha = 1$, as in Gilks (1997); we used 30 000 Gibbs draws for the Monte Carlo integration in each cell. While these values are 0·03–0·17 larger than the corresponding values given by Gilks (1997), partly as a result of different $Y_{\rm obs}$, they show the same pattern as Gilks's (1997) table. In particular, for each v, the smallest value of $\mathrm{corr}(\tau^{(t)}, \tau^{(t+1)})$ is given by the α that is the closest to $1/(1+v)$. Although this does not imply that $\alpha = 1/(1+v)$ exactly maximises $E\{\mathrm{var}(\tau \,|\, Y_{\rm aug}, \alpha) \,|\, Y_{\rm obs}, \alpha\}$, it is clear that $\alpha^{\rm EM}_{\rm opt} = 1/(1+v)$ can be used to approximate $a^{\rm DA}_{\rm opt}$ for practical purposes. In § 3, we will show that we actually can achieve the same convergence rate as $Y_{\rm aug}$ with $\alpha = \alpha^{\rm EM}_{\rm opt}$ but without the unpleasant and time-consuming draws from (2·11), which offsets the gain from the

Table 1. *The autocorrelation* $\mathrm{corr}(\tau^{(t)}, \tau^{(t+1)})$ *as a function of* α *and* v. *As the* $\alpha^{\rm EM}_{\rm opt}$ *approximation suggests, the optimal value of* α *is near* $1/(v+1)$; *the corresponding values of the autocorrelation are underlined*

						α					
v	0·0	0·1	0·2	0·3	0·4	0·5	0·6	0·7	0·8	0·9	1·0
1	0·80	0·75	0·70	0·65	0·61	$\underline{0{\cdot}60}$	0·61	0·65	0·70	0·75	0·79
2	0·69	0·62	0·56	$\underline{0{\cdot}53}$	0·54	0·58	0·64	0·70	0·76	0·80	0·84
4	0·56	0·48	$\underline{0{\cdot}45}$	0·48	0·56	0·65	0·72	0·78	0·83	0·86	0·89
9	0·27	$\underline{0{\cdot}19}$	0·27	0·44	0·60	0·71	0·78	0·84	0·87	0·90	0·92

faster mixing rate in Gilks's (1997) implementation, and thus we provide an efficient algorithm for posterior sampling under the t model.

3. Marginal augmentation

3·1. *Motivating marginal augmentation from conditional augmentation*

As seen in § 2, in order to minimise the lag-1 autocorrelation, the conditional augmentation approach aims to maximise $E\{\text{var}(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{obs}}, \alpha\}$. By viewing α as a random variable, we recognise that there is another way to increase $E\{\text{var}(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{obs}}, \alpha\}$. Suppose we assign a proper 'working' prior distribution, $p(\alpha)$, to α and define the joint distribution of $(\theta, \alpha, Y_{\text{aug}})$ as

$$p(\theta, \alpha, Y_{\text{aug}}) = p(Y_{\text{aug}} \mid \theta, \alpha) p(\theta) p(\alpha). \tag{3·1}$$

Since α is a working parameter, the posterior distribution of θ given Y_{obs} implied by (3·1) is proportional to $p(Y_{\text{obs}} \mid \theta)p(\theta)$, our original model. Note that (3·1) assumes θ and α are a priori independent, but, if we replace $p(\alpha)$ by $p(\alpha \mid \theta)$ in (3·1), the implied conditional distribution $p(\theta \mid Y_{\text{obs}})$ is unchanged. For simplicity of presentation, we do not pursue the dependent case in this paper.

Under the joint distribution (3·1), it is easy to verify that

$$\begin{aligned} E\{\text{var}(\theta \mid Y_{\text{aug}}) \mid Y_{\text{obs}}\} &= E[E\{\text{var}(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{aug}}\} \mid Y_{\text{obs}}] + E[\text{var}\{E(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{aug}}\} \mid Y_{\text{obs}}] \\ &= E[E\{\text{var}(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{obs}}, \alpha\} \mid Y_{\text{obs}}] + E[\text{var}\{E(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{aug}}\} \mid Y_{\text{obs}}] \\ &\geqslant E[E\{\text{var}(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{obs}}, \alpha\} \mid Y_{\text{obs}}]. \end{aligned}$$

Therefore, if we use the marginal augmentation induced by (3·1), that is

$$p(Y_{\text{aug}} \mid \theta) = \int p(Y_{\text{aug}} \mid \theta, \alpha) p(\alpha)\, d\alpha, \tag{3·2}$$

then on average, with respect to α, the Bayesian fraction of missing information will be no larger than that from the conditional augmentation, $p(Y_{\text{aug}} \mid \theta, \alpha)$. When $E\{\text{var}(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{obs}}, \alpha\}$ does not depend on α, we have the following stronger result.

Lemma 1. *If* $E\{\text{var}(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{obs}}, \alpha\}$ *does not depend on the working parameter,* α, *then* $\rho(\mathscr{F}_B) \leqslant \rho(\mathscr{F}_B(\alpha))$ *for each* $\alpha \in \mathscr{A}$, *where* $\mathscr{F}_B = \{\text{var}(\theta \mid Y_{\text{obs}})\}^{-1} \text{var}\{E(\theta \mid Y_{\text{aug}}) \mid Y_{\text{obs}}\}$ *and* $\mathscr{F}_B(\alpha) = \{\text{var}(\theta \mid Y_{\text{obs}})\}^{-1} \text{var}\{E(\theta \mid Y_{\text{aug}}, \alpha) \mid Y_{\text{obs}}, \alpha\}$.

In other words, the data augmentation algorithm under the marginal augmentation $p(Y_{\text{aug}} \mid \theta)$ will produce lag-1 autocorrelations over linear functions of θ no greater than those under the conditional augmentation $p(Y_{\text{aug}} \mid \theta, \alpha)$. In view of the second expression in (2·7), we also have an analogous inequality between the geometric rate of convergence of the data augmentation algorithm under marginal augmentation, λ^{DA}, and that under conditional augmentation.

Lemma 2. *If for any scalar-valued function* h, $E[\text{var}\{h(\theta) \mid Y_{\text{aug}}, \alpha\} \mid Y_{\text{obs}}, \alpha]$ *does not depend on* α, *then* $\lambda^{\text{DA}} \leqslant \lambda^{\text{DA}}(\alpha)$ *for all* $\alpha \in \mathscr{A}$.

When the original augmentation scheme can be written as $\tilde{Y}_{\text{aug}} = (\tilde{Y}_{\text{mis}}, Y_{\text{obs}})$, the supposition of Lemmas 1 and 2 is satisfied when the mapping $\mathscr{D}_{\alpha,\theta}$ defined in (2·5) does not depend on θ, that is when we create a conditional augmentation via $Y_{\text{mis}} = \mathscr{D}_{\alpha}(\tilde{Y}_{\text{mis}})$. This is summarised in the following result.

THEOREM 1. *Given an augmentation scheme, $\tilde{Y}_{\text{aug}} = \{\tilde{Y}_{\text{mis}}, Y_{\text{obs}}\}$, and a one-to-one mapping $\mathscr{D}_\alpha(.)$ indexed by a working parameter $\alpha \in \mathscr{A}$, the class of conditional augmentations created by $\{Y_{\text{mis}}, Y_{\text{obs}}\} = \{\mathscr{D}_\alpha(\tilde{Y}_{\text{mis}}), Y_{\text{obs}}\}$ are equivalent, that is $\lambda^{\text{DA}}(\alpha) \equiv \lambda$. Furthermore, for any given proper prior $p(\alpha)$, the geometric rate of convergence of the data augmentation algorithm under the corresponding marginal augmentation*

$$p(Y_{\text{mis}}, Y_{\text{obs}} \mid \theta) = \int p(Y_{\text{mis}}, Y_{\text{obs}} \mid \theta, \alpha) p(\alpha)\, d\alpha$$

cannot exceed λ.

Proof. Under the joint distribution (3·1), $(\theta, \tilde{Y}_{\text{mis}}, Y_{\text{obs}})$ is jointly independent of α, where $\tilde{Y}_{\text{mis}} = \mathscr{D}_\alpha^{-1}(Y_{\text{mis}})$. This is because, by (2·6) and (3·1),

$$p(\theta, \tilde{Y}_{\text{mis}}, Y_{\text{obs}}, \alpha) = \{\tilde{p}(\tilde{Y}_{\text{mis}} \mid Y_{\text{obs}}, \theta) p(Y_{\text{obs}} \mid \theta) p(\theta)\} p(\alpha).$$

Consequently, θ is independent of α given $(\tilde{Y}_{\text{mis}}, Y_{\text{obs}})$ and $\tilde{Y}_{\text{mis}}$ is independent of α given Y_{obs}. It follows then that, for any $h(\theta)$,

$$\begin{aligned} E[\text{var}\{h(\theta) \mid Y_{\text{mis}}, Y_{\text{obs}}, \alpha\} \mid Y_{\text{obs}}, \alpha] &= E[\text{var}\{h(\theta) \mid \tilde{Y}_{\text{mis}}, Y_{\text{obs}}, \alpha\} \mid Y_{\text{obs}}, \alpha] \\ &= E[\text{var}\{h(\theta) \mid \tilde{Y}_{\text{mis}}, Y_{\text{obs}}\} \mid Y_{\text{obs}}], \end{aligned}$$

which is free of α. □

The implication of this result is that we can always try to improve an augmentation scheme $\tilde{Y}_{\text{aug}} = \{\tilde{Y}_{\text{mis}}, Y_{\text{obs}}\}$ by introducing a working parameter via a mapping $Y_{\text{mis}} = \mathscr{D}_\alpha(\tilde{Y}_{\text{mis}})$ and then implementing the data augmentation algorithm under the marginal augmentation with a suitable choice of $p(\alpha)$ which is independent of $p(\theta)$. The two main requirements for choosing the mapping $\mathscr{D}_\alpha$ and the prior $p(\alpha)$ are as follows.

Requirement 1. Given θ, α and Y_{aug} are not independent.

Requirement 2. It is relatively easy to draw from $p(\alpha \mid Y_{\text{aug}})$.

Requirement 1 excludes the trivial choice of α, that is $p(Y_{\text{aug}} \mid \theta, \alpha)$ does not depend on α, and Requirement 2 ensures the savings from faster mixing will not be offset by the computational costs involved with the marginal augmentation approach. Additional requirements, such as ensuring that $p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta, \alpha)$ is easy to draw from, should be a part of the requirements for the original augmentation scheme $\tilde{p}(\tilde{Y}_{\text{mis}} \mid Y_{\text{obs}}, \theta)$.

3·2. *Implementing the Gibbs sampler under marginal augmentation*

Consider the t example of § 2, but now with the following construction:

$$y = \mu + \frac{\sqrt{\alpha\sigma}z}{\sqrt{q}}, \quad z \sim N(0, 1), \quad q \mid \alpha \sim \alpha\chi_\nu^2/\nu, \quad z \perp q. \tag{3·3}$$

In other words, we use $\mathscr{D}_\alpha(\tilde{Y}_{\text{mis}}) = \{\alpha\tilde{q}_1, \ldots, \alpha\tilde{q}_n\}$, where $\{\tilde{q}_1, \ldots, \tilde{q}_n\}$ are the missing data in the original augmentation as in (2·3). This scheme is clearly not suitable for the conditional augmentation approach because $\mathscr{D}_\alpha$ does not depend on θ. However, Liu et al. (1998), who introduced (3·3), have shown that when (3·3) is used with the parameter-expanded EM, that is PXEM, algorithm it leads to a PXEM implementation which is identical to the optimal EM obtained under the conditional augmentation approach using augmentation scheme (2·4) with $\alpha = 1/(1+\nu)$. This suggests that (3·3) may be efficient for implementing the Gibbs sampler for drawing from $p(\mu, \sigma^2 \mid Y_{\text{obs}})$.

To derive a data augmentation implementation under (3·3), we choose $\alpha \sim \beta/\chi^2_\gamma$, where $\beta, \gamma > 0$ are given; the choices of β and γ will be the subject of § 3·4. Under this proper prior for α and the standard improper prior $p(\mu, \log\sigma^2) \propto 1$, it is straightforward to derive that

$$q_i \,\Big|\, \mu, \sigma^2, Y_{\text{obs}}, \alpha \sim \frac{\alpha}{(y_i - \mu)^2/\sigma^2 + \nu} \chi^2_{\nu+1} \tag{3·4}$$

independently for $i = 1, \ldots, n$,

$$\mu \,\Big|\, \sigma^2, Y_{\text{aug}}, \alpha \sim N\left(\hat{\mu}, \frac{\alpha\sigma^2}{\sum_{i=1}^n q_i}\right), \tag{3·5}$$

where $\hat{\mu} = \sum_{i=1}^n q_i y_i / \sum_{i=1}^n q_i$, and

$$\sigma^2 \,\Big|\, Y_{\text{aug}}, \alpha \sim \frac{\sum_{i=1}^n q_i (y_i - \hat{\mu})^2}{\alpha \chi^2_{n-1}}, \tag{3·6}$$

$$\alpha \,\Big|\, Y_{\text{aug}} \sim \frac{\beta + \nu \sum_{i=1}^n q_i}{\chi^2_{\gamma+n\nu}}. \tag{3·7}$$

Given these conditional distributions, we can implement a data augmentation algorithm with the marginal augmentation (3·2) as follows. At the $(t+1)$st iteration, we draw $q^{(t+1)}$ from the marginal augmentation

$$p(q \mid \mu^{(t)}, [\sigma^2]^{(t)}, Y_{\text{obs}}) = \int p(q \mid \mu^{(t)}, [\sigma^2]^{(t)}, Y_{\text{obs}}, \alpha) p(\alpha)\, d\alpha,$$

by first drawing $\tilde{\alpha}^{(t+1)}$ from its prior $p(\alpha)$ and then $q^{(t+1)}$ from $p(q \mid \mu^{(t)}, [\sigma^2]^{(t)}, Y_{\text{obs}}, \alpha = \tilde{\alpha}^{(t+1)})$ as given in (3·4). Given $q = q^{(t+1)}$, draw $(\mu^{(t+1)}, [\sigma^2]^{(t+1)})$ from $p(\mu, \sigma^2 \mid Y_{\text{aug}}) = \int p(\mu, \sigma^2 \mid Y_{\text{aug}}, \alpha) p(\alpha \mid Y_{\text{aug}})\, d\alpha$ by first drawing $\alpha^{(t+1)}$ from the posterior $p(\alpha \mid Y_{\text{aug}})$ given by (3·7), then drawing $[\sigma^2]^{(t+1)}$ from (3·6) given $\alpha = \alpha^{(t+1)}$, and finally drawing $\mu^{(t+1)}$ from (3·5) given $\alpha = \alpha^{(t+1)}$ and $\sigma^2 = [\sigma^2]^{(t+1)}$. As a comparison, the conditional augmentation approach would fix α at a particular value, for example $\alpha = 1$, in (3·4)–(3·6), and ignore (3·7). It is easy to see that when α is fixed at a the actual value of a is irrelevant for $\{(\mu^{(t)}, [\sigma^2]^{(t)}), t \geqslant 0\}$.

The description for the t procedure is again general. The data augmentation algorithm under the marginal augmentation (3·2) is typically implemented according to the following iterative scheme.

Step 1. Draw $\tilde{\alpha}^{(t+1)}$ from the prior $p(\alpha)$, and then draw $Y^{(t+1)}_{\text{aug}}$ from $p(Y_{\text{aug}} \mid Y_{\text{obs}}, \theta^{(t)}, \tilde{\alpha}^{(t+1)})$.

Step 2. Draw $\alpha^{(t+1)}$ from the posterior $p(\alpha \mid Y^{(t+1)}_{\text{aug}})$, and then draw $\theta^{(t+1)}$ from $p(\theta \mid Y^{(t+1)}_{\text{aug}}, \alpha^{(t+1)})$.

To ensure easy drawing from both $p(\alpha)$ and $p(\alpha \mid Y_{\text{aug}})$, it is generally effective to use conditional conjugate working priors with respect to $p(Y_{\text{aug}} \mid \theta, \alpha)$, as we did with the t model.

It is important to distinguish the $\tilde{\alpha}^{(t+1)}$ in the first step from the $\alpha^{(t+1)}$ in the second step. While $\tilde{\alpha}^{(t+1)}$ facilitates the draw from $p(Y_{\text{aug}} \mid \theta)$, $\alpha^{(t+1)}$ facilitates the draw from $p(\theta \mid Y_{\text{aug}})$. Both of these distributions have α integrated out and thus can be much more difficult to draw from directly, as with the t model. Since $\tilde{\alpha}^{(t+1)}$ is independent of $\alpha^{(t)}$, the working parameters are not a part of the Markov chain. The resulting chain is given by $\{(\theta^{(t)}, Y_{\text{aug}}^{(t)}), t \geqslant 0\}$, and all results regarding the standard data augmentation algorithm apply, e.g. both $\{\theta^{(t)}, t \geqslant 0\}$ and $\{Y_{\text{aug}}^{(t)}, t \geqslant 0\}$ are reversible Markov chains; see Liu, Wong & Kong (1994).

The above discussion suggests that there is an alternative implementation that sets $\tilde{\alpha}^{(t+1)} = \alpha^{(t)}$ instead of drawing a new $\tilde{\alpha}^{(t+1)}$ from the prior $p(\alpha)$. This is simply implementing the data augmentation algorithm with $\tilde{\theta} = (\theta, \alpha)$ by iteratively drawing from $p(Y_{\text{aug}} \mid \tilde{\theta}, Y_{\text{obs}})$ and $p(\tilde{\theta} \mid Y_{\text{aug}})$. These two schedules for implementation correspond to Scheme [1] and Scheme [2] of Liu et al. (1994).

Scheme [1]. We iteratively draw from $p(Y_{\text{aug}} \mid \theta, Y_{\text{obs}})$ and $p(\theta \mid Y_{\text{aug}})$, which induces a marginal Markov chain for θ.

Scheme [2]. We iteratively draw from $p(Y_{\text{aug}} \mid \theta, \alpha, Y_{\text{obs}})$ and $p(\theta, \alpha \mid Y_{\text{aug}})$, which induces a joint Markov chain for (θ, α).

Since α is not identifiable given Y_{obs}, the invariant distribution given by Scheme [2] is

$$p(\tilde{\theta}, Y_{\text{aug}} \mid Y_{\text{obs}}) = p(\theta, Y_{\text{aug}} \mid Y_{\text{obs}}, \alpha) p(\alpha), \tag{3·8}$$

and thus the limiting distribution of $\{\theta^{(t)}, t \geqslant 0\}$ under Scheme [2] is our target distribution $p(\theta \mid Y_{\text{obs}})$.

While Scheme [1] and Scheme [2] have the same lag-1 autocorrelation for linear combinations of $\theta^{(t)}$, the geometric rate of convergence of Scheme [1] cannot be bigger than that of Scheme [2] because the maximum correlation between θ and Y_{aug} cannot exceed that of $\tilde{\theta}$ and Y_{aug} (Liu et al., 1994). We note that, when α and Y_{aug} are independent given θ, the two maximum correlations are equal and thus the two methods have the same geometric rate of convergence, but this possibility is excluded by Requirement 1 given in § 3·1.

Another difference between Schemes [1] and [2] is that the induced marginal chain $\{\theta^{(t)}, t > 0\}$ is a Markov chain under Scheme [1], but not necessarily under Scheme [2]. However, with additional assumptions which are met in our applications, we can conclude that $\{\theta^{(t)}, t > 0\}$ is a Markov chain under Scheme [2].

LEMMA 3. *Suppose the conditional augmentation is constructed via a one-to-one mapping* $\mathscr{D}_\alpha(.)$ *indexed by a working parameter* $\alpha \in \mathscr{A}$, *that is* $\{Y_{\text{mis}}, Y_{\text{obs}}\} = \{\mathscr{D}_\alpha(\tilde{Y}_{\text{mis}}), Y_{\text{obs}}\}$, *where* $\tilde{p}(\tilde{Y}_{\text{mis}} \mid Y_{\text{obs}}, \theta)$ *is the original augmentation scheme which does not depend on* α, *and* $p(\alpha)$ *is a prior on* $\mathscr{A}$ *such that Scheme* [2] *is computable. Suppose also that the posterior of* θ *given* $\{Y_{\text{mis}}, Y_{\text{obs}}\} = \{\mathscr{D}_\alpha(\tilde{Y}_{\text{mis}}), Y_{\text{obs}}\}$ *does not depend on the value of* $\alpha \in \mathscr{A}$. *Then the marginal chain* $\{\theta^{(t)}, t > 0\}$ *induced by Scheme* [2] *is a Markov chain.*

Proof. Since $\theta^{(t+1)}$ is a draw from $p(\theta \mid Y_{\text{mis}}^{(t+1)}, Y_{\text{obs}})$ where $Y_{\text{mis}}^{(t+1)} = \mathscr{D}_{\alpha^{(t)}}(\tilde{Y}_{\text{mis}}^{(t+1)})$ and $\tilde{Y}_{\text{mis}}^{(t+1)}$ is a draw from $\tilde{p}(\tilde{Y}_{\text{mis}} \mid Y_{\text{obs}}, \theta)$ which does not depend on $\alpha^{(t)}$, the transition probability $p(\theta^{(t+1)} \mid \theta^{(t)}, \alpha^{(t)})$ does not depend on $\alpha^{(t)}$ under our assumption that $p(\theta \mid \mathscr{D}_{\alpha^{(t)}}(\tilde{Y}_{\text{mis}}^{(t+1)}), Y_{\text{obs}})$ does not depend on $\alpha^{(t)}$. □

3·3. *Potential benefits of nonpositive recurrent Markov chains*

The previous discussion suggests that as long as drawing α from the working prior is relatively simple one should always use Scheme [1]. If an improper prior for α is used, then obviously Scheme [1] cannot be implemented. However, Scheme [2] can be, as long as $p(\alpha | Y_{\text{aug}})$ is proper, which is always the case under Requirement 2 of § 3·1. Of course, since α is not identifiable given Y_{obs}, we see from (3·8) that the invariant distribution of the joint chain produced by Scheme [2] is improper. Consequently, the joint chain $\{(\theta^{(t)}, \alpha^{(t)}, Y_{\text{aug}}^{(t)}), t \geqslant 0\}$ is not positive recurrent; we assume irreducibility throughout.

Using an improper working prior is a nontrivial extension in that it can greatly complicate the convergence behaviour, particularly because the marginal augmentation in (3·2) is not even defined; § 3·4 will use the t example to illustrate this point. The advantage, however, is illustrated in Fig. 1 which shows $\text{corr}(\tau^{(t)}, \tau^{(t+1)})$ under Scheme [1] as a function of γ using the same four datasets as in Table 1. The parameters for $p(\alpha | \beta, \gamma)$ are selected so that $E(\alpha^{-1}) = \gamma\beta^{-1}$ is held constant, $\gamma\beta^{-1} = c$, and thus $\text{var}(\alpha^{-1}) = 2\gamma\beta^{-2}$ is inversely proportional to γ. It is clear from Fig. 1 that for all four datasets the autocorrelation is reduced as $\log(\gamma) \to -\infty$, which suggests the choice $\gamma = 0$; the limiting levels match the underlined values in Table 1 well, a phenomenon discussed in § 4·2. However, in the limit as $\gamma \to 0$, $p(\alpha | \beta = \gamma/c, \gamma) \propto \alpha^{-1}$, an improper prior.

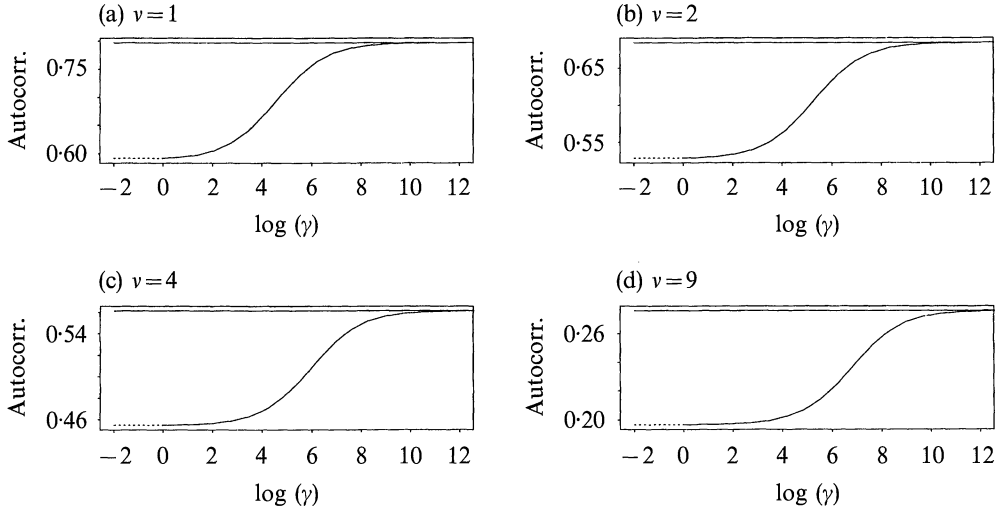

Fig. 1. Reduction in autocorrelation using marginal data augmentation. The plots illustrate the effect of the hyperparameter γ, with $\beta = \gamma/50$, on the autocorrelation under marginal augmentation. The horizontal lines represent the autocorrelation under the standard augmentation.

While the joint chain $\{(\theta^{(t)}, \alpha^{(t)}, Y_{\text{aug}}^{(t)}), t \geqslant 0\}$, where $\theta^{(t)} = (\mu^{(t)}, [\sigma^2]^{(t)})$, does not converge jointly in distribution with $p(\alpha) \propto \alpha^{-1}$, the induced marginal chain $\{\theta^{(t)}, t \geqslant 0\}$ is a positive recurrent Markov chain with $p(\theta | Y_{\text{obs}})$ as its invariant distribution. To see this, we note from (3·5)–(3·7) that, when $\beta = 0$, $p(\theta | \alpha\tilde{q}, Y_{\text{obs}})$ does depend on $\alpha (> 0)$. Consequently, the condition of Lemma 3 is satisfied and thus $\{\theta^{(t)}, t \geqslant 0\}$ is a Markov chain; note that Lemma 3 does not require $p(\alpha)$ be proper. That the target density, $p(\theta | Y_{\text{obs}})$, is the invariant distribution of this chain is a consequence of the following result, which can also be used to verify the Markovian property of $\{\theta^{(t)}, t \geqslant 0\}$.

Theorem 2. *Suppose an improper prior, $p(\alpha)$, is used for implementing Scheme* [2] *that induces an irreducible Markov chain $\{\theta^{(t)}, \alpha^{(t)}, t \geqslant 0\}$ with transition kernel $p(\theta, \alpha | \theta', \alpha')$. Suppose for any pair $\theta, \theta' \in \Theta$ there exists a sequence of proper prior distributions $p_m(\alpha)$ such that the corresponding kernels $p_m(\theta | \theta')$ under Scheme* [1] *converge to $p(\theta | \theta', \alpha') = \int p(\theta, \alpha | \theta', \alpha') \, d\alpha$ when $m \to \infty$. Then $\{\theta^{(t)}, t > 0\}$ is a positive recurrent reversible Markov chain with $p(\theta | Y_{\text{obs}})$ as its unique invariant distribution.*

Proof. Since Scheme [1] is a standard data augmentation algorithm under each proper prior, $p_m(\alpha)$, by Lemma 3.1 of Liu et al. (1994), $p_m(\theta \mid \theta')$ satisfies the detailed balance condition,

$$p_m(\theta \mid \theta')p(\theta' \mid Y_{\text{obs}}) = p_m(\theta' \mid \theta)p(\theta \mid Y_{\text{obs}}). \tag{3·9}$$

When $m \rightarrow \infty$, the conditions of Theorem 2 imply not only that $\{\theta^{(t)}, t > 0\}$ is an irreducible Markov chain because $p(\theta \mid \theta', \alpha') = \lim_m p_m(\theta \mid \theta')$ does not depend on α', but also that the detailed balance condition is satisfied by its transition kernel $p(\theta \mid \theta') \equiv p(\theta \mid \theta', \alpha')$. □

It is clear from the proof that as long as there is a sequence $p_m(\theta \mid \theta')$, satisfying (3·9), that converges to $p(\theta \mid \theta', \alpha')$ as $m \rightarrow \infty$, the result of Theorem 2 holds. That is, $p_m(\theta \mid \theta')$ does not need to be from Scheme [1], although in practice it is most natural to construct $\{p_m(\theta \mid \theta'), m \geqslant 1\}$ from Scheme [1]. For our t problem, this comparison between Scheme [1] and Scheme [2] also allows us to determine what values of the 'hyperparameters' β and γ can be used when $p(\alpha \mid \beta, \gamma)$ is improper; see § 3·4. We start by noting that when $p(\alpha \mid \beta, \gamma)$ is proper, that is when $\beta > 0$ and $\gamma > 0$, Scheme [1] can be represented as follows by a stochastic mapping, $\theta^{(t)} \rightarrow \theta^{(t+1)}$, where $\theta = (\mu, \sigma^2)$; see (3·4)–(3·7) for the derivation.

Step 1. Draw independently $Z \sim N(0, 1)$, χ^2_γ, χ^2_{n-1}, $\chi^2_{\gamma+n\nu}$ and n copies of $\chi^2_{\nu+1}$, denoted by $\{\chi^2_{\nu+1,1}, \ldots, \chi^2_{\nu+1,n}\}$.

Step 2. Compute

$$w_i^{(t)} = \left\{\frac{(y_i - \mu^{(t)})^2}{[\sigma^2]^{(t)}} + \nu\right\}^{-1} \quad (i = 1, \ldots, n),$$

$$\hat{\mu}^{(t+1)} = \frac{\sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i} y_i}{\sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i}}. \tag{3·10}$$

Step 3. Compute

$$\mu^{(t+1)} = \hat{\mu}^{(t+1)} + Z\sqrt{\left\{\frac{\sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i}(y_i - \hat{\mu}^{(t+1)})^2}{\chi^2_{n-1} \sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i}}\right\}}, \tag{3·11}$$

$$[\sigma^2]^{(t+1)} = \frac{\sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i}(y_i - \hat{\mu}^{(t+1)})^2}{\chi^2_\gamma + \nu \sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i}} \times \frac{\chi^2_{\gamma+n\nu}}{\chi^2_{n-1}}. \tag{3·12}$$

Expressions (3·10)–(3·12) provide an explicit stochastic-mapping representation of the transition kernel from $\theta^{(t)} \rightarrow \theta^{(t+1)}$, which we find easier to use in the following investigation than the transition kernel itself.

Under Scheme [2], the mapping for μ, that is (3·11), is unchanged, but the mapping for σ^2 is replaced by

$$[\tilde{\sigma}^2]^{(t+1)} = \frac{\sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i}(y_i - \hat{\mu}^{(t+1)})^2}{\beta/\alpha^{(t)} + \nu \sum_{i=1}^n w_i^{(t)} \chi^2_{\nu+1,i}} \times \frac{\chi^2_{\gamma+n\nu}}{\chi^2_{n-1}}, \tag{3·13}$$

where $\alpha^{(t)}$ is from the tth iteration; recall that Scheme [2] is a joint chain on $\{\mu, \sigma^2, \alpha\}$. If we compare (3·13) with (3·12), we see that when $\beta = \gamma = 0$ the transition kernel under Scheme [2] is the limit of the transition kernel under Scheme [1] as $\gamma \downarrow 0$ with any fixed $\beta > 0$, because χ^2_γ becomes a point mass at zero as $\gamma \downarrow 0$. Consequently, Theorem 2 is applicable. In fact, when Scheme [1] is expressed as (3·10)–(3·12), we can formally allow Scheme [1] to admit the case $\gamma = 0$ by defining $\chi^2_{\gamma=0} = 0$ in (3·12).

The key message from the t example is not just that the marginal chain $\{\theta^{(t)}, t > 0\}$ converges properly when $\beta = \gamma = 0$, but also that it has the fastest mixing rate in the class of algorithms underlying Fig. 1. Figure 2 displays the relative gains offered by this algorithm, Scheme [2] with $\beta = \gamma = 0$, over the commonly used standard algorithm, i.e. with $\alpha = 1$ throughout the iteration. In Fig. 2 the two algorithms are compared using three independent chains each for σ^2 under the Cauchy model, i.e. when $\nu = 1$, starting from three over-dispersed initial values. Figures 2(a) and (b) display the estimated autocorrelations of chain one under each algorithm. The improved algorithm not only reduces the lag-1 autocorrelation from 0·8 to 0·6, but also substantially reduces the number of consecutive iterations between draws that are essentially uncorrelated. Figures 2(c) and (d) show a time-series plot of the realisations from chain one, and Fig. 2(e) and (f) plot the estimated potential scale reduction factor $\hat{R}^{\frac{1}{2}}$ of Gelman & Rubin (1992) based on all three chains as a measure of convergence. The improved algorithm reaches an acceptable $\hat{R}^{\frac{1}{2}}$ much faster than the standard algorithm.

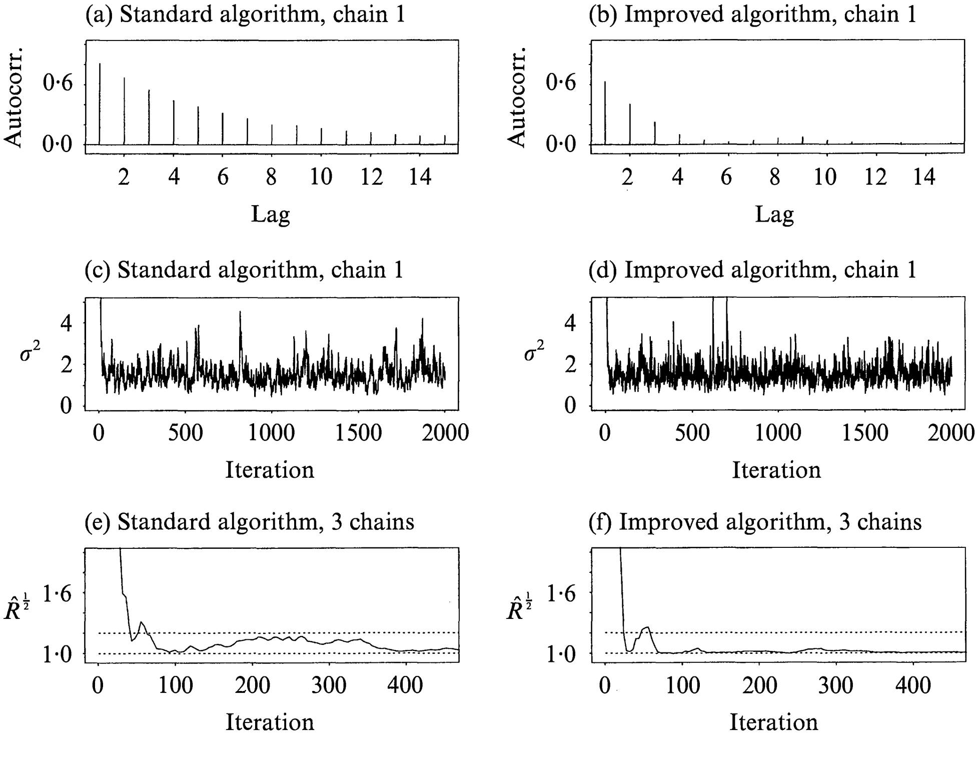

Fig. 2. Comparing the standard and improved algorithms for a univariate Cauchy model. (a) and (b), and (c) and (d), compare, respectively, the autocorrelation and time series plots of σ^2 for one chain. (e) and (f) compare the Gelman–Rubin $\hat{R}^{\frac{1}{2}}$ statistic based on three independent chains, where values near one give evidence of convergence.

The relative improvement becomes more dramatic with a multivariate t model, in parallel to the findings for EM in Meng & van Dyk (1997). Figure 3 repeats Fig. 2 with a four-dimensional Cauchy distribution, using the multivariate counterparts of (3·4)–(3·6). Under the improved algorithm the autocorrelation is effectively zero after lag two, while under the standard algorithm this does not happen until after lag ten. It is remarkable that such striking gains are achieved with a simple addition of the draw of α given by (3·7) with $\beta=\gamma=0$, essentially the same 'free lunch' as in Meng & van Dyk (1997) for the EM implementation.

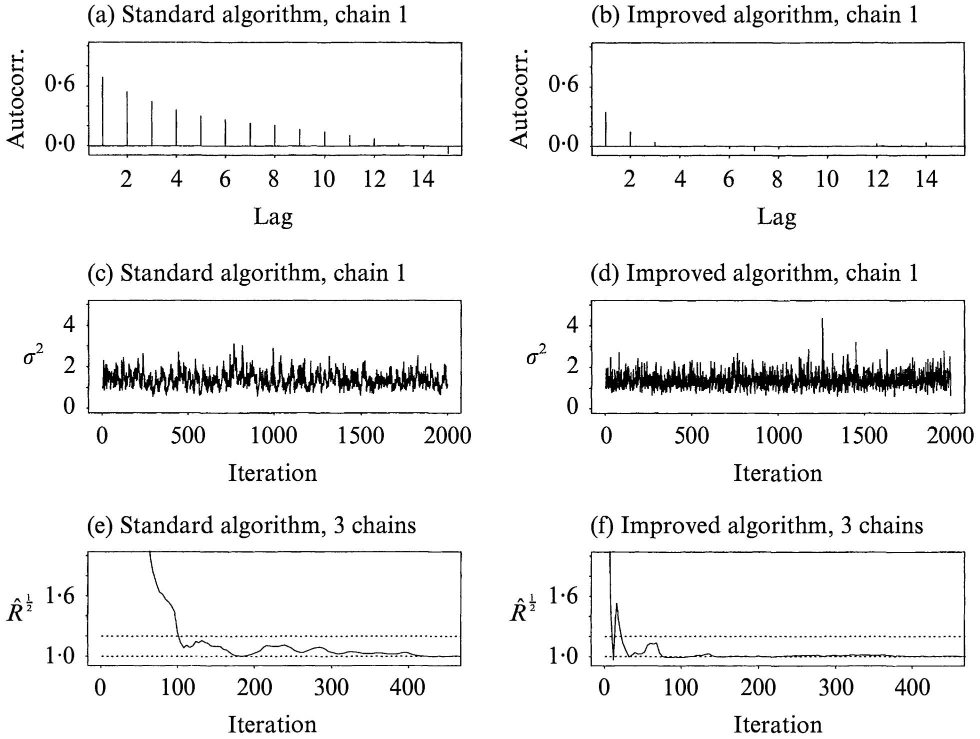

Fig. 3. Comparing the standard and improved algorithms for a four-dimensional Cauchy model. The plots are the counterparts of those in Fig. 2 using one of the diagonal elements of the scale matrix.

These examples illustrate the possibility of obtaining a fast-mixing positive recurrent chain by purposely constructing a larger nonpositive recurrent Markov chain. For the t example, not only is $\{\theta^{(t)}, t>0\}$ a positive recurrent Markov chain with the desired stationary distribution, but also $\{q^{(t)}/\alpha^{(t)}, t>0\}$ is a positive recurrent chain, providing a real example of the phenomenon discussed by George (1996). Recognising which function of the joint chain $\{(\theta^{(t)}, \alpha^{(t)}, q^{(t)}), t>0\}$ converges properly is useful for proper diagnosis of convergence. This recognition is typically easy because it is directly related to the construction of the working parameter; for the t example, q/α is simply the original 'missing data', $\tilde{q}$.

Not all improper working priors lead to Markov chains, and in general these chains may not converge in distribution to the target distribution. For the t problem, by comparing, or coupling, the stochastic mappings of Schemes [1] and [2], we are able to give a definite answer as to the choice of the hyperparameters for $p(\alpha \mid \beta, \gamma)$. As this coupling approach may be useful for other problems, we present key details in § 3·4.

3·4. *Investigating the choice of improper working priors: The t example*

We first consider the case with $\beta=0$ but $\gamma>0$, corresponding to the improper prior $p(\alpha) \propto \alpha^{-(1+\gamma/2)}$. In this case, the induced chain $\{\theta^{(t)}=(\mu^{(t)}, [\sigma^2]^{(t)}), t\geqslant 0\}$ under

Scheme [2], though Markovian, does not converge to the desired target density. This is because if $\beta = 0$ and $\gamma > 0$ and we condition on $\theta^{(t)}$, due to the extra χ^2_γ in the denominator of (3·12), the random variable given by (3·13) is stochastically strictly larger than the one given by (3·12), which corresponds to the correct target density. To see this more clearly, let $F_\gamma(\sigma^2 | Y_{\text{obs}}, \theta')$ be the conditional cumulative distribution function of $[\sigma^2]^{(t+1)}$ given by (3·12) conditioning on $\theta^{(t)} = \theta'$, and let $\tilde{F}_{\gamma,\beta}(\sigma^2 | Y_{\text{obs}}, \theta', \alpha')$ be the conditional cumulative distribution function of $[\tilde{\sigma}^2]^{(t+1)}$ given by (3·13) conditioning on $\theta^{(t)} = \theta'$ and $\alpha^{(t)} = \alpha'$. When $\beta = 0$, $\tilde{F}_{\gamma,\beta}(\sigma^2 | Y_{\text{obs}}, \theta', \alpha')$ can be rewritten as $\tilde{F}_\gamma(\sigma^2 | Y_{\text{obs}}, \theta')$. With this notation, the· stochastic ordering mentioned earlier is represented by

$$\tilde{F}_\gamma(\sigma^2 | Y_{\text{obs}}, \theta') < F_\gamma(\sigma^2 | Y_{\text{obs}}, \theta'), \tag{3·14}$$

for any $\sigma^2 \in R^+$, $\theta' \in R^1 \times R^+$ and $\gamma \in R^+$, where $R^+ = (0, +\infty)$ and $R^1 = (-\infty, +\infty)$. Integrating both sides of (3·14), with respect to the target density $p(\theta' | Y_{\text{obs}})$, yields

$$\tilde{F}_\gamma(\sigma^2 | Y_{\text{obs}}) \equiv \int \tilde{F}_\gamma(\sigma^2 | Y_{\text{obs}}, \theta') p(\theta' | Y_{\text{obs}})\, d\theta' < \int F_\gamma(\sigma^2 | Y_{\text{obs}}, \theta') p(\theta' | Y_{\text{obs}})\, d\theta' = F(\sigma^2 | Y_{\text{obs}}), \tag{3·15}$$

where $F(\sigma^2 | Y_{\text{obs}})$ is the marginal cumulative distribution function for σ^2 under our target density $p(\theta | Y_{\text{obs}})$ and the right-hand equality holds because $p(\theta | Y_{\text{obs}})$ is the invariant distribution of the Markov chain given by (3·10)–(3·12) when $\gamma > 0$. It follows that $p(\theta | Y_{\text{obs}})$ cannot be the invariant distribution for Scheme [2] because, if it were, $\tilde{F}_\gamma(\sigma^2 | Y_{\text{obs}})$ would be the same as $F(\sigma^2 | Y_{\text{obs}})$.

Similarly, when $\beta = 0$ but $-nv < \gamma < 0$, Scheme [2], while computable, also does not produce draws from the target distribution. This can be seen by comparing (3·12) with (3·13) and noting that $\chi^2_{\gamma+nv}$ is stochastically smaller than χ^2_{nv} when $-nv < \gamma < 0$, so we have $\tilde{F}_\gamma(\sigma^2 | Y_{\text{obs}}, \theta') > F_0(\sigma^2 | Y_{\text{obs}}, \theta')$ for any $\sigma^2 \in R^+$ and $\theta' \in R^1 \times R^+$. Here $F_0(\sigma^2 | Y_{\text{obs}}, \theta') = F_{\gamma=0}(\sigma^2 | Y_{\text{obs}}, \theta')$ and

$$\int F_0(\sigma^2 | Y_{\text{obs}}, \theta') p(\theta' | Y_{\text{obs}})\, d\theta' = F(\sigma^2 | Y_{\text{obs}})$$

because (3·12) can be formally extended to $\gamma = 0$, as discussed in § 3·3. Consequently, by analogy with (3·15), we obtain $\tilde{F}_\gamma(\sigma^2 | Y_{\text{obs}}) > F(\sigma^2 | Y_{\text{obs}})$ when $-nv < \gamma < 0$.

Thus, when $\beta = 0$, the only value of γ that will lead to correct sampling is $\gamma = 0$. That is, we must use $p(\alpha) \propto \alpha^{-1}$ as the improper prior within the class $p(\alpha) \propto \alpha^{-(1+\gamma/2)}$; in particular, the constant prior on α is excluded because it corresponds to $\gamma = -2$. However, there is another set of values of $\{\beta, \gamma\}$ for which Scheme [2] is computable, namely when $\beta > 0$ and $-nv < \gamma \leqslant 0$. For this set, the rigorous theory for the convergence behaviour of $\{\theta^{(t)}, t \geqslant 0\}$ is more complicated because it is no longer Markovian, and we need to use properties, e.g. Meyn & Tweedie (1993, p. 454), of the joint Markov chain $\{(\theta^{(t)}, \alpha^{(t)}, q^{(t)}), t \geqslant 0\}$, which is a null chain. Intuitively speaking, since $\alpha^{(t)}$ will eventually drift to infinity and thus $\beta/\alpha^{(t)} \to 0$ in probability, the limiting distribution of (3·13) is the same regardless of whether $\beta = 0$ or $\beta > 0$. In fact, when $-nv < \gamma \leqslant 0$, our simulated chains under $\beta = 0$ and $\beta > 0$ effectively coincide before the end of the 'burn-in' period because of excessively large values of $\alpha^{(t)}$; the largest value we observed was $6{\cdot}7e^{118}$.

This can be seen in Fig. 4 which displays quantile–quantile plots between the empirical quantiles of $\log \sigma^2$ obtained from the output of Scheme [2] under various choices of β and γ and the quantiles from the target density $p(\log \sigma^2 | Y_{\text{obs}})$. Here the empirical quantiles are based on 5000 draws, the first 100 being discarded, and the target quantiles are based on 100 000 draws, the first 6000 being discarded. The fact that the top and bottom plots are indistinguishable in the two right-hand columns is caused by the aforementioned 'coincidence' phenomenon, and not because $\beta=0{\cdot}0001$ is so close to $\beta=0$. These two values of β are chosen to highlight the singularity at $\beta=0$ in Fig. 4(a), (b), (e) and (f). Figures 4(e) and (f) show that, when $\beta=0$ but $\gamma>0$, Scheme [2] produces simulated $\log\sigma^2$'s that are stochastically too large; it appears there is a location shift, i.e. a multiplicative factor on the σ^2 scale, the magnitude of which depends on the magnitude of γ. However, once $\beta>0$, that is $p(\alpha\,|\,\beta,\gamma)$ is proper, Scheme [2] produces the correct invariant distribution, as can be seen in Fig. 4(a) and (b). In other words, for any $\gamma>0$ there is a singularity at $\beta=0$ in terms of the behaviour of the invariant distribution of $\{\theta^{(t)}, t\geqslant 0\}$ under Scheme [2], as is particularly visible in Fig. 4(a) and (e), when $\gamma=20$.

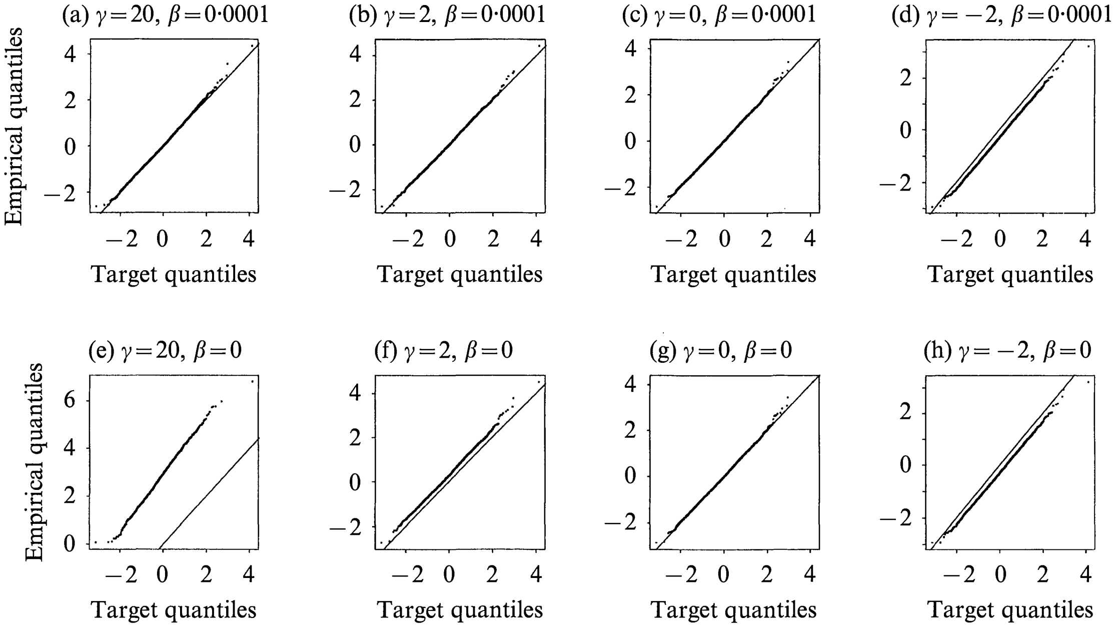

Fig. 4. The impact of the hyperparameters on Scheme [2]. The plots show quantile–quantile plots comparing the target distribution with draws of $\log\sigma^2$ using a variety of priors for α, $p(\alpha\,|\,\beta,\gamma)$.

When $\gamma=0$, Scheme [2] provides the correct limiting distribution regardless of whether $\beta>0$ or $\beta=0$, as can be seen from Fig. 4(c) and (g). When $-n\nu<\gamma<0$, Scheme [2] provides draws of $\log\sigma^2$'s that are stochastically too small regardless of the value of β, as indicated by Fig. 4(d) and (h), where we choose $\gamma=-2$ because this results in the constant working prior when $\beta=0$. Table 2 summarises our findings and indicates whether or not the choice of hyperparameter gives a limiting distribution that is the same as, or stochastically larger/smaller than the target distribution. Our conclusion is that the best hyperparameter value is $\gamma=\beta=0$, not only because this choice provides the fastest algorithm but also because it is the simplest to implement; e.g. it avoids potential numerical problems caused by excessively large values of $\alpha^{(t)}$ under the choice of $\beta>0$ and $\gamma=0$. Since $\alpha\sim\beta\chi_\gamma^{-2}$ is the conditional conjugate prior for $p(Y_{\text{aug}}\,|\,\theta,\alpha)$ and any other family would be likely to make implementation more complicated, this essentially establishes the best choice of an independent working prior for α, at least for practical purposes.

Table 2. *Impact of the choice of β and γ in $p(\alpha \mid \beta, \gamma)$ on the behaviour of $\{(\mu^{(t)}, [\sigma^2]^{(t)}), t \geqslant 0\}$ under Scheme [2]. First line in each cell indicates whether or not the chain is Markovian; second line indicates whether or not the choice has the correct limiting distribution, where 'too small' and 'too large' mean that the marginal limiting distribution of σ^2 is stochastically smaller or larger than the target marginal distribution of σ^2*

	$\gamma > 0$	$\gamma = 0$	$-n\nu < \gamma < 0$
$\beta > 0$	Non-Markovian Correct	Non-Markovian Correct	Non-Markovian Too small
$\beta = 0$	Markovian Too large	Markovian Correct	Markovian Too small

4. Discussion

4·1. *Connections between marginal augmentation and some other methods*

The marginal augmentation approach suggests that we can extend (2·2) to

$$\int_{\mathscr{M}(Y_{\text{aug}})=Y_{\text{obs}}} \left\{ \int p(Y_{\text{aug}} \mid \theta, \alpha) p(\alpha)\, d\alpha \right\} \mu(dY_{\text{aug}}) = p(Y_{\text{obs}} \mid \theta),$$

or, using the notation of (3·2),

$$\int_{\mathscr{M}(Y_{\text{aug}})=Y_{\text{obs}}} p(Y_{\text{aug}} \mid \theta) \mu(dY_{\text{aug}}) = p(Y_{\text{obs}} \mid \theta). \qquad (4·1)$$

The fact that (4·1) appears to be completely identical to the standard augmentation identity (2·1) is both correct and deceptive. It is correct because the marginal augmentation given by (3·2) is a legitimate augmentation in the sense of (2·1), and thus (2·1), as a general definition, is applicable. It is deceptive because, once we realise that a specific $f(Y_{\text{aug}} \mid \theta)$ in (2·1) is in fact a conditional density conditioning on a specific value of an invisible variable α, then (4·1) is a marginalisation of (2·1) via (3·2). In other words, once we identify a working parameter, it is better to write (2·1) as (2·2). This reflects an important difference between marginal augmentation and the usual form of the auxiliary variable method.

As described in Besag & Green (1993) and Green (1997), standard auxiliary variable methods first enlarge the parameter θ to (θ, ξ), without changing the marginal posterior $p(\theta \mid Y_{\text{obs}})$ for the given data Y_{obs}, and then implement a Markov chain Monte Carlo algorithm on (θ, ξ). In other words, ξ plays the same role as $\tilde{Y}_{\text{mis}}$ in our notation. The marginal augmentation approach operates on a different level, in that we assume we have already chosen $\tilde{Y}_{\text{mis}}$ but realise that $p(\tilde{Y}_{\text{mis}} \mid Y_{\text{obs}}, \theta)$ is really $p(Y_{\text{mis}} \mid Y_{\text{obs}}, \theta, \alpha = \alpha_0)$, and consequently that we can try to refine the augmentation scheme, or equivalently refine an auxiliary variable, via the marginalisation of α. A consequence of this difference is that for a standard auxiliary variable, ξ, (ξ, θ) are not independent given Y_{obs}, since otherwise the use of ξ offers no help. However, for a working parameter α, if (α, θ) are a priori independent they are also a posteriori independent because the data contain no information about α.

The marginal augmentation approach also has an interesting connection with simulated tempering (Marinari & Parisi, 1992; Geyer & Thompson, 1995), which runs a Markov chain Monte Carlo algorithm on a sequence of distributions $\{p_i(x),\ i = 1, \ldots, m\}$, but typically only one of them, $p_1(x)$, the 'cold' distribution, is of real interest. Here we can view i as a model parameter which is used as a working parameter, and each model, $p_i(x)$, is the conditional model $p(x \mid i)$ implied by the joint distribution on the (x, i) space, $p(x, i) = p_i(x)p(i)$. We say i is a model parameter because only samples corresponding to $i = 1$ are from the target density, but it is also a working parameter because $\{p_i, i \geqslant 2\}$ are introduced to improve mixing. In contrast, marginal augmentation allows continuous α and all samples are from the target density upon convergence. Note that for simulated tempering the p_i's can even be densities on spaces with different dimensions, e.g. Green (1995) and Richardson & Green (1997), and marginal augmentation can be used to improve mixing within each p_i.

Finally, the marginal augmentation approach also has an intrinsic connection with the collapsed Gibbs sampler method of Liu (1994b), which uses the fact that, when Y_{mis} can be decomposed into several components, one can 'collapse θ down' in $p(Y_{\text{mis}} \mid \theta, Y_{\text{obs}})$ by implementing a nested Gibbs sampler to sample from $p(Y_{\text{mis}} \mid Y_{\text{obs}})$. Similarly, the marginal augmentation method can be viewed as collapsing down the originally implicit α in $p(Y_{\text{mis}} \mid \alpha, \theta, Y_{\text{obs}})$ to produce $p(Y_{\text{mis}} \mid \theta, Y_{\text{obs}})$.

4·2. *Comparison between conditional augmentation and marginal augmentation*

Under the condition of Lemma 2, conditional augmentation is useless while the marginal augmentation can produce dramatic gains using the same working parameter. However, this says nothing about the comparison of the two approaches when they use different working parameters. For instance, in the t model, the data augmentation algorithm resulting from the optimal conditional augmentation, optimal over the value of the working parameter α introduced via $\sigma^{-2\alpha}\tilde{q}$, although more difficult to implement, produces the same empirical convergence rate as the data augmentation algorithm which uses the optimal marginal augmentation, optimal over the value of the hyperparameter in $p(\alpha \mid \beta, \gamma)$ with α introduced via $\alpha\tilde{q}$. This can be seen by comparing the underlined values in Table 1 with the values in the plots of Fig. 1 as $\gamma \to 0$.

The key to this equivalence lies in comparing two minimum maximum correlations:

(i) the maximum correlation between $Y_{\text{mis}} \equiv \mathscr{D}_{\alpha,\theta}(\tilde{Y}_{\text{mis}})$ and θ under the conditional model $p(Y_{\text{mis}}, \theta \mid Y_{\text{obs}}, \alpha)$ minimised over α, and

(ii) the maximum correlation between $Y_{\text{mis}} \equiv \mathscr{D}_{\alpha}(\tilde{Y}_{\text{mis}})$ and θ under the marginal model

$$p(Y_{\text{mis}}, \theta \mid Y_{\text{obs}}) = \int p(Y_{\text{mis}}, \theta \mid Y_{\text{obs}}, \alpha) p(\alpha)\, d\alpha$$

minimised over the choice of a class of $p(\alpha)$, for example $p(\alpha \mid \beta, \gamma)$ indexed by the hyperparameters β and γ.

The t example suggests that it is possible to find different $\mathscr{D}_{\alpha,\theta}$ and $\mathscr{D}_{\alpha}$ such that these two minimum maximum correlations are equal. It also motivates the view of marginal augmentation as conditional augmentation that conditions on some hyperparameters, and of conditional augmentation as marginal augmentation with point-mass prior on the working parameter. The t example suggests first using marginal augmentation via $\mathscr{D}_{\alpha}(\tilde{Y}_{\text{mis}})$ and then conditional augmentation to find the optimal value of the hyperparameter to determine the optimal prior for α. More applications of this strategy will be reported in a subsequent paper.

The t model also illustrates that, when we view both missing data Y_{mis} and θ as parameters, the conditional augmentation method is equivalent to, possibly unusual, reparameterisations, and the marginal augmentation is a form of overparameterisation. Thus, in general, if it is true that marginal augmentation can always achieve what conditional augmentation can, it carries a revolutionary message: overparameterise rather than reparameterise.

Acknowledgement

X.-L. Meng's research was supported in part by a National Science Foundation grant and in part by a National Security Agency grant, and D. A. van Dyk's research was supported in part by a National Science Foundation grant and in part by the U.S. Census Bureau. We thank A. Gelman, J. Liu and Y. Wu for insightful exchanges. The manuscript was prepared using computer facilities supported in part by several National Science Foundation Grants awarded to the Department of Statistics at the University of Chicago, and by The University of Chicago Block Fund.

References

Amit, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Mult. Anal.* **38**, 82–9.

Berger, J. (1996). Discussion of paper by G. Casella. *Test* **5**, 293–5.

Besag, J. & Green, P. J. (1993). Spatial statistics and Bayesian computation. *J. R. Statist. Soc.* B **55**, 25–37.

Casella, G. (1996). Statistical inference and Monte Carlo algorithms (with Discussion). *Test* **5**, 249–344.

Casella, G. & George, E. I. (1992). Explaining the Gibbs sampler. *Am. Statist.* **46**, 167–74.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood estimation from incomplete-data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B **39**, 1–38.

Gelfand, A. E. (1997). Gibbs sampling. In *Encyclopedia of Statistical Science*, Update Vol. 1, Ed. S. Kotz, C. Read and D. Banks, pp. 283–92. New York: Wiley.

Gelman, A. & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with Discussion). *Statist. Sci.* **7**, 457–511.

George, E. I. (1996). Discussion of paper by G. Casella. *Test* **5**, 303–5.

Geyer, C. J. & Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Assoc.* **90**, 909–20.

Gilks, W. R. (1997). Discussion of paper by X.-L. Meng and D. van Dyk. *J. R. Statist. Soc.* B **59**, 543–5.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (Ed.). (1995). *Markov Chain Monte Carlo in Practice.* London: Chapman and Hall.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–32.

Green, P. J. (1997). Discussion of paper by X.-L. Meng and D. van Dyk. *J. R. Statist. Soc.* B **59**, 554–5.

Liu, C., Rubin, D. B. & Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**, 755–70.

Liu, J. S. (1994a). Fraction of missing information and convergence rate of data augmentation. In *Computationally Intensive Statistical Methods: Proc. 26th Symp. Interface*, Ed. J. Sall and A. Lehmann, pp. 490–7. Fairfax Station, VA: Interface Foundation of North America.

Liu, J. S. (1994b). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Statist. Assoc.* **89**, 958–66.

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

Marinari, E. & Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **19**, 451–8.

McLachlan, G. J. & Krishnan, T. (1997). *The EM Algorithm and Extensions.* New York: John Wiley.

Meng, X.-L. (1997). The EM algorithm. In *Encyclopedia of Statistical Science*, Update Vol. 1, Ed. S. Kotz, C. Read and D. Banks, pp. 218–27. New York: Wiley.

Meng, X.-L. & van Dyk, D. (1997). The EM algorithm—an old folk song sung to a fast new tune (with Discussion). *J. R. Statist. Soc.* B **59**, 511–67.

Meng, X.-L. & van Dyk, D. (1998). Fast EM-type implementations for mixed-effects models. *J. R. Statist. Soc.* B **60**, 559–78.

Meyn, S. P. & Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* New York: Springer-Verlag.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with Discussion). *J. R. Statist. Soc.* B **59**, 731–92.

Roberts, G. O. & Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.* B **59**, 291–317.

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

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with Discussion). *J. Am. Statist. Assoc.* **82**, 528–50.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with Discussion). *Ann. Statist.* **22**, 1701–62.

van Dyk, D. A. (2000). Nesting EM algorithms for computational efficiency. *Statist. Sinica* **10**. To appear.

[*Received September* 1997. *Revised October* 1998]