Bayesian Computation and Stochastic Systems

Julian Besag, Peter Green, David Higdon and Kerrie Mengersen

Abstract. Markov chain Monte Carlo (MCMC) methods have been used extensively in statistical physics over the last forty years, in spatial statistics for the past twenty and in Bayesian image analysis over the last decade. In the last five years, MCMC has been introduced into significance testing, general Bayesian inference and maximum likelihood estimation. This paper presents basic methodology of MCMC, emphasizing the Bayesian paradigm, conditional probability and the intimate relationship with Markov random fields in spatial statistics. Hastings algorithms are discussed, including Gibbs, Metropolis and some other variations. Pairwise difference priors are described and are used subsequently in three Bayesian applications, in each of which there is a pronounced spatial or temporal aspect to the modeling. The examples involve logistic regression in the presence of unobserved covariates and ordinal factors; the analysis of agricultural field experiments, with adjustment for fertility gradients; and processing of low-resolution medical images obtained by a gamma camera. Additional methodological issues arise in each of these applications and in the Appendices. The paper lays particular emphasis on the calculation of posterior probabilities and concurs with others in its view that MCMC represents a fundamental breakthrough in applied Bayesian modeling.

Key words and phrases: agricultural field experiments, Bayesian inference, conditional distributions, deconvolution, gamma camera imaging, Gibbs sampler, Hastings algorithms, image analysis, logistic regression, Markov chain Monte Carlo, Markov random fields, Metropolis method, prostate cancer, simultaneous credible regions, spatial statistics, time reversibility, unobserved covariates, variety trials.

1 INTRODUCTION

Let $\{\pi(x) : x \in \mathcal{X}\}$, where $x = (x_1, \ldots, x_n)^T$, denote a specific multivariate distribution, sufficiently complex that important properties of π cannot easily be studied by standard analytical, numerical or simulation methods. In this introduction, we suppose for simplicity

Julian Besag is Professor in the Department of Statistics and Director of the Center for Spatial Statistics, University of Washington GN-22, Seattle, Washington 98195. Peter Green is Professor of Statistics and Head of the Department of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom. David Higdon is PhD candidate and Research Assistant in the Department of Statistics, University of Washington GN-22, Seattle, Washington 98195. Kerrie Mengersen is Lecturer in the School of Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia.

that π is discrete. Often π is known only up to scale, as is typical both in spatial statistics and in Bayesian inference. In the latter, π is generally a posterior distribution, $\pi(x) \equiv \pi(x|y)$, for parameters x, given fixed data y, but we will usually suppress y from our notation. We shall be interested in computing properties of π that can be represented in terms of expectations $\mathbf{E}_\pi g$ of functions g, under π; that is

$$\mathbf{E}_\pi g = \sum_{x \in \mathcal{X}} g(x)\pi(x).$$

In particular, such expectations include probabilities of specified events under π, when g is an indicator function. We emphasize this because one of the main features of the *Markov chain Monte Carlo* (MCMC) methods described in this paper is that they provide direct approximations to probabilities, rather than the more usual indirect ones found, for example, by fitting asymptotic distributions (e.g. Tierney and Kadane, 1986; Tierney, Kass and Kadane 1989). Which of these approaches is more fruitful will depend on context but it is perhaps the flexibility of MCMC that is its greatest virtue. This will be illustrated, for Bayesian inference, in the various applications considered later in the paper. As instances, MCMC can usually cope with non-standard priors and likelihoods, with missing data, with systems that can only be observed indirectly, and, in a comparative experiment, with the posterior probability that any particular treatment is best or group of treatments contains the best. Further, the accuracy of the approximations can be assessed.

MCMC can be summarized as follows. Let P denote the transition kernel of a Markov chain with state space $\mathcal{X}$, so that $P(x \to x')$ is the probability associated with a move from x to x'. Suppose that P is chosen to be ergodic with limit distribution π. Then if $x^{(1)}, x^{(2)}, \dots$ denotes a realization of the chain, the ergodic theorem implies that, for any seed $x^{(1)}$, the random sequence,

$$\bar{g}_m = \frac{1}{m} \sum_{t=1}^{m} g(x^{(t)}) \tag{1.1}$$

converges almost surely to $\mathbf{E}_\pi g$ as $m \to \infty$. (We often do not distinguish in our notation between random variables or vectors and their values but, when necessary, identify the former by upper case). Thus, the central idea is to construct such a P and then to approximate $\mathbf{E}_\pi g$ by the empirical average $\bar{g}_m$, obtained from a long, partial realization $x^{(1)}, \dots, x^{(m)}$ of the chain.

Ideally, $x^{(1)}$ should have distribution π, so that *general balance*,

$$\pi^T P = \pi^T,$$

ensures that the marginal distribution of each subsequent $x^{(t)}$ is also π but, of course, this is generally not feasible in situations where MCMC is under consideration. Instead, the chain is run from some fairly arbitrary state in $\mathcal{X}$ but an initial "burn-in" period is allowed before collecting samples, so that $x^{(1)}$ should then have distribution rather close to π, in an appropriate norm. Thus, there are five main issues in MCMC: the choice of P, the length of the burn-in period, the value of m, the possibility of estimators alternative to (1.1) and the estimation of the errors due to simulation. Methodological aspects of this paper are concerned primarily with the first of these. For the other four, we cite Diaconis and Stroock (1991), Fishman (1991, 1992a,b), Geyer (1992), Liu, Wong and Kong (1991), Marinari and Parisi (1992), Mykland, Tierney and Yu (1993), Roberts and Polson (1992), Rosenthal (1992), Diaconis and Saloff-Coste (1993), Frigessi, di Stefano, Hwang and Sheu (1993), Meyn and Tweedie (1993, Ch.16), Tierney (1994), Johnson (1994a), Mengersen and Tweedie (1994), Roberts and Tweedie (1994), and the references therein; however, this is an area of very rapid development. Here, we merely comment that MCMC is now well within the power of a typical workstation for a very wide range of hitherto intractable practical applications, including those where a simple formulation is perturbed in a non-standard fashion.

The paper is organized as follows. In Section 2, we first establish some notation and then discuss general methodology for the construction of Markov chains for Monte Carlo calculations, particularly in Bayesian inference. We emphasize the role played by conditional distributions, adopt a spatial point of view in understanding dependence, and focus on methods relevant to the range of applications that follow later in the paper. We also note the importance and intuitive appeal of the Gibbs sampler but suggest that the practical problems that arise in its use can often be countered by embracing the broader class of Hastings algorithms.

Section 3 describes a class of pairwise–difference prior distributions that have been found useful for factors that are spatially or temporally indexed, Particular examples occur in the applications that are presented in the next three sections of the paper. Thus, Section 4 considers an area of general statistical interest, logistic regression with ordinal factors and supposed additional unobserved covariates. This is described in the context of an observational study concerned with mortality from prostate cancer, classified by age group, period and cohort. In making predictions over subsequent time periods, we avoid the usual MCMC "missing data" approach and promote an alternative, more efficient procedure. Then, Section 5 discusses the analysis of crop trials, where Bayesian modeling must allow for the usually substantial variation in fertility across the experimental plots. The methodology is applied to data from a variety trial on spring barley. Some aspects of sensitivity analysis are considered and a wider "hierarchical–t" formulation is introduced. Lastly, Section 6 describes the modeling and analysis of a low–level computer vision problem arising in gamma camera imaging. Here, the true image has been degraded by both blur and noise. This section also discusses the construction of credible regions for two–dimensional functions.

In Section 7, we summarize the main points in the paper, concluding with others before us that MCMC already greatly expands the horizons of Bayesian computing and promises more for the future. The paper concludes with three appendices, on the use of proposal distributions generated by some stochastic mechanism, on new MCMC methods based on less than full conditionals, and on sensitivity analysis.

Spatial applications can be thought of as home ground for MCMC, particularly in statistical physics, where there is a vast literature of which a small sample of particular interest to statisticians could include Fosdick (1963), Swendsen and Wang (1987), Binder (1988), Sokal (1989), Gidas (1992) and Marinari and Parisi (1992). There is also a rapidly expanding list of spatial applications, adopting a Bayesian or neo-Bayesian viewpoint and using MCMC as a computational tool. Medical imaging provides the most common theme but, among a wide range of other topics, we note examples in agriculture, archeology, astronomy, biogeography, computer vision, food technology, geographical epidemiology, remote sensing and texture analysis; we do not itemize papers here but some have been added to the references. Useful collections of papers include those edited by Barone, Frigessi and Piccioni (1992), by Possolo (1991) and by Mardia and Kanji (1993). The pioneering ideas of Ulf Grenander have been especially influential, including his early advocacy of MCMC as a Bayesian "*inference machine*" with the maxim, "*Pattern analysis = Pattern synthesis*" (Grenander, 1983, Section 5.3). The incorporation of hyperparameters into the Gibbs sampler to produce a fully Bayesian analysis is set out in Besag (1989), following a suggestion by David Clayton.

Of course, spatial applications of MCMC are now greatly outnumbered by those in more conventional areas of Bayesian inference, stimulated especially by the landmark papers of Gelfand, Hills, Racine-Poon and Smith (1990) and Gelfand and Smith (1990). Additionally, there have been some recent non-Bayesian applications of MCMC in statistical inference. These also originate in spatial statistics and include maximum likelihood estimation in complicated model formulations, with the possible inclusion of constraints (Penttinen, 1984; Geyer and Thompson, 1992, including the Discussion; Geyer, 1991a, 1994); and MCMC calculation of exact p-values, particularly in preliminary significance testing (Besag and Clifford, 1989, 1991), which provides perhaps the only situation in which $x^{(1)}$ *can* be assumed to be from π! Non-Bayesian applications are not considered further in this paper.

2 CONSTRUCTION OF MARKOV CHAINS FOR MONTE CARLO CALCULATIONS

2.1 Notation

We write $\pi(x)$ for the density of the r.v. $X = (X_1, \ldots, X_n)^T$ with respect to a σ-finite product measure ν. We will freely use ν also to denote the corresponding measure on any subcollection of coordinate subspaces. The corresponding marginal density of X_i is written $\pi(x_i)$; in practice, X_i is usually but not always univariate. The minimal sample spaces for X and X_i are denoted by $\mathcal{X} = \{x : \pi(x) > 0\}$ and $\mathcal{X}_i = \{x_i : \pi(x_i) > 0\}$, respectively. For any subset S of $\mathcal{N} = \{1, 2, \ldots, n\}$, we write $x_S = \{x_i : i \in S\}$ and $x_{-S} = \{x_i : i \notin S\}$; in particular $x_{-i} = \{x_j : j \neq i\}$. Also we extend this notation in an obvious manner, so that, for example, $\pi(x_S)$ is the marginal density for X_S and $\pi(x_S|x_{-S})$ is the density of X_S, given $X_{-S} = x_{-S}$. For appropriate Borel sets $B_1, \ldots, B_n$ associated with $X_1, \ldots, X_n$, we write $B_S = \prod_{i \in S} B_i$ and $B = B_{\mathcal{N}}$. Note that, throughout this section, B will only be used to denote such *product sets*. We use I for the usual indicator function. In Bayesian applications, where $\pi(x)$ is typically a posterior density $\pi(x|y)$, the dependence on the data y will often be subsumed in the notation. Irrelevant sets of measure zero will be ignored throughout. For a more rigorous theoretical treatment of MCMC, especially with regard to continuous state spaces, see, for example, Tierney (1994).

2.2 The role of conditional distributions in multivariate simulation

MCMC is directed at simulation from multivariate distributions, usually of non-standard form, with components which are not independent. Evidently, simulation of a process that follows such a distribution, whether exactly or in the limit, must make use of the conditional distributions of some components given others. Of course, we can always write

$$\pi(x) = \prod_{i=1}^{n} \pi(x_i|x_{<i}), \tag{2.1}$$

where $x_{<i} = \{x_j, j < i\}$. The straightforward case is where, possibly after re-ordering the components $x_1, x_2, \ldots, x_n$, the densities $\pi(x_i|x_{<i})$ are *available* for simulation. In this case, static simulation from $\pi(x)$ is possible, with the components being generated sequentially using (2.1), and MCMC is unnecessary.

Much more commonly, in the sorts of complex stochastic model now being built in many application fields and demonstrated in later sections of the paper, there is *no* re-ordering of the variables for which the "one-sided" conditional distributions are all available. Either to construct them would require expensive numerical integration, or the resulting densities would be awkward to simulate from, or both.

A pair of simple examples may help to illustrate this point. Consider first a multivariate Gaussian distribution $\pi(x)$, with known mean and dispersion matrix. Then $\pi(x_i|x_{<i})$ has a known Gaussian form for each i and static simulation is straightforward; indeed, this is the stochastic interpretation of Cholesky decomposition. In contrast, consider the trivariate distribution (a special case of the auto-exponential model in Besag, 1974; see also Casella and George, 1992),

$$\pi(x) \propto \exp\{-(x_1 + x_2 + x_3 + \theta_{12}x_1x_2 + \theta_{13}x_1x_3 + \theta_{23}x_2x_3)\}, x_1, x_2, x_3 \geq 0,$$

where the θ's are positive constants. The conditional densities $\pi(x_i|x_{-i})$ are just univariate exponential; for example (cf. the general formula (2.2)),

$$\pi(x_3|x_{-3}) \propto \exp\{-x_3(1 + \theta_{13}x_1 + \theta_{23}x_2)\}, x_3 \geq 0.$$

However, the density $\pi(x_2|x_1)$, obtained by elementary integration, is already non-standard and the marginal density $\pi(x_1)$ involves the exponential integral (Abramowitz and Stegun, 1970, Ch.5), even up to scale, so that static simulation is clearly problematical. Looking ahead to Section 4 for a more interesting practical example, equation (4.2) gives the joint posterior/predictive density arising in a Bayesian formulation of logistic regression with factorial structure, extended to allow for unmeasured covariates. It is completely unrealistic in such a model to hope to be able to integrate out variables and calculate marginal posteriors, or to achieve a factorization like (2.1). The MCMC methods described and used in this paper make use only of conditional densities of the form $\pi(x_i|x_{-i})$ and are required only up to scale. They apply to both the simple examples and the real application above, with equal facility.

Areas which might seem more amenable to direct simulation using (2.1) are graphical modeling and pedigree analysis, where formulations often proceed along the lines of this factorization; such models are described by *directed acyclic graphs* (e.g. Whittaker, 1990). However, in the usual situation in which calculations are needed, some of the variables are observed, and so conditioned upon, and this typically destroys the simple sequential structure.

MCMC methods are most conveniently built upon conditional distributions of the form $\pi(x_T|x_{-T})$, for various subsets T of $\mathcal{N}$. Note that *all* variables are present in this expression, within the condition or otherwise. These distributions are called *local characteristics* in statistical physics and spatial statistics but, more recently, the term *full conditionals* has emerged from the Bayesian literature and so we shall adopt it here. The description of a stochastic system via its full conditionals provides an intuitive approach in modeling spatial interaction (Besag, 1974), where there is no natural ordering of the variables. It is therefore not surprising that MCMC methods developed earlier and further in spatial applications, before being more widely used in other areas. In non-spatial contexts, directionality or causality is often natural and models are built accordingly, so that the full conditionals need to be derived subsequently from the model assumptions. We see several examples of this in later sections. When these distributions cannot be conveniently obtained, all is not necessarily lost; see Appendix 2 for an approach based on *partial conditioning.* As we see there, however, some reference to full conditionals must still be retained. We find the spatial mode of thinking about complex systems of random variables helpful in any case. This is reflected in some of our choice of terminology; for example, we write of *visiting* sites, rather than of *looping* over subscripts.

In deriving full conditionals, we note the simple but powerful result that for any $x \in \mathcal{X}$ and $S \subset \mathcal{N}$,

$$\pi(x_S|x_{-S}) \propto \pi(x), \tag{2.2}$$

where only the terms involving components of x_S in any product formula for π need be retained. In particular,

$$\pi(x_i|x_{-i}) \propto \pi(x). \tag{2.3}$$

Equivalently, if x, $x' \in \mathcal{X}$, with $x'_{-S} = x_{-S}$, then

$$\pi(x'_S|x'_{-S})/\pi(x_S|x_{-S}) = \pi(x')/\pi(x). \tag{2.4}$$

These observations are important for two main reasons. First, product formulae are very common. They arise in posterior distributions, in combining the likelihood with priors and hyperpriors and usually within the likelihood itself; in spatial statistics, whenever π is a locally dependent Markov random field; and, in graphical models, as a result of conditional independence (see, for example, Whittaker, 1990, Cox and Wermuth, 1993, and the references therein). Thus, there is often considerable simplification in (2.2–2.4). Second, π needs to be known only up to scale, which is typically the case both in Bayesian and in spatial formulations. As a welcome bonus, MCMC often requires the conditionals themselves only up to scale, so that (2.2) and (2.3) do not need to be normalized.

In a general formulation, we denote the observed data by y and write $x = (\theta, \phi, z)$, where θ, ϕ and z are vectors of parameters, hyperparameters and missing observations, respectively. Then in an obvious notation,

$$\pi(x|y) \propto L(y, z|\theta)\pi(\theta|\phi)\pi(\phi), \tag{2.5}$$

representing a hybrid between the posterior density of θ and ϕ and the predictive density of z. If the missing data factor out, they can be ignored but, otherwise, the standard MCMC procedure is to include z as additional components to be updated, resulting in a (dependent) set of samples from the joint distribution of (θ, ϕ, z) given y. Any marginalization is carried out at the final stage, simply by ignoring the uninteresting components of x. As a bonus, samples from the predictive distribution of z are available. Note that, even when analytical marginalization over z in (2.5) can be carried out, this may be computationally inadvisable because of complications it creates in the new conditional distributions. The full conditionals corresponding to (2.5) are

$$\begin{aligned}
\pi(\theta_i|\theta_{-i}, \phi, z, y) &\propto L(y, z|\theta)\pi(\theta_i|\theta_{-i}, \phi), \\
\pi(\phi_j|\theta, \phi_{-j}, z, y) &\propto \pi(\theta|\phi)\pi(\phi_j|\phi_{-j}), \\
\pi(z_k|\theta, \phi, z_{-k}, y) &\propto L(y, z|\theta).
\end{aligned}$$

In practice, there is often considerable further simplification in these formulae.

For the most part, we will only use MCMC methods that visit single sites or subsets of sites T, in turn, possibly changing the values of the corresponding variables x_T, under the control of the full conditionals $\pi(x_T|x_{-T})$. Then the minimal requirements will be that $\pi(x)$ is always maintained and that the visitation schedule ensures irreducibility and aperiodicity of the algorithm as a whole. There are two issues:

- how to update the variables x_T, and
- how to determine which set T of sites to visit.

We address these two matters separately in the next two subsections; see Appendix 2 for partial conditioning.

2.3 Updating using full conditionals

2.3.1 The Gibbs Sampler

MCMC is driven by conditional probability. In particular, *Gibbs samplers* (Geman and Geman, 1984; and, by other names, Creutz, 1979; Ripley, 1979; Grenander, 1983) depend on the following intuitively obvious result.

Suppose X has density π and that T is a fixed subset of $\mathcal{N}$. Given $X = x$, define the r.v. $X' = (X'_1, \ldots, X'_n)$ such that $X'_{-T} = x_{-T}$ and X'_T has density $\pi(x'_T|x_{-T})$. Then X' has marginal density π. This is proved by observing that

$$\begin{aligned} Pr(X' \in B) &= \int_B \pi(x_{-T})\pi(x_T|x_{-T})\,d\nu(x) \\ &= \int_B \pi(x)\,d\nu(x) = \pi(B) \end{aligned}$$

The above procedure defines a Markov chain transition kernel

$$P_T(x \to B) = I[x_{-T} \in B_{-T}] \int_{B_T} \pi(x_T|x_{-T})\,d\nu(x_T). \tag{2.6}$$

What the observation above implies is that π is a stationary distribution for P_T.

The simplest special case is the *single site Gibbs sampler*, for which

$$P_i(x \to B) = I[x_{-i} \in B_{-i}] \int_{B_i} \pi(x_i|x_{-i})\,d\nu(x_i).$$

This involves only univariate sampling, which is a major attraction. Note that each time x_i is updated, the other variables may have different values, so random variate generation methods with high set-up costs are inappropriate. In particular, when direct methods of simulation are unavailable, standard rejection sampling is likely to be very inefficient as an alternative. For continuous univariate densities, the ratio method (e.g. Ripley, 1987, Ch. 3) is sometimes useful and the ARS (Adaptive Rejection Sampling) algorithms of Gilks and Wild (1992) and Gilks (1992) are available for densities that are log concave; see also Appendix 1.

2.3.2 Time reversibility

A stationary stochastic process is *time reversible* if its finite dimensional distributions are invariant under time reversal. For a Markov chain, a necessary and sufficient condition for reversibility is that, for every pair of successive states X and X',

$$Pr(X \in B, X' \in C) = Pr(X \in C, X' \in B), \tag{2.7}$$

for all (product sets) B and C (and hence all measurable $B, C \subset \mathcal{X}$). If π denotes the corresponding (invariant) marginal density of X and $P(x \to B)$ is the transition kernel of the chain then (2.7) becomes

$$\int_B \pi(x)P(x \to C)\,d\nu(x) = \int_C \pi(x)P(x \to B)\,d\nu(x). \tag{2.8}$$

Equation (2.8) is referred to as *local* or *detailed balance* and implies general balance, putting $B = \mathcal{X}$. The MCMC algorithms in this paper do not all have time reversible kernels but they all involve the concept of reversibility, either explicitly or implicitly. For instance, it is easily shown that the Gibbs kernel (2.6) is reversible. Such kernels are also more amenable to theoretical investigations; for example, the central limit theorems of Kipnis and Varadhan (1986) and the Monte Carlo error variance estimates of Geyer (1992) apply. Thus, there are advantages in modifying an irreversible MCMC algorithm, so that it becomes reversible, at least if this can be done at negligible cost.

2.3.3 Hastings Algorithms

Hastings algorithms (Hastings, 1970) provide a general class of alternatives to P_T defined in (2.6). The construction, given the current state x and a set of sites T, is as follows. First, a potential new value or *proposal* x' is generated from a transition kernel with a density $R_T(x_T \to x'_T; x_{-T})$ and which has the properties, (i) $x'_{-T} = x_{-T}$, (ii) $R_T(x_T \to x'_T; x_{-T}) > 0 \Leftrightarrow R_T(x'_T \to x_T; x_{-T}) > 0$; the definition of R_T is otherwise arbitrary, provided that the chain that is ultimately constructed is aperiodic and irreducible. Then x' is accepted as the new state with probability

$$A_T(x_T \to x'_T; x_{-T}) = \min\left\{1, \frac{\pi(x')R_T(x'_T \to x_T; x_{-T})}{\pi(x)R_T(x_T \to x'_T; x_{-T})}\right\} \tag{2.9}$$

or else x is retained as the next state; because of (2.4), massive cancellations may occur in (2.9). Thus, writing

$$Q_T(x_T \to x'_T; x_{-T}) = R_T(x_T \to x'_T; x_{-T})A_T(x_T \to x'_T; x_{-T}),$$

we obtain

$$\begin{aligned} P_T(x \to B) &= I[x_{-T} \in B_{-T}]\left[\int_{B_T} Q_T(x_T \to x'_T; x_{-T})\, d\nu(x'_T)\right] \\ &\quad + I[x \in B]\left[1 - \int_{\mathcal{X}_T} Q_T(x_T \to x'_T; x_{-T})\, d\nu(x'_T)\right]. \end{aligned}$$

The time reversibility condition (2.7) is met because

$$\pi(x)Q_T(x_T \to x'_T; x_{-T}) = \pi(x')Q_T(x'_T \to x_T; x_{-T})$$

for all $x, x' \in \mathcal{X}$, by (2.9).

There is great flexibility in the choice of the proposal distribution R_T for a given set T. However, it is clearly desirable that proposals can be generated very quickly and that acceptance probabilities are easy to calculate and are typically quite large, though this must not be at the expense of mobility. Usually such conditions suggest that the X_T's should be low-dimensional and most often they are chosen to be univariate, though the only requirement is that the proposal distribution should be easy to sample. Indeed, we use vector proposals in both Sections 4 and 5 of the paper.

Gibbs samplers correspond to the choice

$$R_T(x_T \to x'_T; x_{-T}) = \pi(x'_T | x_{-T}),$$

independent of x_T. Such proposals are always accepted, by (2.9), but this property is not exclusive to Gibbs; see Barone and Frigessi (1989) for other (Gaussian) examples. In practice, Gibbs samplers require univariate updates unless $\pi(x_T|x_{-T})$ is Gaussian or possibly if x_T is binary. This can be a major disadvantage.

For *Metropolis algorithms* (Metropolis, Rosenbluth, Rosenbluth, Teller and Teller, 1953; Hammersley and Handscomb, 1964, Ch. 9), $R_T(x_T \to x'_T; x_{-T})$ is chosen to be symmetric in x_T and x'_T, so that

$$A_T(x_T \to x'_T; x_{-T}) = \min\{1, \pi(x'_T|x_{-T})/\pi(x_T|x_{-T})\}.$$

Note that higher-density proposals are always accepted and that the only new function evaluation at each successive stage is that of the non-cancelling terms in the numerator of the odds ratio, at the single point x'.

With single-site updating, $P_T(x \to B)$ becomes $P_i(x \to B)$, leaving x_{-i} unchanged. For unrestricted real variables, a simple Metropolis proposal, obtained from a distribution that is symmetric about the current x_i and that has a spread similar to that of the marginal posterior for that variable, is usually effective. Rectangular and Gaussian distributions provide the obvious choices here. When x_i is restricted to an interval, a similar strategy can be applied to a suitably transformed variable: see, for example, Section 6. The validity of some of these points assumes perfect random variate generation and perfect floating-point arithmetic. In reality, these are not available, and it is advisable to adopt algorithms that are not too sensitive to this; thus, routine use of proposal distributions with bounded support may be safer than using the heavy-tailed distributions that are sometimes advocated. Fixing on appropriate spreads usually requires a few short pilot runs and can be automated or carried out in an *ad hoc* manner. At this stage, we have no definite recommendations but experience suggests that an acceptance rate between about 30% and 70% for each variable often produces satisfactory results.

2.3.4 Choice between samplers

As yet, there seem to be few hard and fast rules determining which Hastings algorithm is best in any particular situation. Computational, as well as statistical, efficiency should be taken into account and it should be noted that different considerations apply before and after burn-in. However, often there will be no need for a near-optimal choice. When $\pi(x_T|x_{-T})$ can be easily sampled for each T, Gibbs has a natural appeal and usually performs adequately. Hastings is almost always easy to program, easier than Gibbs if $\pi(x_T|x_{-T})$ does not admit a simple simulation method. If high efficiency is crucial, Gibbs may not be ideal, partly because it may run very much slower per cycle than Metropolis (say) and partly because it may be less efficient even cycle for cycle. In statistical physics, where interest typically centers on very large (the larger the better) systems of often binary parameters (site values), single-component MCMC was used for more than 30 years, until the advent of auxiliary-variable methods. Yet, despite the triviality of sampling from the componentwise full conditionals, the general choice was Metropolis, rather than Gibbs (known there as the heat bath method and, for binary variables, Barker's algorithm). A good reason for this in binary systems is that Metropolis allows one always to propose the opposite of the current value at each site and thereby increase the probability of a change, with the general aim of achieving greater mobility around the state space and eventually more efficient estimation for a given equilibrium run length; see Peskun (1973), for example. There is also an advantage in the rate of convergence to equilibrium, provided the interactions between the parameters are strong, which is the case of interest; see Frigessi et al (1993) for detailed discussion.

In continuous parameter spaces, the situation is more subtle but non-Gibbs algorithms still have the advantage that their proposal distributions can take account of x_T, as well as x_{-T}. Examples include the antithetic variable methods in Barone and Frigessi (1989), Green and Han (1992) and Besag and Green (1993), which can be thought of as continuous analogues of the binary Metropolis scheme mentioned above. It may also be prudent to employ a different algorithm before and after burn-in, where the possibly-conflicting goals are, respectively, fast convergence to π and efficient estimation. There is much uncharted territory here but the notion that Hastings is merely to be used when Gibbs is difficult to implement is surely false. Note that Metropolis generally does not seek even to approximate the full conditionals in its proposal distributions.

However, that is not to deny that one of the most common applications of Hastings in Bayesian MCMC is that of "correcting" any crude version of a Gibbs sampler. For example, as suggested in Tierney (1994), a Hastings step can be combined with a discrete histogram approximation to $\{\pi(x_T|x_{-T})\}$, so as to maintain π exactly. In Appendix 1, we extend this idea to adaptive rejection sampling. Here, we provide a simple recipe for simulating from a wide range of continuous multivariate distributions using *vector* proposals.

Suppose we write $\pi(x) \propto \exp\{-u(x)\}, x \in \mathbf{R}^n$, and assume the existence of $\nabla u(x)$, the vector of partial derivatives of u. Then, in general, the stochastic differential equation,

$$dx(t) = -\nabla u(x(t))dt + \sqrt{2}dw(t),$$

where $w(t)$ is standard n-dimensional Brownian motion. defines a continuous time Langevin diffusion which has stationary distribution π; see, for example, Gidas (1992) for detailed discussion. This has led to the use of discrete time Markov chain approximations (e.g. Amit et al, 1991) in which the current state x is replaced at the next stage by

$$x' \sim \mathbf{N}(x - \tau\nabla u(x), 2\tau I_n),$$

where τ is a small positive constant. However, if x' is used merely as a Hastings proposal for the next state, then the acceptance probability obtained from (2.9) ensures that π is maintained *exactly* by the modified Markov chain. For a fixed run length, τ should be chosen not so small as to mimic the continuous time process, but rather to produce appreciable proposal increments, accepted moderately often. Such a Langevin-Hastings step could be employed on subsets rather than all of the variables and τ might be given a distribution of its own instead of remaining fixed. Also, there is no necessity for the proposals to be Gaussian.

2.4 Visiting schedules

We have seen in Section 2.3 that the Hastings family provides a convenient means of constructing Markov transition kernels P_T with a prescribed stationary distribution π, where P_T has positive probability of changing each component of x_T but which leaves x_{-T} unaltered. At least if the set T is small, a transition according to such a kernel may be readily simulated. However, such a kernel cannot be irreducible, unless $T = \mathcal{N}$. We thus need to combine the kernels for a family of sets T, not necessarily disjoint, that cover $\mathcal{N}$ in such a way that the result is irreducible, and also aperiodic, without losing the property that π is stationary. There is no need at all to use the same prescription to construct each P_T; thus one can "mix-and-match" by combining Gibbs and other Hastings samplers as convenient and efficient for each set T.

2.4.1 Random scans

Random scan algorithms are constructed by first choosing a subset T of $\mathcal{N}$ according to a probability distribution $\{p_T\}$ and then generating a new value of X_T from the corresponding kernel $P_T(x \to B)$. That is,

$$P(x \to B) = \sum_T p_T P_T(x \to B) \tag{2.10}$$

for all B. Reversibility is preserved:

$$\begin{aligned} \int_B \pi(x) P(x \to C)\, d\nu(x) \quad &= \quad \sum_T p_T \int_B \pi(x) P_T(x \to C)\, d\nu(x) \\ = \sum_T p_T \int_C \pi(x) P_T(x \to B)\, d\nu(x) \quad &= \quad \int_C \pi(x) P(x \to B)\, d\nu(x). \end{aligned}$$

As usually implemented, the sets T with $p_T \neq 0$ are taken to be disjoint, but it is evident that all that is needed is that their union is $\mathcal{N}$.

2.4.2 Semi-regular scans

In general, alternatives to random scan impose some regularity on the order in which sites are visited. Thus, Amit (1991) investigates a version of the Gibbs sampler in which successive visits to the same site are prohibited. The resulting chain is Markov on the augmented state space $\mathcal{X} \times \mathcal{N}$ and, marginalizing over the update site I, must maintain π. The chain is irreversible.

2.4.3 Systematic scans

The most common variation of (2.10) is to visit each site in turn over a single *cycle*. Cycle by cycle, this produces a Markov chain which, with appropriate indexing of sites, has transition kernel

$$P = P_1 P_2 \dots P_n. \tag{2.11}$$

Since each P_i maintains π, so does P. Equally, given subsets $T_1, T_2, \dots, T_k$ of $\mathcal{N}$, not necessarily disjoint, with $\cup T_j = \mathcal{N}$, we could visit $T_1, T_2, \dots, T_k$ in turn, so that over a complete cycle we have

$$P = P_{T_1} P_{T_2} \dots P_{T_k}.$$

Reversibility of P_i or P_{T_j} does not imply that of P, but there are several means by which reversibility can be restored. For example, instead of the fixed scan in (2.11), a random order can be used, but this is cumbersome when n is large. Alternatives include the forwards–backwards scan, for which

$$P = P_1 P_2 \dots P_{n-1} P_n^2 P_{n-1} \dots P_2 P_1,$$

or making a random choice between the two scans on each cycle. One demerit of a systematic scan is the potential for significant artificial "drift" among the variables, which may in some situations hinder the mixing of the chain, and produce visible directional effects in spatial problems when the order of visiting the sites follows their spatial arrangement.

2.4.4 Coding sets

In many spatial applications, the components of X can be partitioned into a few "coding sets" (Besag, 1974), within each of which, conditionally, the X_i's are mutually independent. For such a set S,

$$\pi(x_S | x_{-S}) = \prod_{i \in S} \pi(x_i | x_{-i})$$

and sequential and simultaneous updating of the corresponding X_i's are indistinguishable. This is used in Sections 4 and 5. It also opens up the possibility of partially parallel processing and the availability of almost instantaneous results, even in very large problems. Some parallel implementations already exist in Bayesian image analysis; see, for example, Grenander and Miller (1994). The use of coding sets can be made time reversible by choosing sets at random or at random within a cycle.

2.4.5 Grouping variables

Of course, single site Gibbs sampling is at the opposite extreme from what one would ideally like to do, namely sample from π itself. Sometimes it will be worthwhile to make a compromise and simultaneously update small groups of conditionally dependent components, chosen randomly or deterministically, with the aims of improving the speed of convergence to π and the efficiency of estimation.

From the point of view of statistical efficiency, some guidance on the merits of such grouping can be gained by considering the very special case in which $\pi(x)$ is a multivariate Gaussian density. All updating schemes that (a) partition the variables into disjoint groups, (b) visit the groups in a deterministic order, and (c) use a Gibbs kernel on each, can be expressed as first-order vector autoregressions, over a complete cycle. This representation permits an explicit evaluation of the asymptotic variance matrix of the vector of empirical averages of components of x. A straightforward generalization of the Theorem in Green and Han (1992) shows that this variance is $(1/m)V\,\text{blockdiag}(V^{-1})V$. Here, V is the variance of $\pi(x)$, and $(\text{blockdiag}(V^{-1}))_{ij} = (V^{-1})_{ij}$ if variables i and j are in the same group, and is otherwise 0. This result may be used to assess the advantages of grouping, in estimating linear combinations of the components of $\mathbf{E}_\pi X$. For example, if all coefficients in the linear combination are non-negative, and the variables are strongly positively associated (in the sense that all off-diagonal elements of V are non-negative and all those of V^{-1} non-positive), then the more the variables are grouped the better. Another interesting consequence of the result is that the asymptotic variance is independent of the order in which the groups are visited. These observations are suggestive of good strategies in more general models. Adaptations of such updating schemes that replace Gibbs by Hastings steps with an antithetic property – there is a close analogy with over-relaxation in matrix iterative methods – can be analyzed in a similar way. Of course, the benefits of grouping must be weighed against the extra effort and computer time per update involved in sampling from more complicated conditional distributions; see Section 5, for example.

Sometimes, especially when the X_i's are constrained, there is no alternative to implicit or explicit grouping. For example, suppose the X_i's are binary, indicating absence or presence of a particle at each site i of a finite regular lattice. If the total number of particles is known, then at least two X_i's at a time must be considered for any change to take place.

In simultaneous updating using Hastings algorithms, proposal distributions will need scaling down, so as to maintain appreciable blockwise acceptance probabilities. Block updating is sometimes a worthwhile proposition, if only to reduce programming effort. On the other hand, it is rarely practicable for Gibbs samplers, unless the relevant components are conditionally independent, or are Gaussian, or the state space is very small and discrete.

3 PAIRWISE DIFFERENCE PRIORS

In each of the applications in the subsequent sections, we have need for prior distributions for variables ordered in time or space. In this section, we describe classes of *pairwise interaction* MRF's that can fulfill this role. For a random vector $\psi = (\psi_1, \ldots, \psi_n)$, these MRF's take the form

$$\pi(\psi|\gamma) \propto \exp\{-\sum_{i\sim j} w_{ij}\Phi(\gamma(\psi_i - \psi_j))\}, \tag{3.1}$$

where γ is a scale parameter, $\Phi(u) = \Phi(-u)$, the w_{ij}'s are specified non-zero weights and the summation is over neighboring sites $i \sim j$. A wide range of different Φ's, for both discrete and continuous distributions, can be found in the spatial literature; for some basic choices, see Geman and McClure (1985, 1987), Besag (1986, 1989), Künsch (1987, 1994), Green (1990), Geman and Reynolds (1992) and Geman, McClure and Geman (1992). In almost any practical context, π is improper, unless the minimal state space is bounded; however, we demand that the full conditionals,

$$\pi(\psi_i|\psi_{-i}) \propto \gamma \exp\{-\sum_{j\in\partial i} w_{ij}\Phi(\gamma(\psi_i - \psi_j))\},$$

are well defined, so that π is informative about some or all contrasts among the ψ_i's. Note that, if Φ is convex and the weights w_{ij} are positive, then π is log-concave. Often the weights are set to unity, though one might specify them in accordance with an appropriately defined distance between neighbors. In detailed modeling, there may also be negative weights, as we discuss later in the section.

First we consider some of the basic choices for Φ. The two simplest, $\Phi(u) = u^2$ and $\Phi(u) = |u|$ for $u \in \mathbf{R}$, with positive weights, are contrasted in Besag (1989). They are used in two examples in Besag, York and Mollié (1991) to estimate the risk from a rare disease over a set of contiguous geographical regions, via the Gibbs sampler. The sole impropriety in such priors is that of an arbitrary level and is removed from the corresponding posterior distribution by the presence of any informative data. If $\Phi(u) = u^2$, then

$$\psi_i|\psi_{-i} \sim \mathbf{N}(\sum_{j\in\partial i} w_{ij}\psi_j/w_{i+}, 1/\gamma^2 w_{i+}) \tag{3.2}$$

(the + denoting summation over the missing subscript), whereas, if $\Phi(u) = |u|$, then

$$\pi(\psi_i|\psi_{-i}) \propto \gamma \exp\{-\gamma \sum_{j\in\partial i} w_{ij}|\psi_i - \psi_j|\}. \tag{3.3}$$

If the weights are equal, (3.3) has its mode at the median, rather than at the mean, of the neighboring ψ_j's, where here we define the median of an even number of observations as lying anywhere between the two central values in the corresponding ordered sample. This suggests that the L1-based prior is more appropriate when the truth is believed to embody discrete jumps. One interpretation is as a stochastic version of the median filter, which is often used in remote sensing packages.

Green (1990) proposes an alternative, in the context of medical imaging, for which

$$\Phi(u) = \delta(1+\delta)\log\cosh(u/\delta), \tag{3.4}$$

where $\delta \in (0, \infty)$ is a tuning parameter. If the weights are positive, π is log-concave and differentiable everywhere. This generates (3.2) as $\delta \to \infty$ and (3.3) as $\delta \to 0$, and in general treats discontinuities suggested by the data in a manner intermediate between these extremes.

The other references cited above include examples in which Φ is chosen deliberately to be non-convex, with the intention that there should be even less resistance to the formation of discontinuities in scenes from the posterior. Thus, Geman and McClure (1985, 1987) and Geman and Reynolds (1992) adopt

$$\Phi(u) = -1/(1 + |u|^{\delta}),$$

where $\delta = 2$ and 1, respectively. In the latter case, Φ is concave on $(0, \infty)$ and supports images that consist of flats and jumps, rather than exhibit smooth variation.

Returning to a general Φ, the simplest useful neighborhood system is the linear one for which $i = 1, \ldots, R$, say, and i and j are neighbors if and only if $|i - j| = 1$. If the w_{ij}'s are unity, this defines a random walk with independent increments $\psi_i - \psi_{i+1}$ having distribution determined by Φ. We use versions of such prior distributions in Section 5 to model one-dimensional fertility variation in agricultural experiments; see also Besag and Higdon (1993). Another variation is to replace the independent first differences by independent second differences, $\psi_{i-1} - 2\psi_i + \psi_{i+1}$, again with distribution determined by Φ. Generally, this does not define a pairwise difference prior, though it does in the Gaussian case $\Phi(u) = u^2$, for which

$$\mathbf{E}(\psi_i|\psi_{-i}) = \frac{2}{3}(\psi_{i-1} + \psi_{i+1}) - \frac{1}{6}(\psi_{i-2} + \psi_{i+2}),$$

for $i = 3, \ldots, R-2$, with appropriate modifications otherwise. This is the conditional expectation that corresponds to a locally quadratic rather than a locally linear stochastic interpolant.

The above types of prior are also relevant in some applications that are not obviously spatial, such as one-way and higher-dimensional tables in which there are ordered categories. For example, in Section 4, we use a logistic regression formulation to analyze data on deaths from prostate cancer, with a linear predictor that includes three known factors, age-group, period and cohort. For the corresponding vectors of parameters, we adopt independent random walk priors that link together successive age-groups, successive periods and successive cohorts, without making strong structural assumptions. An easily-handled modification would be to constrain these random walks to have mean zero, or to have positive increments if this was believed to be appropriate. Another, which we illustrate in Section 4 would be to replace the independent first differences by second differences, as above; see also Berzuini, Clayton and Bernardinelli (1993).

When (3.1) refers to a finite two-dimensional array, with sites identified by integer-pairs of cartesian coordinates $i = (r, c)$, there are many different options available. The simplest non-degenerate choice is (3.2), with i and j deemed to be neighbors if they are unit distance apart. Yet, even here, one needs to take care in specifying the weights over the array, since a naive choice can produce serious edge effects and so provide a poor approximation to the finite restriction of the corresponding spatially invariant infinite lattice process; see Besag and Kooperberg (1993) for a method of adjustment and Besag and Higdon (1993) for an agricultural example in which lateral asymmetry is an additional ingredient. Fortunately, edge effects are irrelevant in many imaging applications, where the focus of interest is well removed from the boundary, and, in that case, there is also no reason to make a Gaussian assumption. Thus, in Section 6, we adopt the corresponding four-neighbor version of the log cosh prior (3.4) in the context of gamma-camera imaging.

The above 2-d and corresponding higher-dimensional formulations may again be useful in multi-way tables. Indeed, the prostate cancer data suggest the existence of further unmeasured covariates, which we accommodate in Section 4 by including an additional term in the linear predictor for each cell. Our prior for these terms is a Gaussian density with independent components having common variance, but an alternative that might be useful elsewhere would allow for dependence between terms.

There is an interesting degenerate form of Gaussian prior on 2-d arrays. Suppose that in (3.2), directly and diagonally adjacent sites are neighbors, with $w_{ij} = 2$ and -1, respectively. This distribution annihilates, not only an overall level, but also arbitrary row and column effects, because its "density" depends only on contrasts of the form $\psi_{r,c} - \psi_{r+1,c} - \psi_{r,c+1} + \psi_{r+1,c+1}$, despite fitting into the pairwise-interaction framework. Furthermore, there are no edge effects with respect to the corresponding infinite lattice process. It can easily be shown that the second-order properties are equivalent to those of $\xi_{r,c} = \theta_r \phi_c$, where θ and ϕ are the unrestricted random walks described above but with equal scale parameters; see Besag and Kooperberg (1993) for further details. There is a trivial extension to unequal scales. Distributions for which such one-dimensional decompositions exist are called *separable* and have received considerable attention in the literature on non-Bayesian 2-d fertility adjustment (Martin, 1990; Cullis and Gleeson, 1991; Kempton, Seraphin and Sword 1993), in which it may be reasonable to expect fertility effects to run along rows and columns because of management practice.

We conclude this discussion by considering an example of detailed modeling on a 2-d grid. We have in mind imaging applications where edge effects can be ignored. First, recall the four-neighbor equally-weighted and the eight-neighbor separable priors described above. It is easily seen that the conditional mean in (3.2) is the corresponding predictor of $\psi_{r,c}$, obtained from the least squares fit of a plane to the nearest four site values or a quadratic surface to the nearest eight. The degeneracy in the second case is partly the result of fitting six parameters to only eight data-points. If one believes in the locally quadratic surface, then a simple remedy is to include additional neighbors in the formulation. For example, the least squares fit to the nearest 12 values produces weights $w_{ij} = 2$, 1 and -1 for the neighbors at distance 1, $\sqrt{2}$ and 2, respectively. Note that, because of symmetry, these weights also give the best cubic fit. For a simulated example, see Kooperberg (1993), though this has since been extended to include the effects of blurring, as well as white noise, in forming the data y. The above notions are not restricted to local polynomial fits and could presumably be extended to non-Gaussian priors, rather along the lines shown by Geman, McClure and Geman (1992). We believe there is considerable potential in these ideas.

4 LOGISTIC REGRESSION WITH ADDITIONAL UNOBSERVED COVARIATES

Logistic regression is very widely used as a means of modeling binomial data, especially in biostatistics. However, in the analysis of disease prevalence, for example, the inclusion of known covariates may not be sufficient to explain the observed variability, with the consequence that interval estimates of the logits are inappropriately narrow. Williams (1982) describes a frequentist approach to this problem in terms of *extra-binomial variation*, which replaces the usual variance by one that is suitably inflated. Breslow (1984) discusses the corresponding modification of the Poisson distribution and re-analyzes the data in Holford (1983) on mortality from prostate cancer among non-white males in the USA.

Here we describe a Bayesian formulation of logistic regression in which unexplained variability is catered for by the introduction of additional unmeasured covariates into the model; see also Zeger and Karim (1991) and, in the context of geographical epidemiology, Besag, York and Mollié (1991). In illustrating the methodology, again on Holford's data, we encounter sets of temporally-ordered parameters for which the pairwise-difference distributions of Section 3 represent our initial beliefs more plausibly than exchangeable priors; see also Berzuini et al (1993). Moreover, since the formulation can lead to significant bimodality, we suggest simple mode-jumping steps that can be included in the algorithm, while still maintaining the correct stationary distribution for the chain. Computationally, our approach also reinforces the point that Bayesian MCMC copes routinely with substantial amounts of missing data. Indeed, the prediction of future events can be handled rigorously in this way, though here we promote a more efficient alternative using traditional simulation.

4.1 Prostate cancer and an extended logistic regression

Holford (1983) analyzes data on mortality from prostate cancer among the non-white male population of the USA, classified by 7 five-year age groups and 7 five-year time periods covering 1935 to 1969. Table 4.1 provides the raw data, together with those for 3 subsequent five-year periods, and also identifies the corresponding 16 birth-cohort groups. Only the data from the first 7 periods will be used to fit our models.

We adopt the following formulation, in which $i = 1, \ldots, I$ denotes age group, $j = 1, \ldots, J$ refers to period, and $k = [ij] = 1, \ldots, K$ is the corresponding cohort. Thus, $k = I - i + j$ and initially, $I = J = 7$ and $K = 13$. Let n_{ij} denote the number of individuals at risk and y_{ij} the number of "respondents" (here deaths from prostate cancer) in cell (i, j). Let p_{ijr} be the probability of response for the rth individual of the n_{ij}. We assume that the p_{ijr}'s can be regarded as a simple random sample from a distribution indexed by (i, j). Then, conditional on the value of $p_{ij} = \mathbf{E}p_{ijr}$, it follows that $y_{ij} \sim \text{bin}(n_{ij}, p_{ij})$. Alternatively, if the p_{ijr}'s are very small, the binomial assumption follows as an approximation, even when the p_{ijr}'s are known individually. We allow for the possibility that there could have been some missing data, distinguishing between observed and unobserved cells by means of superscripts, + and

$-$, respectively. Unobserved cells (i, j) must be assigned a non-zero number n_{ij}^- of individuals at risk. Any choice is valid and the most convenient one is usually $n_{ij}^- = 1$. We have two main objectives: first, to produce posterior distributions for the p_{ij}'s in the first 7 time periods and for associated quantities such as odds or odds ratios; and, second, to provide predictive distributions for the following 3 periods.

Our initial model is a logistic regression with age group, period and cohort as explanatory variables. However, this proves to be inadequate (see also Breslow, 1984) and so we include further covariates z_{ij}, which are unobserved. In a frequentist setting, this would be referred to as "extra-binomial variation" (Williams, 1982) but here the term is ambiguous, so inappropriate. We represent the z_{ij}'s as a random sample from a $\mathbf{N}(0, \delta^{-1})$, so that the p_{ij}'s are related via a logistic-Normal model (Aitchison and Shen, 1980),

$$\xi_{ij} = \ln\{p_{ij}/(1 - p_{ij})\} = \mu + \theta_i + \phi_j + \psi_{[ij]} + z_{ij}. \tag{4.1}$$

We mention a possible "spatial" alternative for the z_{ij}'s in Section 4.4. Note that fitting such a completely confounded model in a frequentist framework requires considerable concessions in the form of assumptions or constraints and is not usually pursued in practice; the Bayesian formulation, however, avoids this difficulty through the adoption of mildly informative priors.

Thus, for the age, period and cohort effects θ, ϕ and ψ, we anticipate similarities between components that are adjacent in time, so that exchangeable priors are inappropriate. Instead, we adopt Gaussian pairwise difference distributions, as described in Section 3. Specifically, we follow Berzuini et al (1993) in opting for priors based on independent *second* differences. Thus, the (unconstrained) prior "density" for θ is

$$\pi(\theta|\kappa) \propto \kappa^{\frac{1}{2}I} \exp\{-\frac{1}{2}\kappa \sum_{i=2}^{I-1} (\theta_{i-1} - 2\theta_i + \theta_{i+1})^2\},$$

and those for ϕ and ψ take similar forms with different scale parameters λ and ν respectively. For μ, we adopt a diffuse prior (uniform on the whole real line) and, for κ, λ, ν and δ, independent highly dispersed but proper gamma distributions with specified parameters (a, b), (c, d), (e, f) and (g, h), respectively. Other choices might be more appropriate but, as in any hierarchical model, one needs to take some care in specifying priors for scale parameters, so as to avoid improper posteriors.

If we also include any missing values y^- as parameters in the model and if the components are all treated independently, the posterior/predictive density, given the observations y^+, is

$$\begin{aligned}\pi(\mu, \theta, \phi, \psi, z, \kappa, \lambda, \nu, \delta, y^-|y^+) \propto \prod_{i,j} \frac{e^{\xi_{ij} y_{ij}}}{(1 + e^{\xi_{ij}})^{n_{ij}}} \times \kappa^{\frac{1}{2}I} e^{-\frac{1}{2}\kappa \sum_{i=2}^{I-1} (\theta_{i-1} - 2\theta_i + \theta_{i+1})^2} \\ \times \lambda^{\frac{1}{2}J} e^{-\frac{1}{2}\lambda \sum_{j=2}^{J-1} (\phi_{j-1} - 2\phi_j + \phi_{j+1})^2} \times \nu^{\frac{1}{2}K} e^{-\frac{1}{2}\nu \sum_{k=2}^{K-1} (\psi_{k-1} - 2\psi_k + \psi_{k+1})^2} \\ \times \delta^{\frac{1}{2}IJ} e^{-\frac{1}{2}\delta \sum_{i,j} z_{ij}^2} \times \kappa^{a-1} e^{-b\kappa} \times \lambda^{c-1} e^{-d\lambda} \times \nu^{e-1} e^{-f\nu} \times \delta^{g-1} e^{-h\delta}.\end{aligned} \tag{4.2}$$

Note that individual location parameters θ_i, ϕ_j and ψ_k are *unidentifiable* in the prior and in the likelihood, so that posterior probability statements can only be made about quantities such as the log-odds ratios ξ_{ij} and the corresponding probabilities p_{ij}. One of the major computational benefits of MCMC is that marginalization over parameters, whether they are identifiable or not, is carried out automatically, merely by ignoring them in the output.

The full conditionals follow immediately from (4.2), up to scale, but the MCMC procedure can be simplified somewhat by first transforming from the z_{ij}'s to the ξ_{ij}'s, in which case the conditional distributions for μ, θ, ϕ and ψ are Gaussian. We sample from the last three of these using Cholesky decompositions, rather than running a componentwise algorithm. Thus, with z_{ij} and p_{ij} defined in terms of μ, θ, ϕ, ψ and ξ as in (4.1),

$$\begin{aligned}
\pi(\mu|\ldots) &\propto e^{-\frac{1}{2}\delta\sum\sum z_{ij}^2}, \\
\pi(\theta|\ldots) &\propto e^{-\frac{1}{2}\kappa\sum(\theta_{i-1}-2\theta_i+\theta_{i+1})^2-\frac{1}{2}\delta\sum\sum z_{ij}^2}, \\
\pi(\phi|\ldots) &\propto e^{-\frac{1}{2}\lambda\sum(\phi_{j-1}-2\phi_j+\phi_{j+1})^2-\frac{1}{2}\delta\sum\sum z_{ij}^2}, \\
\pi(\psi|\ldots) &\propto e^{-\frac{1}{2}\nu\sum(\psi_{k-1}-2\psi_k+\psi_{k+1})^2-\frac{1}{2}\delta\sum\sum z_{ij}^2}, \\
\pi(\xi_{ij}|\ldots) &\propto \frac{e^{\xi_{ij}y_{ij}-\frac{1}{2}\delta z_{ij}^2}}{(1+e^{\xi_{ij}})^{n_{ij}}}, \\
\kappa|\ldots &\sim \Gamma(a+\frac{1}{2}I, b+\frac{1}{2}\sum(\theta_{i-1}-2\theta_i+\theta_{i+1})^2), \\
\lambda|\ldots &\sim \Gamma(c+\frac{1}{2}J, d+\frac{1}{2}\sum(\phi_{j-1}-2\phi_j+\phi_{j+1})^2), \\
\nu|\ldots &\sim \Gamma(e+\frac{1}{2}K, f+\frac{1}{2}\sum(\psi_{k-1}-2\psi_k+\psi_{k+1})^2), \\
\delta|\ldots &\sim \Gamma(g+\frac{1}{2}IJ, h+\frac{1}{2}\sum z_{ij}^2), \\
y_{ij}^-|\ldots &\sim \mathrm{bin}(n_{ij}^-, p_{ij}).
\end{aligned}$$

We use Metropolis to update each of the ξ_{ij}'s, with a symmetric proposal distribution centered on the current parameter value, but it would be almost as easy to implement a corresponding Gibbs step. To achieve reversibility at negligible cost in algebra and programming, we form seven blocks, $[\mu]$, $[\theta]$, $[\phi]$, $[\psi]$, $[\xi]$, $[\kappa, \lambda, \nu, \delta]$ and $[y^-]$, within the last three of which the components are conditionally independent, so that updates can be carried out simultaneously. One cycle of the MCMC algorithm consists of a visit to each block, in random order. Partially parallel updating also reduces the computer time in one of the languages (APL) in which the algorithm was coded.

A possible consequence of the above formulation is that the inclusion of the z_{ij}'s can produce a multi-modal posterior distribution. Often such modes can be located by hill-climbing algorithms with starting points near the maxima in the likelihood and in the prior. We borrowed the ICM (iterated conditional modes) algorithm (Besag, 1986) from spatial statistics and image analysis for this purpose. However, the locations and the corresponding densities (up to scale) do not identify whether there is significant *probability* in more than one mode. If the modes are at all "sticky", one needs to include additional *mode-jumping* steps in the algorithm. Here, for example, differences between age groups might be explained either by variability in θ or by between-row variability in z. In particular, the likelihood is invariant to the transformation from $(\mu, \theta, \phi, \psi, z, \kappa, \lambda, \nu, \delta)$ to $(\mu, \theta^*, \phi, \psi, z^*, \kappa^*, \lambda, \nu, \delta^*)$, where

$$\theta_i^* = z_{i.} \quad , \quad z_{ij}^* = z_{ij} + \theta_i - \theta_i^* \quad , \quad \kappa^* = \delta \quad , \quad \delta^* = \kappa$$

where $z_{i.}$ is the average over j of the z_{ij}'s.

This suggests using the transformation to make a *deterministic* proposal into the other mode on each cycle. Since the transformation is symmetric, it provides a Metropolis step and is accepted with probability

$$\min\{1, \frac{\pi(\mu, \theta^*, \phi, \psi, z^*, \kappa^*, \lambda, \nu, \delta^*, y^-|y^+)}{\pi(\mu, \theta, \phi, \psi, z, \kappa, \lambda, \nu, \delta, y^-|y^+)}\} ,$$

from which most terms cancel out. If the proposal is made within the $[\phi]$ or $[\psi]$ block, reversibility is maintained. A corresponding procedure can be used for ϕ and ψ swaps.

In fact, there seems to be no significant multi-modality in this particular example but that was not the case in a rather similar meta-analysis formulation, concerned with passive smoking (ETS) and lung cancer. There, data for three cohorts and two risk groups produced a significantly bimodal posterior distribution for which an ordinary Gibbs or Metropolis algorithm mixed extremely slowly. The addition of the above mode-swapping proposals provided a simple remedy, with a substantial acceptance rate (Besag, Higdon and Mengersen, 1994). Incidentally, pairwise difference priors were not appropriate in the ETS application.

We close this subsection with a few comments on alternative methods of constructing the basic MCMC algorithm. Initially, our model omitted the z_{ij}'s, in which case we updated all the location parameters via componentwise Metropolis steps, though the Gibbs sampler provides a ready alternative since the full conditionals are log-concave. When the need to include the z_{ij}'s became apparent, it was natural to add further Metropolis steps, with a Gibbs step for δ. It was only with hindsight that we noticed the simplification produced by the transformation from z to ξ. The results from the two runs agree very closely and could of course have been combined. However, a feature common to both formulations is that there is pronounced drift among individual unidentifiable parameters, though the ξ_{ij}'s remain stable. The drift could lead to numerical problems, so we note that it is equivalent here to re-center θ, ϕ and ψ at the end of each cycle, with corresponding adjustment to μ. We also experimented with constrained formulations which require abandoning componentwise Gibbs or Metropolis but for which vector Metropolis updates are entirely straightforward. We sometimes incorporate vector Metropolis updates even in unconstrained models and note that it can be useful to revamp an unbalanced table into a balanced one with missing values. These comments serve merely to emphasize the variety of approaches available through MCMC.

4.2 Results for the first seven periods

Here we describe the results of fitting the model (4.1) and (4.2) to the prostate cancer data in the first seven time periods. Thus, the first seven columns of Table 4.2 provide the observed negative log-odds and summarize the corresponding marginal posterior distributions for the risk of death among non-white males in the USA, classified by age group, period and, implicitly, cohort. The fit is the product of a single Gibbs/Metropolis run of length 275,000 cycles, discarding the first 25,000 and storing every 50th sample thereafter. Acceptance rates for the ξ_{ij}'s vary between 40% and 56%. The Monte Carlo standard errors for the posterior means, assessed via the initial sequence estimators in Geyer (1992), all lie between 1% and 2% of the corresponding posterior standard deviations. In this case, because we make extensive use of large-block Gibbs, via Cholesky, so that the successive samples are close to independent, almost the same accuracy could be achieved from a considerably shorter run length by decreasing the sampling interval.

Not surprisingly, given the number of parameters in the model, there is good agreement between the observed values and the posterior means in Table 4.2. In general, the agreement is much closer than that in Breslow (1984), since the fit there demands an additive row + column decomposition and also ignores the (admittedly rather inconsequential) period effects. Note that the frequentist fit must provide exact fits to the (1,7) and (7,1) cells, which correspond to the single observations on cohorts 1 and 13.

The first seven columns of Table 4.3 present the observed mortality rates and compare them with the corresponding 80% pointwise credible intervals. All the intervals, except that for age group 7 and period 3, straddle the observed values. As regards the necessity for including the z_{ij}'s, the posterior probability that all fall in the range $(-0.1, 0.1)$ is only 0.06, even though the corresponding posterior means all lie in this interval. As an informal comparison, if the z_{ij}'s are omitted from the model, 30 of the 49 observed mortality rates fall outside the corresponding 80% credible intervals. The improved coverage of the extended model is partly because the means necessarily provide a better fit but importantly also because of an increase in the posterior standard deviations.

4.3 Prediction

We now consider predictions for the three additional five-year periods 8, 9 and 10. One possibility is to treat the corresponding 21 cells as missing data and to proceed exactly as in Section 4.1. A drawback of this approach is that it is necessary to decide in advance on the set of cells U for which predictions are required. An alternative, which we first describe in terms of a single run, is to entirely ignore the U-cells when updating the remainder T and then, at the end of each cycle, produce predictions for the U-cells given the current values in the T-cells. This process employs standard forward prediction on each cycle, which must of course be valid. Formally, it happens also to fit into the framework of Appendix 2, with the same use of T and U and with $T \cup U = S = \mathcal{N}$. A major advantage of this procedure is that it can be replaced by two runs, the first of which employs standard MCMC updates and storage for the T-cells, to be followed by a U-cell simulation, with forward prediction from the stored T-cell values on each cycle. This avoids the need to specify U in advance and also usually reduces the computational load. In addition, it is statistically more efficient because sampling for the dataless U-cells is now carried out from the exactly correct conditional distribution given the rest of the parameters. The point here is that one should always use direct simulation, as in (2.1), whenever it is easy to carry out.

In the present context, the aim is to predict over the three periods $J+1$, $J+2$ and $J+3$. Thus, we take the tth stored cycle from the basic $I \times J$ simulation and, for $i = 1, \ldots, I$, $j = J, J+1, J+2$ and $k = K, K+1, K+2$, generate

$$\begin{array}{ll} \phi_{j+1}^{(t)} \sim \mathbf{N}(2\phi_j^{(t)} - \phi_{j-1}^{(t)}, 1/\lambda^{(t)}), & \psi_{k+1}^{(t)} \sim \mathbf{N}(2\psi_k^{(t)} - \psi_{k-1}^{(t)}, 1/\nu^{(t)}), \\ z_{i,j+1}^{(t)} \sim \mathbf{N}(0, 1/\delta^{(t)}), & y_{i,j+1}^{(t)} \sim \mathrm{bin}(n_{i,j+1}, p_{i,j+1}^{(t)}), \end{array}$$

where $n_{i,j+1}$ is the relevant number at risk and $p_{i,j+1}^{(t)}$ is calculated from the corresponding $\xi_{i,j+1}^{(t)}$. Note the linear extrapolation implied by the locally quadratic prior and the way in which additional variability is produced as time progresses. Also, the $y_{i,j+1}^{(t)}$'s need to be computed only if predictions of *numbers* of deaths are required, rather than merely predicted risks. This contrasts with "missing data" implementations.

The observed values and summaries of the predictive distributions for periods 8, 9 and 10 are given in the final three columns of Tables 4.2 and 4.3. There are large increases in the posterior standard deviations of the log-odds, as expected, but, despite this, 9 of the 21 observed values lie outside the corresponding 80% credible intervals. The denominator values are sufficiently large that this remains the case even if one predicts numbers of deaths, rather than risks.

In fact, the above summary is perhaps a little unfair. Only 2 of the 21 observed values fall outside the 90% intervals and, additionally, one should not expect too much from many highly dependent individual forecasts. Instead, one can make *simultaneous* credibility statements by constructing pairs of "surfaces" within which one believes the 7×3 table of risks to lie. We describe one technique in Section 6 and here we merely report some results. In particular, Table 4.3 includes the 80% surfaces and these embrace all the observed mortality rates. The same holds for the 70% surfaces but those at 60% fail in age group 3, period 9.

4.4 Concluding remarks

We close this section with a few remarks that may have wider methodological ramifications, beyond this dataset. First, we have not discussed the sensitivity of our results to changes in the prior distribution and in the likelihood. Here, we merely note that investigations can be carried out along the same lines as in Section 5.5, either by importance sampling or, when this is seen to fail, by re-running the algorithm. For example, one could replace the Gaussian components in the prior by heavier-tailed distributions or, as a more radical structural modification, substitute mixtures of t_f distributions, with hyperpriors on the f's.

Both variants have been used on the passive smoking data (Besag et al, 1994), acknowledging the strong conflicts between the results from some of the studies. In Section 5.6, we illustrate corresponding modifications in both the prior and the likelihood but in the context of agricultural field experiments.

Another possible modification is to replace the locally quadratic priors by the locally linear ones used in the next section. In fact, this produces a slightly better fit to the data for the first seven periods but is unappealing for prediction, because it entirely ignores any trends in the period and cohort effects in making forecasts. Note that this is not the case when the locally linear prior has an interpolative role. In the overall context of the paper, temporal prediction requires more ambitious models than spatial interpolation.

There is sometimes interest in Bayesian image analysis in reconstructing a scene at a coarser or finer pixel resolution than the data (e.g. Gidas, 1989; Jubb and Jennison, 1991). The corresponding procedure is of some methodological interest in the present context; for example, one might replace the 5-year averages by annual populations. In datasets for which annual mortality data are not available, one could still carry out an MCMC analysis involving a multinomial missing values procedure. There are some delicate issues here, concerning approximate self-consistency of prior distributions at different scales.

We make two final points. First, we suspect that many readers would prefer a somewhat different formulation from the one we have adopted. Second, we anticipate that analytical approximations should work well on our model and on others similar to it, especially for the present data where there appears not to be any significant multimodality in the posterior distribution. Nevertheless, we surmise that, in the first instance, MCMC will still provide a ready means of analysis and, in the second, that it is able to address questions that would otherwise remain unanswered.

5 BAYESIAN FORMULATION AND ANALYSIS OF VARIETY TRIALS

5.1 Introduction

Variety trials enable comparisons to be made between different varieties of the same or similar crops and embrace a substantial proportion of agricultural field experiments. They are used by plant breeders in developing new lines and by statutory authorities in drawing up recommendations to the farming community. Experiments consist of a large number of rectangular plots (of land), each of which is devoted to a particular variety. For example, Figure 5.1 shows the spatial layout and the plot yields (standardized to have unit crude variance) in a final assessment trial for 75 varieties of spring barley. It is fairly typical in that the plots are laid out in a few complete replicates, each of which forms a column of long, narrow plots, abutting one another along the longer side. The variety allocated to each plot is chosen at random, within the constraints of the blocking structure of the experimental design. This structure may be quite sophisticated, as in two-dimensional balanced lattice squares

(Kempthorne, 1952, Ch. 24), still widely used in South Africa, or in the less demanding α-designs of Patterson and Williams (1976), exemplified by the layout in Figure 5.1; but often nothing more complicated than a randomized complete blocks design is chosen. In the randomization framework espoused by R. A. Fisher, the design and the analysis should mirror one another. Below, we enlarge on some issues of methodological concern, particularly those that involve variations in fertility.

The usual measurement in each plot is the yield at harvest, as in Figure 5.1. Yield may be influenced by several external factors, particularly weather and "plot fertility". Provided the former can be assumed to have a uniform effect, it need not be considered further in making comparisons between varieties. Otherwise, there is a need for repeat experiments, possibly at different sites. Plot fertility represents the notion that the same variety harvested on different plots would not return the same yield, quite apart from any variation in the seed itself. Fertility effects are not measured directly but are generally acknowledged to be substantial and inherently spatial. Fisher (1928, p.229) states, "the peculiarity of agricultural field experiments lies in the fact, verified in all careful uniformity trials, that the area of ground chosen for the experimental plots may be assumed to be markedly heterogeneous, in that its fertility varies in a systematic, and often a complicated manner from point to point." Here, a uniformity trial is one in which the same variety is assigned to every plot, so that patterns of fertility can be more easily discerned. In fact, patterns are generally evident even when varieties are different; the graphs of raw yields for the spring barley trial in the top panel of Figure 5.2 are by no means untypical.

Consequently, it is highly desirable to employ a sophisticated design or an analysis that takes explicit account of fertility, or some combination of these two. For example, α-designs subdivide replicates into small blocks, within each of which fertility is assumed to be constant. The basic analysis then seeks to eliminate fertility effects by only using intra-block differences in yields. This provides one reason for adopting long, narrow plots, the other being ease of management practice, which also precludes the use of very small plots. Additionally, in an α-design, varieties are assigned to blocks so as to approximately equalize the variance of variety contrasts; it is generally impracticable to ensure exact balance, because this would require far too high a level of replication.

In the above discussion, emphasis rests on spatial *design*, on which there is an enormous literature. In the last 15 years, there has been acknowledgment, particularly in the UK and in Australia, that the adoption of model-based spatial *analysis*, even in a randomized complete blocks experiment, can recover most, if not all, of the accuracy and precision obtained classically from a very sophisticated design. In a linear layout, the simplest plausible model for plot fertilities is the Gaussian random walk of Section 3, first implemented implicitly by J. N. Papadakis, a distinguished Greek agronomist and soil scientist, in the 1920's and explicitly by Besag and Kempton (1984, 1986). In such a formulation, the conditional expectation of fertility ψ_i on plot i, given the ψ_j's on all other plots, is the mean of those ψ_j's on plots adjacent to it. Note here that Papadakis subsequently adopted an attenuated version of the model (Bartlett, 1938, 1978) but eventually reverted to the original form; see Papadakis

(1984) for a history and further references. An alternative, proposed by Green, Jennison and Seheult (1985), is to assume that second, rather than first, differences in fertility are uncorrelated, again as described in Section 3; see also Section 4. Note that, in either model, the only unknown parameter is the variance of the first or second differences.

A more radical approach, adopted by Gleeson and Cullis (1987), is to employ Box-Jenkins ARIMA methodology for identification of a particular fertility model. This can result in quite complicated formulations and the need to estimate several additional parameters. Our view is that this shifts the emphasis too far towards a model-fitting exercise and may lead to overfitting, in that genuine variety effects can be "washed out" into the fertility model. Our preference, in the context of a Bayesian analysis, is to permit flexibility in the distributional assumptions, as described in Sections 5.5 and 5.6.

Other references to spatial analysis of field experiments include Wilkinson, Eckert, Hancock and Mayo (1983), Wilkinson (1984), Green (1985), Williams (1986) Besag and Seheult (1989), Martin (1990), Baird and Mead (1991), Zimmerman and Harville (1991), Cullis, McGilchrist and Gleeson (1991), Kempton, Seraphin and Sword (1993). All of these adopt a frequentist or data analytic approach to the inferential tasks. Bayesian viewpoints are discussed briefly by Smith (1978) and Besag and Green (1993, Section 6) and at length, with examples, by Besag and Higdon (1993) and in the forthcoming Taplin and Raftery (1994). The second list is surprisingly short since the Bayesian paradigm seems ideally suited to such important areas as ranking and selection in variety trials. For example, one can produce a posterior probability distribution for the variety that is best, or determine the smallest subset that contains the best variety with prescribed probability; and the difficulties of making multiple comparisons, which bedevil any frequentist approach, do not arise.

In Sections 5.2 and 5.3, we develop a quite widely applicable Bayesian formulation for the analysis of field experiments, implemented via MCMC. Since this extends to more complicated experiments that have a treatment structure, such as factorial designs, it is convenient initially to address a wider class of problems and then specialize to the case in which treatments and varieties coincide. Section 5.4 provides a basic analysis of the spring barley trial and Sections 5.5 and 5.6 consider sensitivity analysis and extensions of the basic formulation.

5.2 Basic Bayesian formulation

We assume that the experimental layout is in r separate columns, as for example in Figure 5.1, with plots i indexed in some convenient manner. The observed yield and the (fixed) unknown fertility in plot i are denoted by y_i and ψ_i, respectively. We write τ for the vector of m treatment effects, among which particular contrasts are the main focus of the experiment, and T for the corresponding design matrix. Then the simplest model for the random vector y is, perhaps after transformation,

$$y|\tau, \psi, \lambda_y \sim \mathcal{N}(\psi + T\tau, I_n/\lambda_y), \tag{5.1}$$

where n is the total number of plots, I_n is the $n \times n$ identity matrix and λ_y is the unknown precision of the y_i's. Of course, one might wish to replace (5.1) by a robust/resistant alternative (cf. Section 5.6) or, in particular applications, by a binomial or Poisson distribution.

The formulation is completed by specifying (presumably) independent prior densities for λ_y, τ and ψ. Obvious choices are $\lambda_y \sim \Gamma(a, b)$, where a and b are specified, and, in experiments such as variety trials that have no special treatment structure, $\tau \sim \mathcal{N}(0, I_m/\lambda_\tau)$, with $\lambda_\tau \equiv 0$ or $\lambda_\tau \sim \Gamma(c, d)$. However, there will often be information from previous experiments that can be incorporated into these distributions. At the basic level, we represent our views about the plot fertilities in separate columns by independent Gaussian random walks, each with precision parameter $\lambda_\psi \sim \Gamma(g, h)$, so that

$$\pi(\psi|\lambda_\psi) \propto \lambda_\psi^{\frac{1}{2}n} \exp\left\{-\tfrac{1}{2}\lambda_\psi \sum_{i\sim j}(\psi_i - \psi_j)^2\right\}, \tag{5.2}$$

where the summation is over pairs of adjacent plots $i \sim j$; thus, each ψ_i occurs in n_i terms, where $n_i = 1$ or 2 is the number of plots adjacent to i. We mention two-dimensional fertility adjustment in Section 5.6 but usually there is little information to connect the columns, because of their physical separation and/or the standard use of long, narrow plots; indeed, one might want to allow separate λ_ψ's in some contexts. Note that the impropriety in (5.2) permits an arbitrary level of fertility in each column and corresponds to the usual frequentist practice of removing replicate effects.

Of course, it is unrealistic to expect a spatial "model" to provide anything more than a crude representation of the true fertility process. However, this is generally not critical, firstly because it is contrasts between the replicated treatment effects rather than the individual plot fertilities themselves that are of primary concern, and secondly because the purpose of such a model is one of interpolation rather than extrapolation. Again, we note that spatial formulations can often be less ambitious than those in time-series analysis.

5.3 Computation

It is convenient to rewrite the summation in (5.2) as $\psi^T W \psi$, where W is the $n \times n$ matrix with (i, j) element,

$$W_{ij} = \begin{cases} n_i, & i = j, \\ -1, & i \sim j, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, W is block diagonal, with r identical blocks. It then follows that the posterior density for τ, ψ and λ, given y, is proportional to

$$\begin{aligned} &\lambda_y^{\frac{1}{2}n} \exp\left\{-\tfrac{1}{2}\lambda_y(y - \psi - T\tau)^T(y - \psi - T\tau)\right\} \times \lambda_\tau^{\frac{1}{2}m} \exp\left\{-\tfrac{1}{2}\lambda_\tau \tau^T\tau\right\} \\ &\quad \times \lambda_\psi^{\frac{1}{2}n} \exp\left\{-\tfrac{1}{2}\lambda_\psi \psi^T W \psi\right\} \times \lambda_y^{a-1} e^{-b\lambda_y} \times \lambda_\tau^{c-1} e^{-d\lambda_\tau} \times \lambda_\psi^{e-1} e^{-f\lambda_\psi}, \end{aligned} \tag{5.3}$$

with appropriate modification if $\lambda_\tau \equiv 0$. Hence,

$$\tau|\psi, \lambda, y \sim \mathcal{N}(\lambda_y Q_\tau^{-1} T^T (y - \psi), Q_\tau^{-1}), \tag{5.4}$$

where $Q_\tau = \lambda_y T^T T + \lambda_\tau I_m$, and

$$\psi|\tau, \lambda, y \sim \mathcal{N}(\lambda_y Q_\psi^{-1} z, Q_\psi^{-1}), \tag{5.5}$$

where $z = y - T\tau$ and $Q_\psi = \lambda_y I_n + \lambda_\psi W$. The conditionals for the precisions are mutually independent, with (when $\lambda_\tau \not\equiv 0$)

$$\begin{aligned} \lambda_y|\ldots &\sim \Gamma\left(a + \tfrac{1}{2}n, b + \tfrac{1}{2}(z-\psi)^T(z-\psi)\right), & (5.6)\\ \lambda_\tau|\ldots &\sim \Gamma\left(c + \tfrac{1}{2}m, d + \tfrac{1}{2}\tau^T\tau\right), & (5.7)\\ \lambda_\psi|\ldots &\sim \Gamma\left(g + \tfrac{1}{2}n, h + \tfrac{1}{2}\psi^T W \psi\right). & (5.8) \end{aligned}$$

For simulating from the posterior distribution, the results (5.4 – 5.8) are ideally suited to a time-reversible, cyclic Gibbs sampler, with three components τ, ψ and λ updated in random order within each cycle. The updating of τ and ψ can be carried out via Cholesky decompositions, with the block diagonal structure of Q_ψ easing the ψ-step. Note that any missing y_i's merely add a fourth component to the Gibbs sampler, in which corresponding values are generated from (5.1). The variability in these y_i's is inherited by the relevant treatment or variety effects without any further change to the algorithm.

For the prior (5.2), fertilities can also be considered individually. It follows immediately from (5.3) that

$$\psi_i|\ldots \sim \mathcal{N}\left(\frac{\lambda_\psi n_i \bar{\psi}_i + \lambda_y z_i}{\lambda_\psi n_i + \lambda_y}, \frac{1}{\lambda_\psi n_i + \lambda_y}\right), \tag{5.9}$$

where $\bar{\psi}_i$ is the current mean fertility of the plots adjacent to i. The distribution (5.9) is intuitively appealing, having a mean that combines the information in plot i with that in the neighbors, according to the values of n_i, λ_y and λ_ψ. Also, the ψ_i's in alternate plots are conditionally independent, so that, instead of (5.5), one can carry out two steps, in each of which the samples are drawn from independent Gaussian distributions. However, the advantage of (5.5) is that, if τ and λ have the correct limit distribution, then so does ψ, whereas this is not necessarily the case in (5.9), which depends also on $\bar{\psi}_i$. Thus, though an algorithm based on (5.9) runs much faster per cycle, it may take more cycles to converge, as indeed we have found in practice. We can suggest no clear winner at this stage.

Note that, when $\lambda_\tau \equiv 0$, the mean in (5.4) is the ordinary least squares estimate of τ, based on the current value of $y - \psi$. The analysis is then *additive*; that is, if y is replaced by $y + T\tau_0$, for any fixed τ_0, then the posterior density of $\tau - \tau_0$, ψ and λ is proportional to (5.3), omitting terms that involve λ_τ. This permits a rigorous assessment of performance to be made, model-free except for the assumption that treatments act additively on yields, by using data from uniformity trials with superimposed dummy treatments; see, for example, Besag and Kempton (1986) and Zimmerman and Harville (1991), in a non-Bayesian context. A non-zero λ_τ provides the usual shrinkage of τ towards zero. When $T^T T$ is diagonal, the components of τ in (5.4) are conditionally independent, so that, for example, in variety trials, the full conditional for the effect of variety k takes the form,

$$\tau_k | \ldots \sim \mathcal{N}\left(\frac{\lambda_y r_k \bar{u}_k}{\lambda_\tau + \lambda_y r_k}, \frac{1}{\lambda_\tau + \lambda_y r_k}\right),$$

where $\bar{u}_k$ is the current mean value of $y_i - \psi_i$ for the r_k plots receiving variety k; when $\lambda_\tau \equiv 0$, the mean reduces to $\bar{u}_k$ itself. Usually, $r_k = r$ but sometimes there is additional replication for a standard variety, with which the others are to be compared.

5.4 Analysis of a spring barley trial

The Scottish Agricultural Colleges are responsible for many of the final assessment trials on spring barley and winter wheat in the U.K. The data in Figure 5.1 are extracted from a catalog of ten such trials, kindly supplied to us with the intention of covering a wide range of experimental and environmental conditions. Each trial conformed to an α-design, with the number of varieties between 17 and 75. All but one were in three replicates; two had a missing value. We have applied exactly the same method of analysis to each of the experiments, after first rescaling the yields to have crude variance unity. The data in Figure 5.1 are perhaps the most challenging in the catalog: there are 75 centered variety effects and 225 fertilities to be estimated; there is no clear winner, so that probability statements are especially important; variety 27 has a missing value in the third replicate; and, as is also true in the other trials, there is strong evidence of substantial variation in fertility within replicates.

Figure 5.2 shows three different decompositions of the yields into posterior means of variety effects, fertilities and residuals. Here, we concentrate on the topmost graphs, which correspond to the basic formulation (5.3). The results were obtained using a reversible Cholesky version of the Gibbs sampler, with a run length of 6000 cycles, of which the first 1000 were discarded. The run was initiated by a simple randomized complete blocks (RCB) decomposition of the yields, which imposes constant fertility effects within columns. Figure 5.3 shows the progress of the log posterior density (up to a constant) for the entire run, split into successive batches of 500. Note that the RCB estimates have much larger posterior density than typical samples from (5.3). In practice, we look at many other similar graphs in assessing convergence.

Table 5.1 provides estimates of centered variety effects under several models and methods of analysis, with the RCB and RevG columns referring to the two analyses above. The MCMC standard errors for the RevG posterior means are about 0.007 (slightly larger for variety 27). The 90% credible intervals for the λ's are

$$\hat{\lambda}_y : (5.1, 12.8), \hat{\lambda}_\psi : (4.9, 11.5), \hat{\lambda}_\tau : (2.8, 5.8).$$

All point estimates were within acceptable Monte Carlo errors of a previous run of length 12000 cycles, discarding the first 2000, but based on (5.9) rather than (5.5).

In practice, interest centers on variety differences. Here the typical width of a 90% credible interval is about 1.2, with negligible Monte Carlo standard error. The top 15 varieties under RevG, ranked by posterior means, can be found in Table 5.2 but, in fact, 90% credible intervals for differences among any of the top 11 straddle zero. Thus, the experiment is somewhat inconclusive. If we wish to select the best variety, under the conditions of the experiment, this is variety 56 with posterior probability 0.32 and either 56 or 35 with probability 0.51; in each case, the Monte Carlo standard error is 0.01. Alternatively, variety 56 is among the top 5 with posterior probability 0.82. Were there an intention to perform a subsequent, smaller, comparative trial, one would need to carry over the top 6, 8 or all 15 varieties in the RevG column of Table 5.2 to have posterior probability 0.90, 0.95 or 0.99 of including the best. Each of the above probabilities is the relative frequency of the corresponding event in the MCMC run. Incidentally, not all the datasets in the catalogue are so vague; for example, in one of the winter wheat trials, there is a single variety that is best with posterior probability 0.99.

5.5 Sensitivity analysis

The above statements all assume that the formulation is correct. Thus, we need to address some aspects of sensitivity analysis, following the ideas described in Appendix 3 and here focusing on assumptions in the prior. First, there is the choice of constants a, b, c, d, g, h. In our base model, we have always employed $a = c = g = 1$ and $b = d = h = 0.005$, after first standardizing the yields. The value 0.005 is a potential source of instability and was first changed to 0.0005. Importance sampling, with respect to the original run, produced negligible changes in the variety estimates and an acceptable maximum weight of 0.00023 against the average 0.0002. In the event, we did check this conclusion by re-running at the new setting. Secondly, we changed the values of b, d and h to 0.05, again with negligible effect.

A second type of modification is structural and concerns the choice of prior for the fertilities. Instead of (5.2), one alternative is (3.3), with $w_{ij} = 1$, for adjacent plots $i \sim j$, and $\gamma = \lambda_\psi^{\frac{1}{2}}$. Then,

$$\pi(\psi_i|\psi_{-i}, \lambda_\psi) \propto \lambda_\psi^{\frac{1}{2}} \exp\{-\lambda_\psi^{\frac{1}{2}} \sum_{j \in \partial i} |\psi_i - \psi_j|\}, \tag{5.10}$$

which is symmetric about $\frac{1}{2}(\psi_{i-1} + \psi_{i+1})$ and uniform between ψ_{i-1} and ψ_{i+1}. This prior is able to accommodate occasional large jumps in fertility. Note that scaling by $\lambda_\psi^{\frac{1}{2}}$ rather than λ_ψ maintains common units for the λ's but destroys the conjugacy with respect to a Γ distribution in the posterior for λ_ψ. In implementing (5.10), we update fertilities individually, using a Hastings algorithm, and use standard rejection sampling for λ_ψ. The results are shown in the middle graphs of Figure 5.2 and in the L1MH columns of Tables 5.1 and 5.2, all of which provide satisfactory agreement with the corresponding RevG estimates. This is not always the case, as we discuss briefly in Section 5.6.

We also carried out an MCMC run with a vague prior (i.e. $\lambda_\tau = 0$) for variety effects. The results appear in the RevU columns of the two tables and can be compared with those for the Gaussian prior and also with the frequentist BK extended first-differences analysis in Besag and Kempton (1986) which, like RevU, imposes no shrinkage on the variety effects. The contrast with RevG is quite substantial, caused by the rather small variety effects, and does not occur in all of the trials in the catalogue.

5.6 Extensions

MCMC allows one to explore a very wide range of different formulations. One possibility is to replace the Gaussian assumptions in the likelihood and the fertility increments by t_{ν_y} and t_{ν_ψ} distributions, with specific choices of ν_y and ν_ψ. Outliers, as well as sudden, large jumps in fertility, are accommodated when the degrees of freedom in the t–distributions are small. This is illustrated in Besag and Higdon (1993, Section 2) on an awkward randomized complete blocks trial for wheat (Wilkinson, 1984), carried out in El Batán, Mexico, by the International Center for Improvement of Maize and Wheat (CIMMYT).

As a further extension, one can treat ν_y and ν_ψ as additional parameters in the model, with their own priors. This *hierarchical–t* formulation is most easily implemented using the definition of t in terms of Gaussian and chi-square distributions. Formally, this part of the algorithm is a Gibbs sampler construction, via an auxiliary variable. The bottom graphs in Figure 5.2 and the $t_? - t_?$ columns in Tables 5.1 and 5.2 show the results for the spring barley trial, with priors for ν_y and ν_ψ that are independent and uniform on the integers 1, 2, 4, 8, 16, 32 and 64, though almost any other choice could be made. Table 5.3 approximates the joint posterior distribution of ν_y and ν_ψ.

The spring barley trial provides an example of a well-designed experiment, so the robustness of the analysis to changes in the likelihood and in the fertility prior is not surprising. For the El Batán data, the results are less consistent, with the estimates for two of the varieties differing markedly under the Gaussian and the hierarchical–t formulations. This is because the latter analysis detects a large jump in fertility and an apparent outlier, both of which materially affect the two variety estimates. A standard non-spatial analysis of these data would be wholly unreliable because of the large fertility gradients and the lack of an appropriate blocking structure. Of course, the results would still be unbiased under randomization.

One modification of the basic formulation in Section 5.2 is to experiments that have a factorial design, in which case one must ensure that the priors and the MCMC honor the treatment structure. For an example, we refer to Besag and Higdon (1993, Section 3) in which a single replicate of a 2×3^3 experiment on triticale is analyzed.

In some trials that have a two-dimensional layout, linear fertility adjustment may be insufficient. For a brief discussion of some pairwise difference priors on the rectangular lattice, see Section 3, though it is usually necessary to allow asymmetry between the plot axes and to include finite-sample edge corrections (cf. Besag and Kooperberg, 1993) unless the distribution satisfies separability. Besag and Higdon (1993, Section 4) discuss a 10×15 balanced lattice square experiment on 25 varieties of spring wheat, in which a particular variety appears on the boundary of all six replicates and is shown to be vulnerable to spatial effects in a standard analysis.

Finally, the basic formulation can be extended to combine information from several different but related experiments. For example, in developing new lines, plant breeders conduct a sequence of trials, with successive reduction of the number of varieties at each stage. In a Bayesian framework, the posterior distribution from the first trial should become the prior for the second, and so on. This is achieved if, at each point in the sequence, an overall MCMC analysis is carried out simultaneously on the set of completed experiments.

6 GAMMA CAMERA IMAGING

In this section, we describe an application of MCMC to image reconstruction. Bayesian approaches to image analysis, originating in Grenander (1983), Geman and Geman (1984) and Besag (1983, 1986), can be described briefly in terms of an unobservable true image x, which is subject to degradation in producing an observed version y. The loss of information may be extremely severe, as in tomography, where the data are degraded projections of the true scene, or in high-level computer vision tasks such as object recognition, where more subtly the truth is a construct in image space. Thus, the Bayesian paradigm is especially attractive in imaging, because it allows the user great flexibility both in specifying the prior distribution $\pi(x)$ for x and the likelihood $L(y|x)$. The former is important because there is always contextual information about an image, the latter because different sensing devices have their own peculiar characteristics, which can usually be assessed experimentally.

One basic use of Bayesian image analysis is to produce point estimates of pixel images or image functionals, either via MCMC, such as Gibbs sampling or simulated annealing, or by deterministic methods, such as ICM or modified EM. In object recognition, deformable templates and models from stochastic geometry have been the focus of much recent research (e.g. Grenander and Keenan, 1989; Ripley and Sutherland, 1990; Amit et al, 1991; Aykroyd and Green, 1991; Grenander et al, 1991; Phillips and Smith, 1994; Baddeley and van Lieshout, 1994; Grenander and Miller, 1994; Mardia and Hainsworth, 1994, and references therein, including the influential Bookstein, 1989, 1991). In the former, each category of object is assigned a fixed geometrical prototype, which is then allowed to deform stochastically to cater for the variability of objects within each class, which may be partly a product of natural variation and partly due to perspective, magnification, occlusion and so on. If gross deformations are permitted, categories will be ill-defined, whereas too stringent a definition will exclude some objects from their true classes. Thus, the definition of the prior is more critical than in low-level computer vision. In addition, it is usually necessary to specify more mundane aspects of the image prior, relating to pixel intensities, for example. Incidentally, the usual updating procedure for stochastic templates is an approximation to Langevin diffusion, which can sometimes be rigorized as in Section 2.3.4, but a promising alternative is a Metropolis algorithm based on proposals that involve simple vertex-displacements.

In our application, we are concerned only with low-level imaging but nevertheless the task involves deconvolution and is methodologically similar to problems in tomographic reconstruction. In keeping with the remainder of the paper, we shall place some emphasis on measures of uncertainty, which is unusual in image analysis, and perhaps controversial.

6.1 Modeling nuclear medicine images

Gamma-camera imaging is a "nuclear medicine" technique, used in modern medical diagnosis. While it is incapable of the high spatial resolution of anatomic detail that characterize digital X–ray and magnetic resonance imaging, for example, it nevertheless has an increasingly important role through its ability to study function rather than form.

A radioactively labeled substance is introduced into the patient, usually though injection or inhalation. The substance is chosen as one known to become concentrated in the organ of interest, in a manner related to the phenomenon under study, and of course the labeling allows the pattern of concentration to be observed non-invasively. Photon emission occurs in the organ, at a rate varying spatially according to the concentration, and indirect measurements of this concentration can thus be made by counting emitted photons with a suitable instrument.

The detector in use in most hospitals is the *gamma camera*; physical and operational properties are described by Larssen (1980). Incoming photons have to traverse the narrow parallel bores in a lead collimator; those unable to do so are absorbed. The photons successful in passing through impact on a fluorescing crystal, causing a fluorescence whose location and energy are measured by means of photo-multiplier tubes and electronic circuitry.

Ultimately, a spatially discretized map of the photon arrivals, within some prescribed energy range, is available to the clinician in the form of an image. In simple use of the device, such images are presented "raw", but the camera can also be used in other modes. These include the production of sequences of images recording the dynamic progress of the radioactive isotope through body tissue, as a result of metabolization, for example, and the technique of single photon emission computed tomography (SPECT). This is the reconstruction of three-dimensional information about the patient from data obtained by rotating the gamma camera around the patient to obtain a sequence of projections. Geman, Manbeck and McClure (1993) build on their earlier work to produce a comprehensive framework for reconstruction of SPECT data. They use maximum likelihood, followed by ICM (Besag, 1986), to converge to a local maximum of the posterior distribution. Estimation of the interaction parameter in the robust prior is carried out off-line, using MCMC. Their approach to modeling the likelihood has been further developed by Weir and Green (1994). Green (1990) takes a different path, via a "one-step late" adaptation of the EM algorithm (Dempster, Laird and Rubin, 1977) to approximate the maximum *a posteriori* reconstruction. This has the computational advantage of not requiring the maximum likelihood solution en route, to which convergence is extremely slow.

Here, we are concerned with inference based only on a single image frame obtained with a gamma camera. The data consist of a rectangular array of counts of detected photons, which we denote by $\{y_t, t \in T\}$. The appropriate statistical model is essentially that which applies to all emission data, including both PET (Shepp and Vardi, 1982) and SPECT. If $\{x(\mathbf{u}), \mathbf{u} \in \mathbf{R}^3\}$ denotes the isotope concentration at spatial location $\mathbf{u}$, and $a_t(\mathbf{u})$ the mean rate of arrival at detector location t of photons emitted at $\mathbf{u}$, per unit concentration at $\mathbf{u}$, then we find

$$y_t \sim \text{Poisson}\left(\int a_t(\mathbf{u}) x(\mathbf{u}) d\mathbf{u}\right)$$

independently for all $t \in T$. This model captures the principal physical effects of absorption and scattering in the body (due to nuclear interactions), and the collimation, all of which lead to the thinning expressed by $a_t(\mathbf{u})$, while the integral corresponds to the superposition of emissions from different point sources $\mathbf{u}$.

The model will apply to data from a gamma camera used in any of the modes mentioned above: here, with a single frame, the integral above represents simply a pixelated, blurred, attenuated projection. Because of the narrow angle of view imposed by the collimator and the high level of attenuation per unit distance within the body, there is very little information in the data about the variation in $x(\mathbf{u})$ in the third dimension, orthogonal to the plane of the projection. As in Aykroyd and Green (1991), we therefore drop one dimension and, after discretization of body space on the same grid as that on which the data are recorded, assume the model

$$y_t \sim \text{Poisson}\left(\sum_s h_{ts} x_s\right),$$

in which x_s represents the discretized attenuated projection of $x(\mathbf{u})$ onto the plane of the camera, and h_{ts} is the discretized point spread function. Our present analysis does not use information about physical units, so we assume that $\sum_t h_{ts} = 1$, that is, that y and x are dimensionless, and on a comparable scale.

In practice, medical physicists routinely form gamma camera images of phantoms involving line and point sources in order to calibrate the device, and to estimate the point spread function; thus h_{ts} can be assumed known. Here we will assume that h_{ts} is a discretization of a circular Gaussian function, with standard deviation 2.0 in pixel units.

We observe y and wish to make inference about x. Our prior on x is again a non-Gaussian pairwise difference distribution. It is convenient to work on a logarithmic scale, because values of x are necessarily non-negative, and we assume

$$\pi(x) \propto \exp\left\{-\sum_{r \sim s} \Phi\{\gamma(\log x_r - \log x_s)\}\right\} \Bigg/ \prod_r x_r, \tag{6.1}$$

where $\Phi(u) = \delta(1+\delta)\log\cosh(u/\delta)$, as in (3.4), and γ and δ are regarded as known positive constants. The summation in (6.1) is over all pairs of directly and diagonally adjacent pixels

$r \sim s$, and the product in the denominator is a Jacobian term arising from adoption of the logarithmic scale for x. In the example below, we set $\gamma = 1.8$ and $\delta = 2.0$ but the results are quite robust. In some imaging tasks, it is necessary to treat such constants as patient- and context-dependent, in which case an extra layer in the hierarchical formulation is required. Pragmatic approaches are available but we briefly describe a fully Bayesian approach in Section 7.

6.2 MCMC for gamma camera imaging

The posterior density $\pi(x|y)$ has full conditionals,

$$\pi(x_s|x_{-s}, y) \propto \pi(x_s|x_{-s}) \prod_{t:h_{ts}\neq 0} p_t(y_t|x),$$

where $p_t(y_t|x)$ denotes the relevant Poisson probability, so that

$$\pi(x_s|x_{-s}, y) \propto \exp\left\{-\sum_{r\in\partial s} \Phi\{\gamma(\log x_r - \log x_s)\} + \sum_{t:h_{ts}\neq 0} (y_t \log\{\sum_r h_{tr}x_r\} - \sum_r h_{tr}x_r) - \log x_s\right\}.$$

This conditional density has no mathematically aberrant behavior but it is clearly non-standard and is neither log-concave in x_s nor in $\log x_s$. Values from it might be generated by the ratio method, for example, but since that would incur a considerable setup cost for each random value sampled, not to mention algebraic complications, we prefer a sitewise Hastings algorithm, using a proposal density that corresponds to a uniform distribution for $\log x_s{}'$, centered on the current value $\log x_s$. Thus,

$$R_s(x_s \to x_s{}') \propto (x_s{}')^{-1}$$

on an interval $|\log(x_s{}'/x_s)| < c$. Such a density satisfies the positivity requirement (ii) of Section 2.3.3, and results in an acceptance function calculated from (2.9):

$$\begin{aligned} A_s(x_s \to x_s{}'; x_{-s}) \quad = \quad & \min\left[1, \exp\left\{-\sum_{r\in\partial s} (\Phi\{\gamma \log x_r/x_s{}'\} - \Phi\{\gamma \log x_r/x_s\})\right.\right. \qquad (6.2) \\ & \left.\left. + \sum_{t:h_{ts}\neq 0} (y_t \log\{1 + h_{ts}(x_s{}' - x_s)/\sum_r h_{tr}x_r\} - h_{ts}(x_s{}' - x_s))\right\}\right]. \end{aligned}$$

Note that the Jacobian terms have canceled.

The performance of the MCMC method is influenced by the value of the constant c, determining the spread of the proposal distribution for $x_s{}'$. In the example below, a small-scale pilot experiment was carried out to assess the effect of c on the autocorrelation times for individual x_s, and on statistics of the Hastings changes. As c was increased, the mean acceptance probability decreased but the root mean square of the resulting changes increased.

The autocorrelation times, estimated by the truncated periodogram method in Sokal (1989), apparently achieved a minimum around $c \approx 0.35$, corresponding to an interval for $x_s{}'$ with endpoints differing by a factor of 2, and this was the value adopted for the main runs. One factor that was not considered, that might be relevant in other analyses, was that low average probabilities of acceptance mean many ties in the MCMC sample, which may cause artifacts in exploratory analysis of data from such runs. In drawing $x_s{}'$ and deciding whether to accept it, according to (6.2), there are computational advantages in maintaining the values $\mu_t = \sum_r h_{tr} x_r$ up-to-date as x varies. We do this incrementally, but then recompute all μ_t from scratch every 50 sweeps to mitigate the effect of accumulated rounding error.

We routinely compute pixelwise means and variances along the MCMC run, but formal assessment of joint posterior variability of the whole of x is more problematic. A simple technique we find useful is to display the sequence of $x^{(t)}$ on the screen during the simulation; such informal dynamic graphics give a valid impression of joint variability provided that dependence along the sequence is not excessive. We have previously applied the same idea in geographical epidemiology and suggest that such spatial displays may also be helpful in non-spatial applications.

6.3 Simultaneous credible regions based on order statistics

If a substantial sample from the post-burn-in MCMC realization is stored, then it is possible to construct simultaneous credible regions for the whole vector x, without making parametric assumptions. Since these bounds will be based on order statistics, they will be exactly equivariant to (strictly) monotone componentwise transformations of the variables, for continuous distributions $\pi(x)$.

Denoting the stored sample by $\{x_i^{(t)} : i = 1, 2, \dots, n;\ t = 1, 2, \dots, m\}$, order $\{x_i^{(t)} : t = 1, 2, \dots, m\}$ separately for each component i, to obtain order statistics $x_i^{[t]}$ and ranks $r_i^{(t)}$, $t = 1, 2, \dots, m$. For fixed $k \in \{1, 2, \dots, m\}$, let t^* be the smallest integer such that $x_i^{[m+1-t^*]} \leq x_i^{(t)} \leq x_i^{[t^*]}$ for all i, for at least k values of t. It is equal to the k^{th} order statistic from the set $\{\max(\max_i r_i^{(t)},\ m + 1 - \min_i r_i^{(t)}) : t = 1, 2, \dots, m\}$.

Then $\{[x_i^{[m+1-t^*]}, x_i^{[t^*]}] : i = 1, 2, \dots, n\}$ are a set of simultaneous credible regions containing at least $100k/m\%$ of the empirical distribution. The ordering and ranking at each component i can be done at the same time, so the procedure requires only ordering $n + 1$ vectors of length m.

Even for continuously distributed x, these bounds will be slightly conservative, to an extent increasing in n, because of ties between the ranks over different components. For such x, however, the bounds will be consistent as $m \to \infty$, for fixed n. One-sided and other asymmetric bounds can be constructed analogously.

6.4 An example

Data from one gamma-camera study, based on a 256×256 image of a pair of hands, are shown in Figure 6.1. This image shows the raw data: the photon counts in individual pixels vary between 0 and 93. The MCMC simulation was organized as described above and run for 8000 sweeps, of which the first 500 were discarded. The image in Figure 6.2(a) represents the MCMC estimate of the posterior mean for a 64×64 sub-image, and that in Figure 6.2(b) shows the corresponding pixelwise posterior standard deviations. It is difficult to gain a clear impression of the scale of this variability in image form, so results for a left-to-right cross-section through the wrists, passing through the hot spot evident in the left wrist, are displayed in Figure 6.3 and include the raw data y_t, and MCMC estimates of the posterior mean $\mathbf{E}(x_s|y)$, 80% pixelwise credibility bands for x_s, 80% simultaneous credibility bands for x_s, and the blurred posterior mean, $\mathbf{E}(\sum_s h_{ts}x_s|y)$. The credibility calculations are based on a stored subset of 2500 samples, equally spaced along the MCMC realization.

Note that noise and blur have been removed, without destroying definition at the edges of the features. As discussed in Green (1990), use of the $\log\cosh()$ function has countered the effect of a Gaussian prior to smooth over steep gradients in x at tissue boundaries, to an extent controlled by γ and δ. The formulation has respected the underlying Poisson process and the non-negativity of the image itself. Credible intervals and simultaneous credible regions are obtained as a by-product of the MCMC and posterior probability statements about other functionals of x could have been made. The extension of MCMC to some different imaging modalities is relatively straightforward, as in synthetic MRI (Glad and Sebastiani, 1994; Besag and Maitra, 1994), or is problematic merely for want of current raw computing power, as perhaps for a full three-dimensional treatment of emission tomography.

7 DISCUSSION

In conclusion, we draw together some of the key issues in the application of Markov chain Monte Carlo methods in Bayesian computation. It is evident from a rapidly expanding literature that MCMC is applicable to a very wide range of complex Bayesian models, many of which are at present well beyond the reach of other computational methods. Quite apart from philosophical differences, such formulations often have no obvious frequentist counterparts (as in Section 4, for example) or there are no existing frequentist computational procedures (as in Section 5.6), though here MCMC maximum likelihood might prove useful. We have found the Bayesian paradigm especially persuasive in spatial applications where there are known to be local contextual regularities in the true scene, that cannot be modeled plausibly by a physical process but for which one can (crudely) represent one's beliefs via a Markov random field. Thus, the true scene is considered to be fixed, rather than sampled from a process, but our views about it are represented stochastically. In image analysis, the paradigm provides a unified approach to many different problems, so that the application to deconvolution in Section 6 is part of a general framework and there is no need continually to invent *ad hoc* techniques as each new problem comes along. Other applications might involve mosaic priors in classification, edge sites in segmentation, and stochastic templates in object recognition. Likewise, changes in the imaging system are immediately accommodated by appropriate modification of the likelihood, so that in MRI, for example, the Poisson model is replaced by a two- or three-component Gaussian distribution at each pixel.

The equivalence between any Markov random field formulation and the Gibbs sampler explains why the origins of such methods are to be found in spatial applications. One could describe Bayesian computation via MCMC abstractly as the restoration of hidden Markov random fields. We hope also to have shown that a spatial perspective may be useful even in contexts that are not overtly spatial. For example, Section 4 uses pairwise difference priors to represent associations between factors ordered in time and it is suggested that higher-dimensional "spatial" priors may be useful in modeling interactions in factorial experiments.

Of course, MCMC should not be implemented if any direct method of simulation exists. We illustrate this in Section 4.3, where MCMC estimation from current data is followed by ordinary simulation for future prediction. This, and the use of Cholesky decompositions in Sections 4 and 5, provide straightforward examples of the methodology in Appendix 2. However, much more subtle uses of partial conditioning can be envisaged in high-dimensional multi-modal applications, where it can be helpful to engage the posterior distribution at different scales. In image analysis, these scales are associated with coarsenings or sometimes refinements of a pixel lattice; in other contexts, such as pedigree analysis, the different levels may correspond to abstractions of the physical reality.

MCMC has its uses even for models in which standard analytical or numerical methods are applicable. Thus, it can provide a (computationally intensive) check on the accuracy of other methods; its introduction may be necessary in carrying out sensitivity analysis; and it can be used to calculate complicated functionals of the posterior distribution, to which other methods may not relate (for example, the probability that a particular treatment is best, as in Section 5, or the construction of simultaneous credible regions, as in Sections 4.2 and 6.3).

Indeed, one of the most appealing features of MCMC is the ease with which estimates of probabilities and associated quantities, such as credible intervals, are obtained directly from the corresponding empirical distributions, rather than via moment-based approximations. In this sense, we see MCMC as "putting the probability back into statistics" and have quoted posterior means and standard deviations because it is usual to do so, rather than out of conviction. Furthermore, functionals of the posterior distribution whose estimators are sensitive to small changes in the MCMC sample need to be handled with great care; and this can include moments as well as probabilities of rare events. We have an example in which, admittedly for an untransformed scale parameter, the sample means in two long runs differed by a factor of two and yet the medians were almost identical.

Arguably, the most important aspect of MCMC in Bayesian inference is its flexibility, as referred to by Smith (1992): "... for many of us one of the most exciting consequences of the combination of Markov chain Monte Carlo methodology and ever-increasing computer power should be a 'model liberation movement'!". Accordingly, one can invoke models that are considered most appropriate to the data, often involving non-standard likelihoods and non-conjugate priors. We hope that our examples in Sections 5 and 6 illustrate this point and that, where readers find our model formulations unappealing, they will at least agree they could easily substitute their own, without creating computational difficulties. MCMC also deals rigorously with missing values, in properly representing the additional variability involved in the model. Whereas sensitivity to departures from a basic formulation may be assessed via importance sampling or by re-running with a suitably modified prior and/or likelihood, an alternative is to model uncertainty explicitly by adding appropriate additional layers to the existing hierarchy, as illustrated in Section 5. Of course, the corresponding MCMC algorithm must be designed so as to move freely around those parts of the model space that have substantial support in the posterior distribution.

In the methodological sections of the paper we set out the basic ingredients for building MCMC algorithms. We challenged the assumption that the Gibbs sampler should be the automatic choice, even where convenient to code, and indicated where simultaneous updating of several variables is convenient and desirable. We focused on the construction of general frameworks which embrace a family of algorithms: for example, Hastings, Metropolis and Gibbs samplers in Section 2, and the ARMS procedure in Appendix 1. Above all, we hope we have shown the opportunities for flexibly combining samplers of different types to deal with different variables, and for using other hybrid samplers.

We have avoided making any firm recommendations about how to choose between MCMC algorithms. Though methodological development and practical experimentation will provide useful guidelines, as will probabilistic research into Markov chain simulation, the eventual choice will often depend on hardware and software environment, and whether the context is routine production on many similar data-sets or is a one-off analysis.

Finally, what of future methodological developments in MCMC? One of the most promising directions is in the construction of new algorithms based on auxiliary variables and processes. These usually concern severely multi-modal distributions, for which the intention is to speed up the very slow movement of standard samplers across modes. Here we provide a partial update on the general review in Besag and Green (1993, Section 5).

As discussed there, the most successful auxiliary variables method has been the Swendsen and Wang (1987) algorithm designed specifically for the c-color Potts model (and hence, when $c = 2$, the Ising model) in the absence of an external magnetic field. If such a field is present, the distribution may still be severely multi-modal but the performance of the algorithm can deteriorate dramatically (e.g. Marinari and Parisi, 1992) because the clusters it forms are too large. One remedy is to replace the total decoupling of Swendsen-Wang by *partial decoupling*, based on the external field; see Higdon (1993) for an example with the Ising model. This method may also be relevant for some classification problems in spatial statistics and image analysis, where Potts models are sometimes used as exchangeable prior distributions, with each color identifying a particular class. The effect of the observed data is then equivalent to that of applying an inhomogeneous external magnetic field.

A development with wider implications has been the introduction of *simulated tempering* into the literature on auxiliary processes (Marinari and Parisi, 1992; Geyer and Thompson, 1994). Recall that the original idea of auxiliary processes (Geyer, 1991a) was also to speed up the mixing of an awkward multi-modal sampler, by now devising an ordered sequence of chains, which, at one extreme, has the target distribution as its limit and, at the other, a rapidly mixing chain. The chains are run in parallel, with intermittent proposals to swap states of adjacent chains and with acceptance probabilities that ensure each maintains its own limiting distribution. The occurrence of swaps can substantially increase mixing.

In simulated tempering, a corresponding hierarchy of chains exists but only one is running at any particular time. Moves from the current chain to the adjacent chain(s) are proposed periodically, again such that the individual limit distributions are maintained. Thus, at one extreme, information can be collected on the target chain. Now suppose that, at the other, we can use an exact method of simulation, in which case visits to this chain form *regeneration points* for the overall chain and the benefits of regenerative simulation can be exploited (Ripley, 1987, Section 6.4; Mykland, Tierney and Yu, 1993). One of these is that, if the MCMC is initiated at the easy extreme and run for a fixed number of regenerations, then sample means calculated from the target chain must have exactly the correct expectation or else long-run biases would accumulate; this leads into the use of standard ratio estimators (since a normalizing constant must be estimated) and central limit theorems based on independent observations.

Although the above description does not explain how the hierarchy should be selected, a parametric family may be suggested by the formulation itself. Thus, in the above classification problem, with an exchangeable Potts prior, it is necessary to specify the value β_0 of an interaction parameter β, where $\beta = 0$ corresponds to independence. If multi-modality is a problem, an alternative to partial decoupling is to use simulated tempering with β taking a finite set of values in $[0, \beta_0]$. Unfortunately, this introduces a new complication, because the normalizing constants in the *prior* distributions are unknown as a function of β! The simplest remedy is to pre-compute the constants (up to scale) by running a simulation for each prior and then combining the results via Geyer's reverse logistic regression or otherwise. This will be particularly appropriate if the same constants are to be used in many future tasks; for further details, see Marinari and Parisi (1992), Geyer and Thompson (1994) and Higdon (1994). Alternatively, several on-line implementations are currently being developed, using stochastic approximation or Fisher scoring, for example.

The computational effort required for simulated tempering may not seem worthwhile and, in Higdon's example, partial decoupling even proves to be more effective. Nevertheless, it has an interesting consequence that one can use exactly the same approach to implement a fully Bayesian procedure when β is a parameter with a corresponding finite-support hyperprior. For examples, see Higdon (1994) and, in a different context, where the normalizing constants were pre-computed without MCMC, Besag and Higdon (1993, Section 4). The same reasoning could have been applied in Section 6, treating δ as a parameter rather than a fixed constant, had this been thought necessary. We expect this to be a fruitful area for future research.

8 Acknowledgements

We particularly thank David Clayton, Arnoldo Frigessi, Charles Kooperberg, Allan Seheult, Richard Tweedie and Iain Weir for stimulating discussions. The data in Section 5 were kindly supplied by Tony Hunter of the Scottish Agricultural Statistics Service, and those in Section 6 by Chris Gibson of the Regional Medical Physics Department, Dryburn Hospital, Durham, U.K. The work was supported in part by the U.K. SERC Complex Stochastic Systems Initiative, first at the University of Durham (PJG and JEB), and subsequently at the University of Bristol (PJG); by NSF Grants DMS-91-14402 and DMS-92-16158 (JEB and DMH); and by Colorado State University (KLM).

References

Abramowitz, M. and Stegun, I. (1970). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,* U.S. Government Printing Office, Washington.

Aitchison, J. and Shen, S. M. (1980). "Logistic-normal distributions: some properties and uses." *Biometrika,* **67,** 261-272.

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

Amit, Y., Grenander, U. and Piccioni, M. (1991). "Structural image restoration through deformable templates." *J. Amer. Statist. Assoc.,* **86,** 376-387.

Aykroyd, R. G. and Green, P. J. (1991). "Global and local priors, and the location of lesions using gamma camera imagery." *Phil. Trans. R. Soc. Lond. A,* **337,** 323-342.

Baddeley, A. J. and van Lieshout M. N. M. (1993). "Stochastic geometry models in high-level vision." In *Statistics and Images,* volume 1, (K. V. Mardia and G. K. Kanji, eds.) 231-256. Carfax, Abingdon.

Baird, D. and Mead, R. (1991). "The empirical efficiency and validity of two neighbour models." *Biometrics,* **47,** 1473-1487.

Barone, P. and Frigessi, A. (1989). "Improving stochastic relaxation for Gaussian random fields." *Probab. Engng. Inform. Sci.,* **4,** 369-389.

Barone, P., Frigessi, A. and Piccioni, M., eds. (1992). *Stochastic models, Statistical Methods and Algorithms in Image Analysis.* Springer-Verlag, Berlin.

Bartlett, M. S. (1938), "The approximate recovery of information from field experiments with large blocks," *J. Agric. Sci.,* **28,** 418-427.

Bartlett, M. S. (1978), "Nearest neighbour models in the analysis of field experiments (with discussion)." *J. R. Statist. Soc. B,* **40,** 147-174.

Berzuini, C., Clayton, D. and Bernardinelli, L. (1993). "Bayesian inference on the Lexis diagram." *Bull. Int. Statist. Inst.,* **55,** no.1, 149-164.

Besag, J. E. (1974). "Spatial interaction and the statistical analysis of lattice systems" (with Discussion). *J. R. Statist. Soc. B,* **36,** 192-236.

Besag, J. E. (1986). "On the statistical analysis of dirty pictures" (with Discussion). *J. R. Statist. Soc. B,* **48,** 259-302.

Besag, J. E. (1989). "Towards Bayesian image analysis." *Appl. Statist.,* **16,** 395-407.

Besag, J. E. (1992). Contribution: "Constrained Monte Carlo maximum likelihood for dependent data" by C. J. Geyer and E. A. Thompson. *J. Roy. Statist. Soc. Ser. B,* **54,** 657-700.

Besag, J. E. and Clifford, P. (1989). "Generalized Monte Carlo significance tests." *Biometrika,* **76,** 633-642.

Besag, J. E. and Clifford, P. (1991). "Sequential Monte Carlo p-values." *Biometrika,* **78,** 301-304.

Besag, J. E. and Green, P. J. (1993). "Spatial statistics and Bayesian computation" (with Discussion). *J. R. Statist. Soc. B,* **55,** 25-37.

Besag, J. E. and Higdon, D. M. (1993), "Bayesian inference for agricultural field experiments." *Bull. Int. Statist. Inst.,* **55,** no.1, 121 - 136.

Besag, J. E. and R. A. Kempton (1984), "Spatial methods in the analysis of agricultural field trials," In *Proc. 12th Int. Biometrics Conf., Tokyo,* 80-88.

Besag, J. E. and Kempton, R.A. (1986). "Statistical analysis of field experiments." *Biometrics,* **78,** 301-304.

Besag, J. E. and Kooperberg, C. (1993). "On conditional and intrinsic autoregressions."

Besag, J. and A. H. Seheult (1989), Contribution: "Leave-k-out diagnostics for time series" by A. G. Bruce and R. D. Martin, *J. R. Statist. Soc. B,* **51,**405-406.

Besag, J. E. and York, J. C. (1989). "Bayesian restoration of images." In *Analysis of Statistical Information* (T. Matsunawa, ed.) 491-507. Institute of Statistical Mathematics, Tokyo.

Besag, J. E., York, J. C. and Mollié, A. (1991). "Bayesian image restoration, with two applications in spatial statistics" (with Discussion). *Ann. Inst. Statist. Math.,* **43,** 1-59.

Besag, J. E., Higdon, D., and Mengersen K. (1994). "Meta-analysis via Markov chain Monte Carlo: combining information in random-effects logistic regression." to appear.

Binder, K. (1988). *Monte Carlo Simulation in Statistical Physics: an Introduction.* Springer-Verlag, Berlin.

Breslow, N. (1984). "Extra-Poisson variation in log-linear models." *Appl. Statist.,* **33,** 38-44.

Bookstein, F. L. (1989). "Principal warps: thin-plate splines and the decomposition of deformations." *IEEE Trans. Pattern Anal. Machine Intell.,* **11,** 567-585.

Bookstein, F. L. (1991). "Thin-plate splines and the atlas problem for biomedical images." In *Information Processing in Medical Imaging,* (A. C. F. Colchester and D. Hawkes, eds.) 326-342. Springer-Verlag, Berlin.

Casella and George (1992). "Explaining the Gibbs sampler." *American Statistician,* **46,** 167-174.

Clayton, D. G. and Bernardinelli, L. (1992). "Bayesian methods for mapping disease risk." In *Small Area Studies in Geographical and Environmental Epidemiology,* (J. Cuzick and P. Elliott, eds.). Clarendon, Oxford.

Clifford, P. and Middleton, R. D. (1989). "Reconstruction of polygonal images." *Appl. Statist.,* **16,** 409-422.

Cohen, F. S., Fan, Z. and Patel, M. A. (1991). "Classification of rotated and scaled textured images using Gaussian Markov random field models." *IEEE Trans. Pattern Anal. Machine Intell.,* **13,** 192-202.

Cox, D. R. and Wermuth, N. (1993). "Linear dependencies represented by chain graphs." *Statistical Science,* **8,** 204-218.

Creutz, M. (1979). "Confinement and the critical dimensionality of space-time." *Phys. Rev. Lett.,* **43,** 553-556.

Cullis, B. R. and Gleeson, A. C. (1991). "Spatial analysis of field experiments: an extension to two dimensions." *Biometrics,* **47,** 1449-1460.

Cullis, B. R., McGilchrist, C. A. and Gleeson, A. C. (1991). " Error model diagnostics in the general linear model relevant to the analysis of repeated measurements and field experiments." *J. R. Statist. Soc. B,* **53,** 409-416.

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

Diaconis, P. and Saloff-Coste, L. (1993). "Comparison theorems for reversible Markov chains." *Ann. Appl. Probab.,* **3,** 696-730.

Diaconis, P. and Stroock, D. (1991). "Geometric bounds for eigenvalues of Markov chains." *Ann. Appl. Probab.,* **1,** 36-61.

Fisher, R. A. (1928). *Statistical Methods for Research Workers,* 2nd edition. Oliver and Boyd, Edinburgh.

Fishman, G. S. (1991). "Choosing warm-up interval and sample size when generating Monte Carlo data from a Markov chain." Technical Report UNC/OR/TR 91–11.

Fishman, G. S. (1992a). "Markov chain sampling and the product estimator." Technical Report UNC/OR/TR–92/1.

Fishman, G. S. (1992b). "Choosing sample path length t and number of sample paths n when starting in steady state." Technical Report UNC/OR/TR–92/14.

Fosdick, L. D. (1963). *Monte Carlo Computations on the Ising Model."* Academic Press, New York.

Frigessi, A., di Stefano, P., Hwang, C-R. and Sheu, S-J. (1993). "Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-site updating dynamics." *J. R. Statist. Soc. B,* **55,** 205-220.

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). "Illustration of Bayesian inference in normal data models using Gibbs sampling." *J. Amer. Statist. Assoc.,* **85,** 972-985.

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-based approaches to calculating marginal densities." *J. Amer. Statist. Assoc.,* **85,** 398-409.

Geman, D. (1991). "Random fields and inverse problems in imaging." *Lect. Notes Math.,* **1427,** Springer-Verlag, New York.

Geman, D. and Geman, S. (1986). "Bayesian image analysis." In NATO ASI Series, vol F20, *Disordered Systems and Biological Organization* (E. Bienenstock *et al.,* eds.). Springer-Verlag, Berlin.

Geman, D., Geman, S., Graffigne, C. and Dong, P. (1990). "Boundary detection by constrained optimization." *IEEE Trans. Pattern Anal. Mach. Intell.,* **12,** 609-628.

Geman, D. and Reynolds, (1992). "Constrained restoration and the recovery of discontinuities." *IEEE Trans. Pattern Anal. Machine Intell.,* **14,** 367-383.

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." *IEEE Trans. Pattern Anal. Machine Intell.,* **6,** 721-741.

Geman, S., Manbeck, K. M. and McClure, D. E. (1993). "A comprehensive statistical model for single-photon emission tomography." In *Markov Random Fields* (R. Chellappa and A. Jain, eds.) 93-130, Academic Press, Boston.

Geman, S. and McClure, D. E. (1985). "Bayesian image analysis: an application to single photon emission tomography." In *1985 Proceedings of the Statistical Computing Section, American Statistical Association,* 12-18.

Geman, S. and McClure, D. E. (1987). "Statistical methods for tomographic image reconstruction." *Bull. Int. Statist. Inst.,* **52,** no.4, 5-21.

Geman, S., McClure, D. E., and Geman, D. (1992). "A nonlinear filter for film restoration and other problems in image processing." *CVGIP: Graphical Models and Image Proces.,* **54,** 281-289.

Geyer, C. J. (1991a). "Markov chain Monte Carlo maximum likelihood." In *Computer Science and Statistics: Proc. 23rd Symp. Interface.* (E.M. Keramidas, ed.), 156-163.

Geyer, C. J. (1991b). "Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo." Technical Report No. 568, School of Statistics, University of Minnesota.

Geyer, C. J. (1992). "Practical Markov chain Monte Carlo" (with Discussion). *Statistical Science,* **7,** 473-511.

Geyer, C. J. (1994). "On the convergence of Monte Carlo maximum likelihood calculations." *J. R. Statist. Soc. B,* **56,** 261-274.

Geyer, C. J. and Thompson, E. A. (1992). "Constrained Monte Carlo maximum likelihood for dependent data." (with Discussion). *J. R. Statist. Soc. B,* **54,** 657-699.

Geyer, C. J. and Thompson, E. A. (1994). "Annealing Markov chain Monte Carlo with applications to ancestral inference." *J. Amer. Statist. Assoc.,* to appear.

Gidas, B. (1989). "A renormalization group approach to image processing problems." *IEEE Trans. Pattern Anal. Mach. Intelligence,* **11,** 164-180.

Gidas, B. (1992). "Metropolis-type Monte Carlo simulation algorithms and simulated annealing." In *Trends in Contemporary Probability* (Doyle and E. J. Snell, eds.).

Gilks, W. R. (1992). "Derivative-free adaptive rejection sampling for Gibbs sampling." In *Bayesian Statistics* 4 (Bernardo, J., Berger, J., Dawid, A. P. and Smith, A. F. M., eds.). Oxford University Press.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1994). "Adaptive rejection Metropolis sampling." to appear.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D. and Kirby, A. J. (1993). "Modeling complexity: Applications of Gibbs sampling in medicine" (with Discussion). *J. R. Statist. Soc. B,* **55,** to 39-52.

Gilks, W. R. and Wild, P. (1992). "Adaptive rejection sampling for Gibbs sampling." *Appl. Statist.,* **41,** 337-348.

Glad, I. K. and Sebastiani, G. (1994). "A Bayesian approach to synthetic magnetic resonance imaging." Dept. of Mathematical Sciences, Norwegian Institute of Technology, Trondheim, Norway.

Gleeson and Cullis (1987). "Residual maximum likelihood (REML) estimation of a neighbour model for field experiments." *Biometrics,* **43,** 277-288.

Green, P. J. (1985). "Linear models for field trials, smoothing and cross-validation." *Biometrika* **72,** 527-537.

Green, P. J. (1990). "Bayesian reconstructions from emission tomography data using a modified EM algorithm" *IEEE Trans. Med. Imaging,* **9,** 94-93.

Green, P. J. and Han, X-L. (1992). "Metropolis methods, Gaussian proposals, and antithetic variables." In *Stochastic models, Statistical Methods and Algorithms in Image Analysis, Lect. Notes Statist.,* **74,** 142-164, Springer-Verlag, Berlin.

Green, P. J., Jennison, C. and Seheult, A. H. (1985). "Analysis of field experiments by least squares smoothing." *J. R. Statist. Soc. B,* **47,** 299-315.

Grenander, U. (1983). "Tutorial in pattern theory." Report, Division of Applied Mathematics, Brown University, Providence.

Grenander, U., Chow, Y. and Keenan, D. M. (1991). "Hands: a pattern theoretic study of biological shapes." *Res. Notes Neural Comput.,* **2,** Springer-Verlag, New York.

Grenander, U. and Keenan, D. M. (1989). "Towards automated image understanding." *Appl. Statist.,* **16,** 207-221.

Grenander, U. and Manbeck, K. M. (1992). "Abnormality detection in potatoes by shape and color." Division of Applied Mathematics, Brown University, Providence.

Grenander, U. and Miller, M. (1994). "Representations of knowledge in complex systems (with Discussion)." *J. R. Statist. Soc. B,* **56** to appear.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods,* Wiley, New York.

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika,* **57,** 97-109.

Heikkinen, J. and Högmander, H. (1992). "Bayesian image restoration from sparse data, and an application in biogeography." Dept. of Statistics, University of Jyväskylä, 2.

Higdon, D. (1993). Contribution: Meeting on MCMC methods. *J. R. Statist. Soc. B,* **55,** 78.

Higdon, D. (1994). "Incorporating uncertainty in Bayesian inference for spatial data." Ph.D. dissertation, Dept of Statistics, University of Washington.

Högmander, H. and Møller, J. (1993). "Classification of atlas maps using methods of statistical image analysis." Dept. of Statistics, University of Jyväskylä, 8.

Holford, T. R. (1983). "The estimation of age, period and cohort effects for vital rates." *Biometrics,* **39,** 311-324.

Johnson. V. E. (1994a). "Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths." Report 94-07, ISDS, Duke University.

Johnson, V. E. (1994b). "A model for segmentation and analysis of noisy images." *J. Amer. Statist. Assoc.,* **89,** 230-241.

Jubb, M. and Jennison, C. (1991). "Aggregation and refinement in binary image restoration." In *Spatial Statistics and Imaging* (Possolo, A., ed.) 150-162, IMS Lecture Notes, Haywood, California.

Kempthorne, O. (1952), *The Design and Analysis of Experiments,* Wiley, New York.

Kempton, R. A., Seraphin, J. C. and Sword, A. M. (1993), "Statistical analysis of two dimensional variation in variety yield trials." to appear.

Kipnis, C. and Varadhan, S. R. S. (1986). "Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions." *Comm. Math. Phys.,* **104,** 1-19.

Kooperberg(1993). Contribution: Meeting on MCMC methods. *J. R. Statist. Soc. B,* **55,** 79-81.

Künsch, H. R. (1987). "Intrinsic autoregressions and related models on the two-dimensional lattice." *Biometrika,* **74,** 517-524.

Künsch, H. R. (1994). "Robust priors for smoothing and image restoration." *Ann. Inst. Statist. Math.,* **46,** 1-19.

Larssen, S.A. (1980). "Gamma camera emission tomography." *Acta Radiologica,* Supplementum 363, Stockholm.

Liu, J., Wong, W. H. and Kong, A. (1991). "Correlation structure and convergence rate of the Gibbs sampler: Applications to the comparisons of estimators and augmentation schemes." Technical Report 299, Dept. of Statistics, University of Chicago.

Mardia, K. V., Hainsworth, T. J. and Haddon, J. (1992). "Deformable templates in image sequences." *IAPR Conference, The Hague.*

Mardia, K. V. and Kanji, G. K., eds. (1993). *Statistics and Images,* volume 1, Carfax, Abingdon.

Marinari, E. and Parisi, G. (1992). "Simulated tempering: a new Monte Carlo scheme." *Europhys. Lett.,* **19** 451-458.

Marroquin, J., Mitter, S. and Poggio, T. (1987). "Probabilistic solution of ill-posed problems in computational vision." *J. Amer. Statist. Assoc.,* **82,** 76-89.

Martin, R. (1990). "The use of time-series models and methods in the analysis of agricultural field trials." *Comm. Statist.,* **19,** 55-81.

Mengersen, K. L. and Tweedie, R. L. (1994). "Rates of convergence of the Hastings and Metropolis algorithms." Dept. of Statistics, Colorado State University.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953). "Equations of state calculations by fast computing machines." *J. Chemical Physics,* **21,** 1087-1091.

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

Miller, M. I., Roysam, B., Smith, K. and O'Sullivan, J. A. (1991). "Representing and computing regular languages on massively parallel networks." *IEEE Trans. Neural Networks,* **2,** 56-72.

Mykland, P., Tierney, L. and Yu, B. (1993). "Regeneration in Markov chain samplers." *J. Amer. Statist. Assoc.,* to appear.

Papadakis, J. S. (1984), "Advances in the analysis of field experiments," ΠΡΑΚΤΙΚΑ ΤΗΣ ΑΚΑΔΗΜΙΑΣ ΑθΗΝΩW *(Proceedings of the Academy of Athens),* **59,** 326-342.

Patterson, H. D. and Williams, E. R. (1976). "A new class of resolvable incomplete block designs." *Biometrika,* **63,** 83-92.

Penttinen, A. (1984). "Modelling interaction in spatial point patterns: Parameter estimation by the maximum likelihood method." *Jyväskylä Studies in Computer Science, Economics and Statistics,* 7.

Peskun, P. H. (1973). "Optimum Monte-Carlo sampling using Markov chains." *Biometrika* **60,** 607-612.

Phillips D. B. and Smith A. F. M. (1993). "Dynamic image analysis using Bayesian shape and texture models." In *Statistics and Images,* volume 1, (K. V. Mardia and G. K. Kanji, eds.) 299-322. Carfax, Abingdon.

Possolo, A., ed. (1991). *Spatial Statistics and Imaging,* IMS Lecture Notes, Haywood, California.

Ripley, B. D. (1979). "Algorithm AS 137: Simulating spatial patterns: dependent samples from a multivariate density." *Appl. Statist.,* **28,** 109-112.

Ripley, B. D. (1987). *Stochastic Simulation,* Wiley, New York.

Ripley, B. D. and Sutherland, A. I. (1990. "Finding spiral structures in images of galaxies." *Phil. Trans. R. Soc. Lond. A,* **332,** 477-485.

Roberts, G. O. and Polson, N. G. (1994). "On the geometric convergence of the Gibbs sampler." a *J. R. Statist. Soc. B,* **56,** 377-384.

Roberts, G. O. and Tweedie, R. L. (1994). "Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms." to appear.

Rosenthal, J. S. (1992). "Rates of convergence for the Gibbs sampler and other Markov chains." Dept. of Mathematics, Harvard University.

Sheehan, N. and Thomas, A. W. (1992). "On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme." *Biometrics,* **49,** 163-175.

Shepp, L. A. and Vardi, Y. (1982). "Maximum likelihood reconstruction in positron emission tomography." *IEEE Trans. Medical Imaging,* **1,** 113-122.

Smith, A. F. M. (1978). Contribution: "Nearest neighbour models in the analysis of field experiments" by M. S. Bartlett. *J. R. Statist. Soc. B,* **40,** 167-168.

Smith, A. F. M. (1992). Contribution: "Constrained Monte Carlo maximum likelihood for dependent data" by C. J. Geyer and E. A. Thompson. *J. Roy. Statist. Soc. Ser. B,* **54,** 657-700.

Smith, A. F. M. and Roberts, G. O. (1993). "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods" (with Discussion). *J. R. Statist. Soc. B,* **55,** 3-23.

Sokal, A.D. (1989). "Monte Carlo methods in statistical mechanics: foundations and new algorithms." *Cours de Troisiéme Cycle de la Physique en Suisse Romande.* Lausanne.

Srivastava, A., Miller, M. I. and Grenander, U. (1991). "Jump diffusion processes for object tracking and direction finding." In *Proc. 29th Annual Allerton Conference on Communication, Control and Computing,* 563-570. University of Illinois.

Swendsen, R. H. and Wang, J. S. (1987). "Nonuniversal critical dynamics in Monte Carlo simulations." *Phys. Rev. Lett.,* **58,** 86-88.

Taplin, R. and Raftery, A. E. (1994). "Analysis of agricultural field trials in the presence of outliers and fertility jumps." *Biometrics,* to appear

Tierney, L. (1994). "Markov chains for exploring posterior distributions" (with Discussion), *Ann. Statist.,* **22,** to appear.

Tierney, L. and Kadane, J. B. (1986). "Accurate approximations for posterior moments and marginal densities." *J. Am. Statist. Assoc.,* **81,** 82-86.

Tierney, L., Kass, R. E. and Kadane, J. B. (1989). "Fully exponential Laplace approximations to expectations and variances of nonpositive functions." *J. Amer. Statist. Assoc.,* **84,** 710-716.

Weir, I. S. and Green, P. J. (1994). "Modelling data from single photon emission computed tomography." University of Bristol Mathematics Research Report S-93-01, to appear in *Statistics and Images*, volume 2, (K.V. Mardia, ed.) Carfax, Abingdon.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Wiley, Chichester.

Wilkinson, G. N. (1984). "Nearest neighbour methodology for design and analysis of field experiments." *Proc. 12th Int. Biometrics Conf., Tokyo,* 64-79.

Wilkinson, G. N., Eckert, S. R., Hancock, T. W. and Mayo, O. (1983). "Nearest neighbour (NN) analysis of field experiments" (with Discussion). *J. R. Statist. Soc. B,* **45,** 151-211.

Williams, D. A. (1982). "Extra-binomial variation in logistic linear models." *Appl. Statist.,* **31,** 144-148.

Williams, E. R. (1986). "A neighbour model for field experiments", *Biometrika ,* **73,** 279-87.

Wright, W. A. (1989). "A Markov random field approach to data fusion and colour segmentation." *Im. Vis Comput.,* **7,** 144-150.

Zeger, S. L. and Karim, M. R. (1991). "Generalized linear models with random effects; a Gibbs sampling approach." *J. Amer. Statist. Assoc.,* **86** 79-86.

Zimmerman, D. L. and Harville, D. A. (1991). "A random field approach to the analysis of field-plot experiments and other spatial experiments" *Biometrics,* **47** 223-239.

Appendix 1: Random proposal distributions

We referred in Section 2.3.1 to the ARS algorithms of Gilks and Wild (1992) and Gilks (1992), proposed for use in Gibbs sampling, which rely on the relevant full conditionals being log-concave. This condition is relaxed in the Adaptive Rejection Metropolis Sampling (ARMS) method of Gilks, Best and Tan (1994) by the introduction of a two-stage procedure involving a Hastings step. All three algorithms are open-ended, in the sense that an indefinite number of random variates are generated until a particular condition is satisfied. This stimulated us to look for a broader framework embracing ARMS, and which also allows the use of curtailment rules in both ARS and ARMS.

We restrict attention to a single step of a sampler which, like most of those in Section 2, updates the variables x_T while holding x_{-T} fixed. Suppose there is available a class of appropriate kernels $\{P_T^\alpha(x \to B)\}$ indexed by an abstract parameter α.

Updating x_T proceeds by first drawing α at random from a distribution $\mu(\alpha; x_{-T})$ parameterized by x_{-T}, and then using $P_T^\alpha(x \to B)$. If all $\{P_T^\alpha\}$ have detailed balance, so does the mixture:

$$\begin{aligned}
& \int_B \pi(x) \int P_T^\alpha(x \to C) d\mu(\alpha; x_{-T}) d\nu(x) \\
= & \int_{B_{-T} \cap C_{-T}} \pi(x_{-T}) \int \int_{B_T} \pi(x_T | x_{-T}) P_T^\alpha(x \to C) d\nu(x_T) d\mu(\alpha; x_{-T}) d\nu(x_{-T}) \\
= & \int_{B_{-T} \cap C_{-T}} \pi(x_{-T}) \int \int_{C_T} \pi(x_T | x_{-T}) P_T^\alpha(x \to B) d\nu(x_T) d\mu(\alpha; x_{-T}) d\nu(x_{-T}) \\
= & \int_C \pi(x) \int P_T^\alpha(x \to B) d\mu(\alpha; x_{-T}) d\nu(x),
\end{aligned}$$

with a similar result assuming only global balance.

For a concrete but generic example, $P_T^\alpha(x \to B)$ might be a Hastings sampler using a proposal distribution specified by α; what the equality above shows is that it is legitimate to use a *random* proposal distribution, even one depending on x_{-T}, and compute the acceptance probability ignoring this random mechanism.

The ARMS algorithm is an instance of this. Repeatedly, piecewise exponential approximations to $\pi(x_T|x_{-T})$ are constructed as in ARS, until a value $x_T{}'$ generated from the latest one passes a rejection test. This value is then used in (2.9) as if the proposal density R_T was the pointwise minimum of the final approximation and the full conditional.

It is evident from the argument above, firstly that the algorithm is valid for any sequence of approximating densities, and secondly that the procedure can be curtailed at any stage independent of the value of the final $x_T{}'$. Whether the full conditional is log-concave, so that ARS is available, or not, such a strategy can be used to speed up the algorithm; it remains to be seen if this is effective in practice.

Appendix 2: MCMC based on partial conditioning

All of the MCMC algorithms discussed up to this point have been based on the full conditionals $\pi(x_T|x_{-T})$ for some collection of sets T. We now explore methods based on only partial conditioning. Thus, let $T \subset S \subset \mathcal{N}$, where S is fixed and T is chosen from a probability distribution $\{p_T\}$, as in the random scan algorithms of Section 2.4.1. Write $U = S\backslash T$. We consider transition kernels P of the form

$$P(x_S \to B_S) = \sum_T p_T I[x_U \in B_U] P_T(x_T \to B_T; x_U), \tag{A2.1}$$

where P_T leaves x_U unaltered and is independent of x_{-S}; indeed, we ignore x_{-S} without loss. Now suppose P_T is time reversible with respect to $\pi(x_T|x_U)$; that is,

$$\int_{B_T} \pi(x_T|x_U) P_T(x_T \to C_T; x_U)\, d\nu(x_T) = \int_{C_T} \pi(x_T|x_U) P_T(x_T \to B_T; x_U)\, d\nu(x_T), \tag{A2.2}$$

for all B_T, $C_T \subset \mathcal{X}_T$ and $x_U \in \mathcal{X}_U$. Then we have the following result.

Theorem 1 *Suppose X_R has density $\pi(x_R)$, where $R \supset S$, and X'_S is generated from X_S via (A2.1). If the P_T's satisfy (A2.2), then P is reversible with respect to $\pi(x_S)$, so that, in particular, X'_S also has density $\pi(x_S)$.*

Proof

$$\begin{aligned}
& \int_{B_S} \pi(x_S) P(x_S \to C_S)\, d\nu(x_S) \\
= & \sum_T p_T \int_{B_U \cap C_U} \pi(x_U) \int_{B_T} \pi(x_T|x_U) P_T(x_T \to C_T; x_U)\, d\nu(x_T)\, d\nu(x_U) \\
= & \int_{C_S} \pi(x_S) P(x_S \to B_S)\, d\nu(x_S).
\end{aligned}$$

using (A2.1), (A2.2) and (A2.1) again.

We remark in passing that again the weaker "global" version of this result is also true: if (A2.2) holds, for all T, with $B_S = \mathcal{X}_S$, then $\int \pi(x_S)P(x_S \to C_S)d\nu(x_S) = \pi(C_S)$.

Note that after such a transition only variables X_S indexed by a subset S of R can be guaranteed to be distributed as π; if S is a proper subset, $\pi(x_R)$ can never be retrieved by this mechanism. However, if we use a Gibbs kernel (which in this appendix will be interpreted as meaning that variables are drawn from a conditional distribution determined by the target distribution, even if this is not the full conditional), this shrinking of the domain of equilibrium can be reversed, as the following theorem shows.

Theorem 2 *Suppose X_R has density $\pi(x_R)$ and that $\{p_T\}$ ensures that $U \subset R$. Let X'_S be generated from X_S via (A2.1), in which the P_T's are* Gibbs kernels, *i.e.*

$$P_T(x_T \to B_T; x_U) = \int_{B_T} \pi(x_T | x_U)\, d\nu(x_T),$$

and hence X'_T is conditionally independent of X_T. Then X'_S has density $\pi(x_S)$.

The difference between the kernels is that, in a Gibbs step, it is not required that the updated r.v. already has the correct conditional distribution. In short, under both schemes, each updating X_T while holding x_U fixed, X_R distributed as $\pi(x_R)$ leads to X'_S distributed as $\pi(x_S)$, where $S = T \cup U$. With non-Gibbs kernels, it is necessary that $S \subset R$, but with Gibbs kernels, we only require $U \subset R$. The following theorem summarizes the implications when such kernels are used in sequence.

Theorem 3 *Consider the sequence*

$$(\mathcal{N} = R) \to (S = R') \to \ldots \to (S^{(K-1)} = R^{(K)}) \to (S^{(K)} = R^{(K+1)}) \to \ldots \to (S^{(L)} = \mathcal{N}),$$

where the arrow linking $R^{(K)}$ and $S^{(K)}$ denotes update of $x_{T^{(K)}}$. The sequence supports a valid cycle of an MCMC algorithm if, for each K, either $S^{(K)} \subset S^{(K-1)}$, in which case a Gibbs or non-Gibbs step can be made, or $U^{(K)} \subset S^{(K-1)}$, where $U^{(K)} = S^{(K)} \backslash T^{(K)}$, in which case a Gibbs step must be made.

Example Suppose X has $n = 5$ components and that successive kernels provide

(a) a new X_1 given current X_1, X_2, X_4, X_5,
(b) a new X_2 and X_4 given current X_1, X_2, X_4,
(c) a new X_2 and X_3 given current X_2, X_3, X_4,
(d) a new X_4 given current X_2, X_3, X_4,
(e) a new X_1 and X_5 given current X_1, X_2, X_3, X_4, X_5.

Then steps (c) and (e) must be Gibbs, whereas (a), (b) and (d) need only satisfy Theorem 1.

To obtain full generality, suppose valid sequences $(S, S' \ldots, S^{(L)})$ are generated stochastically. Theorem 3 ensures that, conditional on any sequence and hence marginally, the cycle is valid. Thus, a valid $(T^{(K)}, U^{(K)})$ combination can be chosen probabilistically at each successive step. Note that mid-cycle marginal statements cannot be made and that marginalization only works because $S^{(L)} = \mathcal{N}$ is a fixed set. For an auxiliary variable, multigrid, Swendsen-Wang application in spatial statistics, see Besag and Green (1993, Section 5). The validity of the algorithm, which was not given explicitly in the paper, is an immediate consequence of the general result. There are other much more straightforward examples: for instance in any distribution that can be represented via the sequential factorization (2.1), and in particular whenever a multivariate Gaussian random vector is generated via the Cholesky decomposition of its variance matrix.

Appendix 3: Sensitivity analysis via MCMC

In sensitivity analysis, the effects of plausible changes in the basic formulation are studied. Though important in any brand of statistical inference, this is especially so under the Bayesian paradigm, where modifications both to the prior and to the likelihood need to be considered. MCMC is ideally suited to this previously difficult task. There are two general approaches, both of which were illustrated in Sections 4 and 5. Obviously, the first is to re-run the MCMC for each new model of interest. This can be very time consuming, though one might often use rather shorter runs than for the basic formulation. An alternative, more satisfactory approach is often available through *importance sampling*, as suggested independently by Besag (1992) and by Smith (1992).

Let $\{\pi(x) : x \in \mathcal{X}\}$ denote the basic model and $\{\tilde{\pi}(x) : x \in \tilde{\mathcal{X}}\}$ some other formulation, where both densities are with respect to a single measure ν; as usual, we suppress dependence on data y from the notation. Suppose that we are interested in the expectation of some function g under $\tilde{\pi}$. Then,

$$\begin{aligned}
\mathbf{E}_{\tilde{\pi}} g &= \int_{\tilde{\mathcal{X}}} g(x)\tilde{\pi}(x) d\nu(x) \\
&= \int_{\mathcal{X}} \frac{g(x)\tilde{\pi}(x)}{\pi(x)} \pi(x)\, d\nu(x) \\
&= \mathbf{E}_{\pi} g\tilde{\pi}/\pi, \qquad \text{(A3.1)}
\end{aligned}$$

provided π dominates $\tilde{\pi}$; that is, $\mathcal{X} \supset \tilde{\mathcal{X}}$, implying no zero divisors occur in (A3.1). In practice, we may only know $\pi(x) \propto h(x)$ and $\tilde{\pi}(x) \propto \tilde{h}(x)$, in which case we must replace (A3.1) by

$$\mathbf{E}_{\tilde{\pi}} g \propto \frac{\mathbf{E}_{\pi} g\tilde{h}/h}{\mathbf{E}_{\pi} \tilde{h}/h}.$$

It follows that we can apply the approximation,

$$\mathbf{E}_{\tilde{\pi}} g \approx \sum_{t=1}^{m} a^{(t)} g(x^{(t)}), \tag{A3.2}$$

where $x^{(1)}, \ldots, x^{(m)}$ is the already existing MCMC sample, from π itself, and

$$a^{(t)} = \frac{\tilde{h}(x^{(t)})/h(x^{(t)})}{\sum_{t=1}^{m} \tilde{h}(x^{(t)})/h(x^{(t)})}. \tag{A3.3}$$

Note that this is particularly simple in that the weights $a^{(t)}$ do not depend on g.

Importance sampling will not work well when π and $\tilde{\pi}$ are very different, because too much weight will be placed on just a few $g(x^{(t)})$'s. Nevertheless, this can be monitored by examining the values of the weights, which of course sum to unity, and resorting to MCMC under $\tilde{\pi}$ when a problem arises. In turn, this suggests that one might run several MCMC simulations, covering a range of models, and base inferences for any particular $\tilde{\pi}$ on a corresponding mixture distribution. This procedure is not entirely straightforward in practice, assuming the various densities are known only up to scale. For further details, see Geyer (1991b). The above development is analogous to that given for general MCMC maximum likelihood estimation in the response to discussion in Geyer and Thompson (1992). Finally, note that importance sampling can also be useful for approximating functionals of the *base* density π, when it is much easier to simulate from a similar density π_0; see, for example, Sheehan and Thomas (1992).

Another approach is to examine sensitivity by changing π infinitesimally in the direction of $\tilde{\pi}$. Consider evaluating expectations under the distribution corresponding to

$$h_\epsilon(x) = (h(x))^{1-\epsilon}(\tilde{h}(x))^{\epsilon},$$

for which the weights $a_\epsilon^{(t)}$ to be used in (A3.2) are just those in (A3.3), raised to the power ϵ and then renormalized. For small enough ϵ, these weights are arbitrarily close to $1/m$, and so the instability in MCMC estimates resulting from grossly unequal weights does not arise.