STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:2279–2294

Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models

Lynn E. Eberly and Bradley P. Carlin*,†

Division of Biostatistics, School of Public Health, University of Minnesota, Box 303, Mayo Memorial Building, Minneapolis, Minnesota 55455-0392, U.S.A.

SUMMARY

The marked increase in popularity of Bayesian methods in statistical practice over the last decade owes much to the simultaneous development of Markov chain Monte Carlo (MCMC) methods for the evaluation of requisite posterior distributions. However, along with this increase in computing power has come the temptation to fit models larger than the data can readily support, meaning that often the propriety of the posterior distributions for certain parameters depends on the propriety of the associated prior distributions. An important example arises in spatial modelling, wherein separate random effects for capturing unstructured heterogeneity and spatial clustering are of substantive interest, even though only their sum is well identified by the data. Increasing the informative content of the associated prior distributions offers an obvious remedy, but one that hampers parameter interpretability and may also significantly slow the convergence of the MCMC algorithm. In this paper we investigate the relationship among identifiability, Bayesian learning and MCMC convergence rates for a common class of spatial models, in order to provide guidance for piror selection and algorithm tuning. We are able to elucidate the key issues with relatively simple examples, and also illustrate the varying impacts of covariates, outliers and algorithm starting values on the resulting algorithms and posterior distributions. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

The marked increase in popularity of Bayesian methods in statistical practice over the last decade owes much to the simultaneous development of Markov chain Monte Carlo (MCMC) methods for the evaluation of requisite posterior distributions. This is especially true in hierarchical modelling, where previously foreboding multiple layers of random effects (whose estimation depends on the 'borrowing of strength' across typically independent but similar data components) can now be accommodated fairly easily. However, along with this increase in computing power has come the temptation to fit models larger than the data can readily support, meaning that the propriety of

*Correspondence to: Bradley P. Carlin, Division of Biostatistics, School of Public Health, University of Minnesota, Box 303, Mayo Memorial Building, Minneapolis, Minnesota 55455-0392, U.S.A.
† E-mail: brad@muskie.biostat.umn.edu

Contract/grant sponsor: NIAID; contract/grant numbers: 5-U01-AI42170-07, R01-AI41966
Contract/grant sponsor: NIEHS; contract/grant number: 1-R01-ES07750

the posterior distributions for certain parameters depends on the propriety of the associated prior distributions. In many models it is obvious which parameters are well identified by the data and which are not, thus determining which prior distributions must be carefully specified and which may be left vague. Unfortunately, in complex model settings these identifiability issues may be rather subtle. As such, many authors (for example, Carlin and Louis, [1], p. 188) recommend avoiding such models unless the classification of parameters as 'identifiable' and 'unidentifiable' is well understood.

Models for the analysis of areal data (that is, data which are sums or averages aggregated over a particular set of regional boundaries) are a common example of this problem. Introduced by Clayton and Kaldor [2] in an empirical Bayes context and later expanded to a fully Bayesian setting by Besag, York and Mollié [3], these models express the number of disease events in region i, Y_i, as a Poisson random variable having mean $E_i \exp(\mu_i)$. Here, E_i is an *expected* number of events and μ_i is a log-relative risk of disease, modelled linearly as

$$\mu_i = \mathbf{x}_i'\boldsymbol{\beta} + \theta_i + \phi_i, \quad i = 1, \ldots, I \tag{1}$$

In this equation the $\mathbf{x}_i$ are explanatory spatial covariates, $\boldsymbol{\beta}$ is a vector of fixed effects, and $\boldsymbol{\theta} = \{\theta_i\}$ and $\boldsymbol{\phi} = \{\phi_i\}$ are collections of region-specific random effects capturing regional *heterogeneity* and *clustering*, respectively. Typically these spatial effects are captured by assuming the mixing distributions

$$\theta_i \overset{\text{iid}}{\sim} \text{N}(0, 1/\tau_\text{h}) \quad \text{and} \quad \phi_i \mid \phi_{j \neq i} \sim \text{N}(\mu_{\phi_i}, \sigma^2_{\phi_i}), \quad i = 1, \ldots, I \tag{2}$$

where

$$\mu_{\phi_i} = \frac{\sum_{j \neq i} w_{ij}\phi_j}{\sum_{j \neq i} w_{ij}} \quad \text{and} \quad \sigma^2_{\phi_i} = \frac{1}{\tau_\text{c} \sum_{j \neq i} w_{ij}}$$

and the weights w_{ij} are fixed constants. In practice, one often takes $w_{ij} = 0$ unless areas i and j are adjacent (that is, share a common boundary). If areas i and j are adjacent, we set $w_{ij} = 1$, although other forms of w_{ij} are possible [4, 5]. This distribution for $\boldsymbol{\phi}$ is called a *conditionally autoregressive* specification, which for brevity we typically write in vector notation as $\boldsymbol{\phi} \sim \text{CAR}(\tau_\text{c})$. Note that since this prior is specified conditionally, the parameters are only uniquely determined up to an additive constant. This problem is normally corrected either by insisting that the covariate vector not include an intercept term, or by imposing the sum-to-zero constraint $\sum_{i=1}^{I} \phi_i = 0$. A fully Bayesian model specification is completed by specifying fixed values or prior distributions for each of $\boldsymbol{\beta}, \tau_\text{h}$, and τ_c.

Intuitively, the identifiability issue in equation (1) is obvious: the single data point Y_i cannot possibly provide information about θ_i and ϕ_i individually, but only about their sum, $\eta_i = \theta_i + \phi_i$. This is the usual notion of identifiability, which a Bayesian might refer to as *likelihood identifiability*. From a more formal mathematical point of view, if we consider the reparameterization from $(\boldsymbol{\theta}, \boldsymbol{\phi})$ to $(\boldsymbol{\theta}, \boldsymbol{\eta})$, we have the joint posterior $p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \mathbf{y}) \propto L(\boldsymbol{\eta}; \mathbf{y})\, p(\boldsymbol{\theta})\, p(\boldsymbol{\eta} - \boldsymbol{\theta})$. This means that

$$p(\theta_i \mid \theta_{j \neq i}, \boldsymbol{\eta}, \mathbf{y}) \propto p(\theta_i)\, p(\eta_i - \theta_i \mid \{\eta_j - \theta_j\}_{j \neq i})$$

Since this conditional distribution is free of the data $\mathbf{y}$, we say that θ_i is *Bayesianly unidentified* (Gelfand and Sahu [6]), a condition that can also be shown to hold for the ϕ_i. Note that this does

not mean that the model precludes *Bayesian learning* about θ_i; this would instead require

$$p(\theta_i \mid \mathbf{y}) = p(\theta_i)$$

This is a stronger condition than Bayesian unidentifiability, since it says that the data have no impact on the *marginal* (not just the conditional) posterior for θ_i.

Of course, in some sense identifiability is a non-issue for Bayesian analyses, since given proper prior distributions the corresponding posteriors must be proper as well, hence every parameter can be well estimated. However, Bayesians are often drawn to vague priors, since they typically ensure that the data play the dominant role in determining the posterior. In addition, most standard *reference priors* [7], that is, distributions derived from formal rules, are not only vague but usually improper as well. Posterior impropriety resulting from a (likelihood) unidentified parameter having an improper prior will manifest itself in the MCMC context as convergence failure. Unfortunately, the problem may not be readily apparent, since a slowly converging sampler may produce output that is virtually indistinguishable from one that will *never* converge due to an improper posterior!

Several authors (for example, Besag *et al.* [8]) have pointed out that MCMC samplers running over spaces that are not fully identified are perfectly legitimate, *provided* that their samples are used only to summarize the components of the *proper embedded posterior*, that is, a lower-dimensional parameter vector having a unique integrable posterior distribution (in our case, the η_i). In the context of our spatial model (1), however, the unidentified components θ_i and ϕ_i are actually interesting in their own right, since they capture the impact of missing spatial covariates which vary globally (heterogeneity) and locally (clustering), respectively. For instance, Best *et al.* [9] define the quantity

$$\psi = \frac{\mathrm{SD}(\boldsymbol{\phi})}{\mathrm{SD}(\boldsymbol{\theta}) + \mathrm{SD}(\boldsymbol{\phi})} \tag{3}$$

where $\mathrm{SD}(\cdot)$ is the empirical marginal standard deviation of the random effect vector in question. A posterior for ψ concentrated near 1 suggests most of the excess variation (that is, that not explained by the covariates $\mathbf{x}_i$) is due to spatial clustering, while a posterior concentrated near 0 suggests most of this variation is mere unstructured heterogeneity. This genuine interest in the trade-off between $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ forces these authors into a search for proper yet vague priors for these two components – a task complicated by the fact that the prior for the former is specified marginally, while that for the latter is specified conditionally (see equation (2)). We return to these issues in Section 3.

In the case of the partially identified Gaussian linear model, $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with X less than full column rank, Gelfand and Sahu [6] provide a surprising MCMC convergence result. They show that under a flat prior on $\boldsymbol{\beta}$, the Gibbs sampler for the full parameter vector $\boldsymbol{\beta}$ is divergent, but the samples from the identified subset of parameters (say, $\boldsymbol{\delta} = X_1\boldsymbol{\beta}$) form an *exact* sample from their (unique) posterior density $p(\boldsymbol{\delta} \mid \mathbf{y})$. That is, such a sampler will produce identically distributed draws from the true posterior for $\boldsymbol{\delta}$, and convergence is immediate. The authors then consider a logistic growth curve model (an example clearly outside the Gaussian family), and observe steadily decreasing lag 1 sample autocorrelations in the Gibbs samples for $\boldsymbol{\delta}$ as the prior variance for $\boldsymbol{\beta}$ increases. This leads them to postulate that such monotonic and continuous improvement in the convergence rate for estimable functions will occur quite generally as the prior on $\boldsymbol{\beta}$ becomes more and more vague. In subsequent work, Gelfand, Carlin and Trevisani [10] in fact show that, for a broad class of Gaussian models with or without covariates, $\mathrm{corr}(\boldsymbol{\delta}^{(t)}, \boldsymbol{\delta}^{(t+1)})$ approaches 0 as the prior variance for $\boldsymbol{\beta}$ goes to infinity once the chain has converged.

Unfortunately, in the case of our spatial models (1) and (2), one cannot choose such an 'arbitrarily vague' prior, since in this case the parameters would not be identified, and thus the entire joint posterior (and hence the marginal posteriors for $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ and ψ) would become improper. As such, guidance is needed on how to choose appropriate priors that will still produce acceptable MCMC convergence behaviour. We investigate several such specifications, and show that a variety of factors (such as the value of the data point Y_i and the starting point of the MCMC chain for θ_i and ϕ_i) can have a significant effect on the observed convergence rate and estimated posterior distributions.

The remainder of our paper is organized as follows. Section 2 investigates a very simple model that uses Gaussian distributions but is analogous in many ways to our spatial model (1). In this case, we demonstrate significant non-monotonicity in the convergence rate when the starting points for the chain are sufficiently far from the equilibrium distribution. Section 3 considers the more advanced spatial setting, in the context of the oft-analysed Scottish lip cancer data [2]. Here convergence behaviour is even more difficult to predict; however, we also explore the Bayesian learning within the system through the distribution of ψ, as defined in equation (3). Finally, Section 4 discusses our findings and suggests directions for future research.

2. MOTIVATING EXAMPLE

To begin with, we consider perhaps the simplest example of a Bayesianly unidentified model:

$$y \mid \theta, \phi \sim \mathrm{N}(\theta + \phi, 1) \tag{4}$$

where we place $\mathrm{N}(0, \sigma_1^2)$ and $\mathrm{N}(0, \sigma_2^2)$ priors on θ and ϕ, respectively. Obviously this model would rarely if ever be considered in practice, but we select it since it does fit into the Gaussian framework of Gelfand and Sahu [6] and Gelfand, Carlin and Trevisani [10], and since the unidentifiability in its mean structure is of the same form as that in our spatial model of interest (1). A previous analysis of this model by Carlin and Louis (Reference [1], p. 203), shows that moderate values of σ_1^2 and σ_2^2 coupled with poor starting values for the sampler lead to slow convergence for θ, ϕ and $\eta = \theta + \phi$, but that this problem is not apparent from plots of η sample traces alone. That is, even if posterior summaries are desired only for the well-identified parameter η, the convergence of the unidentified parameters θ and ϕ must be monitored as well (say, using sample traces and lag 1 sample autocorrelations). Here we focus on the convergence rate for η.

Let $\sigma_1^2 = \infty$, and define $\varepsilon = \sigma_2^2/(\sigma_2^2 + 1)$. Thinking of ε as a fixed tuning constant, the full conditional distributions necessary for the Gibbs sampler are

$$\theta \mid y, \phi \sim \mathrm{N}(y - \phi, 1) \quad \text{and} \quad \phi \mid y, \theta \sim \mathrm{N}(\varepsilon(y - \theta), \varepsilon) \tag{5}$$

The result of Gelfand and Sahu [6] ensures that as $\varepsilon \to 1$ (that is, $\sigma_2^2 \to \infty$), the samples for θ and ϕ diverge but those for η converge immediately, while that of Gelfand, Carlin and Trevisani [10] ensures that in running a Gibbs sampler using the full conditionals (5), $\mathrm{corr}(\eta^{(t)}, \eta^{(t+1)} \mid y) = 0$ for *every* t and *all* values of ε, once the sampler has converged. Note that the latter result does not contradict the Gelfand and Sahu claim of monotonically improving convergence as $\varepsilon \to 1$, since it shows the convergence rate is in fact constant (and perfect, since the lag 1 autocorrelation is 0) for all ε. However, the result pertains only to draws $\eta^{(t)}$ from the stationary distribution of the Markov chain; the draws are uncorrelated *once the sampler has converged*. It says nothing

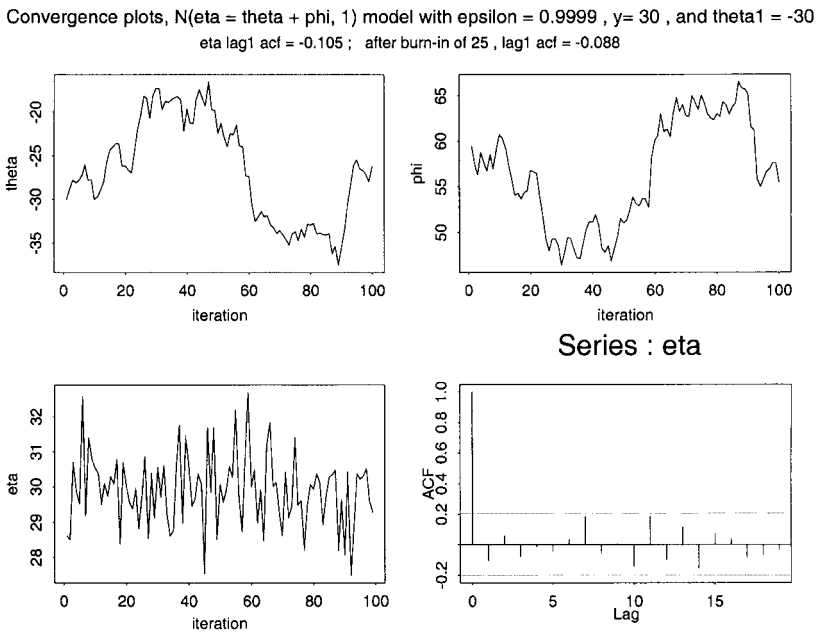

Figure 1. Convergence plots, N(η, 1) model with $\varepsilon = 0.9999$.

about what may happen if the algorithm's starting value $\theta^{(1)}$ is not a plausible draw from the true stationary distribution of θ.

To investigate the behaviour of this sampler during a pre-convergence 'burn-in' period, we ran a version in which we set $y = 30$ but $\theta^{(1)} = -30$. In this case, unless ε is quite large the starting value for θ will be quite far from its true posterior distribution. Figure 1 summarizes the first 100 draws from this sampler when $\varepsilon = 0.9999$, corresponding to an essentially flat prior on ϕ (recall the prior on θ is also flat). The first row of the figure shows the anticipated divergence of the θ and ϕ chains, while the second row shows immediate convergence for η (lag 1 sample autocorrelation = -0.105), also as expected. Figure 2 then considers the effect of setting $\varepsilon = 0.0001$, so that the ϕ prior is now tightly centred around 0, in contrast to the flat prior for θ. Now convergence for ϕ (hence θ) is immediate, and so η convergence is as well (lag 1 sample autocorrelation = -0.106), as in the previous case.

The interesting situation arises in Figure 3, wherein we set $\varepsilon = 0.9$, a moderate value leading to a weakly informative prior for ϕ, but again maintaining the flat prior for θ. Now the poor starting value for θ leads to a burn-in period covering roughly the first 25 iterations, and noticeably slow convergence for η (lag 1 sample autocorrelation = 0.471). While again this does not contradict the aforementioned theoretical results (the lag 1 sample autocorrelation based only on the final 75 post-burn-in iterations is an insignificant -0.144), it does illustrate a practical difficulty in their implementation. Specifically, if good starting values are not available for every parameter in the MCMC algorithm, observed convergence behaviour for the identified subset need *not* improve monotonically as the prior variance on the full parameter vector approaches infinity (that is, as $\varepsilon \to 1$ in our case). Difficulties like the one illustrated in Figure 3 may well be the reason for

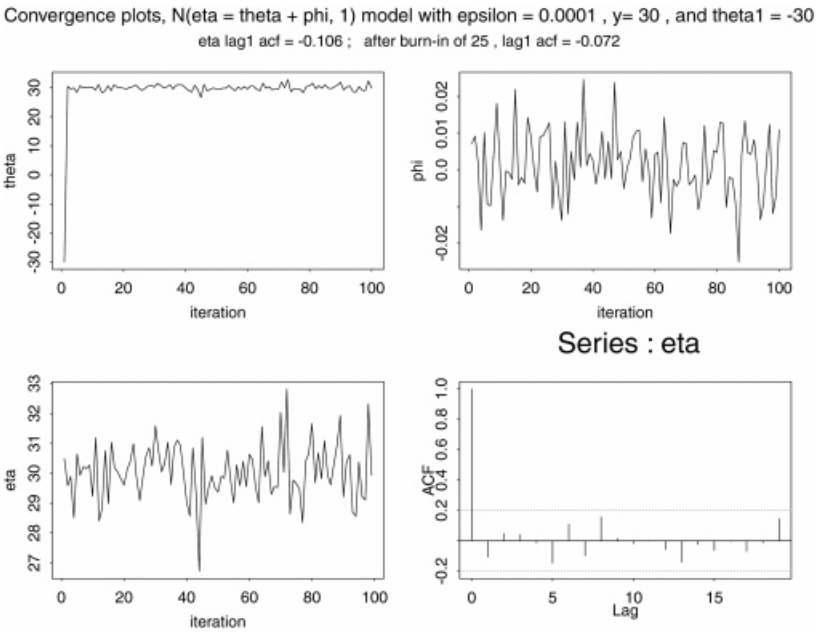

Figure 2. Convergence plots, N(η, 1) model with $\varepsilon = 0.0001$.

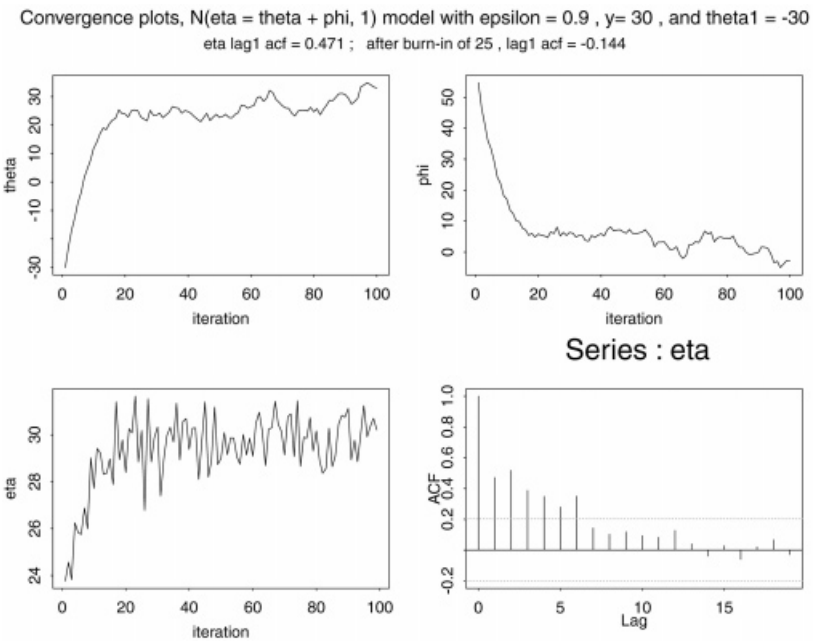

Figure 3. Convergence plots, N(η, 1) model with $\varepsilon = 0.9$.

the past recommendations to 'play it safe' in such situations and use tightly informative priors for all parameters not in the identified subset. This after all is the Bayesian equivalent of the usual frequentist approach to this problem; namely, imposing constraints on the parameter space. Figure 2 essentially sets $\phi = 0$, analogous to the 'corner point' constraints used in ANOVA modelling; an obvious alternative would be the summation constraint $\theta + \phi = k$ for some constant k.

3. CONVERGENCE FOR SPATIAL MODELS

Perhaps the most worrisome thing about the results of the previous section is that the troublesome case summarized in Figure 3 is precisely the one most often used in the analogous spatial model (1). That is, one typically selects values for τ_h (the marginal heterogeneity precision) and τ_c (the conditional clustering precision) that are neither very large nor very small. If instead *hyperpriors* are specified on τ_h and τ_c, they also tend to favour moderate values. We choose to proceed with fixed prior values for the precision parameters; this will enable us to track changes in convergence diagnostics and posterior values as the prior precisions change. Use of prior distributions instead would require the specification of at least two hyperprior control parameters, further muddying the interpretation of changes. We focus on three different values for each of τ_h and τ_c: $1, 10^{-3}$ and 10^{-6}, which produce a fairly wide range in the corresponding variances.

As alluded to above, the choices of prior values (or hyperpriors) are made to preserve the interpretability of the θ_i and ϕ_i random effects: a very large τ_h (or τ_c) effectively constrains these parameters to all be the same, while very small values for both will preclude their convergence altogether. Convergence of the identifiable η_i will still occur if the Markov chain begins from the stationary distribution (that is, if good starting values for these random effects are available), but of course this may not be the case, especially if the number of regions is at all large.

To investigate the degree of this problem, we consider a data set originally presented by Clayton and Kaldor [2] and reanalysed by many others since. This data set provides observed and expected cases of lip cancer in the 56 districts of Scotland for 1975–1980; the expected cases were calculated using the method of Mantel and Stark [11], that is, they are based on MLEs of the age effects in a simple multiplicative risk model. For each district i we also have one covariate x_i (the percentage of the population engaged in agriculture, fishing or forestry) and a list of which other districts j are adjacent to i. We thus apply models (1) and (2) with adjacency weights $w_{ij} = 1$ if regions i and j are adjacent, and 0 otherwise.

Since Gibbs sampler code for analysing these data and model is readily available as an example in the `BUGS` software package [12], we use this language to carry out our investigation. The newest version of `BUGS` for Windows, `WinBUGS 1.2`, automatically imposes the sum-to-zero constraint $\sum_{i=1}^{I} \phi_i = 0$ numerically by recentring the ϕ_i samples around their own mean at the end of each iteration [9]. All older versions of the program (including the one we used) do not, which in turn prohibits the inclusion of an intercept term in the log-relative risk model. Note that neither approach solves the Bayesian identifiability problem with the ϕ_i due to the continued presence of the covariate coefficient β and the θ_i. The sample implementation in the `BUGS` manual also initializes β and all the random effects to 0. Our investigations suggest that this value is not far from the bulk of the posterior mass for all of these parameters, so any convergence rate patterns we observe in the μ_i or η_i as functions of τ_h and τ_c should not be significantly affected by our use of these starting values. We first examine the pattern of the lag 1 autocorrelations for β and the η_i across a range of τ_h and τ_c values, and then examine the behaviour of ψ for evidence of

Bayesian learning across the prior scenarios specified by the precision parameters. All simulations are carried out with the same random seed to ensure comparability.

3.1. Convergence results

The data set lists the 56 counties in order of decreasing standardized mortality ratio, $\mathrm{SMR}_i = Y_i/E_i$, from a high of 6.52 to a low of 0. As such we consider convergence behaviour for $\eta_1, \eta_{27}, \eta_{47}$ and η_{56}, corresponding to districts with observed rates much higher, about the same, and much lower than expected. Running the Gibbs sampler for 10 000 iterations, Figure 4 plots the lag 1 sample autocorrelations for these four parameters and the covariate effect β under our three choices for each of τ_h and τ_c, $1, 10^{-3}$ and 10^{-6}.

Using the BUGS parameterization, the well-identified parameter in this setting is

$$\mu_i = \log E_i + \beta x_i + \theta_i + \phi_i$$

(recall the E_i are assumed known). Our simulations suggest near-immediate convergence for all μ_i (all within-chain lag 1 sample autocorrelations less than 0.2 in magnitude), in concert with the aforementioned theoretical work. However, convergence for the sum of the random effects, $\eta_i = \theta_i + \phi_i$, is rather more difficult to predict. For η_{56}, convergence behaviour is more or less as predicted by the Gelfand and Sahu [6] result for the corresponding μ_i, in that having both priors tighter leads to generally poorer convergence, as measured by the lag 1 ACF. When the heterogeneity prior is very tight ($\tau_h = 1.0$), a vague clustering prior ($\tau_c = 10^{-6}$) is able to compensate and produce a small lag 1 ACF; similarly when the clustering prior is tight, a vague heterogeneity prior also produces good convergence behaviour. However, for η_1, η_{27} and η_{47} (top three panels) the pattern is reversed, with somewhat poor to very poor convergence for all nine prior combinations *except* the most restrictive one ($\tau_h = \tau_c = 1$). Apparently the SMR of 0 in districts 55 and 56 encourages a posterior log-relative risk of $-\infty$, creating a conflict with the tighter priors, hence slower convergence, but for regions whose data encourage more moderate SMR values, identifiability of the η_i weakens, so that the tighter priors offer a relative convergence benefit.

Finally, the situation for β is particularly discouraging. With the exception of the tightest prior combination ($\tau_h = \tau_c = 1$), all prior settings produce lag 1 autocorrelations greater than 0.998, a value so large that acceptable posterior summaries cannot be produced even with our 10 000 samples. Indeed, trace plots (not shown) using poorer starting values (for example, $\theta_i = \phi_i = -2$ for all i) have not yet located the bulk of the posterior mass by iteration 10 000. Again, to the extent that only the well-identified parameters (the μ_i) are of interest, this sampler remains legitimate; however, it would not be appropriate in settings where we have genuine interest in the impact of the AFF covariate on lip cancer risk.

Incidentally, we were somewhat surprised by this poor performance for β, especially since most previous analyses of this data set do not mention any particular difficulty in estimating it via MCMC. However, these analyses generally place hyperprior distributions on τ_c and τ_h, which in turn afford the opportunity to learn about these parameters from the data. For instance, Conlon and Waller [5] use gamma hyperpriors having mean 1 for both τ_h and τ_c, the first vague (variance 1000) and the second weakly informative (variance 10). These authors report resulting posterior medians of 55.2 and 2.95 for τ_h and τ_c, respectively – much larger than any of the fixed values we tried and suggestive of significant learning for these two hyperparameters. Since our setting instead fixes (often quite small) values for τ_h and τ_c, this learning is precluded, and the excess uncertainty seems to propagate through the system to the fixed effect β.

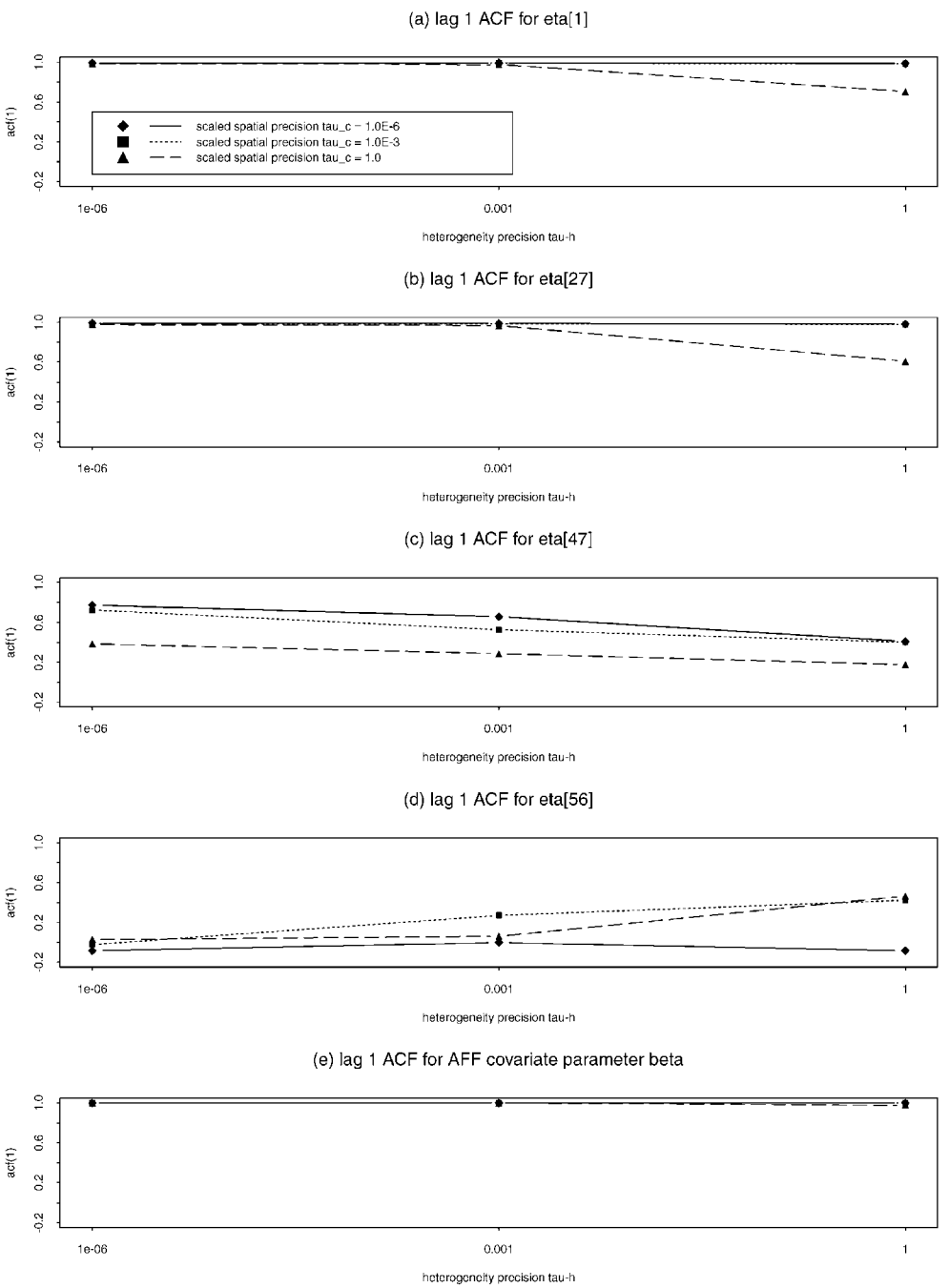

Figure 4. Lag 1 sample autocorrelations for η_1, η_{27}, η_{47}, η_{56} and β, Scottish lip cancer data, various priors for τ_h and τ_c.

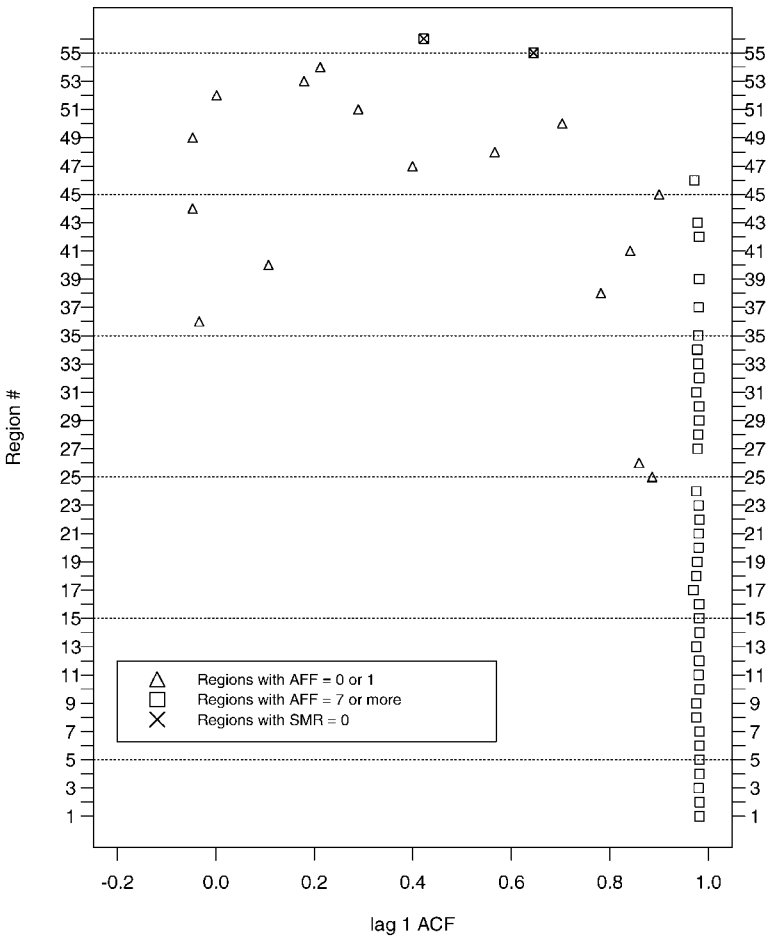

Figure 5. Lag 1 sample autocorrelations for all η_i, when $\tau_c = 10^{-3}$ and $\tau_h = 1$.

Focusing on a single prior for the moment ($\tau_c = 10^{-3}$ and $\tau_h = 1$), interesting differences in the η_i convergence rate emerge for regions having differing values of the covariate x_i. Figure 5 plots the lag 1 ACF values by region where the plotting character indicates the value of the covariate. Note that the regions having $x_i = 0$ or 1 (which incidentally includes region 47 from Figure 4) consistently exhibit convergence behaviour similar to the two regions with an SMR of 0 (regions 55 and 56). This convergence is better than all other regions, and sometimes vastly better. The other regions (all having covariate values of 7 or more) have lag 1 ACF values which are essentially 1; even under the tightest prior combination ($\tau_h = \tau_c = 1$, not shown), the lag 1 ACF for these regions hovers around a disappointing 0.7. A possible explanation for this behaviour is similar to the one given above for the regions having SMR = 0; namely, that when the covariate x_i equals 0 or 1, the covariate is essentially 'not in the model', and thus the burden of explaining the total Poisson variability falls on θ_i and ϕ_i. This in turn leads to enhanced identifiability for η_i, hence more rapid convergence. For the remaining regions, however, the strong covariate and SMR

values leave relatively little in the data for θ_i and ϕ_i to explain, hence rapid convergence only for the tightest of priors. A more algorithmic explanation is that β's slow convergence adversely affects that for those η_i's corresponding to large x_i, but not those having small x_i.

3.2. Results for Bayesian learning

If vague priors for the θ_i and ϕ_i are indeed inappropriate, this forces us to use proper priors instead. As mentioned above, the usual strategy is to attempt to specify priors for τ_h and τ_c which are 'fair' (that is, that assume excess Poisson variability in the data is due in equal parts to heterogeneity and to clustering); again, this is difficult due to the conditional structure of the $\boldsymbol{\phi}$ prior and the marginal structure of the $\boldsymbol{\theta}$ prior. The posterior proportion of excess variability due to the clustering parameters is then summarized by the posterior distribution of ψ, given in equation (3). However, recently some authors have doubted whether this approach is sensible, given that the parameters in question are not identified. Is any 'Bayesian learning' (from marginal prior to marginal posterior) really occurring for this proportion ψ?

To investigate this, we must first make a connection between the prior distribution of ψ and the prior precision parameters τ_c and τ_h. This is difficult, since ψ itself is an empirical quantity, defined only in terms of the marginal distribution of the actual random effects, while τ_c is part of a conditional, theoretical variance expression. To reconcile the two quantities, we first note the observation of Bernardinelli *et al.* [13] that the prior marginal standard deviation of ϕ_i is roughly equal to the prior conditional standard deviation divided by 0.7, that is

$$\mathrm{SD}(\phi_i) \doteq \frac{1}{0.7\sqrt{(n_i \tau_c)}} \tag{6}$$

where $n_i = \sum_j w_{ij}$, the number of neighbours for the ith region. To account for the varying number of neighbours across regions, we replace n_i by $\bar{n} = 264/56 = 4.71$, the average number of neighbours across the Scotland map.

Since $\psi = \mathrm{SD}(\boldsymbol{\phi})/(\mathrm{SD}(\boldsymbol{\theta}) + \mathrm{SD}(\boldsymbol{\phi}))$ from equation (3), ψ has a non-degenerate prior distribution even when τ_c and τ_h are fixed, due to the variability in $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. We can thus use equation (3) to approximate the prior mean of ψ for fixed τ_c and τ_h. In order to explore a range of ψ values, we consider each ψ value in the set $\{\frac{8}{9}, \frac{6}{7}, \frac{4}{5}, \frac{2}{3}, \frac{1}{2}, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \frac{1}{9}\}$, a reasonably good covering of the range of possible values from 0 to 1. We do this by selecting τ_h so that our approximate prior $\mathrm{SD}(\theta_i)$ value is a multiple of our approximate prior $\mathrm{SD}(\phi_i)$ value, that is, we set $\mathrm{SD}(\theta_i) \equiv 1/\sqrt{\tau_h} = c\,\mathrm{SD}(\phi_i)$ for each c in $\{\frac{1}{8}, \frac{1}{6}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 6, 8\}$. We then ran 27 simulations, corresponding to the nine ψ values above arising from each of our three τ_c values (1, 10^{-3} and 10^{-6}), monitoring the convergence and final posterior distribution of ψ in each case.

Figure 6 shows the posterior medians and 95 per cent equal-tail credible intervals arising from our 27 MCMC runs of 10 000 iterations each, with the figure's three panels corresponding to the three possible values of τ_c. Within each panel, the dotted line connects points having equal prior and posterior ψ. For the $\tau_c = 10^{-6}$ case shown in Figure 6(a), there is very little Bayesian learning when ψ is small (that is, when τ_h is small, so the priors are both vague), with broad intervals and posterior medians closely tracking the prior values. However, for larger ψ values the posterior medians move towards values smaller than the corresponding prior values, with the confidence intervals actually excluding the prior values for the two largest ψ's (that is, there is 'significant' Bayesian learning about ψ under the least vague heterogeneity priors).

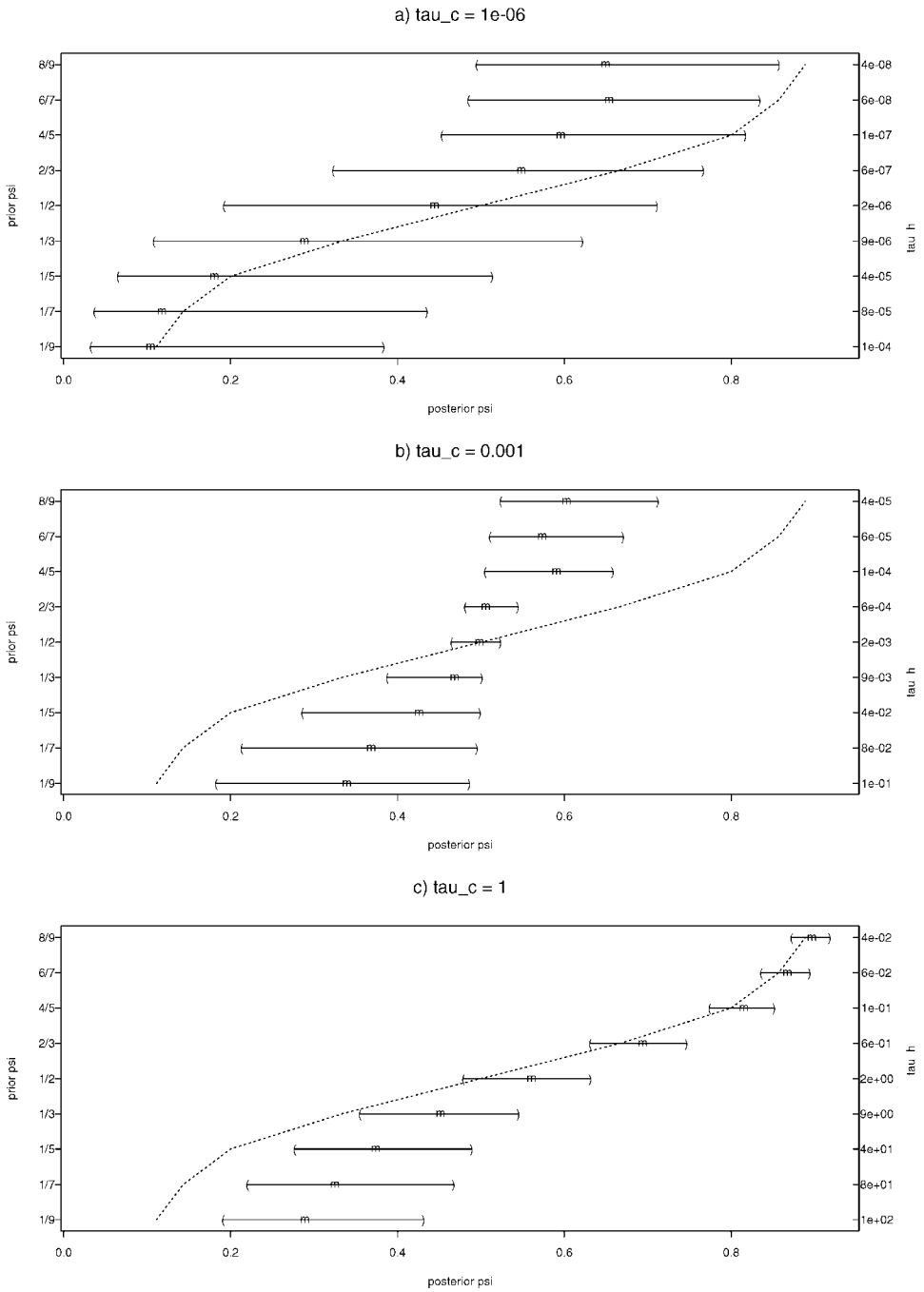

Figure 6. Posterior medians and 95 percent equal-tail credible intervals for ψ, three different τ_c values. Points of prior and posterior equality are connected by the dashed lines.

Much more prior-to-posterior movement is evident in the $\tau_c = 10^{-3}$ case shown in Figure 6(b). This panel shows the familiar Bayesian shrinking of the posterior away from the prior and towards the data-supported value; the only 95 per cent credible interval failing to exclude its prior mean is the one using the 'fair' prior value $\psi = 0.5$. Finally, Figure 6(c) shows the setting where $\tau_c = 1$, a rather informative prior on the amount of excess clustering. Here when ψ is small we see prior-posterior disagreement, with the data suggesting more clustering than the prior. Increasing ψ apparently brings the prior into agreement with the data, to the point where the posterior credible intervals become very narrow and do not exclude the prior ψ value.

The main message from Figure 6 is that, despite the non-identifiability of the θ_i and ϕ_i parameters, using ψ in equation (3) to judge the relative presence of excess clustering and heterogeneity in a spatially referenced data set is indeed sensible, since Bayesian learning about ψ can still take place. However, the issue of selecting an appropriate scale for τ_c and τ_h is critical, since as the figure shows, the learning pattern can change markedly with this scale. It is also important to remember that the range of prior τ_h values induced by our chosen ψ pattern varies greatly with our three chosen τ_c values; for instance, $\tau_c = 10^{-6}$ implies $3.6 \times 10^{-8} < \tau_h < 1.5 \times 10^{-4}$, while $\tau_c = 1$ implies $3.6 \times 10^{-2} < \tau_h < 1.5 \times 10^{2}$. Best *et al.* [9] investigate this issue by comparing several gamma(ε, ε) hyperpriors (having mean 1 and variance $1/\varepsilon$) for τ_c and τ_h on several *simulated* Scotland data sets, and then seeing which combination most accurately reproduces the true values in the resulting posterior. While this approach appears quite sensible, further experimentation with other regional patterns is certainly warranted to see if prior selection recommendations can be made in more general contexts.

Of course, strictly speaking Figure 6 does not really measure 'Bayesian learning' for ψ, since it does not have a proper prior; our approach using approximation (6) does not explicitly incorporate the sum-to-zero constraint needed to make the CAR prior proper. Moreover, our approach provides only a 'best guess' for ψ *a priori*, and thus we really obtain only a rough idea of the prior-to-posterior movement of ψ. An alternative, more direct approach to measuring Bayesian learning for ψ is possible using the latest release of the Windows-based version of the `BUGS` program, `WinBUGS 1.2`, freely available over the web at `http://www.mrc-bsu.cam.ac.uk/bugs/`. This program allows direct sampling from the centred version of the CAR prior (that is, the version incorporating the sum-to-zero constraint) via its `car.normal` function, and so a direct investigation of Bayesian learning for ψ is possible in this sense. That is, we simply rerun our Gibbs sampling code *without* the data, and compare the resulting prior draws for ψ to the posterior draws already obtained.

Figure 7 summarizes some results from this alternative approach. For one of our τ_c values ($\tau_c = 10^{-3}$), the three columns of this figure compare the histograms of the prior (top row) and posterior (bottom row) ψ samples for the largest, middle and smallest τ_h values considered in Figure 6. The 'best prior guesses' for ψ ($\frac{8}{9}$, $\frac{1}{2}$ and $\frac{1}{9}$, respectively) are marked by vertical reference lines; notice that in all three cases these lines are quite close to the middle of the prior histograms. This suggests that our rough approach making use of the Bernardinelli *et al.* [13] 'rule of thumb' in equation (6) works rather well, justifying its use in Figure 6. The vertical alignment of the prior and posterior histograms also makes it easy to identify the Bayesian learning that is taking place; its magnitude and direction are also in concert with that indicated in Figure 6(b).

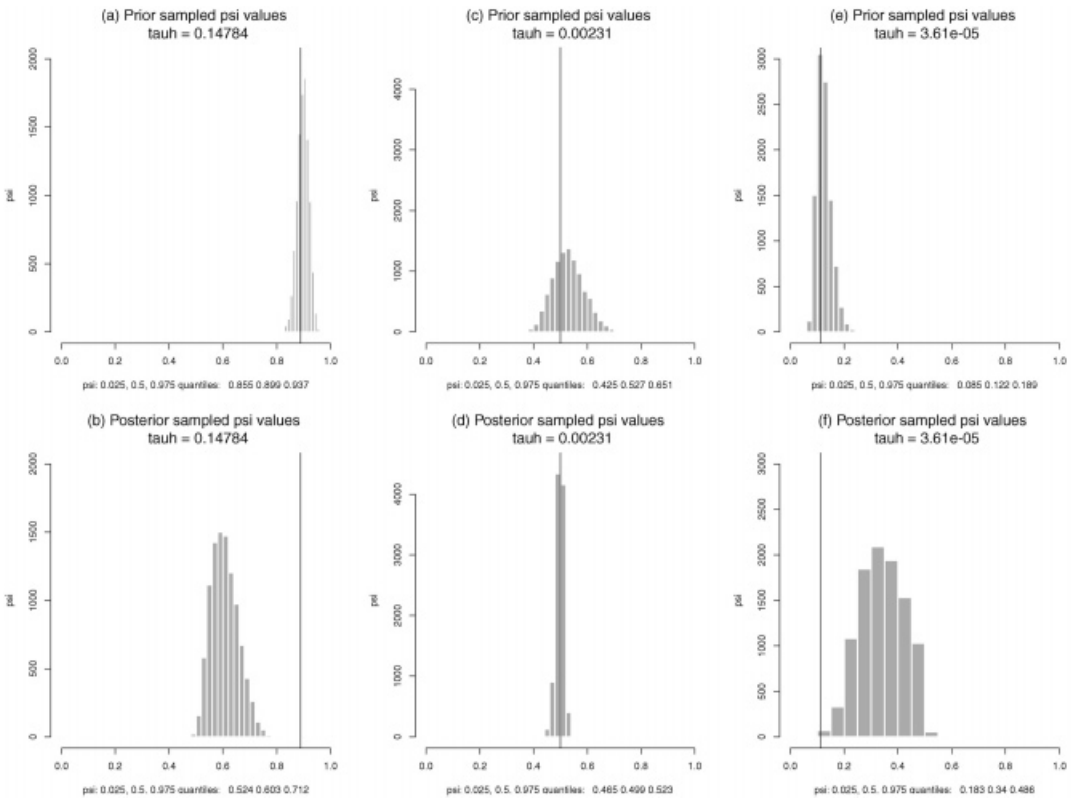

Figure 7. Histograms of prior and posterior ψ samples, $\tau_c = 10^{-3}$. Columns correspond to the three indicated choices of τ_h, which in turn correspond to 'best prior guesses' for ψ (indicated by vertical reference lines) of $\frac{8}{9}$, $\frac{1}{2}$ and $\frac{1}{9}$, respectively.

4. SUMMARY AND DISCUSSION

In summary, our results indicate that the convergence rates of parameters of interest in spatial models can be affected by the starting values of the chains, the precise prior values chosen, and even the values of the response variables and covariates themselves. In the specific context of spatial models like (1), we also find that Bayesian learning about ψ (the equation (3) measure of the proportion of excess variability due to clustering) is indeed possible, and that formula (6) for converting the conditional CAR precision parameter τ_c into a marginal standard deviation seems to lead to a reasonable method for calibrating the heterogeneity and clustering priors (in particular, for obtaining a setting in which $\psi \approx \frac{1}{2}$ *a priori*). Figure 7 suggests that a way to check this rough calibration would be to sample directly from the resulting prior and see if the resulting ψ samples are indeed centred at the desired value. We hasten to add, however, that since our key findings are specific to a single data set (the over-analysed and oft-maligned Scotland data), our recommendations should be tested under a variety of spatial data sets, featuring different adjacency and covariate patterns.

In the light of our Section 3.1 results, one might wonder if using different priors for different random effects would be a sensible approach for improving convergence. Such an approach would be hard to justify, however, since it nullifies the exchangeability in the spatial prior and implicitly uses the data to help make the prior choice. The use of vague or ultravague priors for the θ_i and ϕ_i and subsequent monitoring and summarization of only the well-identified log relative risks μ_i might also be contemplated. However, this approach would seem to forfeit the interpretability of the random effects so prized by many spatial statisticians, and in any case need not produce good convergence for all parameters (as seen in the first, second and last panels of Figure 4). In the presence of covariates x_i, this approach also seems to greatly weaken the identifiability of the corresponding covariate effect β, which may often be of substantive interest.

Given the occasionally high autocorrelations present in such systems, one might worry about the starting values used in the MCMC algorithm. We experimented with four different yet plausible sets of starting values for the random effects: $(\phi_i, \theta_i) = (0,0), (-2,2), (-2,0)$ and (–2,–2), with all regions initialized identically. Again using 10 000 iterations, we found little posterior dependence on the starting values for the μ_i, η_i and ψ, but marked dependence for the ϕ_i, θ_i and β. Initializing the ϕ_i and θ_i random effect to 0 is certainly plausible (since they are additive adjustments to internally standardized log-relative risks), but very long runtimes may still be required to achieve an adequate 'effective sample size' (Kass *et al.*, [14], p. 99) for β.

As such, spatial models do not seem to provide a good example of a setting where prior selection could plausibly be based on a strategy to improve the convergence of the computational algorithm. Since our results in Figures 6 and 7 suggest that ψ is a sensible summary of the relative contribution of the clustering parameters to the spatial model (1), future work looks to developing more general guidelines for proper prior selection (for example, determining an appropriate scale) and calibration (for example, encouraging prior ψ values near 0.5, or in some other way insisting on an equal *a priori* allocation to heterogeneity and clustering). In this paper we deliberately avoided using hyperpriors for τ_c and τ_h, due to the resulting increase in the difficulty of the calibration problem, but since such hyperpriors are commonly used in practice, investigation of Bayesian learning for ψ in their presence is clearly warranted.

ACKNOWLEDGEMENTS

The research of the first author was supported in part by National Institutes of Allergy and Infectious Diseases (NIAID) grant 5-U01-AI42170-07. The research of the second author was supported in part by National Institute of Environmental Health Sciences (NIEHS) grant 1-R01-ES07750 and NIAID grant R01-AI41966. The authors thank Alan Gelfand and Erin Conlon for helpful discussions and editorial assistance.

REFERENCES

1. Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC Press: Boca Raton, FL, 1996.
2. Clayton DG, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**:671–681.
3. Besag J, York JC, Mollié A. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 1991; **43**:1–59.
4. Cressie N, Chan NH. Spatial modelling of regional variables. *Journal of the American Statistical Association* 1989; **84**:393–401.
5. Conlon EM, Waller LA. Flexible neighborhood structures in hierarchical models for disease mapping. Research Report 98-018, Division of Biostatistics, University of Minnesota, 1998.
6. Gelfand AE, Sahu SK. Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association* 1999; **94**:247–253.

7. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *Journal of the American Statistical Association* 1996; **91**:1343–1370.
8. Besag J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic systems (with discussion). *Statistical Science* 1995; **10**:3–66.
9. Best NG, Waller LA, Thomas A, Conlon EM, Arnold RA. Bayesian models for spatially correlated disease and exposure data (with discussion). In *Bayesian Statistics 6*, Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Oxford University Press: Oxford, 1999; 131–156.
10. Gelfand AE, Carlin BP, Trevisani M. On computation using Gibbs sampling for multilevel models. Research Report 2000–004, Division of Biostatistics, University of Minnesota, 2000.
11. Mantel N, Stark CR. Computation of indirect-adjusted rates in the presence of confounding. *Biometrics* 1968; **24**:997–1005.
12. Spiegelhalter DJ, Thomas A, Best N, Gilks WR. BUGS examples, Version 0.50. Technical report, Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge University, 1995.
13. Bernardinelli L, Clayton DG, Montomoli C. Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine* 1995; **14**:2411–2431.
14. Kass RE, Carlin BP, Gelman A, Neal R. Markov chain Monte Carlo in practice: a roundtable discussion. *American Statistician* 1998; **52**:93–100.

