DISCUSSION

DONALD GEMAN[1] AND STUART GEMAN[2]

[1] *Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003, U.S.A.*
[2] *Division of Applied Mathematics, Brown University, Providence, RI 02912, U.S.A.*

First of all, it is a pleasure to congratulate the authors for contributing this insightful study to the growing literature on the use of Markov random fields for analyzing spatial data. Indeed, this trend was largely inspired by the first author (Besag (1974)), who has always maintained that some of the most substantial applications of the methodology associated with "Bayesian Image Analysis" may in fact be *outside* image analysis per se, involving neither light intensities nor regular lattices.

We should also like to point out that this paper sustains a related trend: applying that same methodology, particularly certain algorithmic tools (Gibbs Sampler, etc.), to problems in conventional statistics in which progress had been stalled by the inability actually to *compute* anything. Going further still, many researchers have advocated the importance of the posterior distribution *itself*, revealing the likely and unlikely states of the target attributes, rather than merely summary statistics, such as the (posterior) mode or mean. In practice, however, this has rarely been done. Thus, we endorse the emphasis here on *interval estimates* and related analyses.

Turning to specifics, there are several points which require clarification. One is the dependence of the conditional variance of the u_i component of the log relative risk on the size η_i of its neighborhood in the underlying graph. As it stands, the variance decreases as η_i increases; see (4.3). Hence there is an implicit assumption that, by whatever method these adjacencies are determined, i.e., how one population center is "hard-wired" to another, more links means smaller variance, all other factors being the same. Is this designed? For example, does one expect more (conditional) variation in outlying regions and so arrange for it by assigning fewer neighbors? We were also somewhat confused by the remarks about the accuracy of the Bayesian estimates in comparison with those (presumably just the MLE's) calculated directly from the data y_i. How exactly was this ascertained?

Finally, the latter remark leads us to a larger point, really a cautionary note. These models are certainly complex in comparison to non-spatial ones or ad hoc smoothers of standard estimators. There are lots of choices to be made: the decomposition of the relative risk into spatial, "noise," and covariate terms; the marginal distribution of u (choice of ϕ and neighborhoods); and various associated (hyper)parameters. Consequently, there is plenty of room for imagination. Now all of these steps are quite well-reasoned and the resulting apparatus looks solid,

but in the absence of "ground truth" of the sort usually available in image analysis (i.e., the ability synthetically to degrade images, thus having the "original" for comparisons; or simply having the actual digits or road maps for checking algorithms for optical character recognition or automated cartography), how does one *demonstrate* that the Bayesian approach is truly more accurate than more elementary methods? Evidently, the authors believe it is, and so do we, but it would be comforting to have that "smoking pistol" for applications of such evident social importance as the ones treated here.

PETER J. GREEN

Faculty of Science, University of Bristol, Bristol BS8 1TH, U.K.

I welcome this paper, especially for the two-way exchange of ideas that it promotes between statistical approaches to image analysis and other areas of statistics. Both sides should benefit, and the authors should be congratulated.

I want to make a few comments about the application to the mapping of risk from disease, treated in Section 4. This is a very appealing formulation, although one suspects that the Poisson variation and the unstructured variables v_i are close to being "confounded". Had the model instead specified that y_i was *Normally* distributed with mean $x_i = u_i + v_i$, and (for correspondence with the Poisson assumption) prescribed variance σ^2, then the distinction between v_i and the Normal errors would be seen to be artificial; there might indeed be problems of estimation if the variance of v_i was rather smaller than σ^2, and due to sampling fluctuations y_i were less variable than expected.

An innovation here is to take the Bayesian formulation one layer higher than usual in image analysis, by imposing a prior on the parameters κ and λ, corresponding to the reciprocals of the interaction parameters in the usual Gibbs distributions. Obviously, some experiments were needed before the prior specification (4.6) was adopted. The text rather suggests that the choice of a precise value for ε is not too important, but I note that, in the examples, posterior means of κ and λ are often as small or even smaller than the selected value $\varepsilon = 0.01$. Doesn't this imply that ε has greater significance than simply as a convenient factor to make the algorithm work, since the prior is not locally uniform around values supported by the likelihood?

In current work in image analysis at Bristol, we are making much use of Markov chain simulation methods to estimate functionals of the posterior distribution of the true scene (on a moderately large scale: current workstations can cope with the computations, contrary to the impression given in the abstract). Except in purely Gaussian models (where normalising constants are known) or when

using small pallettes of discrete "colours" (where they can be calculated quickly), we generally use the Metropolis method, rather than the Gibbs sampler. Possible disadvantages in terms of reduced rate of convergence, counting in sweeps, can easily be offset by reduction in computing time per sweep when normalising constants are not needed. The form of Metropolis we use is the generalisation discussed by Hastings (1970) which allows a virtually arbitrary "proposal" distribution. In the context of Section 4, such an algorithm could take the following form. When visiting site i, draw a new value for u_i not from the local characteristic given in the paper (the display following (4.6)), but from a Normal distribution with mean $(1+\theta)\mu - \theta u_i$ and variance $(1-\theta^2)\sigma^2$, where θ, μ and σ are yet to be determined. This new value is not automatically adopted, but regarded as a "proposal" u_i'. It is only accepted with probability

$$\alpha = \min(1, \exp(g(u_i') - g(u_i)))$$

where the function g is defined by

$$g(u_i) = \frac{1}{2\sigma_2}(u_i - \mu)^2 - \frac{n_i}{2\kappa}(u_i - \bar{u}_i)^2 + u_i y_i - c_i e^{u_i + v_i}.$$

With probability $1-\alpha$, the current value u_i is left unchanged, and another site or variable addressed. It is quite easy to check that detailed balance holds for any θ, μ and σ, and these can legally depend on all variables except u_i and u_i'. It is now straightforward to select values for μ and σ that make g reasonably constant over a range of u values around the middle of the proposal distribution, so that the acceptance probability α is close to 1 with high probability. A very similar method would be used to resample v_i. Unlike the "carefully designed rejection methods" that are needed in the presence of the Poisson likelihood when using the Gibbs sampler, there is no need here to ensure that the true density is bounded above by a multiple of an approximating one, and so the awkward tail in the conditional density is no objection to the use of the Normal distribution for a proposal. My guess is that this simulation method would be much faster.

In the Gaussian case, μ and σ can always be chosen to make g identically constant, so $\alpha = 1$. With $\theta = 0$, this is exactly the Gibbs sampler, otherwise (note that g does not involve θ), this becomes the proposal of Barone and Frigessi (1990), which they show can give faster convergence for some $\theta > 0$ in cases of positive interaction.

Finally, I would like to draw attention to the different possible requirements for speed of convergence when using these Markov chain methods. This usually seems to be considered in terms of the rate of weak convergence, given by $\max\{|\lambda_j|\}$, required to be small, where $\{\lambda_j\}$ are the non-unit eigenvalues of the transition matrix. But often in practice, the purpose of the simulation is to estimate some functional of the form

$$E_{post}(F(u)) = \int F(u)p(u|data)du.$$

When this is estimated by an empirical average

$$\hat{E}_{post}(F(u)) = (t_2 - t_1)^{-1} \sum_{t=t_1+1}^{t_2} F(u^{(t)}),$$

the relevant measure of convergence is the asymptotic variance, and better precision is obtained by choosing the sampling method to make the ratios $\{(1+\lambda_j)/(1-\lambda_j)\}$ small: λ_j negative is better than zero. Details are given in Peskun (1973): the idea is roughly analogous to the use of antithetic variables in Monte Carlo, and seems to give additional support to using $\theta > 0$ above.

References

Barone, P. and Frigessi, A. (1990). Improving stochastic relaxation for gaussian random fields, Quaderno 9/89, Istituto per le Applicazioni del Calcolo, Rome.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains, and their application, *Biometrika*, **57**, 77–109.

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

SHIGERU MASE

Faculty of Integrated Arts and Sciences, Hiroshima University, Hiroshima 730, Japan

Authors pointed us in this paper the applicability of the Bayesian image restoration method to several important problems in spatial statistics. I completely agree with them in that this method will be a very flexible and strong tool in future. There is a good possibility that it leads to a whole branch of spatial statistics. If so, it is important to realize merits and demerits of the method. Authors discussed several merits of this method in the paper. I want to point out two problems in the following.

After repeating Gibbs samplers sufficient times, we can get some *restored images* as shown in Fig. 1. How can we be certain that the method has worked good and we have a good image? In the original image restoration problem we can fairly easily judge whether the method works good or bad. Just "Seeing is believing". But, I think, this is not the case in the present pseudo-image restoration problem.

One popular justification of the Bayesian approach in estimation problem is that resulting estimators are asymptotically independent on the choice of priors. But it seems that realizations which are results of Gibbs sampler are strongly dependent on the prior used. If this is true we should be very careful to select a prior. We must take account of underlying spatial structures and relevant knowledges carefully and can not choose a prior because of its mere simplicity and tractability. Let us consider the prior used in the archaeological problem. As authors remarked, this prior encourages neighboring sites to have the same values.

As a result, resulting restored images tend to consist of isolated convex-like components as shown in Fig. 1. (Also it may cause the instability of boundary sites of each objects as well as sites near the frame boundary.) Do these components have proper meanings? Are there no possibilities that they only appear because we used such a prior? The final criterion is, of course, archaeological evidences. But also we need to know to what extent this prior is consistent with archaeological knowledges.

GOODNESS-OF-FIT OF BAYESIAN MODELS BY THE MONTE CARLO SIMULATION

YOSIHIKO OGATA

Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106, Japan

The main contribution of the paper, I feel, is the suggestion and implementation of the posterior mean estimate for the regulation parameters κ and λ by the Gibbsian Monte Carlo simulation. This will be very useful for an objective Bayesian analysis based on the Monte Carlo method. In addition, the authors introduced the prior

$$\text{prior}(\kappa, \lambda) \propto e^{-\epsilon/2\kappa} e^{-\epsilon/2\lambda} \tag{1}$$

which, referring to the authors, is "to avoid the absorbing state of the Markov chain invalidating the Gibbs sampler". According to the authors ϵ is a small positive, and they choose $\epsilon = 0.01$ in an ad hoc manner. However, I am concerned in the fact that the posterior mean estimates of either κ or λ are as small as $\epsilon = 0.01$. Then I must suspect that the *absorbing state* itself has a statistical significance such that the optimal estimate of either $u = \{u_i\}$ or $v = \{v_i\}$ is constant (i.e., independent of i), so that κ or λ tend to zero. If this is the case, the introduction of the prior for avoiding the absorbing state, such as the one in (1), should not be, or at least should be carefully posed. I consider that the Jeffreys' ignorance prior or alternatively the uniform prior, as described by the authors, may still be preferable in their procedure. Also, I think that it may be helpful to discuss the goodness-of-fit of Bayesian models for the suitable prior selection. Indeed, the authors considered two priors, the Gaussian and double exponential types, as well as the terms t, u and v in the likelihood. Let me argue these issues by the use of related data sets and models to the authors'.

Considering the last example analyzed by the authors, I have to start with the processed data $z = \{z_i\}_{i=1,2,\ldots,n}$, $n = 216$, given in Fig. 14 by the authors

(or Fig. F1 for the comparison with the patterns shown hereafter), showing the observed mortality rate of cancers in Great Manchester, since I do not know the original data for the observed number of deaths from cancer. Also, I have to assume the Gaussian model instead of the Poisson likelihood. Further, the prior function of the Gaussian intrinsic auto-regression is considered, so that the posterior is given by

$$(2) \qquad P(\theta \mid \sigma, \tau;\ z) = \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\theta_i - z_i)^2 \right\} \times \frac{C}{\tau^n} \exp\left\{ -\frac{1}{2\tau^2} \sum_{i \sim j} (\theta_i - \theta_j)^2 \right\} \times \text{prior}(\sigma^2, \tau^2),$$

where C is the normalizing constant of the prior for θ: this can be arbitrarily a fixed one for the Monte Carlo simulation. At first, I simulated the above posterior distribution by the Gibbs sampler with the ignorance prior, i.e. $\text{prior}(\sigma^2, \tau^2) \propto \sigma^{-2}\tau^{-2}$, as well as the uniform prior, $\text{prior}(\sigma^2, \tau^2) \propto 1$. Then, the posterior mean estimate of (σ, τ) for the former case converges to (0.410, 0.000), while the one for the latter case converges to (0.391, 0.188). The convergence of τ to zero can be avoided by assuming the similar prior to the one in (1) for τ in conjunction with the ignorance prior, but ϵ needs to be large enough to obtain a similar posterior mean of θ to that in Fig. 15 of the authors: this leads me to the difficulty in determining a suitable ϵ. On the other hand, I had no problem of such singularity in the case of the uniform prior. Its posterior mean of θ (Fig. F2) provides a rather similar pattern to the authors' estimate (Fig. 15), while the one for the ignorance prior provides the estimate of constant risk 1.07 throughout the region. Which one, then, gives a better fit to the data?

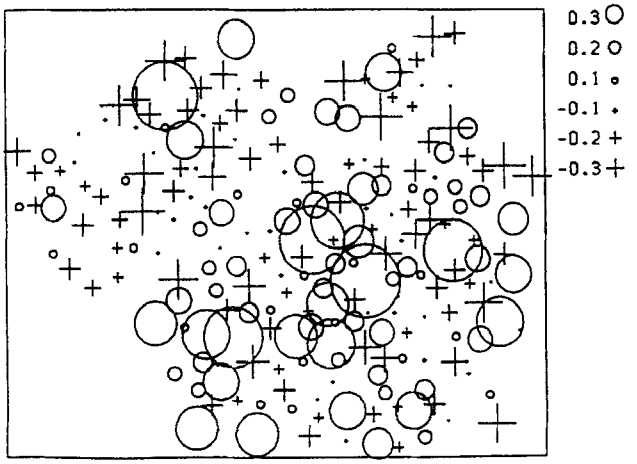

Fig. F1. The mortality rate data $\theta = \{\theta_i\}$ identical to Fig. 14 of the authors. Size of signs shows the deviations from the average $\hat{\theta} = 1.07$.

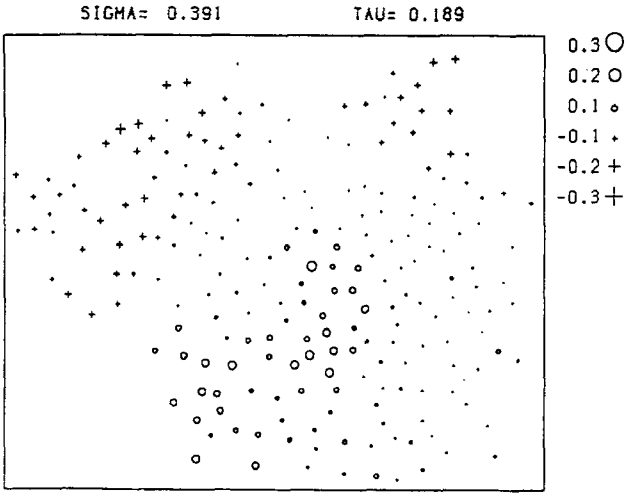

Fig. F2. Posterior mean estimate of θ for the uniform prior of σ and τ.

Consider no prior for σ^2 and τ^2 in (2): i.e., formally, prior$(\sigma^2, \tau^2) = 1$. Define the Bayesian likelihood

$$\mathcal{L}(\sigma, \tau) = \int P^*(\theta \mid \sigma, \tau;\ z)d\theta, \tag{3}$$

where P^* is the same P as in (2) except that the prior for θ is a probability distribution with the normalized constant C^* which is calculated somehow avoiding the improper prior, e.g., in a Gaussian case as the above, the degenerated Hessian matrix should be avoided somehow: see Akaike (1979), Ogata and Katsura (1988), Ogata (1990) and Ogata *et al.* (1991). The type II maximum likelihood method is suggested by Good (1965) to find the optimal hyperparameters, such as σ and τ, which maximizes the Bayesian likelihood. To compare the goodness-of-fit of Bayesian models, Akaike (1979) defined

$$\text{ABIC} = (-2)\max\{\log \text{Bayesian likelihood}\} + 2 \cdot \{\text{number of hyperparameters}\}$$

which can be comparable with the AIC which is used in case of the ordinary maximum likelihood models (Akaike (1987)). A model with a smaller ABIC or AIC indicates a better fit. Akaike justified such selection procedure based on the *entropy maximization principle* (Akaike (1977, 1978)).

A computation method for the log Bayesian likelihood by the Metropolis' Monte Carlo procedure is suggested by Ogata (1989, 1990), which assessed the high dimensional integrations for the normalizing constants of the posterior as well as the prior. Figure F3 summarizes the values for the log Bayesian likelihood of a number of pairs (σ, τ). This indicated a local maximum around $(\sigma, \tau) = (0.375, 0.30)$. The posterior mean of θ in this case is shown in Fig. F4, which appears very similar to Fig. 15 by the authors. The standard deviations of the

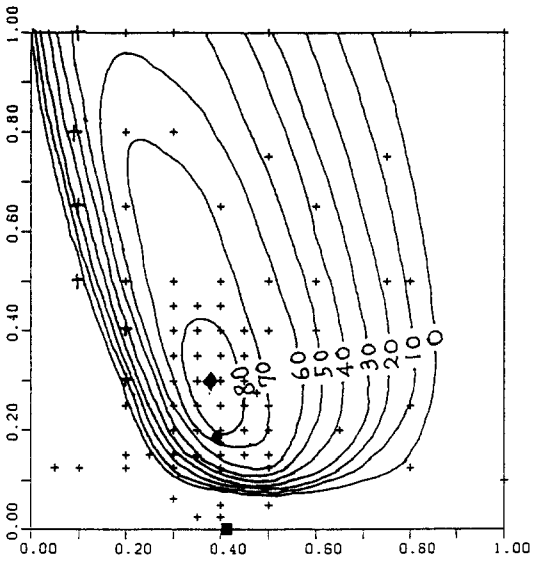

Fig. F3. Contour map of the Bayesian log likelihood with respect to (τ, σ). The contours of the minus value are not shown. Signs of closed diamond, closed circle, and closed squares respectively indicate the location of the local maximum, posterior means with respect to the uniform and ignorance priors.

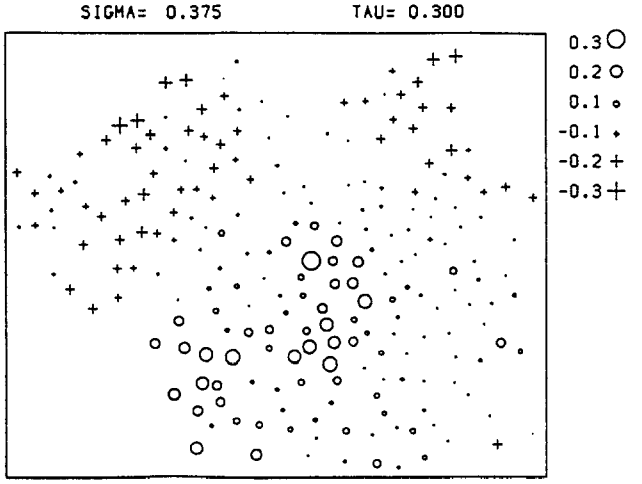

Fig. F4. Posterior mean estimate of θ for $(\tau, \sigma) = (0.375, 0.30)$ which locally maximizes the log Bayesian likelihood.

posterior marginals, with respect to θ_i for each i, was about $0.15 \sim 0.19$ so that its 95% confidence bands entirely includes the constant risk 1.07, the mean of the processed data $z = \{z_i\}$, throughout the region. It is found by the Monte Carlo computation that the local maximum value of the log Bayesian likelihood with the standard error is $\log \mathcal{L}(0.375,\ 0.30) = 82.32 \pm 1.19$. Incidentally, at the

posterior mean estimate of (σ, τ), we have $\log \mathcal{L}(0.391,\ 0.188) = 75.64 \pm 1.75$. Then, these are to be compared with the same quantity for the case of $\tau = 0$ and $\sigma = 0.410$. Unfortunately, the Monte Carlo integration of $\log \mathcal{L}(\sigma, \tau)$ for a very small σ or τ is not so reliable due to the large estimation errors (see Ogata (1989, 1990)). Nevertheless, we expect that the following relation between the Bayesian and ordinary likelihoods holds by assuming the exchange of signs between the limit and the integral:

$$\begin{aligned}
\max_{\sigma} \log \mathcal{L}(\sigma, 0) &\equiv \max_{\sigma} \lim_{\tau \to 0} \log \mathcal{L}(\sigma, \tau) = \max_{\sigma} \log \int \lim_{\tau \to 0} P^*(\theta \mid \sigma, \tau;\ z) d\theta \\
&= \max_{\theta_0, \sigma} \log \int \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\theta_i - z_i)^2 \right\} \prod_{i=1}^{n} \delta_{\theta_0}(\theta_i) d\theta \\
&= \max_{\theta_0, \sigma} \log \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\theta_0 - z_i)^2 \right\} \\
&= \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2} = 85.62,
\end{aligned}$$

where $\delta_{\theta_0}(\theta_i)$ is the Dirac's delta function such that $\delta_{\theta_0}(\theta_i) = 1$ for $\theta_i = \theta_0$, otherwise 0. Eventually, the last term is nothing but the maximum log likelihood assuming $\theta_0 = \theta_i$ for all i, and the maximum likelihood estimate is given by $\hat{\theta} = 1.07$ with the standard error $\hat{\sigma} = 0.41$.

Now we compared the goodness-of-fit of the Bayesian and ordinary parametric models by the ABIC and AIC. In the present case, $\mathrm{ABIC}_1 = -160.6 \pm 2.4$ and $\mathrm{ABIC}_2 = -151.3 \pm 3.5$ corresponding to the local maximum and posterior mean of the Bayesian likelihood, respectively, and $\mathrm{ABIC}_0 = \mathrm{AIC} = -169.2$ to the ordinary maximum likelihood model. Thus, it is seen that the model with equal mortality rate throughout the whole region is superior to the other estimates, as far as the Gaussian likelihood is concerned. Next, for an alternative Bayesian model with the normalized prior

$$(D/\tau^n) \exp\left\{ -\frac{1}{\tau} \sum_{i \sim j} |\theta_i - \theta_j| \right\}$$

of θ for some D, a similar local maximum of the log Bayesian likelihood is attained around $(\sigma, \tau) = (.40,\ .40)$ with $\mathrm{ABIC}_3 = -155.1 \pm 2.6$. Again, the simplest model of constant mortality rate shows the better fit. It was also suggested that the Bayesian log likelihood for these models is at least bimodal: another peak was seen on σ-axis but in a singular manner against the direction of τ-axis, and the corresponding local maximum was attained at the point $(\hat{\sigma},\ 0)$, where $\hat{\sigma}$ was the ordinary maximum likelihood estimate.

Does the log Bayesian likelihood at such singular point (i.e., log maximum likelihood) always have a higher value than anywhere else? Let us consider artificial mortality rates (see Fig. F5) with $\mu_i = 1.07 + 0.5(i - n/2)/n$, $i = 1, 2, \ldots, n$; $n = 216$, on the same contiguities as the electoral wards of Great Manchester in Fig. 15. The artificial data, shown in Fig. F6, is generated according to the

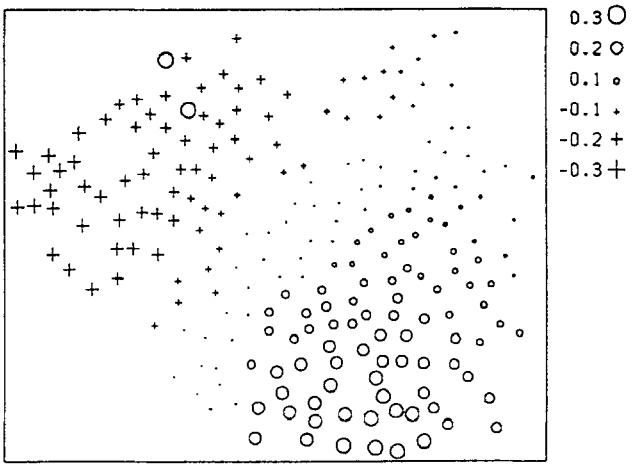

Fig. F5. Pattern of artificial mortality rates with $\mu_i = 1.07 + 0.5(i - n/2)/n$, $i = 1, 2, \ldots, n$; $n = 216$.

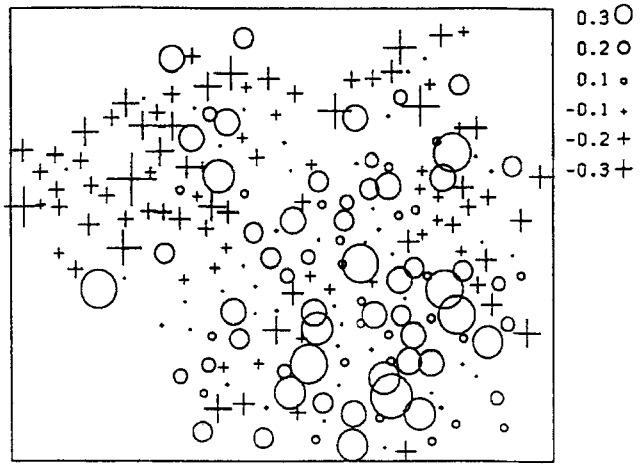

Fig. F6. Artificially generated data according to $\mathcal{N}(\mu_i, \sigma^2)$ with $\sigma = 0.3$: see Fig. F5 for the pattern of $\{\mu_i\}$.

normal distribution $\mathcal{N}(\mu_i, \sigma^2)$ with $\sigma = 0.30$. For this data, I implemented the Gibbs sampler of the posterior in (2) with both the ignorance and uniform priors of the hyperparameters. The respective posterior means of the hyperparameters were $(\sigma, \tau) = (0.259, 0.229)$ and $(\sigma, \tau) = (0.262, 0.206)$. In the present data, I had no problem with the absorbing state, at least up to 100,000 cycles by the Gibbs sampler procedure, even when I started from the possible singular states of the Markov chain. The standard deviation of the posterior marginal of $\theta = \{\theta_i\}$ for each i ranged from 0.09 to 0.14. Then, I had the following values for the log Bayesian likelihood, $\log \mathcal{L}(0.259, 0.229) = 151.40 \pm 2.24$ and $\log \mathcal{L}(0.262, 0.206) = 148.97 \pm 2.29$, while $\log \mathcal{L}(0.311, 0.0) = 144.52$ for the maximum likelihood estimate

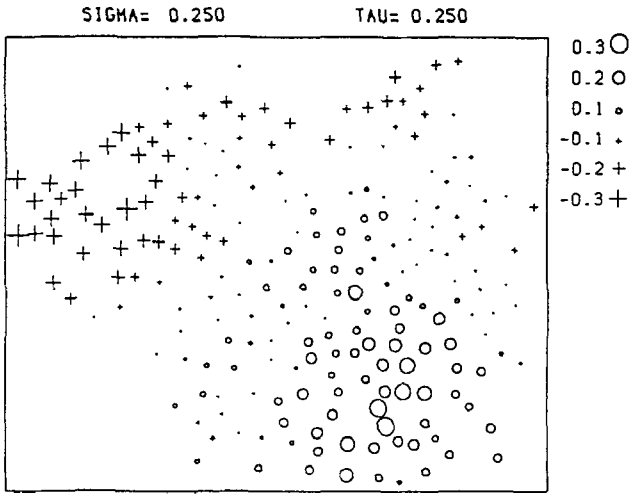

Fig. F7. The mean estimate by the Gibbs sampler of the posterior distribution with the maximum Bayesian estimate.

$\hat{\theta} = 1.063$ of the flat image model. This time, the posterior means for both priors were shown to have better fits than the flat image by comparing the corresponding AIC and ABIC's. Further, the maximum Bayesian likelihood was almost attained by $\log \mathcal{L}(0.250,\ 0.250) = 156.95 \pm 1.93$. Thus, the posterior mean of θ with the last value of hyperparameters is shown in Fig. F7, which is very similar to the true image in Fig. F5. Incidentally, the log likelihood value of the true image was 165.09 which provides, of course, the smallest AIC, compared to the above ABIC's.

Addendum

After having sent the above comments to the author, Professor Besag kindly sent me the original Manchester data of $(y_i,\ c_i)$ $(i = 1, \ldots, n;\ n = 216)$ so that I can analyze the Poisson model in comparison with the Gaussian model discussed above. Now, I have made a tentative Monte Carlo evaluation for the Bayesian likelihood of the same posterior as in equation (4.5) by the authors except assuming $v = 0$ and prior$(\kappa) = 1$. The Monte Carlo evaluation was made for every κ with 0.05 unit, and $\kappa = 0.20$ attained the maximum with $\log \mathcal{L}(0.20) = -556.86 \pm 0.90$. On the other hand, the ordinary maximum log likelihood value with the constant risk $\hat{u}$ was -556.46, showing about the same fit with the Bayesian model. The posterior mean estimate was similar to the one shown in Fig. F2.

Further the Gaussian likelihood, to be compared with the Poisson model on the same original data $(y_i,\ c_i)$ $(i = 1, \ldots, n)$, should be rewritten by

$$\prod_{i=1}^{n} \left\{ \sqrt{2\pi}\sigma c_i e^{\hat{u}} \right\}^{-1} \exp\left\{ \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\theta_i - \frac{y_i}{c_i e^{\hat{u}}} \right)^2 \right\} \tag{4}$$

up to the normalizing constant of the joint probability density with respect to the data $\{y_i\}$. Comparing this with the likelihood function for the processed data

$z_i = y_i/c_i e^{\hat{u}}$ in equation (2) of the present contribution, *the difference of the constant term* $= 649.48$ must be subtracted from the logarithm of the maximum Gaussian ordinary and Bayesian likelihoods obtained above. Then, AIC and ABIC of the Manchester data shows that the Poisson models perform clearly better fits than any of the Gaussian models by the difference about 15.

REFERENCES

Akaike, H. (1977). On entropy maximization principle, *Application of Statistics* (ed. P. R. Krishnaiah), 27–41, North Holland, Amsterdam.

Akaike, H. (1978). A new look at the Bayes procedure, *Biometrika*, **65**, 53–59.

Akaike, H. (1979). Likelihood and Bayes procedure, *Bayesian Statistics* (eds. J. M. Bernardo, M. H. Degroot, D. V. Lindley and A. F. M. Smith), University Press, Valencia, Spain.

Akaike, H. (1987). Factor analysis and AIC, *Psychometrika*, **52**, 317–332.

Good, I. J. (1965). *The Estimation of Probabilities*, M. I. T. Press, Cambridge, Massachusetts.

Ogata, Y. (1989). A Monte Carlo method for high dimensional integration, *Numer. Math.*, **55**, 137–157.

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure, *Ann. Inst. Statist. Math.*, **42**, 403–433.

Ogata, Y. and Katsura, K. (1988). Likelihood analysis of spatial inhomogeneity for marked point patterns, *Ann. Inst. Statist. Math.*, **40**, 29–39.

Ogata, Y., Imoto, M. and Katsura, K. (1991). Three-dimensional spatial variation of b-values of magnitude frequency distribution beneath the Kanto District, Japan, *Geophys. J. Int.*, **10** (in press).

STOPPING THE GIBBS SAMPLER, THE USE OF MORPHOLOGY, AND OTHER ISSUES IN SPATIAL STATISTICS*

ADRIAN E. RAFTERY[1] AND JEFFREY D. BANFIELD[2]

[1] *Department of Statistics and Sociology, GN-22, University of Washington, Seattle, WA 98195, U.S.A.*

[2] *Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, U.S.A.*

1. Introduction

It is a pleasure to congratulate Julian Besag, Jeremy York and Annie Mollié on a superb paper that will surely take its place as yet another of Julian Besag's greatest hits, and as a first hit for the other two authors!

They argue that many spatial statistics problems can appropriately be viewed

* This research was supported by the Office of Naval Research under contracts N-00014-88-K-0265 and N-00014-89-J-1114.

as problems in image restoration, and that image restoration problems are best solved by postulating a Markov random field model, and then calculating the posterior distribution of the quantities of interest using the Gibbs sampler. This is an appealing argument and the examples are encouraging. One possible difficulty arises from the fact that the models may not have the same large-scale properties as the data they are used to analyze, and this raises some questions about the status of the resulting inferences (see Section 3 below).

For the practical implementation of the Bayesian image restoration approach it is important to know how many iterations of the Gibbs sampler are required, and we propose a method for determining this in Section 2. In Section 3 we consider an alternative to the Bayesian image restoration approach for the archeology example, based on mathematical morphology. In Section 4 we discuss several issues in the modeling that underlies the Bayesian image restoration approach: the modeling of spatial dependence, allowing for model uncertainty, the improper posterior distributions that arise in hierarchical Bayes modeling, and the modeling of local dependence between counts when it cannot be assumed that the y_i's are independent given x.

2. How many iterations in the Gibbs sampler?

The authors point out that the Bayesian image restoration approach is not yet feasible for typical images containing 10^5 or 10^6 pixels, although it can be implemented for the problems they consider, involving 100–300 "pixels". The main reason for this is the large number of iterations required by the Gibbs sampler. For instance, in the disease risk example, the authors ran the Gibbs sampler for 11,000 iterations, discarding the first 1,000, and storing every 10th or 20th value thereafter; these numbers were fairly arbitrarily picked initially, although they appeared to give reasonable results. As a practical matter, it would seem desirable to run the Gibbs sampler for the smallest number of iterations necessary to attain a required level of accuracy, and we now outline an approximate way of determining what that is.

The validity of the Gibbs sampler stems from the fact that each cycle of the algorithm corresponds to one step of a Markov chain with stationary transition probabilities and that an ergodic theorem applies for functions of x under certain regularity conditions (Geman, S. and Geman, D. (1984)). This suggests that one generate a single long realization of the Markov chain and base inference on it, which is what the authors have done. By contrast, several authors who have recently applied the Gibbs sampler to more standard statistical problems (Gelfand and Smith (1990), Gelfand *et al.* (1990)) have instead adopted the following algorithm: (i) choose a starting point; (ii) run the Gibbs sampler for T iterations and store only the last iterate; (iii) return to (i). The relationship of this latter algorithm to the underlying theory seems problematical, and here we consider only the case of a single long realization.

We consider the specific problem of producing results such as those in the authors' Figs. 7 and 8, namely the calculation of particular quantiles of the posterior distribution of a function of x. We formulate the problem as follows. Suppose that we want to estimate $P[U \leq u \mid y]$ to within $\pm r$ with probability s, where

U is a function of x. We will find the approximate number of iterations required to do this when the correct answer is q. For example, if $q = .025$, $r = .005$ and $s = .95$, this corresponds to requiring that the cumulative distribution function of the .025 quantile be estimated to within $\pm.005$ with probability .95. This might be a reasonable requirement if, roughly speaking, we wanted reported 95% intervals to have actual posterior probability between .94 and .96. We run the Gibbs sampler for an initial M iterations that we discard, and then for a further N iterations of which we store every k-th (in their Section 4 the authors use $M = 1,000$, $N = 10,000$ and $k = 10$ or 20). Our problem is to determine M, N, and k.

We first calculate U_t for each iteration t, and then form $Z_t = \delta(U_t > u)$, where $\delta(\cdot)$ is the indicator function. $\{Z_t\}$ is a binary 0-1 process that is derived from a Markov chain by marginalization and truncation, but it is not itself a Markov chain. Nevertheless, it seems reasonable to suppose that the dependence in $\{Z_t\}$ falls off fairly rapidly with lag, and hence that if we form the new process $\{Z_t^{(k)}\}$, where $Z_t^{(k)} = Z_{1+(t-1)k}$, then $\{Z_t^{(k)}\}$ will be approximately a Markov chain for k sufficiently large. In what follows, we draw on standard results for two-state Markov chains (see, for example, Cox and Miller (1965)).

Assuming that $\{Z_t^{(k)}\}$ is indeed a Markov chain, we now determine $M = mk$, the number of "burn-in" iterations, to be discarded. Let

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

be the transition matrix for $\{Z_t^{(k)}\}$. The equilibrium distribution is then $\pi = (\pi_0, \pi_1) = (\alpha+\beta)^{-1}(\beta, \alpha)$, and the l-step transition matrix is

$$P^l = \begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix} + \frac{\lambda^l}{\alpha+\beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix},$$

where $\lambda = (1-\alpha-\beta)$. Suppose that we require that $P[Z_m^{(k)} = i \mid Z_0^{(k)} = j]$ be within ε of π_i for $i, j = 0, 1$. If $e_0 = (1, 0)$ and $e_1 = (0, 1)$, then $P[Z_m^{(k)} = i \mid Z_0^{(k)} = j] = e_i P^m$, and so the requirement becomes

$$\lambda^m \leq \frac{\varepsilon(\alpha+\beta)}{\max(\alpha, \beta)},$$

which holds when

$$m = m^* = \frac{\log\left(\dfrac{\varepsilon(\alpha+\beta)}{\max(\alpha, \beta)}\right)}{\log \lambda}.$$

Thus $M = m^*k$.

To determine N, we note that the estimate of $P[U \leq u \mid D]$ is $\bar{Z}_n^{(k)} = (1/n)\sum_{t=1}^{n} Z_t^{(k)}$. For n large, $\bar{Z}_n^{(k)}$ is approximately normally distributed with mean

q and variance $(1/n)\alpha\beta(2-\alpha-\beta)/(\alpha+\beta)^3$. Thus the requirement that $P[q-r \leq \bar{Z}_n^{(k)} \leq q+r] = s$ will be satisfied if

$$n = n^* = \frac{\dfrac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^3}}{\left\{\dfrac{r}{\Phi((1+s)/2)}\right\}^2},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Thus we have $N = kn^*$.

To determine k, we form the series $\{Z_t^{(k)}\}$ for $k = 1, 2, \ldots$. For each k, we compare the first-order Markov chain model with the second-order Markov chain model, and choose the smallest value of k for which the first-order model is preferred. We compare the models by first recasting them as (closed-form) log-linear models for a 2^3 table (Bishop *et al.* (1975)), and then using the BIC criterion, $G^2 - 2\log n$, where G^2 is the likelihood ratio test statistic. This was introduced by Schwarz (1978) in another context and generalized to log-linear models by Raftery (1986); it provides an approximation to twice the logarithm of the Bayes factor for the second-order model. One could also use a non-Bayesian test, but the choice of significance level is problematic in the presence of large samples.

We applied the suggested method to series of 11,000 iterations of the Gibbs sampler for u and v for each of 12 départements based on the data of the authors' Fig. 4; the Gibbs sampler output was kindly supplied to us by Jeremy York. We first give illustrative results with $q = .025$, $r = .005$, $s = .95$, and $\varepsilon = .001$. For all 24 parameters considered, k was either 1 or 2, M was never more than 6, and N was always 9,034 or less. However, for the spatial smoothness parameter κ, the situation was quite different and the requirements of the Gibbs sampler were larger: $k = 5$, $M = 65$ and $N = 42,500$.

The authors' Fig. 6 implicitly requires that the .1 quantile of $e^x = e^{u+v}$ be correct to one decimal place with high probability. This implies, approximately, that for each u and v we specify $q = .1$, $r = .012$ and $s = .95$, which yielded $k \leq 3$, $M \leq 12$ and $N \leq 8,300$ for all 24 parameters considered. In practice, the method would be implemented by first running, say, 1,000 iterations and then deciding on k, M and N on the basis of those. In the present case, this appeared to work quite well.

One conclusion is that the number of iterations required can vary considerably depending on what is being estimated. Here, far more iterations are required for the overall spatial smoothness parameter κ than for the relative risk at an individual node. It does not seem necessary to use only every 10th or 20th iterate, and, indeed, doing so is probably quite wasteful. Indeed, it is not clear that discarding *any* iterates is advantageous, although it does simplify the calculations here. Also, it does not seem necessary to discard the first 1,000 iterates, or anything like it; our calculations never indicated it to be necessary to discard more than the first 65.

We hope that the suggestion made here will allow the Gibbs sampler to be used more efficiently, and hence to make Bayesian image restoration feasible for larger

problems. The computer code used to carry out these calculations is available from Adrian Raftery by electronic mail at *raftery@stat.washington.edu.*

3. Using morphology to locate archeological sites: The EP algorithm

The problems of locating archeological sites in Section 3 can be regarded as one of locating and finding the boundaries of objects in the image, in this case sites of previous activity. For comparative purposes, we apply a different technique based on mathematical morphology, known as the EP algorithm, that was originally developed for locating ice floes in satellite images (Banfield and Raftery (1989)).

The EP algorithm consists of two parts: erosion and propagation. The erosion part of the algorithm, which identifies the potential edge elements, is a standard application of ideas in mathematical morphology (Serra (1982)). The algorithm is iterative and operates on a binary image consisting of objects (sites of activity) on a contrasting background. At the first iteration, if a pixel is classified as "active" and a specified subset of its neighbors is inactive, the pixel is "deactivated" and becomes inactive. At the second iteration, the same operation is performed on the image resulting from the first iteration, and so on. The edge elements consist of the pixels "deactivated" at the first iteration. The propagation part of the EP algorithm keeps track of the site to which an edge pixel belongs by locally propagating the information about edge elements into the interior of the object as it is eroded.

We started the EP algorithm from the naive classification given in the authors' Fig. 1(a), which is, in fact, simple thresholding. The results are shown in Fig. 1. They are quite similar to those obtained from the Bayesian image restoration method, perhaps strikingly so given the noisy appearance of the naive classification in the authors' Fig. 1(a). The pixels where the classifications disagree are pixels where the uncertainty is, in any event, considerable. For almost all these pixels, the posterior probabilities in the authors' Fig. 2 are well away from 0 or 1, and many of them are border pixels for which, as the authors observe, any spatial procedure is necessarily of doubtful value. Note that the EP algorithm uses only the naive classification, and does not, unlike the Bayesian image restoration method, use the full original data.

The EP algorithm has advantages and disadvantages compared to the Bayesian image restoration method: it is much faster but yields less information. The EP algorithm involves only about 10 iterations here, each of which consists only of small integer additions, while the Bayesian image restoration method uses 15,000 iterations each of which involves one exponentiation per pixel. Thus we estimate that the Gibbs iterations take at least 1,000 times, and perhaps 10,000 times as much CPU time as the EP iterations. On the other hand, the Bayesian image restoration method does have the important property of providing a statement of uncertainty in the form of posterior probabilities at each pixel.

However, we do wonder about the precise status of this statement of uncertainty. Markov random field models such as that on which the analysis is based often have a substantial probability of producing infinite one-color patches, in which case typical realizations of $\{p(x)\}$ will not resemble the true scene. This

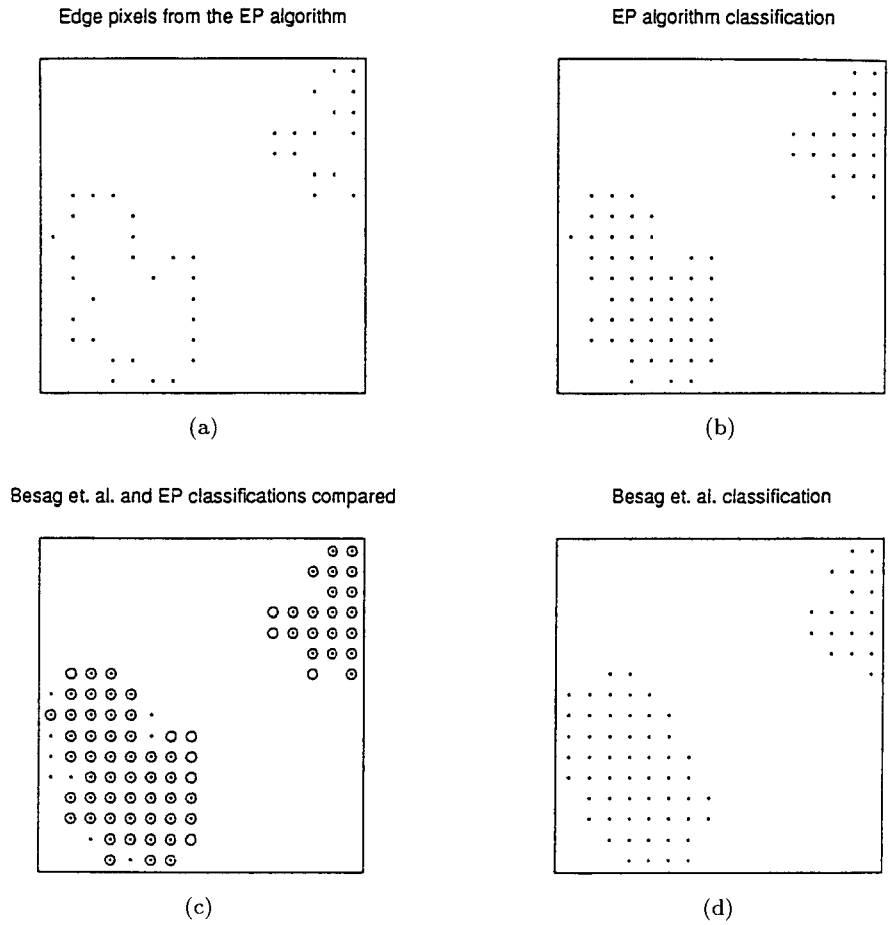

Fig. 1. The EP algorithm applied to the archeology data: (a) The edge pixels identified by the EP algorithm; (b) The classification by the EP algorithm; (c) The EP and Bayesian image restoration classifications superimposed; (d) The Bayesian image restoration classification.

is known as the phase transition phenomenon and is discussed, for example, by Besag (1986). One consequence is that the prior may be heavily concentrated on uniform images, and one might expect this to bias the posterior towards too much uniformity. We would welcome the authors' views on these points.

4. Modeling issues

4.1 *Modeling the spatial dependence*

In the disease mapping example, the authors model the spatial dependence using equation (4.1). This seems sensible in the case of a spatial array that is not too dissimilar to a rectangular array of pixels, such as the French départements. As a historical footnote, the regularity of the administrative map of France is due

to Napoléon, who laid it out in the early nineteenth century in such a way that a man on horseback could reach any part of a département in a day's ride.

However, we wonder whether the specification (4.1) would be as satisfactory for much more irregularly spaced arrays. One example is the Standard Statistical Metropolitan Areas (SMSAs) of the United States, where the "neighbors" are close together in the North-East, but much further apart in the rest of the country.

An alternative but related specification has been developed in geostatistics as the basis for the so-called "kriging" method (Journel and Huijbregts (1978)). This implements the idea that dependence decreases with distance. The form of the dependence is described by the semivariogram, $\gamma(\boldsymbol{h}) = (1/2)\,\mathrm{Var}[u(\boldsymbol{s}) - u(\boldsymbol{s}+\boldsymbol{h})]$, where $u(\boldsymbol{s})$ denotes the value of u at a location $\boldsymbol{s}$. If the covariance function, $C(\boldsymbol{h})$, exists, then $\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h})$. If V is the resulting covariance matrix of the u_i's, and the u_i's are assumed to be jointly Gaussian, then $(u_i \mid u_{-i}) \sim N(\hat{u}_i, \sigma_i^2)$, where $\hat{u}_i = \sum_j a_{ij} u_j$ is the best linear predictor of u_i and σ_i^2 is its variance.

This may provide a more systematic basis for the choice of the quantities $\{a_{ij}\}$, which play a role similar to that of the $\{w_{ij}\}$ in equation (4.1). Another feature is that when, as in the disease risk example, the data correspond to areas rather than to points, the spatial dependence can take account of this explicitly. This is done by postulating a semivariogram for points, as above, and then integrating over areas to provide the corresponding values for the areas (Journel and Huijbregts (1978)). One would then proceed as before.

At first sight, it may seem that such an approach would be computationally prohibitive for even moderate data sets, since, in principle, it requires the inversion of n matrices, each of which is $(n-1) \times (n-1)$. However, if $\gamma(\boldsymbol{h})$ is modeled by a function with "sill", such as the "Mathéron", or spherical, semivariogram,

$$\gamma(\boldsymbol{h}) = \begin{cases} \sigma^2 \left\{ \dfrac{3}{2}\left(\dfrac{|\boldsymbol{h}|}{a}\right) - \dfrac{1}{2}\left(\dfrac{|\boldsymbol{h}|}{a}\right)^3 \right\} & |\boldsymbol{h}| \le a \\ \sigma^2 & |\boldsymbol{h}| > a, \end{cases}$$

then many of the entries in V will be zero, and this can be used to reduce the computation involved in calculating the $\{a_{ij}\}$. Also, most of the $\{a_{ij}\}$ will be close to zero, and they could be set to zero without bad consequences, leading to an effective set of neighbors for each pixel, not necessarily restricted to the contiguous zones. In addition, the $\{a_{ij}\}$ have to be calculated only once for each value of (κ, λ) considered, remaining the same for each iteration of the Gibbs sampler. This suggests advantage to the strategy adopted by the authors for the archeological example, where the parameters of the prior were updated much less frequently than the values at the individual nodes.

these are tentative and untested ideas. However, the notion that the spatial modeling methods developed in geostatistics could be combined with the Bayesian image restoration methods proposed in the present paper may be a potentially fruitful one.

4.2 *Model uncertainty*

Several modeling choices are made in the authors' examples. These include the form of $\phi(z)$, namely whether it should be proportional to z^2 or to $|z|$, which covariates should be included in $t = A\theta$, the way the $\{w_{ij}\}$ are defined, and whether u and v should both be present. The authors, in common with most statistical modelers, have chosen a single model for each data set, and drawn conclusions conditionally on the selected model. This ignores the uncertainty associated with the model selection exercise itself. Analyses conditional on a single selected model fail to take account fully of uncertainty about structure, and so may well underestimate the uncertainty associated with their conclusions, thus, for example, biasing policy choices in favor of policies that rely on more certain information (Hodges (1987)).

Suppose that $m + 1$ models $M_0, M_1, \ldots, M_m$ are being considered. In the present context, these might correspond, for example, to different choices of $\phi(\cdot)$, $\{w_{ij}\}$ and covariates. Then, if Δ is a quantity of interest in the analysis, we can take account of model uncertainty quite simply by basing inference on the unconditional posterior distribution of Δ,

$$p(\Delta \mid y) = \sum_{k=0}^{m} p(\Delta \mid y, M_k) p(M_k \mid y), \tag{1}$$

where $p(M_k \mid y)$ is the posterior probability of model M_k. This is a weighted average of the posterior densities of Δ under each of the models individually, weighted by their posterior probabilities. It will be well approximated by $p(\Delta \mid y, M_{k^*})$, i.e. by conditioning on a single selected model M_{k^*}, only if $p(M_{k^*} \mid y) \approx 1$, or if the posterior distributions of Δ from the models with non-negligible posterior probability are similar.

To calculate the posterior probabilities $p(M_k \mid y)$ we note that

$$p(M_k \mid y) \propto p(y \mid M_k) p(M_k). \tag{2}$$

In equation (2), $p(M_k)$ is the prior probability of M_k and

$$p(y \mid M_k) = \int p(y \mid \theta_k, M_k) p(\theta_k \mid M_k) d\theta_k, \tag{3}$$

where θ_k is the possibly vector parameter of M_k and $p(\theta_k \mid M_k)$ is its prior density. In the present context, this can be implemented by noting that x can also be included in equation (3), yielding

$$p(y \mid M_k) = \iint p(y \mid x, \theta_k, M_k) p(\theta_k \mid M_k) dx d\theta_k. \tag{4}$$

This can be approximated by

$$p(y \mid M_k) \approx \frac{1}{T} \sum_{t=1}^{T} p(y \mid x^{(t)}, \theta_k^{(t)}, M_k), \tag{5}$$

where $\{x^{(t)}, \theta_k^{(t)}\}$ is the result of running the Gibbs sampler to obtain a sample from the *prior* distribution of (x, θ_k). A different approach to finding posterior probabilities using the Gibbs sampler is to include a model indicator as an additional parameter (Carlin *et al.* (1990)).

The implementation of the suggested approach to model uncertainty using equations (1), (2), (4) and (5) does not seem computationally prohibitive. At most, the computation is linear in the number of models that are fully analyzed, multiplying the required CPU time by about $2(m+1)$. However, there are several possible ways of reducing this. For example, the Gibbs sampler could be run in parallel on all the models. Also, an initial short run of equation (5) could be used to identify those models with substantial posterior probability, and a longer run restricted to those models then done to evaluate $p(\Delta \mid y)$ more precisely.

4.3 *Improper posteriors in hierarchical Bayes modeling*

In the authors' equation (4.5), the use of the obvious "non-informative" or scale-invariant prior for κ and λ, $p(\kappa, \lambda) \propto \kappa^{-1}\lambda^{-1}$, leads to an improper posterior distribution. As the authors point out, this is a common feature of Bayesian hierarchical models in general. It arises, for example, even in the simplest normal empirical Bayes model (Morris (1983)) where

$$(y_j \mid \theta_j, V) \sim N(\theta_j, V) \tag{6}$$

$$(\theta_j \mid \mu, A) \sim N(\mu, A) \quad (j = 1, \ldots, N). \tag{7}$$

Then with the standard vague prior, $p(\mu, V, A) \propto V^{-1}A^{-1}$, the posterior $p(\theta_j \mid y)$ is improper. The authors mention the available remedy, in their case, of banning a neighborhood of $\kappa = \lambda = 0$, but instead use the improper prior (4.6), which is intended to approximate an improper uniform prior, but modified to be equal to zero at $\kappa = \lambda = 0$. The use of a uniform prior for a variance-like parameter seems somewhat unsatisfactory intuitively, as it has the disadvantages of an improper prior, without the advantages of scale invariance. Of course, it is not clear that this is really a serious problem in the present application.

Kahn (1990) analyzed this problem in the context of the normal empirical Bayes model specified by equations (6) and (7). He reparameterized the model, setting $S = V + A$ and $T = V/(V + A)$. Then $S = \mathrm{Var}(y_j \mid \mu, S, T)$, and the prior $p(\mu, S, T) \propto S^{-1}$ leads to a proper posterior while retaining the desirable scale-invariant property of the standard prior.

By analogy, this suggests that in the present context we consider $\mathrm{Var}(y_i \mid u_{-i}, \kappa, \lambda)$, which is approximately equal to $(1/c_i + \kappa/n_i + \lambda)$ when κ and λ are small and c_i is large, as here. This suggests specifying the prior in terms of $\sigma = 1/\bar{c} + \kappa/\bar{n} + \lambda$ and $\tau = \lambda/\sigma$, where an overbar denotes the average over all pixels. The natural choice is $p(\sigma, \tau) \propto \sigma^{-1}$, corresponding to $p(\kappa, \lambda) \propto (1/\bar{c} + \kappa/\bar{n} + \lambda)^{-2}$. This is an improper prior which retains, at least roughly, the desired scale-invariance properties, but does not exhibit the behavior near the origin that leads to impropriety. This prior may still lead to the Markov chain defined by the Gibbs sampler having an absorbing state, and one could multiply it by the expression in the authors' equation (4.6) to avoid this.

4.4 *Local dependence between counts*

The authors' model for the disease risk example assumes that, conditionally on the true relative risks x_i, the observed numbers of cases y_i are independent Poisson random variables, arguing that this is usually reasonable when the disease is non-contagious and rare. If the disease is contagious, however, it seems likely that the y_i's will be dependent, even conditionally on x. Even if the disease is non-contagious, it seems possible that the y_i's may be dependent. For example, if a disease is genetically transmitted, this could lead to spatial clustering even when the true risk is constant over space. If such dependence is present, then failing to take account of it seems likely to bias the estimated x_i's away from uniformity and hence, for example, to overstate the size and significance of the effects of covariates.

In the spirit of the authors' paper, the way to take account of such dependence is to model it explicitly. However, how to do this is not immediately obvious. The first possibility that springs to mind is the auto-Poisson model of Besag (1974). The problem with this is that it can represent only negative dependence between neighboring pixels, producing a chessboard-like pattern, which seems unsatisfactory.

We would like to suggest another possible way of representing such spatial dependence between Poisson random variables that draws on ideas first developed in the time series context. The *mixture transition distribution* (MTD) model for a stationary time series $\{Z_t\}$ taking values in an arbitrary space $\boldsymbol{Z}$ is defined as follows (Raftery (1985*a*, 1985*b*), Martin and Raftery (1987)). Suppose that (V_i, W_i) $(i = 1, \ldots, p)$ is a set of bivariate random vectors taking values in $\boldsymbol{Z} \times \boldsymbol{Z}$, with conditional densities $f_i(v \mid w)$ with respect to some measure, where the marginal distribution of V_i is the same as that of W_i for each $i = 1, \ldots, p$. Then the conditional density of Z_t given $Z_{t-1}, \ldots, Z_{t-p}$ is given by

$$p(z_n \mid z_{n-1}, \ldots, z_{n-p}) = \sum_{i=1}^{p} \lambda_i f_i(z_t \mid z_{t-i}), \tag{8}$$

where $\sum \lambda_i = 1$. This can represent time series with arbitrary marginal distributions taking values in arbitrary spaces; in the discrete-valued case it fits data well, is physically motivated and is analogous in several ways to the standard autoregressive model. To specify a Poisson time series model, all that is needed is a bivariate Poisson distribution such as that of Holgate (1964) with mean μ and dependence parameter ζ, which yields

$$f_i(v \mid w) = f(v \mid w) = e^{-(\mu-\zeta)}\mu^{-w} \sum_{h=0}^{\min\{v,w\}} \frac{\binom{w}{h} \zeta^h (\mu-\zeta)^{v+w-2h}}{(v-h)}. \tag{9}$$

When the Poisson means are constant (i.e. the c_i and the x_i are constant) the obvious spatial generalization is just to replace the summation over past values in equation (8) by a summation over the neighbors of pixel n. Then the model is specified in terms of conditional distributions, and the Gibbs sampler machine

can be set in motion as before. One way of generalizing this to the non-stationary situation that we have actually got, where the c_i and the x_i are not constant, is as follows. First postulate the existence of a spatial process $\{z_i^*\}$ defined by equations (8) and (9), corresponding to constant c_i and x_i, and let $F(\cdot)$ be the corresponding Poisson cumulative distribution function. Let $F_i(\cdot)$ be the Poisson cumulative distribution function corresponding to c_i and x_i. Then we model z_i as $z_i = F_i^{-1}(F(z_i^*))$. If the expected counts are very small, then this will not be quite accurate due to the discreteness, and an exact solution may be obtained by allowing the dependence of z_i on z_i^* to be stochastic.

One difficulty with this suggestion is that the conditional distributions defined in this way do not define a valid joint distribution for the y_i's, by the Hammersley-Clifford theorem (Besag (1974)). However, it seems likely that any joint distribution for Poisson random variables that does satisfy the Hammersley-Clifford theorem will not allow a sufficiently broad range of positive dependence. The MTD model suggested here may well have the right *local* conditional dependence structure, while distributions that *do* satisfy the Hammersley-Clifford theorem will often have undesirable large-scale properties as well as unsatisfactory local properties.

Thus one may ask whether conditional distributions such as that specified by the MTD model that do *not* satisfy the Hammersley-Clifford theorem might not, nevertheless, provide useful operational procedures. Besag (1986) refers to this possibility, and we would appreciate the authors' current views on it.

Acknowledgements

The authors are grateful to Julian Besag and Jeremy York for helpful discussions, and also to Jeremy York for computational assistance. Of course, the usual disclaimer applies, as they will be able to make clear in their rejoinder!

References

Banfield, J. D. and Raftery, A. E. (1989). Ice floe identification in satellite images using mathematical morphology and clustering about principal curves, Tech. Report, No. 172, Department of Statistics, University of Washington.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. (1975). *Discrete Multivariate Analysis*, MIT Press, Cambridge, Massachusetts.

Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1990). A Monte Carlo approach to nonnormal and nonlinear state space modeling, Tech. Report, No. 486, Department of Statistics, Carnegie-Mellon University.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*, Chapman and Hall, London.

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

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.

Hodges, J. S. (1987). Uncertainty, policy analysis and statistics (with discussion), *Statist. Sci.*, **2**, 259–291.

Holgate, P. (1964). Estimation for the bivariate Poisson distribution, *Biometrika*, **51**, 241–245.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining Geostatistics*, Academic Press, London.

Kahn, M. J. (1990). Bayes empirical Bayes beta-binomial modeling with covariates, applied to health care policy, Ph. D. dissertation, Department of Statistics, University of Washington (unpublished).

Martin, R. D. and Raftery, A. E. (1987). Robustness, computation and non-Euclidean models, *J. Amer. Statist. Assoc.*, **82**, 1044–1050.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications (with discussion), *J. Amer. Statist. Assoc.*, **78**, 47–65.

Raftery, A. E. (1985*a*). A model for high-order Markov chains, *J. Roy. Statist. Soc. Ser. B*, **47**, 528–539.

Raftery, A. E. (1985*b*). A new model for discrete-valued time series: Autocorrelations and extensions, *Rassegna di Metodi Statistici ed Applicazioni*, **3-4**, 149–162.

Raftery, A. E. (1986). A note on Bayes factors for log-linear contingency tables with vague prior information, *J. Roy. Statist. Soc. Ser. B*, **48**, 249–250.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461–464.

Serra, J. P. (1982). *Image Analysis and Mathematical Morphology*, Academic Press, New York.

B. D. RIPLEY

Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K.

It is very pleasing to see the technology of Bayesian priors development in image analysis being applied by Julian Besag and colleagues for spatial smoothing. I am not sure how widespread such spatial problems are—most of the work I have seen has been interested in explicit spatial covariates rather than a smooth picture. In the one study we did in detail (Mohamed (1988)) when the right covariates were used the spatial effects disappeared. If they had not, Bayesian spatial methods would have provided more believable parameter estimates in the explanatory (Poisson) regression model.

Julian Besag claims that these smaller examples enable computer-intensive methods to be more completely explored. I am not totally convinced. The priors being used often have long-range dependence and, in some cases and to some extent, so do the posteriors. (This is the real message of the examples in Greig *et al.* (1989).) This raises two questions: do we need sizeable examples to reveal all the problems, and how does computation scale with the problem size? The computation *per sweep* of the Gibbs sampler is proportional to the number of sites, but does the rate of convergence depend on the size of the problem? (In a few known cases in statistical physics, the answer is yes.)

We have recently been experimenting with more efficient simulation methods to classify 256×256 images of nematodes into two or three classes (background, nematode, internal organ). The images are noisy and the grey levels of the classes are rather close together. Whereas the site-by-site Gibbs sampler as used here

needs thousands of sweeps to produce (nearly) independent samples, other simulation methods converge much faster without needing much more cpu time per sweep. It is too early to give full details, but some of the methods used are in Ripley (1991). The message is that there is still much to be done both theoretically and in new ideas for simulation methods.

Finally, I would like to endorse the comment in Section 2 on Bayesian methods providing more than point estimates. In my view one of the major attractions of the statistical approach to image analysis/computer vision is the ability to propagate uncertainties to a final decision stage.

REFERENCES

Mohamed, Y. (1988). A study of local area mortality rates in Greater Glasgow, Ph. D. thesis, University of Strathclyde (unpublished).

Ripley, B. D. (1991). Stochastic models for the distribution of rock types in petroleum reserves, *Statistics in the Environmental and Earth Sciences* (eds. A. Walden and P. Guttorp), Griffin (to appear).

D. STOYAN

Section Mathematik, Bergakademie Freiberg, D-9200 Freiberg, Germany

I find the paper quite interesting and stimulating. I agree with the authors that many practical problems can be interpreted as image restoration problems. The idea of the Gibbs sampler seems to have a great practical value. The examples are interesting and convincing; the epidemiological examples demonstrate applications to the practically important case of a non-regular network. I want to ask the authors two questions.

1. You and other statisticians use Markov fields as a helpful tool in image analysis. However, are they really acceptable as stochastic models for random interactions or degradations? In particular, can you report on statistical tests of the goodness-of-fit of the Markov field model in image analysis applications?

2. I was informed by colleagues (Särkkä (1990) and Diggle *et al.* (1990)) that the maximum pseudo-likelihood method (applied to continuous Gibbs point processes) does not give good estimates if the interaction is strong. Probably, the same is true for pixel fields. What can you suggest then? Perhaps a modification of the original pseudo-likelihood method including an improved approximation of spatial dependence?

REFERENCES

Diggle, P. J., Fiksel, T., Grabarnik, P., Ogata, Y., Stoyan, D. and Tanemura, M. (1990). On parameter estimation for pairwise interaction point processes, Tech. Report (submitted).

Särkkä, A. (1990). Applications of Gibbs point processes: pseudo-likelihood estimation method with comparisons, Reports from the Department of Statistics, University of Jyväskylä 10/1990.

REJOINDER

JULIAN BESAG*

Department of Statistics GN-22, University of Washington, Seattle, WA 98195, U.S.A.

I begin by thanking all the discussants for their very valuable comments. Many issues have been raised and I cannot hope to answer them all. I particularly thank those discussants who have provided new analyses of one or both types of example presented in the paper. I must also explain immediately that this reply represents my own views and not necessarily those of my co-authors, both of whom made important contributions to the paper, as part of their graduate studies. I hope this does not appear discourteous but there are a number of logistical constraints, partly brought about by the fact that the three of us are in separate countries and have not met nor worked together for some considerable time now, and partly because of an imminent deadline.

Background

In order to set the paper in context, it may be helpful to acquaint general readers with its background. In the first instance, a version was written for the "Symposium on the Analysis of Statistical Information", held in Tokyo in December 1989, and appears in the proceedings of that meeting. Subsequently, Professor Kitagawa very kindly invited us to submit a modified account, as a discussion paper, to *Ann. Inst. Statist. Math.* The main modification was to be the inclusion of at least one example relating to the mapping of disease. The version that appears in the conference proceedings omits examples from Section 4, though the spoken presentation did include all three. The reason for the omission was that it was not yet clear that Bayesian mapping was at a stage to be put forward as a tool for

* Now at Department of Mathematics and Statistics, University of Newcastle upon Tyne, Newcastle upon Tyne, NE1 7RU, U.K.