Simultaneous Posterior Probability Statements from Monte Carlo Output
Author(s): Leonhard Held
Source: *Journal of Computational and Graphical Statistics*, Vol. 13, No. 1 (Mar., 2004), pp. 20-35
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: https://www.jstor.org/stable/1391142

Simultaneous Posterior Probability Statements From Monte Carlo Output

Leonhard HELD

This article considers the problem of making simultaneous probability statements in multivariate inferential problems based on samples from a posterior distribution. The calculation of simultaneous credible bands is reviewed and—as an alternative—contour probabilities are proposed. These are defined as 1 minus the content of the highest posterior density region which just covers a certain point of interest. We discuss a Monte Carlo method to estimate contour probabilities and distinguish whether or not the functional form of the posterior density is available. In the latter case, an approach based on Rao-Blackwellization is proposed. We highlight that this new estimate has an important invariance property. We illustrate the performance of the different methods in three applications.

Key Words: Contour probability; Highest posterior density region; Monte Carlo; Rao-Blackwell; Simultaneous credible bands; Simultaneous inference.

1. INTRODUCTION

Bayesian hierarchical models are nowadays routinely used to analyze complex problems. Monte Carlo and in particular Markov chain Monte Carlo (MCMC) methods have proven to be the main tool for statistical inference in this class. These methods generate samples from the posterior distribution, which can then be further exploited to calculate posterior summaries.

Typically results are reported using summary statistics of univariate marginal distributions, such as the posterior mean, median, or posterior quantiles. The latter can of course be used to calculate (pointwise) credible intervals. Posterior probabilities may also be reported, for example, one might be interested in the posterior probability that a certain regression parameter is positive.

Summaries of posterior quantities that involve more than one parameter are not as common, but examples exist. For example, one might be interested in the posterior probability

Leonhard Held (formerly Knorr-Held) is Professor, Department of Statistics, University of Munich, Ludwigstrasse 33, 80539 Munich, Germany (E-mail: held@stat.uni-muenchen.de).

Journal of Computational and Graphical Statistics, Volume 13, Number 1, Pages 20–35
DOI: 10.1198/1061860043083

that one parameter is larger than a second one. More generally, some functions of the parameters may be of interest, such as posterior ranks, for example (Goldstein and Spiegelhalter 1996).

However, there seems to be a lack of possibilities to calculate *simultaneous* probability statements relating to a, potentially large, number of parameters. A notable exception was described by Besag, Green, Higdon, and Mengersen (1995, p. 30), a method to calculate *simultaneous credible bands* based on order statistics. Their approach defines such a band as the product of symmetric univariate posterior credible intervals (of the same univariate level) for each parameter; the *simultaneous credible level* is then essentially defined as the proportion of samples which fall *simultaneously* in all intervals. Being based only on ranks, the method is invariant to monotonic transformations of the variables. However, a difficulty is that the simultaneous credible band is defined as a product of univariate intervals, hence is by construction restricted to be hyperrectangular. We will review the method by Besag et al. (1995) in Section 2.1.

Often there is interest in determining if a certain point in the parameter space is supported by the posterior distribution. For example, in hierarchical models with latent Gaussian random effects having prior mode zero, one may ask if the posterior distribution of all random effects is "significantly" different from the zero vector. An initial approach would be to investigate the corresponding univariate credible intervals of some predefined level separately, but it is unclear how this translates to a simultaneous probability statement, if the random effects are dependent. Alternatively, one could calculate simultaneous credible bands on various levels and determine the smallest level, such that the point of interest is contained in the simultaneous credible band. We will test this approach in several applications.

As a promising alternative, we propose to use (multivariate) highest posterior density (HPD) regions to check the support of the posterior distribution for a certain parameter vector of interest. We need some notation to establish our ideas further. Let $\mathbf{y}$ be the observed data and $\boldsymbol{\theta} \in \mathcal{R}^p$ be the unknown parameter vector of interest. The total number of parameters in the model may often exceed p, the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ is then of course obtained by suitable marginalization. Let $\boldsymbol{\eta}$ denote these "nuisance" parameters, with $\boldsymbol{\eta} \in \mathcal{R}^q$, say.

We may now follow Box and Tiao (1973, p. 125) and ask whether a specific parameter point of interest $\boldsymbol{\theta}_0$ lies inside or outside a HPD region with some predefined (simultaneous) credible level $1-\alpha$. This is the case, if and only if

$$P\{p(\boldsymbol{\theta}|\mathbf{y}) > p(\boldsymbol{\theta}_0|\mathbf{y})|\mathbf{y}\} \leq 1-\alpha.$$

Notice that the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ is treated here as a random variable, a function of the random vector $\boldsymbol{\theta}|\mathbf{y}$.

This question can now easily be reversed to define the (posterior) *contour probability* $P(\boldsymbol{\theta}_0|\mathbf{y})$ of $\boldsymbol{\theta}_0$ as 1 minus the content of the HPD region of $p(\boldsymbol{\theta}|\mathbf{y})$ which just covers $\boldsymbol{\theta}_0$:

$$P(\boldsymbol{\theta}_0|\mathbf{y}) = P\{p(\boldsymbol{\theta}|\mathbf{y}) \leq p(\boldsymbol{\theta}_0|\mathbf{y})|\mathbf{y}\}. \tag{1.1}$$

Box and Tiao (1973) did not explicitly define contour probabilities, but made an identical reverse assessment ("what is the posterior evidence against a given point based on HPD regions," Example, 2.12.4, pp. 136–138). An alternative name for a contour probability could be *posterior* or *Bayesian p value*; motivated by well-known analogies between Bayesian and classical inference concepts, as reviewed thoroughly by Box and Tiao (1973). However, Bayesian p values have been defined in many different ways; see, for example, Gelman, Carlin, Stern, and Rubin (1995) or Carlin and Louis (2000), so we stick to our less controversial terminology.

Section 2.2 describes methods for Monte Carlo estimation of contour probabilities based on samples from the posterior distribution. We will distinguish whether or not the functional form of the (unnormalized) density $p(\boldsymbol{\theta}|\mathbf{y})$ is known. In the former case, Monte Carlo estimation is straightforward, but unfortunately this situation is rare in practice. If the functional form is unknown, we propose a method based on Rao-Blackwellization, that is, make use of the corresponding *conditional* density $p(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathbf{y})$ in order to estimate the marginal density $p(\boldsymbol{\theta}|\mathbf{y})$.

Section 3 considers three applications. The first is a simple linear regression problem, where analytic calculation of contour probabilities is possible. With this example we highlight the differences between simultaneous credible bands and contour probabilities. We then illustrate the general applicability of the proposed method to calculate contour probabilities in two nontrivial examples, one on nonparametric trend estimation of a Poisson time series, and the other one taken from spatial epidemiology. In these two examples $\boldsymbol{\theta}$ is of dimension $p = 52$ and $p = 366$, respectively. Finally, Section 4 gives a brief discussion.

2. MONTE CARLO ESTIMATION OF SIMULTANEOUS POSTERIOR PROBABILITY STATEMENTS

Throughout we will assume that we have a sufficiently large sample $(\boldsymbol{\theta}^{(1)}, \boldsymbol{\eta}^{(1)}), \ldots, (\boldsymbol{\theta}^{(n)}, \boldsymbol{\eta}^{(n)})$ from $p(\boldsymbol{\theta}, \boldsymbol{\eta}|\mathbf{y})$, either by Monte Carlo, or Markov chain Monte Carlo (MCMC). Here $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ are samples from the parameters of interest and $\boldsymbol{\eta}^{(1)}, \ldots, \boldsymbol{\eta}^{(n)}$ are samples from additional nuisance parameters. Generally it is not required that the sample size n is the same for the nuisance parameters and the parameters of interest, but we will assume this for simplicity. Also, samples from $\boldsymbol{\theta}|\mathbf{y}$ and $\boldsymbol{\eta}|\mathbf{y}$ may have been obtained independently from separate runs. However, the most common situation in practice is that these samples have been obtained from the same run, and this is the situation we have in mind.

2.1 Monte Carlo Estimation of Simultaneous Credible Bands

The approach proposed by Besag et al. (1995) starts with sorting and ranking the samples separately for each parameter of interest θ_i, $i = 1, \ldots, p$. Let $\theta_i^{[j]}$ denote the corresponding order statistic and $r_i^{(j)}$ the rank of $\theta_i^{(j)}$. Let j^* be the smallest integer such that the hyperrectangular defined by

$$\left[\theta_i^{[n+1-j^*]}, \theta_i^{[j^*]}\right], \quad i = 1, \ldots, p \tag{2.1}$$

contains at least k of the n values $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$. Besag et al. (1995) pointed out that j^* is equal to the kth order statistic of the set

$$\left\{\max\left\{n+1-\min_i r_i^{(j)}, \max_i r_i^{(j)}\right\}, j=1,\ldots,n\right\}. \tag{2.2}$$

By construction, the credible band (2.1) will then contain (at least) $100k/n\%$ of the empirical distribution. We add some further comments:

1. For $p = 1$ and $k \in \{n, n-2, \ldots, 2\}$ (assuming n is even), (2.1) reduces to an ordinary univariate symmetric credible interval; the set (2.2) is then simply (after ordering) $\{\frac{n}{2}+1, \frac{n}{2}+1, \frac{n}{2}+2, \frac{n}{2}+2, \ldots, n-1, n-1, n, n\}$. The same holds (with minor modifications) for n odd.
2. Besag et al. (1995) noted that the method is slightly conservative in the sense that, for n fixed, the credible band (2.1) will typically contain slightly more that $100k/n\%$ of the empirical distribution because of ties in the set (2.2). This problem increases to an extent with p increasing, because the number of ties will then typically increase. However, the method is still consistent as $n \to \infty$.
3. Empirical evidence shows that these credible bands tend to get rather unstable for large levels (larger than 0.95, say), that is, exhibit a lot of Monte Carlo error, even if n is quite large.
4. From a computational point of view, the method requires the storage of all samples from all components of $\boldsymbol{\theta}$ which can be prohibitive is p is large. Furthermore, the sorting and ranking of the samples from each component can be computationally intensive, if n is large.
5. Ranking and sorting still has to be done only once, even if simultaneous credible bands are required on more than one level. Only the set (2.2), the ordered samples $\theta_i^{[j]}$ and the ranks $r_i^{(j)}$ need to be available to calculate simultaneous credible bands at additional levels. The computational effort to calculate these additional simultaneous credible bands is negligible, compared to the initial ranking and sorting.
6. This opens up the possibility to estimate *pseudo contour probabilities* for certain points of interest $\boldsymbol{\theta}_0$ in the parameter space. The *pseudo contour probability* $Q(\boldsymbol{\theta}_0)$ is simply defined as 1—the smallest credible level, for which $\boldsymbol{\theta}_0$ is (still) contained in the corresponding simultaneous credible band. This can be estimated by first estimating simultaneous credible bands for all $k = 1, \ldots, n$. The estimated pseudo contour probability $\hat{Q}(\boldsymbol{\theta}_0)$ is then $1 - k^*/n$, where k^* is the smallest integer such that $\boldsymbol{\theta}_0$ is contained in the corresponding estimated simultaneous credible band.

2.2 MONTE CARLO ESTIMATION OF CONTOUR PROBABILITIES

Suppose now we know the functional form of the (unnormalized) density $p(\boldsymbol{\theta}|\mathbf{y})$. This is not always the case, in particular if marginalization is necessary to calculate $p(\boldsymbol{\theta}|\mathbf{y})$ and we will later relax this assumption. We can then estimate the contour probability (1.1) by the proportion of posterior samples with density value smaller or equal $p(\boldsymbol{\theta}_0|\mathbf{y})$:

$$\hat{P}(\boldsymbol{\theta}_0|\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{p(\boldsymbol{\theta}^{(i)}|\mathbf{y}) \leq p(\boldsymbol{\theta}_0|\mathbf{y})\}. \tag{2.3}$$

This is the obvious Monte Carlo version of Equation (1.1) and was proposed by Wei and Tanner (1990) (see also Tanner 1996) in the related context of data augmentation. It is an unbiased estimate of $P(\boldsymbol{\theta}_0|\mathbf{y})$ with variance $P(\boldsymbol{\theta}_0|\mathbf{y})(1 - P(\boldsymbol{\theta}_0|\mathbf{y}))/n$ if the samples are independent.

Let us now consider the more general case, where the functional form of $p(\boldsymbol{\theta}|\mathbf{y})$ is not available. It is clear that Equation (1.1) is still valid if we replace the density function $p(\boldsymbol{\theta}|\mathbf{y})$ by any arbitrary, but (strictly) monotonic increasing function of $p(\boldsymbol{\theta}|\mathbf{y})$. For example, we could take the log-density. However, such a function is typically not available either. Alternatively, we may try to estimate $p(\boldsymbol{\theta}|\mathbf{y})$ based on some method of density estimation.

Gelfand and Smith (1990) proposed to estimate $p(\boldsymbol{\theta}|\mathbf{y})$ by averaging the conditional densities $p(\boldsymbol{\theta}|\boldsymbol{\eta}^{(j)}, \mathbf{y})$, $j = 1, \ldots, n$. They showed, using the Rao-Blackwell theorem, that

$$\hat{p}(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{n}\sum_{j=1}^{n} p(\boldsymbol{\theta}|\boldsymbol{\eta}^{(j)}, \mathbf{y}) \tag{2.4}$$

is a more efficient density estimate than any method based on the samples $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(n)}$ itself, for example based on the kernel or nearest neighbor method (Silverman 1986). Furthermore the Rao-Blackwell estimate does not require us to choose a smoothing parameter and automatically ensures that the integral over $\hat{p}(\boldsymbol{\theta}|\mathbf{y})$ equals unity.

This suggests that $p(\boldsymbol{\theta}^{(i)}|\mathbf{y})$ and $p(\boldsymbol{\theta}_0|\mathbf{y})$ in (2.3) be replaced with the Rao-Blackwell estimates $\hat{p}(\boldsymbol{\theta}^{(i)}|\mathbf{y})$ and $\hat{p}(\boldsymbol{\theta}_0|\mathbf{y})$ defined as in (2.4). However, we are not really interested in the density function $p(\boldsymbol{\theta}|\mathbf{y})$ evaluated at *all* possible values of $\boldsymbol{\theta}$, all we need is an estimate of the density—or of any arbitrary monotonic function of the density—at the sample points $\boldsymbol{\theta}^{(i)}$, $i = 1, \ldots, n$ and at the reference point $\boldsymbol{\theta}_0$. On the other hand, since we could in fact have chosen any strictly monotonic function of $p(\boldsymbol{\theta}|\mathbf{y})$ rather than $p(\boldsymbol{\theta}|\mathbf{y})$ itself to *define* contour probabilities in Equation (1.1), it seems important to use an estimate which is *invariant* to such arbitrary choices.

We therefore suggest to replace the mean in (2.4) with the median, which leads to the following (modified) Rao-Blackwell estimate of (1.1) if $p(\boldsymbol{\theta}|\mathbf{y})$ is unknown:

$$\hat{P}_{\text{RB}}(\boldsymbol{\theta}_0|\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\text{med}_j\ p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\eta}^{(j)}, \mathbf{y}) \leq \text{med}_j\ p(\boldsymbol{\theta}_0|\boldsymbol{\eta}^{(j)}, \mathbf{y})\right\}; \tag{2.5}$$

here med_j denotes the median of all values indexed by $j = 1, \ldots, n$. This new estimate is now invariant to any such arbitrary choices in the definition of contour probabilities, just as the estimate (2.3) is. This would not be the case for the Rao-Blackwell estimate based on the mean. The price we pay is that the integral over $\hat{p}(\boldsymbol{\theta}|\mathbf{y})$ will in general no longer equal unity, however, this is not important in our context since we are not interested in the whole density function.

3. APPLICATIONS

3.1 LINEAR REGRESSION

For illustration we first consider a (deliberately) simple example taken from Box and Tiao (1973, sec. 2.7.3). The example highlights nicely the differences between the two approaches.

Consider the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\theta}$ is the unknown regression coefficient of dimension $p = 2$ and the error terms ϵ_i are assumed to be independently normal distributed with mean zero and (unknown) precision η. Data are collected for $m = 18$ observations where the response variable y_i corresponds to weight measurements of two specimens (either one of them or both together). The design matrix $\mathbf{X}$ is defined such that the two components of $\boldsymbol{\theta}$ can be interpreted as the true weight of the two specimens.

A noninformative prior $p(\boldsymbol{\theta}, \eta) \propto \eta^{-1}$ is assumed for the unknown parameters and hence the posterior distribution is of the usual normal-gamma form:

$$p(\boldsymbol{\theta}, \eta|\mathbf{y}) = p(\eta|\mathbf{y})p(\boldsymbol{\theta}|\eta, \mathbf{y}),$$

where $p(\eta|\mathbf{y})$ is gamma distributed with parameters $(m-p)/2$ and $s^2 \cdot (m-p)/2$; here s^2 is the classical (unbiased) estimate of the variance η^{-1}. Furthermore, $p(\boldsymbol{\theta}|\eta, \mathbf{y})$ is normal with mean equal to the least squares estimate $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and covariance matrix $\eta^{-1}(\mathbf{X}'\mathbf{X})^{-1}$. We can thus easily sample from this posterior distribution by first sampling $\eta^{(i)}$ from $p(\eta|\mathbf{y})$ and then sampling $\boldsymbol{\theta}^{(i)}$ from $p(\boldsymbol{\theta}|\eta^{(i)}, \mathbf{y})$. Figure 1 displays 10,000 samples from $p(\boldsymbol{\theta}|\mathbf{y})$, which we have used to estimate simultaneous credible bands and contour probabilities. Note that the components are negatively correlated with correlation -0.58.

Of course, due to standard linear regression theory we know that $p(\boldsymbol{\theta}|\mathbf{y})$ follows a multivariate t-distribution with parameters $\hat{\boldsymbol{\theta}}$, $s^2(\mathbf{X}'\mathbf{X})^{-1}$ and $m-p$ degrees of freedom. The contour probability of any point $\boldsymbol{\theta}_0$ can thus also be calculated analytically based on the fact that

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})/(p \cdot s^2),$$

a monotonic function of $p(\boldsymbol{\theta}|\mathbf{y})$, is F-distributed with p and $m-p$ degrees of freedom. We have now estimated pseudo contour probabilities $\hat{Q}(\boldsymbol{\theta}_0)$ based on simultaneous credible bands, contour probabilities $\hat{P}(\boldsymbol{\theta}_0)$ and $\hat{P}_{\mathrm{RB}}(\boldsymbol{\theta}_0)$, based on Equation (2.3) and (2.5), and—for comparison—have also calculated the true contour probabilities $P(\boldsymbol{\theta}_0)$ based on the quantiles of the F-distribution mentioned earlier. Note that (2.3) requires the evaluation of the density of $p(\boldsymbol{\theta}|\mathbf{y})$, the density of a multivariate t-distribution. The Rao-Blackwell estimate (2.5) requires the evaluation of the density $p(\boldsymbol{\theta}|\eta, \mathbf{y})$ which is multivariate normal. All the different quantities have been calculated for 441 different values of $\boldsymbol{\theta}_0$ on a 21×21 grid defined by all combinations of $\theta_1 \in \{80, 82, \ldots, 120\}$ and $\theta_2 \in \{110, 113, \ldots, 140\}$.

The results can be summarized as follows: The contour probability estimates based on

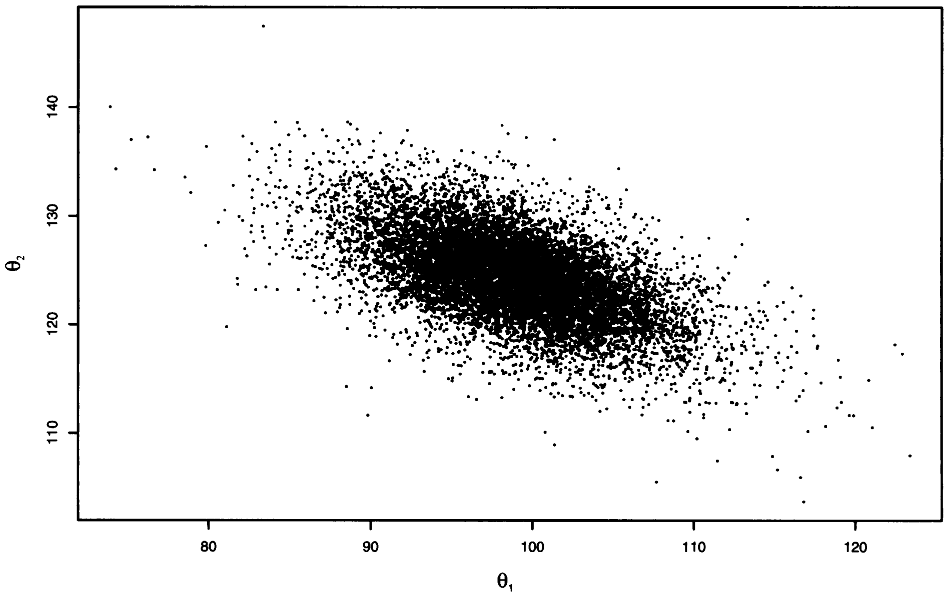

Figure 1. *10,000 samples from* $p(\boldsymbol{\theta}|\mathbf{y})$.

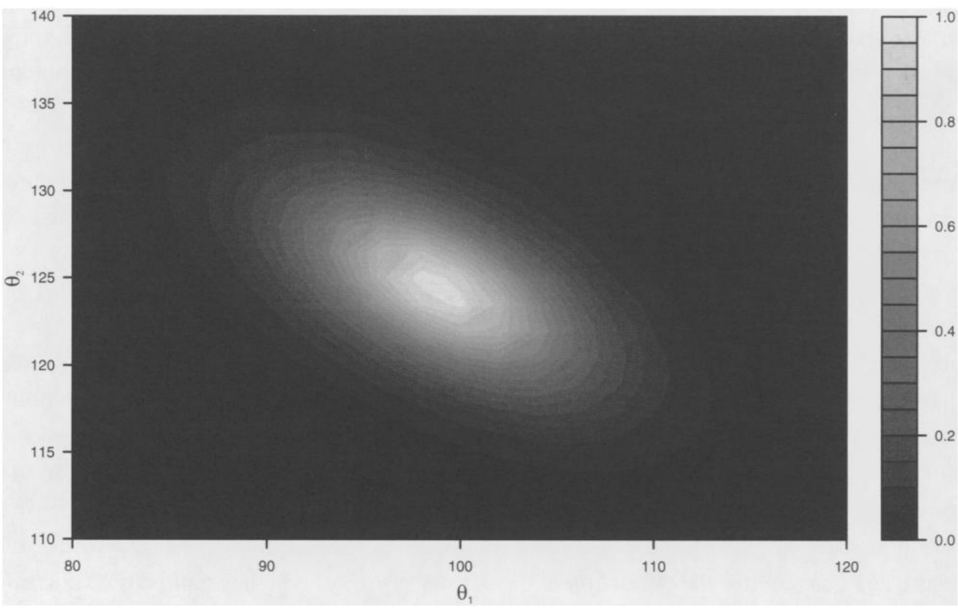

Figure 2. *Contours of marginal posterior distribution* $p(\boldsymbol{\theta}|\mathbf{y})$ *based on estimated contour probabilities on a 21* $\times$ *21 grid.*

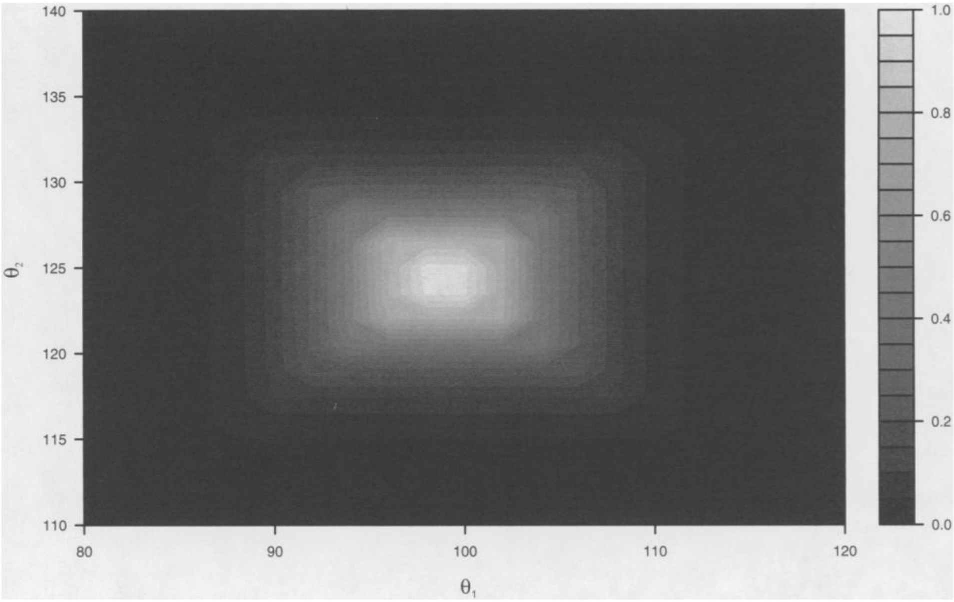

Figure 3. Interpolated pseudo contour probabilities based on a 21 × 21 grid.

Equations (2.3) and (2.5) gave completely identical results for all 441 values of $\boldsymbol{\theta}_0$. However, note that $\hat{P}_{\mathrm{RB}}$ is computationally more demanding, because of the additional sorting required to calculate the median. Nevertheless, this indicates that the Rao-Blackwell estimate is as precise as the original estimate. Incidentally, the Rao-Blackwell estimate based on the mean of the density values also produced virtually identical results.

Figure 2 displays the corresponding contours of $p(\boldsymbol{\theta}|\mathbf{y})$ based on these estimates, using the R-function `filled.contour`. These contours show a very good agreement with the corresponding true contours (not displayed). In contrast, the estimated pseudo contour probabilities shown in Figure 3 are severely biased. Quite obviously from Figure 3, this is a consequence of the fact that simultaneous credible bands are by construction restricted to be rectangular.

Further insight can be gained from Figure 4, which compares the estimated contour and pseudo contour probabilities with the corresponding true contour probabilities in two scatterplots, again for all 441 points. There is a very good agreement between $P(\boldsymbol{\theta}_0)$ and $\hat{P}(\boldsymbol{\theta}_0)$, which is not surprising, because we know that the standard error of the estimates $\hat{P}(\boldsymbol{\theta}_0)$ can never be larger than $\sqrt{0.25/10000} = 0.005$. However, the estimated pseudo contour probabilities are extremely misleading. A comparison with Figures 2 and 3 reveals that underestimation occurs when $\boldsymbol{\theta}_0$ is located in the two tails of the posterior $p(\boldsymbol{\theta}|\mathbf{y})$, while severe overestimation occurs if $\boldsymbol{\theta}_0$ is more orthogonal to the tails. This is a consequence of the rectangular form of the corresponding simultaneous credible bands and suggests that such biases will increase with increasing correlation of the components of $\boldsymbol{\theta}|\mathbf{y}$.

3.2 NONPARAMETRIC MODELING OF TEMPORAL TRENDS

Here we analyze a time series of weekly incidence counts of an infectious disease y_i,

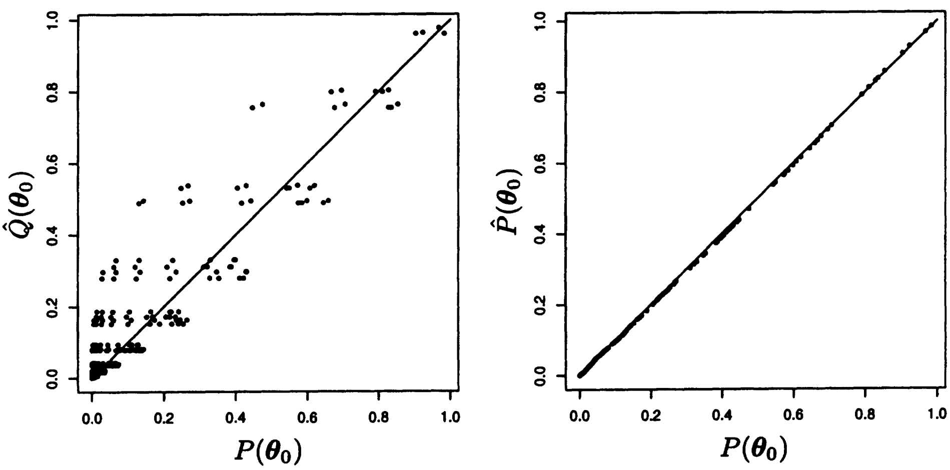

Figure 4. Scatterplots of estimated pseudo contour probabilities (left) and estimated contour probabilities (right) versus true contour probabilities.

$i = 1, \dots, p = 52$, taken from Kashiwagi and Yanagimoto (1992, Fig. 1b). We assume that $y_i|\theta_i$ is conditionally independent Poisson with mean $\exp(\theta_i)$ and assume a second order random walk smoothing prior for the log rate θ_i with flat priors on the initial parameters θ_1 and θ_2, that is,

$$p(\boldsymbol{\theta}|\eta) \propto \eta^{\frac{p-2}{2}} \exp\left(-\frac{\eta}{2}\sum_{i=2}^{p-1}(\theta_{i-1} - 2\theta_i + \theta_{i+1})^2\right).$$

Fahrmeir and Knorr-Held (2000) outlined that this model is the discrete time analogue of a cubic smoothing spline. For the precision parameter η we adopt a gamma prior with parameters $a = 1.0$ and $b = 0.005$. We stored 10,000 samples from the posterior distribution from a long run of length 1,000,000, storing every 100th iteration (following some burn-in period), using conditional prior block proposals as suggested by Knorr-Held (1999).

Figure 5 displays posterior median and 50, 80, and 95% pointwise credible regions for both the rate $\exp(\boldsymbol{\theta})$ and the log rate $\boldsymbol{\theta}$. On the former the data are superimposed. Furthermore, the last panel in Figure 5 displays the corresponding simultaneous credible bands for $\boldsymbol{\theta}$, again on the 50, 80, and 95% level. We note that the posterior correlation between successive components θ_i and θ_{i+1} is between 0.89 and 0.96 (median equal 0.94). In the light of the results in Section 3.1, these simultaneous credible bands have thus to be taken with great care.

Suppose now that we are interested in the simultaneous posterior support for a constant log rate $\boldsymbol{\theta}_c = (c, \dots, c)$. Note that even the 95% simultaneous credible band does not allow any horizontal line to fit in. In fact, further calculations of simultaneous credible bands on all levels on a 0.1% grid showed that only simultaneous credible bands of level 98.3% or larger allow a horizontal line to fit in. For the level 98.3%, the value c must be between 0.742 and 0.796 (hence the estimated pseudo contour probability of $\boldsymbol{\theta}_c$ with $c \in [0.742, 0.796]$ is 1.7%). We note that these numbers may have substantial Monte Carlo error attached to

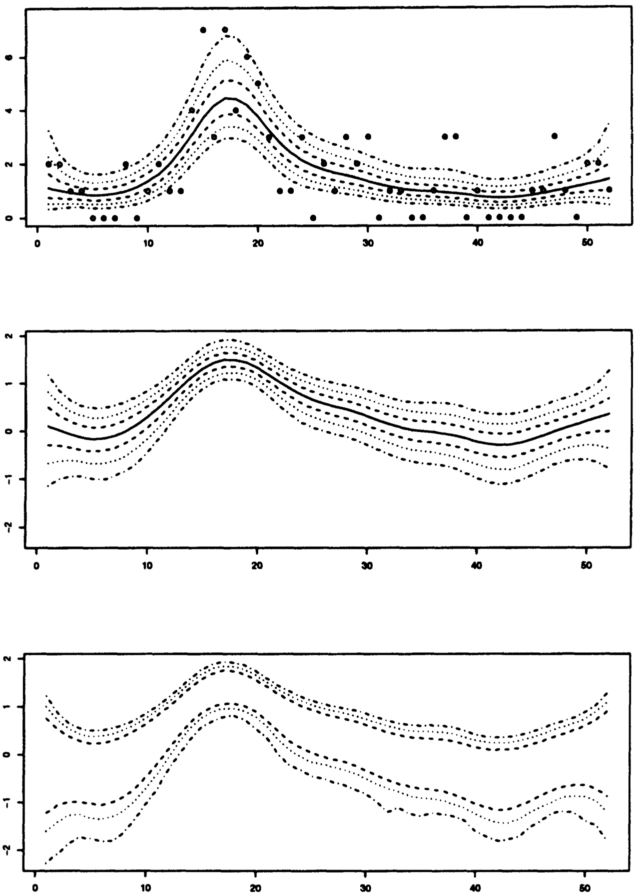

Figure 5. Pointwise credible intervals for exp($\boldsymbol{\theta}$) (top panel) with data superimposed, pointwise credible intervals for $\boldsymbol{\theta}$ (middle), and simultaneous credible bands for $\boldsymbol{\theta}$ (bottom).

it, because the estimated simultaneous credible bands tend to get rather unstable for levels close to 100%.

Alternatively, it may seem more natural to consider samples from the pairwise differences $\theta_i - \theta_{i-1}$, $i = 2, \ldots, p$, calculate the corresponding simultaneous credible bands for all levels, and see if they cover zero. Interestingly, all bands of level 76.2% and larger do indeed cover the zero vector, and hence the estimated pseudo contour probability for this question is 23.8%, quite different from the 1.7% obtained by the approach above.

Finally we have also estimated the corresponding contour probabilities $P(\boldsymbol{\theta}_c|\mathbf{y})$ with our new method. Because the functional form of the marginal posterior density

$$p(\boldsymbol{\theta}|\mathbf{y}) = \int p(\boldsymbol{\theta}, \eta|\mathbf{y}) d\eta$$

is not available, we use the proposed Rao-Blackwell estimate $\hat{P}_{\mathrm{RB}}$ instead. The required conditional density is

$$p(\boldsymbol{\theta}|\eta, \mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta}, \eta)}{p(\eta|\mathbf{y})p(\mathbf{y})} = \frac{\prod_{i=1}^{p} p(y_i|\theta_i)p(\boldsymbol{\theta}|\eta)p(\eta)}{p(\eta|\mathbf{y})p(\mathbf{y})}.$$

Here we can ignore the term $p(\mathbf{y})$, but must retain all other terms, because they either depend on $\boldsymbol{\theta}$, on η, or on both. The logarithm of this density is therefore (ignoring further additive constant, which do not depend on $\boldsymbol{\theta}$ or η):

$$\log p(\boldsymbol{\theta}|\eta, \mathbf{y}) = \sum_{i=1}^{p}(y_i\theta_i - \exp(\theta_i)) - \frac{\eta}{2}\sum_{i=2}^{p-1}(\theta_{i-1} - 2\theta_i + \theta_{i+1})^2 + \frac{p+2a-4}{2}\log(\eta) - b\eta - \log p(\eta|\mathbf{y}). \quad (3.1)$$

Note that the evaluation of $\hat{P}_{\text{RB}}$ only requires the storage of samples of η, $\sum(\theta_{i-1} - 2\theta_i + \theta_{i+1})^2$ and $\sum_{i=1}^{p}(y_i\theta_i - \exp(\theta_i))$, but not of all samples from $\boldsymbol{\theta}$. However, a further problem occurs because we do not know $p(\eta|\mathbf{y})$, which needs to be evaluated at all samples $\eta^{(1)}, \ldots \eta^{(n)}$. We could easily estimate $p(\eta|\mathbf{y})$ based on the samples from $\eta|\mathbf{y}$, but will use the more efficient Rao-Blackwell density estimate

$$\hat{p}(\eta|\mathbf{y}) = \sum_{i=1}^{n} p(\eta|\boldsymbol{\theta}^{(i)}, \mathbf{y}), \quad (3.2)$$

because we know that $p(\eta|\boldsymbol{\theta}, \mathbf{y})$ is the density of a gamma distribution with parameters $a + 0.5 \cdot (p-2)$ and $b + 0.5 \cdot \sum_{i=2}^{p-1}(\theta_{i-1} - 2\theta_i + \theta_{i+1})^2$. The estimation of $p(\eta|\mathbf{y})$ is done in a pre-step, prior to the estimation of the contour probabilities.

Figure 6 displays the estimated contour probabilities $\hat{P}_{\text{RB}}(\boldsymbol{\theta}_c|\mathbf{y})$ as a function of c. These estimates peak for $c = 0.54$ with $\hat{P}_{\text{RB}}(\boldsymbol{\theta}_{c=0.54}|\mathbf{y}) = 0.76$. Incidentally, the maximum likelihood estimate in the simple constant risk model $y_i \sim \text{Poisson}(\exp(\theta))$ is $\hat{\theta} = \log(\sum_i y_i/p) = \log(89/52) = 0.5374$, hence in perfect agreement with the peak of the contour probabilities at $c = 0.54$. This is because in (3.1) the second term equals zero for all $\boldsymbol{\theta}_c$ and therefore the only term remaining, which depends on $\boldsymbol{\theta}_c$, is the log-likelihood. For illustration, Figure 6 also displays the Rao-Blackwell estimates based on the *mean* rather than the median. Although the estimates based on the mean of the log density values are somewhat close to the median-based ones, the estimates based on the mean of the density values are far off. Just to emphasize it again, the median-based estimates are of course the same, whether we use the density or the log density. This comparison highlights two deficiencies of the Rao-Blackwell estimate based on the mean: (a) the problem of outliers in the density values and (b) the lack of invariance.

It is quite striking that the maximized contour probability is rather large, although the corresponding simultaneous credible bands for $\boldsymbol{\theta}$ suggest that a constant risk model is rather unlikely a posteriori. It seems that such severe underestimation of the contour probability occurs, because all points $\boldsymbol{\theta}_c$ are located in long tails of the posterior distribution. This can be best seen by looking at successive parameters θ_i and θ_{i+1}. There will be only minor differences in terms of the posterior mean of these two parameters; therefore the

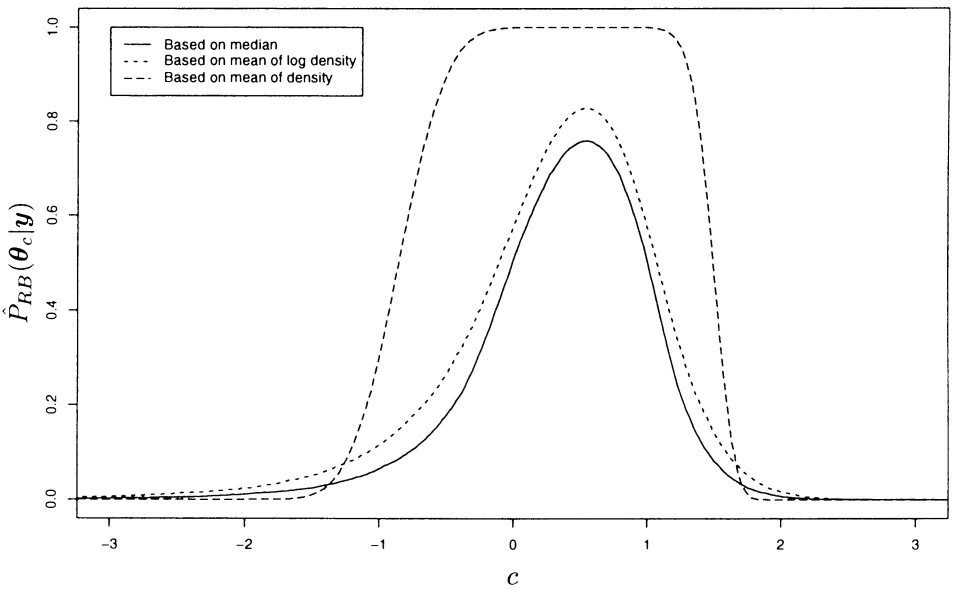

Figure 6. Estimated contour probabilities.

points of interest (c, c) are located along the tails of this bivariate distribution, due to the strong positive correlation mentioned earlier. This explains the severe underestimation, as a similar feature was already observed in Section 3.1. The alternative approach to calculate simultaneous credible bands for the pairwise differences $\theta_i - \theta_{i-1}$ gives a larger pseudo contour probability for the constant risk reference point, but still much smaller than the maximized contour probability with our new approach.

3.3 Disease Mapping

As a final example we now consider a problem in spatial epidemiology taken from Bernardinelli, Pascutto, Best, and Gilks (1997), see also Knorr-Held and Rue (2002). Observed disease counts y_i in each district of Sardinia ($i = 1, \ldots, p = 366$) are conditionally independent Poisson distributed with mean $e_i \exp(\theta_i)$, where e_i are known expected counts and θ_i are unknown log relative risk parameters, which are assumed to follow a Gaussian Markov random field (GMRF). Here we assume a (nonstationary) "intrinsic autoregression"

$$p(\boldsymbol{\theta}|\eta) \propto \eta^{\frac{p-1}{2}} \exp\left(-\frac{\eta}{2}\sum_{i\sim j}(\theta_i - \theta_j)^2\right), \tag{3.3}$$

where $i \sim j$ denotes all pairs of adjacent districts i and j. For the precision parameter η we adopt the usual conjugate gamma prior $\eta \sim \mathrm{G}(a, b)$, with density $p(\eta) \propto \eta^{a-1}\exp(-b\eta)$. We have used three different sets of values, following suggestions by Bernardinelli, Clayton, and Montomoli (1995): $a = 0.25$ and $b = 0.0005$ (Prior 1), $a = 1.0$ and $b = 0.02$ (Prior 2) and finally $a = 5.0$ and $b = 0.25$ (Prior 3).

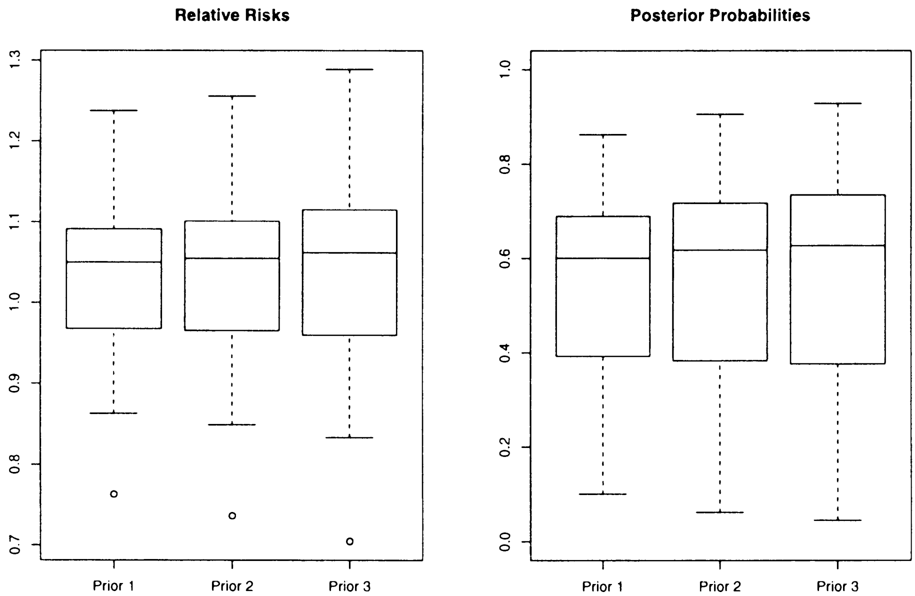

Figure 7. Estimated relative risks and posterior probabilities under three different priors.

We will calculate the contour probabilities of the reference points $\boldsymbol{\theta}_c = (c, \ldots, c)$, which correspond to a constant log relative risk equal to c. In particular we are interested in $\boldsymbol{\theta}_0 = (0, \ldots, 0)$, because the overall log relative risk $\log(\sum y_i / \sum e_i)$ equals zero by design. Note that the corresponding univariate posterior probabilities $P(\theta_i > 0|\mathbf{y})$ vary between 0.10 and 0.86 for Prior 1 (inter quartile range: 0.39 to 0.69), between 0.06 and 0.91 for Prior 2 (IQR: 0.38 to 0.72) and between 0.04 and 0.93 for Prior 3 (IQR: 0.28 to 0.73). Figure 7 displays boxplots of these posterior probabilities as well as posterior mean relative risks under the three different priors. As mentioned before, it is unclear if these univariate posterior summaries indicate that there is evidence for spatial variation in the underlying risk surface. We will calculate contour probabilities for the reference points $\boldsymbol{\theta}_c = (c, \ldots, c)$ to answer this question.

As in the previous example the marginal posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ is not available, so we use the proposed Rao-Blackwell estimate (2.5). The required log conditional density is (ignoring any additive constant):

$$\log p(\boldsymbol{\theta}|\eta, \mathbf{y}) = \sum_{i=1}^{p} (y_i \theta_i - e_i \exp(\theta_i)) - \frac{\eta}{2} \sum_{i \sim j} (\theta_i - \theta_j)^2 + \frac{p + 2a - 3}{2} \log(\eta) - b\eta - \log p(\eta|\mathbf{y}). \quad (3.4)$$

Similar to example 3.2, the evaluation of $\hat{P}_{\text{RB}}$ only requires storage of samples of η, $\sum_{i \sim j} (\theta_i - \theta_j)^2$ and $\sum_{i=1}^{p} (y_i \theta_i - e_i \exp(\theta_i))$, not of all samples from $\boldsymbol{\theta}$. We have generated a file with 10,000 samples of these quantities, taking every 100th iteration, based on the MCMC block updating algorithm described in Knorr-Held and Rue (2002). The samples were virtually independent. Again we had to first estimate $p(\eta|\mathbf{y})$ using the Rao-Blackwell

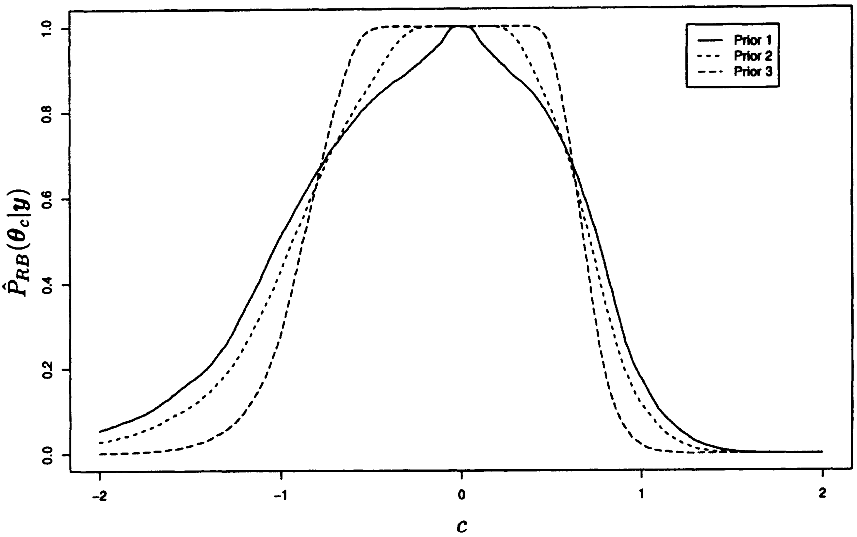

Figure 8. Estimated contour probabilities.

estimate (3.2) where $p(\eta|\boldsymbol{\theta}, \mathbf{y})$ now is the density of a gamma distribution with parameters $a + 0.5 \cdot (p - 1)$ and $b + 0.5 \cdot \sum_{i \sim j}(\theta_i - \theta_j)^2$.

Figure 8 displays the estimated contour probabilities $\hat{P}_{\mathrm{RB}}(\theta_c)$ as a function of c for the three different priors. It is remarkable that—under all three priors—the estimated contour probabilities are exactly 1.0 for c close to zero. However, this is not as surprising as it seems at first sight: the data are extremely sparse with a total of 619 cases. Furthermore, a simple Poisson regression assuming a constant risk gives a deviance value of 344.9 on 365 degrees of freedom, hence no evidence for overdispersion at all.

The contour probabilities for priors with more extreme posterior probabilities (prior 2 and in particular prior 3) approach zero faster for larger values of $|c|$. However, they also have a larger interval where the contour probabilities are virtually 1.0.

We conclude that, under all three priors, the reference points $\boldsymbol{\theta}_c$ with c close to zero are extremely well supported by the posterior distribution. This suggests that there is no evidence against the simple constant risk model.

Incidentally, under prior 1 the Rao-Blackwell estimates based on the mean of the log density values were considerably larger than the ones suggested in this article based on the median. For example, for $c = -1$ the estimate based on the median was 0.51 while the one based on the mean was 0.83. The differences were less pronounced under prior 2 and negligible under prior 3. However, similar to Section 3.2, the estimates based on the mean of the density values were very different (much larger around $c = 0$), even under prior 3.

4. DISCUSSION

The results from Section 3.1 and 3.2 suggest that simultaneous credible bands have to be taken with care. They may give a misleading impression of the support of the posterior distribution and often include areas which are not supported by the posterior at all. Also, they may underestimate the support of the posterior for points of interest which are located in tails of the posterior distribution. Empirical analyses suggest that these problems increase with increasing correlations between parameters of interest.

One may argue that simultaneous credible bands should not be used to check the posterior support for certain points of interest but should be taken more as an exploratory tool. In fact, the main advantage of these bands is that they can easily be visualized (compare, e.g., Figure 5). However, it is difficult to see the benefit of such an exploratory visualization, if one is not allowed to check the posterior support for certain points of interest, not even visually.

As a valuable alternative, we have proposed to use contour probabilities and have discussed methods to estimate these based on samples from a given posterior distribution. These methods are general, easy to use, and provide a useful tool for exploring high-dimensional posterior distributions. The use of a Rao-Blackwell estimate in this context is novel and gives rise to an estimate of the contour probability even if the (marginal) density is not known. From a computational point of view, the method typically requires much less storage space than simultaneous credible bands; the required computing time in the more complex applications of Section 3.2 and 3.3 was only a few minutes and seems negligible, compared to the usual run length necessary to simulate from the posterior via MCMC.

The use of the median instead of the mean ensures that the Rao-Blackwell estimate is (a) invariant with respect to arbitrary choices of the definition of contour probabilities and (b) robust to extreme outliers in the density values. But how much does the choice of the median or the mean matter in practice? There were no virtually differences in the simple example discussed in Section 3.1. However, in the other two applications we found that the estimates based on the mean of the *log* density values were larger than the ones based on the median. The estimates based on the mean of the actual density values (not the log density values) were very different from the median-based estimates, probably due to large outliers in the density values.

The only difficulty we faced in Section 3.2 and 3.3 was that the conditional density $p(\boldsymbol{\theta}|\eta, \mathbf{y})$, required for the Rao-Blackwell estimate, needs to be known including all multiplicative terms depending on η. It was therefore necessary to first estimate the marginal posterior density $p(\eta|\mathbf{y})$. However, this was straightforward to do via Rao-Blackwell again, but may be more involved in other models.

Finally, it is worth pointing out that there are some analogies between the methods discussed in this article and classical approaches for simultaneous inference, in particular in the context of multiple comparisons (e.g., Miller 1985). The Besag et al. (1995) method is somewhat similar to Tukey adjustments of p values for multiple comparisons. Indeed, both approaches are based on products of symmetric univariate intervals at the same level, and both are (nearly) exact under certain conditions. Similarly, the approach based on contour probabilities mirrors in some sense the Scheffé method based on the F-test statistic (compare in particular Section 3.1 where exact contour probabilities were available based on the quantiles of the F-distribution). However, both the Tukey and the Scheffé approach are only valid under normality whereas the two methods discussed in this article are applicable in a much wider context.

ACKNOWLEDGMENTS

The author thanks Isabel Natário for stimulating discussions and access to her function to compute simultaneous credible bands. Thanks also to Luisa Bernardinelli for providing the dataset from Section 3.3. I am grateful to two referees for their constructive comments that led to improvements in the article.

[Received August 2002. Revised January 2003.]

REFERENCES

Bernardinelli, L., Clayton, D., and Montomoli, C. (1995), "Bayesian Estimates of Disease Maps: How Important are Priors?" *Statistics in Medicine*, 14, 2411–2431.

Bernardinelli, L., Pascutto, C., Best, N. G., and Gilks, W. R. (1997), "Disease Mapping with Errors in Covariates," *Statistics in Medicine*, 16, 741–752.

Besag, J. E., Green, P. J., Higdon, D. M., and Mengersen, K. L. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wiley. Reprinted by Wiley in 1992 in the Wiley Classics Library Edition.

Carlin, B. P., and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis* (2nd ed.), London: Chapman and Hall/CRC.

Fahrmeir, L., and Knorr-Held, L. (2000), "Dynamic and Semiparametric Models," in *Smoothing and Regression: Approaches, Computation and Applications*, ed. M. Schimek, New York: Wiley, chap. 18, pp. 513–544.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

Goldstein, H., and Spiegelhalter, D. J. (1996), "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 159, 385–443.

Kashiwagi, N., and Yanagimoto, T. (1992), "Smoothing Serial Count Data Through a State-Space Model," *Biometrics*, 48, 1187–1194.

Knorr-Held, L. (1999), "Conditional Prior Proposals in Dynamic Models," *Scandinavian Journal of Statistics*, 26, 129–144.

Knorr-Held, L., and Rue, H. (2002), "On Block Updating in Markov Random Field Models for Disease Mapping," *Scandinavian Journal of Statistics*, 29, 597–614.

Miller, R. (1985), "Multiple Comparisons," in *Encyclopedia of Statistical Sciences* (vol. 5), eds. S. Kotz and N. L. Johnson, New York: Wiley, pp. 679–689.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Tanner, M. A. (1996), *Tools for Statistical Inference* (3rd ed.), New York: Springer.

Wei, G. C. G., and Tanner, M. A. (1990), "Calculating the Content and Boundary of the Highest Posterior Density Region via Data Augmentation," *Biometrika*, 77, 649–652.