Ann. Inst. Statist. Math.
Vol. 42, No. 3, 403-433 (1990)

A MONTE CARLO METHOD FOR AN OBJECTIVE BAYESIAN PROCEDURE

YOSIHIKO OGATA

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

(Received March 22, 1988; revised December 4, 1989)

Abstract. This paper describes a method for an objective selection of the optimal prior distribution, or for adjusting its hyper-parameter, among the competing priors for a variety of Bayesian models. In order to implement this method, the integration of very high dimensional functions is required to get the normalizing constants of the posterior and even of the prior distribution. The logarithm of the high dimensional integral is reduced to the one-dimensional integration of a certain function with respect to the scalar parameter over the range of the unit interval. Having decided the prior, the Bayes estimate or the posterior mean is used mainly here in addition to the posterior mode. All of these are based on the simulation of Gibbs distributions such as Metropolis' Monte Carlo algorithm. The improvement of the integration's accuracy is substantial in comparison with the conventional crude Monte Carlo integration. In the present method, we have essentially no practical restrictions in modeling the prior and the likelihood. Illustrative artificial data of the lattice system are given to show the practicability of the present procedure.

Key words and phrases: ABIC, Bayesian likelihood, posterior mean, ϕ- and ψ-statistic, Gibbs distribution, hyper-parameters, Metropolis' algorithm, normalizing factor, potential function, type II maximum likelihood method.

1. Objective Bayesian method

Take the case where many parameters $\boldsymbol{\theta} = (\theta_i)$ are required to present a statistical model, such as the one used in the inverse problem. The model is usually described by the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{Y})$ for a given set of data $\boldsymbol{Y}$. If the number of parameters to be estimated is moderate in comparison with the number of data or its resolution, the maximum likelihood method usually provides a sensible and the most accurate estimation. However, in the present case, the likelihood function $L(\boldsymbol{\theta}; \boldsymbol{Y})$ is likely to have many local maxima or possibly to be unbounded. Such a situation is called *ill-*

conditioned. This means that we are not able to have a sensible solution without some restrictions among the parameters. In order to measure the deviations from such restrictions, a penalty function is sometimes required (see Good and Gaskins (1971), for example); consequently, it is then necessary to resolve two conflicting aims which are to produce a good fit to the data whilst also imposing a penalty to avoid any ill-condition. These are described by a trade-off between the log likelihood and the weighted penalty function

$$\log L(\boldsymbol{\theta}; \boldsymbol{Y}) - Q(\boldsymbol{\theta}; \tau) , \tag{1.1}$$

where the function Q represents a set of penalties and $\tau = (\tau_k)$ is the vector of respective weights for the penalties, which we hereafter call the *hyper-parameter.* The crucial point here is the adjustment of the hyper-parameter τ. To obtain the optimal hyper-parameter, we are led to the Bayesian interpretation of the function in (1.1). The exponential of the negative penalty is considered to be proportionate to the prior probability distribution $\pi(\boldsymbol{\theta}|\tau)$, such that

$$\pi(\boldsymbol{\theta}|\tau) = \frac{e^{-Q(\boldsymbol{\theta}; \tau)}}{\int e^{-Q(\boldsymbol{\theta}; \tau)} d\boldsymbol{\theta}} \tag{1.2}$$

is characterized by the hyper-parameter τ, so that the exponential of the penalized log likelihood function in (1.1) is proportionate to the posterior. Then we need to consider its normalizing factor

$$\Lambda(\tau; \boldsymbol{Y}) = \int L(\boldsymbol{\theta}; \boldsymbol{Y}) \pi(\boldsymbol{\theta}; \tau) d\boldsymbol{\theta} , \tag{1.3}$$

to define the posterior probability. The normalizing factor, called the *Bayesian likelihood* of τ, is useful to obtain the optimal hyper-parameter τ which maximizes Λ or its logarithm

$$\log \Lambda(\tau; \boldsymbol{Y}) = \log \int L(\boldsymbol{\theta}; \boldsymbol{Y}) e^{-Q(\boldsymbol{\theta}; \tau)} d\boldsymbol{\theta} - \log \int e^{-Q(\boldsymbol{\theta}; \tau)} d\boldsymbol{\theta} . \tag{1.4}$$

This is called the method of *type II maximum likelihood* by Good (1965). Akaike (1978, 1979) justified and developed this method based on the *entropy maximization principle* and defined

$$\mathrm{ABIC} = (-2) \max_{\tau} \log \Lambda(\tau; \boldsymbol{Y}) + 2 \cdot \dim(\tau) , \tag{1.5}$$

for the comparable use with the Akaike Information Criterion (AIC;

Akaike (1977)), both of which are to be minimized for the comparison of respective statistical models. When both the likelihood and the prior distribution take the Gaussian form, the integration in (1.2)–(1.4) can usually be implemented analytically (see Akaike (1979)).

Since this is not expected in general for non-Gaussian models, a Gaussian approximation method is considered and implemented (Ishiguro and Sakamoto (1983), Ogata and Katsura (1988), Ogata *et al.* (1989)), except in the case when the successive numerical integration is feasible such as in the models of state space representation in time series (Kitagawa (1987)). One problem with the former method, however, is the unestimable bias of ABIC due to the approximation. Therefore, a reasonably accurate and efficient numerical guess of $\log \Lambda(\tau; \boldsymbol{Y})$ for general models is required. The Monte Carlo method developed in the present paper will be useful for this purpose.

2. Monte Carlo integration

2.1 *The method*

Suppose that we wish to estimate the integral

$$Z_N = \int_a^b \int_a^b \cdots \int_a^b f(x_1, x_2, \ldots, x_N)\, dx_1 dx_2 \cdots dx_N \,. \tag{2.1}$$

We shall denote the vector $(x_1, x_2, \ldots, x_N)$ by $\boldsymbol{x}$. Numerical methods for the evaluation of Z_N involve the calculation of $f(\boldsymbol{x})$ at a number N of points $\boldsymbol{x}_i$. The crude Monte Carlo method gives the sum

$$\frac{1}{M} \sum_{i=1}^{M} f(\boldsymbol{x}_i) \tag{2.2}$$

as an estimate for Z_N where the points $\boldsymbol{x}_i$ are chosen at random in the range of integration. Although there are some sophisticated modifications or improvements of the method (see Hammersley and Handscomb (1964) for example), those methods based on (2.2) are not practical for the integral of a large multiplicity N: that is to say, the bias caused by the skewness of the function is significant while the integrated values are usually very small or very large (Ogata (1989)). Thus, in this paper, we are interested in estimating $\log Z_N$ directly, rather than via Z_N itself.

Suppose that $f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_N)$ is a function defined and to be integrated in an N-dimensional cube $V_N = [a, b]^N$. If $f(\boldsymbol{x})$ is defined in the infinite domain, $[a, b]$ is taken sufficiently large for a reasonable approximation of the integral. Assume also that this function is bounded from below, so that we can hereafter assume the non-negativity of the function without any loss of generality. Consider a scaling parameter σ such that

$0 \leq \sigma \leq 1$, and let the vector $\sigma \boldsymbol{x}$ denote $(\sigma x_1, \sigma x_2, \ldots, \sigma x_N)$. Define a family of probability densities $\{g_\sigma(\boldsymbol{x})\}$ on the cube $V_N = [a, b]^N$, which are parameterized by the scale parameter σ in such a way that

$$g_\sigma(\boldsymbol{x}) = \frac{f(\sigma \boldsymbol{x})}{Z_N(\sigma)}, \tag{2.3}$$

where $Z_N(\sigma)$ is the normalizing factor

$$Z_N(\sigma) = \int_{V_N} f(\sigma \boldsymbol{x}) \, d\boldsymbol{x}. \tag{2.4}$$

Under very broad regularity conditions, using Fubini's theorem of change between signs of integral and differential, we can expect

$$\frac{\partial}{\partial \sigma} \log Z_N(\sigma) = \int_{V_N} \left\{ \frac{\partial}{\partial \sigma} \log f(\sigma \boldsymbol{x}) \right\} g_\sigma(\boldsymbol{x}) \, d\boldsymbol{x}. \tag{2.5}$$

For convenience in later description, let us set the equality in (2.5) to $\psi(\sigma)$. Replacing the expectation in the last equality by the time average, we get a consistent and unbiased estimate of $\psi(\sigma)$

$$\hat{\psi}(\sigma) = \frac{1}{M} \sum_{t=1}^{M} \frac{\partial}{\partial \sigma} \log f(\sigma \boldsymbol{X}(t)), \tag{2.6}$$

where $\boldsymbol{X}(t) = (X_1(t), X_2(t), \ldots, X_N(t))$ are vector series of samples following the distribution $g_\sigma(\boldsymbol{x})$. The practical sampling methods include the so-called Metropolis' simulation procedure. For the procedure we define the *potential function* by

$$U_\sigma(\boldsymbol{x}) = -\log f(\sigma \boldsymbol{x}), \tag{2.7}$$

in the present case. See Subsection 2.5 for the review and method of the application of Metropolis' procedure. On the other hand, the eventual estimation of $\log Z_N = \log Z_N(1)$ is written by

$$\log Z_N(1) = \log Z_N(0+) + \int_0^1 \psi(\sigma) \, d\sigma, \tag{2.8}$$

where

$$Z_N(0+) = \lim_{\sigma \to 0} \int_{V_N} f(\sigma \boldsymbol{x}) \, d\boldsymbol{x} = (b - a)^N f(\boldsymbol{0}) \tag{2.9}$$

from (2.4), assuming the right continuity of the function at the origin and that $\mathbf{0} = (0, 0, \ldots, 0)$. Based on this method, Ogata (1989) showed the extremely significant improvement in comparison with the conventional crude Monte Carlo integration.

2.2 *A sophisticated version and its application*

Here, I would like to add some sophistication to the above relations. For that in the last section, the parameterization of the function $f(\boldsymbol{x})$ by σ from the interval [0, 1] need not be in the form of the scaling like in (2.3). Furthermore, the parameterization may be generalized by

$$f_\sigma(\boldsymbol{x}) = f_0(\boldsymbol{x}) \cdot h_\sigma(\boldsymbol{x}) , \tag{2.10}$$

in such a way that $h_0(\boldsymbol{x}) = 1$ and $h_1(\boldsymbol{x}) = f(\boldsymbol{x})/f_0(\boldsymbol{x})$ hold. An example of this parameterization is $h_\sigma(\boldsymbol{x}) = f(\sigma\boldsymbol{x})/f_0(\sigma\boldsymbol{x})$. Here, the function $f_0(\boldsymbol{x})$ should be positive and continuously differentiable and its integral on the given region is known or estimated somehow. Then, for the potential energy

$$U_\sigma(\boldsymbol{x}) = -\log f_\sigma(\boldsymbol{x}) = -\{\log f_0(\boldsymbol{x}) + \log h_\sigma(\boldsymbol{x})\} , \tag{2.11}$$

$$\hat{\psi}(\sigma) = \frac{1}{M} \sum_{t=1}^{M} \frac{\partial}{\partial \sigma} \log h_\sigma(\boldsymbol{X}(t)) \tag{2.12}$$

is the corresponding estimate to that in (2.6). Furthermore, $\log Z_N(0+)$ in (2.8) is given by

$$Z_N(0+) = \int_{V_N} f_0(\boldsymbol{x})\, d\boldsymbol{x} \tag{2.13}$$

instead of that shown in (2.9). A good choice of the function $f_0(\boldsymbol{x})$ would be one in which both $\log h_1(\boldsymbol{x})$ and its derivative are small. For further details, see the forthcoming Section 4 on the implementation of Bayesian integrals and also Ogata (1989) for some experiments with the known integrals.

An interesting example of the choice of $f_0(\boldsymbol{x})$ is related to the Gaussian approximation method suggested by Ishiguro and Sakamoto (1983) (see also Ogata and Katsura (1988) and Ogata *et al.* (1989) for some applications). Let the logarithm of a posterior $T(\boldsymbol{\theta}; \tau) = \log\{L(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\tau)\}$ be approximated by the quadratic form

$$T(\boldsymbol{\theta}|\tau) \cong T(\hat{\boldsymbol{\theta}}|\tau) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\, H(\hat{\boldsymbol{\theta}}|\tau)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t ,$$

where $\hat{\boldsymbol{\theta}}$ is the vector which maximizes T for a fixed τ, and $H(\hat{\boldsymbol{\theta}}; \tau)$ is the Hessian matrix (i.e. second derivatives) of the penalized log likelihood at $\hat{\boldsymbol{\theta}}$.

Using this approximation, we set the exponential of the approximation to $f_0(\boldsymbol{\theta})$ so that the integral of this function is reduced to the computation of the determinant of the Hessian matrix. Thus we have

$$\log Z(0+) = T(\hat{\boldsymbol{\theta}}|\tau) - \frac{1}{2}\log\{\det H(\hat{\boldsymbol{\theta}}|\tau)\} - \frac{N}{2}\log 2\pi ,$$

where N is the dimension of the parameter $\boldsymbol{\theta}$. This approximation seems to be extremely useful when the prior is Gaussian and the likelihood is non-Gaussian. Here, I would like to emphasize that the bias of the approximant from the true is evaluated accurately enough by the integration of $\hat{\psi}$ in (2.12) over [0, 1].

2.3 *Implication of the methods in statistical mechanics*

Consider a system of N particles in a volume V at a specified temperature T, as well as a total potential energy $U_N(\boldsymbol{x}) = \sum_{i<j} \Phi(r_{ij})$, where Φ is a pairwise interaction potential function of the distance between any two points of the state $\boldsymbol{x}$. Then the Gibbsian canonical distribution is given by

$$g(\boldsymbol{x}) = \exp\left\{\frac{1}{kT} U_N(\boldsymbol{x})\right\} \Big/ Z(N, V, T) , \tag{2.14}$$

where k is Boltzmann's constant and $Z(N, V, T)$ is the normalizing factor of a probability density distribution. The quantity ψ defined in (2.5) is related to the *equation of state* which is proportional to $(PV/NkT) - 1$, where P is the *pressure*. Metropolis *et al.* (1953) and some of their followers such as Wood (1968) calculated this quantity by computer simulation in order to investigate the variety of aspects in statistical mechanics, especially the phase transition of liquid.

Given a set of coordinates of N equilibrium points in a region V, Ogata and Tanemura (1981*a*, 1981*b*, 1984*a*, 1984*b*, 1989) developed the evaluation of the normalizing factor $\log Z$ of the Gibbsian distribution in performing the maximum likelihood method to estimate the shape of the pairwise interaction potential function. We estimated the smooth equation of state $\Psi(\tau)$ of $\tau = N\sigma^2/V$ by fitting polynomials or spline functions to a number of evaluated $\Psi(\tau_i)$ of sampled τ_i's obtained from the simulated experimental data of the particles. Then, by a similar relation to (2.8), we obtained the functions $\log Z_N(\sigma)$ of the scale parameter σ for a number of parameterized pairwise potential functions.

Recently, I learned that such an estimation method of $\log Z_N(\sigma)$, which is called *free energy*, by the derivative of a suitable scalar parameter σ has been commonly used in the field of statistical physics since late 1970's

(see Binder (1986) for example). For instance, the reciprocal of the temperature, $\sigma = T^{-1}$, has been used for such a scalar parameter. Incidentally, an increase in the last parameter realizes the so-called *stochastic annealing* (Kirkpatrick *et al.* (1983), Geman, S. and Geman, D. (1984)) of series of equilibrium states. Here, we put $h_\sigma(\boldsymbol{x}) = \{f(\boldsymbol{x})/f_0(\boldsymbol{x})\}^\sigma$ in (2.10). Then, (2.11) and (2.12) imply that

$$\hat{\phi}(\sigma) = \frac{1}{M}\sum_{t=1}^{M} \log \frac{f(\boldsymbol{X}(t))}{f_0(\boldsymbol{X}(t))} \tag{2.15}$$

instead of (2.6) with the total potential energy

$$V_\sigma(\boldsymbol{x}) = -(1-\sigma)\log f_0(\boldsymbol{x}) - \sigma \log f(\boldsymbol{x}) \tag{2.16}$$

instead of (2.7). In the present study, we will use the special case where $f_0(\boldsymbol{x}) = 1$ in the above pair (2.15) and (2.16) to compute (2.8), where ψ is replaced by ϕ as the alternative to those given in (2.6) and (2.7). In this case, $Z_N(0+) = (b-a)^N$ holds in place of (2.9).

2.4 *Numerical approximation and error estimate*

Before carrying out the integration in (2.8), the estimation (2.6) of $\psi(\sigma_j) = (\partial/\partial\sigma)\log Z_N(\sigma_j)$ should be made for many σ_j's sampled from the unit interval $[0, 1]$, preferably with their estimated errors. In Ogata and Tanemura (1981*b*, 1984*a*, 1984*b*, 1989) polynomials or spline functions are fitted to this sort of experimental data to get a smooth and well fitted function, and then to be used for the integrand in (2.8). The alternative but simple method to evaluate the integral in (2.8) will be by the trapezoidal rule, for example,

$$\begin{aligned}&\frac{1}{2}\sum_{j=1}^{J}\{\hat{\psi}(\sigma_{j+1}) + \hat{\psi}(\sigma_j)\}(\sigma_{j+1} - \sigma_j) \\ &= \left(\frac{\sigma_2 - \sigma_1}{2}\right)\hat{\psi}(\sigma_1) + \sum_{j=2}^{J}\left(\frac{\sigma_{j+1} - \sigma_{j-1}}{2}\right)\hat{\psi}(\sigma_j) + \left(\frac{\sigma_{J+1} - \sigma_J}{2}\right)\hat{\psi}(\sigma_{J+1})\,.\end{aligned} \tag{2.17}$$

If the estimates of the $\hat{\psi}(\sigma_j)$ and their errors for respective σ_j are not highly variable or inhomogeneous, the above sum with the equidistant nodes $\{\sigma_j\}_{j=1,2,\ldots,J}$, for example, is expected to provide an accurate estimate of the required integral in (2.8) (see Examples 1 and 2 in Ogata (1989)). This is because the estimate of $\psi(\sigma_j)$ for each σ_j is consistent and unbiased and the random variables $\hat{\psi}(\sigma_j)$ and $\hat{\psi}(\sigma_k)$ are mutually independent for $k \neq j$, provided that a suitable generation of random numbers is achieved. Thus, if the standard error of each $\hat{\psi}(\sigma_j)$ is s_j, then the error variance of $\log Z_N(1)$ estimated by using (2.17) in (2.8) is given by

$$(2.18) \qquad \left(\frac{\sigma_2 - \sigma_1}{2}\right)^2 s_1^2 + \sum_{j=2}^{J} \left(\frac{\sigma_{j+1} - \sigma_{j-1}}{2}\right)^2 s_i^2 + \left(\frac{\sigma_{J+1} - \sigma_J}{2}\right)^2 s_{J+1}^2 .$$

The estimation of s_i^2 can be carried out in the following way. Let us set

$$(2.19) \qquad \eta_\sigma(t) = \frac{\partial}{\partial \sigma} \log f(\sigma \boldsymbol{X}(t)) - \hat{\psi}(\sigma) .$$

Then, utilizing Theorem 18.2.1 in Ibragimov and Linnik (1971), we have an estimate of the variance of $\hat{\psi}(\sigma)$

$$(2.20) \qquad \mathrm{Var}\,(\hat{\psi}(\sigma)) \cong \frac{1}{M^2} \sum_{|\tau| < M} \left(1 - \frac{|\tau|}{M}\right) \sum_{t=1}^{M-|\tau|} \eta_\sigma(t)\eta_\sigma(t + |\tau|) .$$

These evaluation methods will be used for the numerical implementation in Section 4. Finally, similar arguments to those in this section are straightforward for the ϕ-function in (2.15) with the potential in (2.16).

2.5 *Simulation of the Gibbsian field*

For a practical method to get the samples $\{\boldsymbol{X}(t); t = 1, 2, \ldots, M\}$ in (2.6), (2.12) and (2.15) from the distributions in (2.3) and (2.10) or a similar version for the potential in (2.16), let us briefly review a simulation method which uses a particular type of random walk known as a Markov chain. The simulation was originally devised by Metropolis *et al.* (1953) and developed by Wood (1968) and others for the study of atomic systems. Consider a continuous Gibbs random field of a state space V_N (which is $[a, b]^N$ in the present case) whose probability density distribution $g(\boldsymbol{x})$ is of the form

$$(2.21) \qquad g(\boldsymbol{x}) = \frac{1}{Z} \exp\{-U(\boldsymbol{x})\} ,$$

where $U(\boldsymbol{x})$ is the potential of the state $\boldsymbol{x}$ and Z is the normalizing constant called the *partition function.*

The most commonly used simulation algorithm of the density (2.21) is described in the following manner. Assume that, at time t, the state of the N axes is $\boldsymbol{X}(t) = \{(X_n(t); n = 1, \ldots, N) \in V^N\}$. A trial state $\boldsymbol{X}'(t) = \{(X_n'(t); n = 1, \ldots, N)\}$ is then chosen in such a way that the coordinate $X_r'(t)$ of a randomly chosen axis r lies in some neighbourhood (in the present case $[X_r(t) - \delta, X_r(t) + \delta]$, so that $X_r'(t) = X_r(t) + \delta(1 - 2\xi)$ for a uniform random number ξ), while all other $N - 1$ axes have the same position as in state $\boldsymbol{X}(t)$, where $\delta > 0$ is the parameter to be discussed below. We may use a periodic or reflective boundary for the present random walk. The corre-

sponding potential energy, $U(\boldsymbol{X}'(t))$, is then calculated and compared with $U(\boldsymbol{X}(t))$ as follows.

1. If $U(\boldsymbol{X}'(t)) \leq U(\boldsymbol{X}(t))$, then without further ado the next state $\boldsymbol{X}(t+1)$ of the realization is taken as the trial state $\boldsymbol{X}'(t)$.

2. If $U(\boldsymbol{X}'(t)) > U(\boldsymbol{X}(t))$, then we obtain a uniform random number ξ, and (i) if $\xi \leq \exp\{U(\boldsymbol{X}(t)) - U(\boldsymbol{X}'(t))\}$, state $\boldsymbol{X}(t+1)$ is taken to be the trial state $\boldsymbol{X}'(t)$; (ii) otherwise, state $\boldsymbol{X}(t+1)$ is taken to be the previous state $\boldsymbol{X}(t)$.

It should be noticed that the normalizing factor Z in (2.21) has not been used in the simulation. In essence, the Monte Carlo procedure required here is nothing more than to select the transition probabilities

$$q(\boldsymbol{x}, \boldsymbol{dy}) = \text{Prob}\{\boldsymbol{X}(t+1) \in \boldsymbol{dy} | \boldsymbol{X}(t) = \boldsymbol{x}\} \tag{2.22}$$

of the Markov chain $\boldsymbol{X}(t)$ which satisfy $\int p(\boldsymbol{dx}) q(\boldsymbol{x}, \boldsymbol{dy}) = p(\boldsymbol{dy})$ for all states $\boldsymbol{y}$ in V^N and for the equilibrium probability $p(\boldsymbol{dx}) = g(\boldsymbol{x})\boldsymbol{dx}$ in (2.21); furthermore, it is necessary that the n-step transition probability, $q^{(n)}(\boldsymbol{x}, \boldsymbol{dy})$, converges to the given equilibrium probability $p(\boldsymbol{dy})$. Thus, of course, there are many possible algorithms for that can be used to carry out these conditions, other than those stated above (Wood (1968)). Incidentally, there are some special cases where the transition probability is available for the direct simulation of the Gibbsian distribution state without any rejection, unlike the above case of Metropolis' algorithm. For example, Geman, S. and Geman, D. (1984) used such a simulation, called the *Gibbs sampler*, for the discrete state Markov random field on the lattice. Of course, this is also useful for the Gaussian Markov random field, which we are going to describe in an example for the implementation of the current Bayesian method in Section 4.

Back to Metropolis' algorithm, the parameter δ, the maximum single step displacement allowed in passing from one state to the next, ought in principle to be adjusted for the optimum rate of convergence in the Markov chain. Wood (1968), suggests that a reasonable choice of the adjusting parameter δ has been found to be the value leading to the rejection of the trial configuration on about half the time-steps. This is a trade-off between the effective transition of the state and the avoidance of unnecessary repetition of the same state, especially in the case of a highly variable potential. In addition to the selection of δ, in order to attain the equilibrium state in fewer time-steps in the Monte Carlo simulation, the initial configuration should be suitably chosen. Incidentally, Ogata and Tanemura (1981*c*) devised a method of sequentially generating points using the potential function, for the construction of such an initial state.

3. Application of the integration to the Bayesian method

We can now carry out the Monte Carlo method for the high dimensional integrations which appeared in Section 1. Since the first and second integrals in (1.4) are respectively given by

$$\log \int_{V_N} L(\boldsymbol{\theta}; \boldsymbol{Y}) e^{-Q(\boldsymbol{\theta};\tau)} d\boldsymbol{\theta} = \log \{(b-a)^N L(\mathbf{0}; \boldsymbol{Y}) e^{-Q(\mathbf{0};\tau)}\} + \int_0^1 \psi_\tau^{(1)}(\sigma) d\sigma \tag{3.1}$$

and

$$\log \int_{V_N} e^{-Q(\boldsymbol{\theta};\tau)} d\boldsymbol{\theta} = \log \{(b-a)^N L(\mathbf{0}; \boldsymbol{Y}) e^{-Q(\mathbf{0};\tau)}\} + \int_0^1 \psi_\tau^{(2)}(\sigma) d\sigma , \tag{3.2}$$

the log normalizing factor of the posterior in (1.4) is then written by

$$\begin{aligned} \log A(\tau; \boldsymbol{Y}) &= \log L(\mathbf{0}; \boldsymbol{Y}) + \int_0^1 \psi_\tau^{(1)}(\sigma) d\sigma - \int_0^1 \psi_\tau^{(2)}(\sigma) d\sigma \\ &= A_1 + A_2 - A_3 \end{aligned} \tag{3.3}$$

using the relations (2.8) and (2.9). Here, the terms A_1, A_2 and A_3 refer to the tables in the next section. The potential functions to be used in Metropolis' algorithm are, replacing the variable $\boldsymbol{x}$ in Section 2 by $\boldsymbol{\theta}$,

$$U_\sigma^{(1)}(\boldsymbol{\theta}) = -\log L(\sigma\boldsymbol{\theta}; \boldsymbol{Y}) + Q(\sigma\boldsymbol{\theta}; \tau) \tag{3.4}$$

and

$$U_\sigma^{(2)}(\boldsymbol{\theta}) = Q(\sigma\boldsymbol{\theta}; \tau) , \tag{3.5}$$

for the calculation of A_2 and A_3 in (3.3), respectively.

Alternatively, if we use the potential functions

$$V_\sigma^{(1)}(\boldsymbol{\theta}) = \sigma\{-\log L(\boldsymbol{\theta}; \boldsymbol{Y}) + Q(\boldsymbol{\theta}; \tau)\} \tag{3.6}$$

and

$$V_\sigma^{(2)}(\boldsymbol{\theta}) = \sigma Q(\boldsymbol{\theta}; \tau) , \tag{3.7}$$

corresponding to that in (2.16) with $f_0(\boldsymbol{x}) = 1$, then we have

$$\hat{\phi}^{(1)}(\sigma) = \frac{1}{M} \sum_{t=1}^{M} \{-\log L(\boldsymbol{X}(t); \boldsymbol{Y}) + Q(\boldsymbol{X}(t); \tau)\} \tag{3.8}$$

and

$$\hat{\phi}^{(2)}(\sigma) = \frac{1}{M} \sum_{t=1}^{M} Q(\boldsymbol{X}(t); \tau) \tag{3.9}$$

for that in (2.15). Thus, the same log normalizing factor of the posterior is rewritten by

$$\log \Lambda(\tau; \boldsymbol{Y}) = \int_0^1 \phi_\tau^{(1)}(\sigma) d\sigma - \int_0^1 \phi_\tau^{(2)}(\sigma) d\sigma = B_1 - B_2 \,. \tag{3.10}$$

It is worthwhile here to note that we can enjoy an advantage in computing the prior integrals A_3 and B_2, when the hyperparameter τ is a scaling factor of the parameter $\boldsymbol{\theta}$. That is to say, taking the range $[a, b]$ of the integration sufficiently large in comparison with τ, we can use the equality

$$\int_{V_N} e^{-Q(\boldsymbol{\theta}/\tau)} d\boldsymbol{\theta} \cong \tau^N \int_{V_N} e^{-Q(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{3.11}$$

for $V_N = [a, b]^N$, so that we may calculate only the case of τ in (3.5).

Suppose that the optimal hyper-parameter τ in (1.1) is obtained by maximizing (1.4), or by minimizing (1.5). Then maximization of the posterior $L(\boldsymbol{\theta}; \boldsymbol{Y})\pi(\boldsymbol{\theta}; \tau)$ or its logarithm (1.1) with respect to $\boldsymbol{\theta}$ is expected to provide an optimal estimate of $\boldsymbol{\theta}$. This is feasible when the posterior has a Gaussian form or its good approximation (see Akaike (1977), Ishiguro and Sakamoto (1983), Ogata and Katsura (1988) and Ogata *et al.* (1989)). However, this may not be easy to carry out for general models. To obtain the estimation of the posterior mode, the annealing procedure performed by Geman, S. and Geman, D. (1984) and the iterated conditional modes (ICM) by Besag (1986) may be useful.

An alternative estimation of θ suitable for the present procedure is the so-called *Bayes estimator*, or *posterior mean*, $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_N)$ such that

$$\tilde{\theta}_n = \int_a^b \int_a^b \cdots \int_a^b \theta_n g(\boldsymbol{\theta}) d\boldsymbol{\theta}, \qquad n = 1, 2, \ldots, N \,, \tag{3.12}$$

where $g(\boldsymbol{\theta})$ is the posterior probability

$$g(\boldsymbol{\theta}) = \frac{L(\boldsymbol{\theta}; \boldsymbol{Y})\pi(\boldsymbol{\theta}; \tau)}{\Lambda(\tau; \boldsymbol{Y})} \,, \tag{3.13}$$

and Λ is given in (1.3). If $g(\boldsymbol{\theta})$ is defined in the infinite domain of $\boldsymbol{\theta}$, the interval $[a, b]$ is taken sufficiently large for the reasonable approximation

of the integral (see Section 4).

The present sampling of $\{\boldsymbol{\Theta}(t); t = 1, 2, \ldots, M\}$ from $g(\boldsymbol{\theta})$ in (3.13) is carried out by Metropolis' algorithm with the potential $U_1^{(1)}$ or $V_1^{(1)}$ in (3.4) and (3.6), respectively, unless the suitable Gibbs sampler is available. The simulated data set $\{\boldsymbol{\Theta}(t)\}$ includes any information about the posterior distribution such as the average, variance and sample quantiles of marginal distribution of θ_n for any n, as well as any covariances between θ_k and θ_m, etc. For example, we have the time average $\bar{\boldsymbol{\Theta}} = (\bar{\Theta}_1, \bar{\Theta}_2, \ldots, \bar{\Theta}_N)$ such that

$$\bar{\Theta}_n = \frac{1}{M} \sum_{t=1}^{M} \Theta_n(t), \qquad n = 1, 2, \ldots, N, \tag{3.14}$$

for the Bayes estimate $\tilde{\boldsymbol{\theta}}$ in (3.12). Similarly, the estimated variances

$$\mathrm{Var}\,(\Theta_n) \cong \frac{1}{M} \sum_{t=1}^{M} \Theta_n(t)^2 - \bar{\Theta}_n^2, \qquad n = 1, 2, \ldots, N \tag{3.15}$$

are useful for the variability of the posterior marginal of each axis.

4. Implementation

As an illustrative example of our method, we consider an array of data $\{Y_{ij}\}$ on a 20×20 lattice which has been artificially generated from the following:

$$Y_{ij} \sim N(\theta_{ij}, 1.0^2), \qquad 1 \le i, j \le 20,$$

$$\theta_{ij} = \begin{cases} 1 & \text{for} \quad 1 \le i \le 10 \quad \text{and} \quad 1 \le j \le 10, \\ 2 & \text{for} \quad 1 \le i \le 10 \quad \text{and} \quad 11 \le j \le 20, \\ -1 & \text{for} \quad 11 \le i \le 20 \quad \text{and} \quad 1 \le j \le 10, \\ 0 & \text{for} \quad 11 \le i \le 20 \quad \text{and} \quad 11 \le j \le 20. \end{cases} \tag{4.1}$$

Graphs of the contour lines and bird's-eye view of the data are shown in Fig. 1(a). The problem is to estimate the step function θ_{ij} of the field (i, j) as shown in Fig. 1(b). For this type of data, we use the log likelihood

$$\log L(\boldsymbol{\theta}; \boldsymbol{Y}) = -\frac{400}{2} \log s^2 - \frac{1}{2s^2} \sum_{i,j} (Y_{ij} - \theta_{ij})^2. \tag{4.2}$$

We hereafter assume for simplicity that $s^2 = 1.0^2$ is known, which actually can either be another member of $\boldsymbol{\theta}$ to be integrated or one of the hyperparameters to be adjusted for the maximization of (1.3) or (1.4).

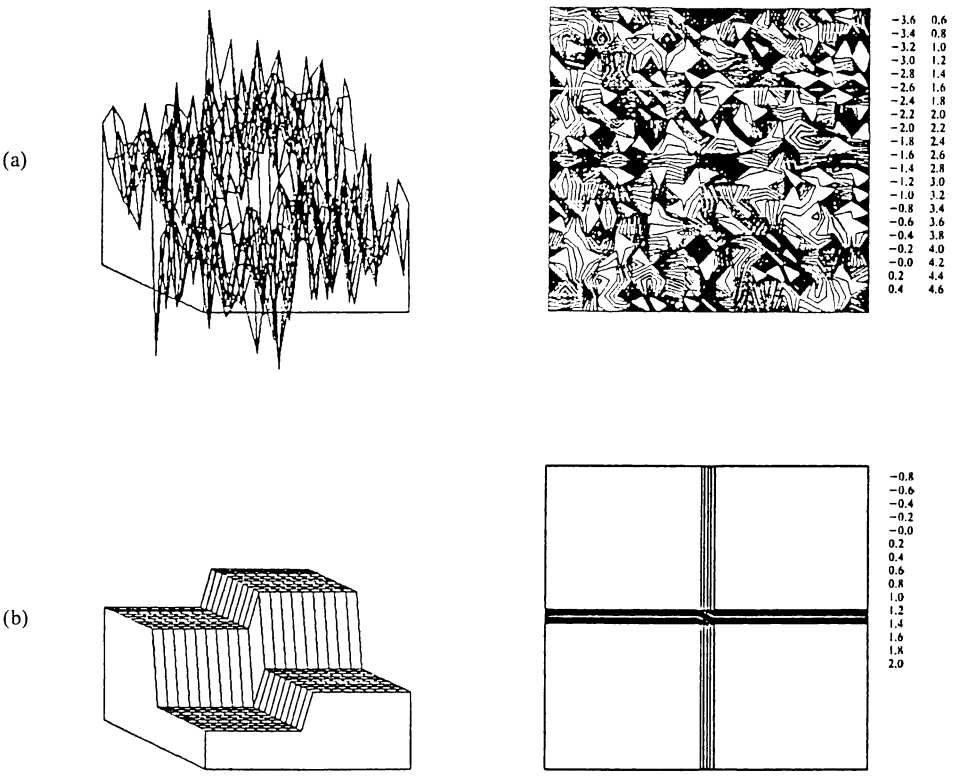


Fig. 1. Graphs of bird's-eye-view and contour lines of (a) the noisy data and (b) the true pattern.

Thus, we have $A_1 = \log L(\mathbf{0}; \boldsymbol{Y}) = 400 \log 2 - (1/2) \sum_{ij} Y_{ij}^2 = -518.51$ and we simply have to evaluate either the two integrals A_2 and A_3 in (3.3) or B_1 and B_2 in (3.10) for the present case.

For the roughness penalty in (1.1) we first use the following sum of the pairwise potential function of the nearest neighbour

$$Q(\boldsymbol{\theta}; \tau) = \frac{1}{2} \sum_{i=1}^{20} \sum_{j=1}^{20} \frac{1}{2\tau^2} \sum_{(k,m) \in R_{ij}} (\theta_{ij} - \theta_{km})^2 , \tag{4.3}$$

where R_{ij} is the nearest neighbourhood of (i, j): that is to say, $R_{ij} = \{(i, j \pm 1), (i \pm 1, j)\}$ when (i, j) is in the interior of the lattice domain, and similarly only the available (k, m) composes the neighbourhood of (i, j) at the edges. The corresponding prior distribution is nothing but Gaussian in this case. There is a natural alternative roughness penalty, called the Laplacian type, such that $(\theta_{ij} - (1/4) \Sigma \theta_{km})^2$ instead $\Sigma (\theta_{ij} - \theta_{km})^2$ in (4.3) (see Ogata (1988) for example). Tanabe and Tanaka (1983) use the same penalty, but they treat the prior distribution in a similar sense to Besag's

pseudo-likelihood (Besag (1974)) with a certain boundary condition so that they have no need to get the normalizing constants in the sense of (1.2). In our Monte Carlo method, we can assume the free boundary by virtue that the above priors are defined on a finite domain such as $V_N = [-20.0, 20.0]^N$ with $N = 20 \times 20$. Incidentally, with the same domain, Ogata (1989) checked the accuracy of the current estimation of the integral of a 1000 dimensional Gaussian distribution whose covariance obeys the inverse of a Toeplitz matrix.

In Metropolis' simulation of all experiments throughout this section, the maximum single step displacement parameter was chosen as $\delta = 20.0$, half the size of a cube, although this may not be optimal, depending on the scale σ and the hyperparameter τ of potentials. Looking at the time series of potentials and the ψ-values, the simulation of the first 400×200 steps, which may not be in equilibrium, was thrown away, and another $M = 400 \times 200$ steps was set for the estimation of the ψ-values.

First, the potentials (3.4) and (3.5) are considered. For a fixed τ, the corresponding functions $\psi_\tau^{(1)}(\sigma)$ and $\psi_\tau^{(2)}(\sigma)$ in (3.3) are calculated for the $\{\sigma_k\}$ equidistant 100 nodal points in $[0, 1]$ and for a further additional equidistant 20 points in $[0, 0.05]$ (thus 116 distinct points altogether). The reason for the extra partitioning is that we found a sharp trough in the ψ-function around the origin $\sigma = 0$ when τ is small. The estimated functions $\hat{\psi}_\tau^{(1)}(\sigma_k)$, $\hat{\psi}_\tau^{(2)}(\sigma_k)$ are plotted for $\tau = 0.25$, 0.5, 1.0, 2.0 and 4.0 in Figs. 2(a) and (b), respectively. Integrating these functions by the trapezoidal rule in (2.17), we have Table 1(a) for integrals $A_1 + A_2$ and A_3 in (3.3) with the standard error from (2.18) with (2.20) for each τ. Then the estimate of the Bayesian log likelihood in (3.3) is obtained with its standard error.

For the alternative estimation of the same Bayesian log likelihood by way of (3.10), I made another simulation using the potentials in the form in (3.6) and (3.7) to obtain $\phi_\tau^{(1)}(\sigma)$ and $\phi_\tau^{(2)}(\sigma)$ in (3.8) and (3.9), respectively. Since it is found that the trough of the ϕ-functions are extremely deep near the origin, I made a geometrical partition with the nodal points of $\{0.9^k; k = 0, 1, \ldots, 115\}$ for the trapezoidal formula of the integration on the unit interval $[0, 1]$. Thus, similarly to the above experiment, the ϕ-functions are obtained. These are shown in Figs. 3(a) and (b). The estimated integrations of B_1 and B_2 and the Bayesian log likelihoods for respective τ's are also shown in Table 1(d) for comparison with results via (3.3).

Furthermore, to see whether the partitioning seriously affects the unbiasedness of the integral, we calculated the integrals A_2 and A_3 in (3.3) with the current geometrical partition, as well as B_1 and B_2 in (3.10) with the above defined equidistant partition. The corresponding values are also listed in Tables 1(b) and (c) for comparison. Thus, these two different estimation methods with their two types of partition for the integrals provide the four respective estimates of the log Bayesian likelihood for a same model. The estimates of $A_1 + A_2$ and A_3 agree reasonably well with

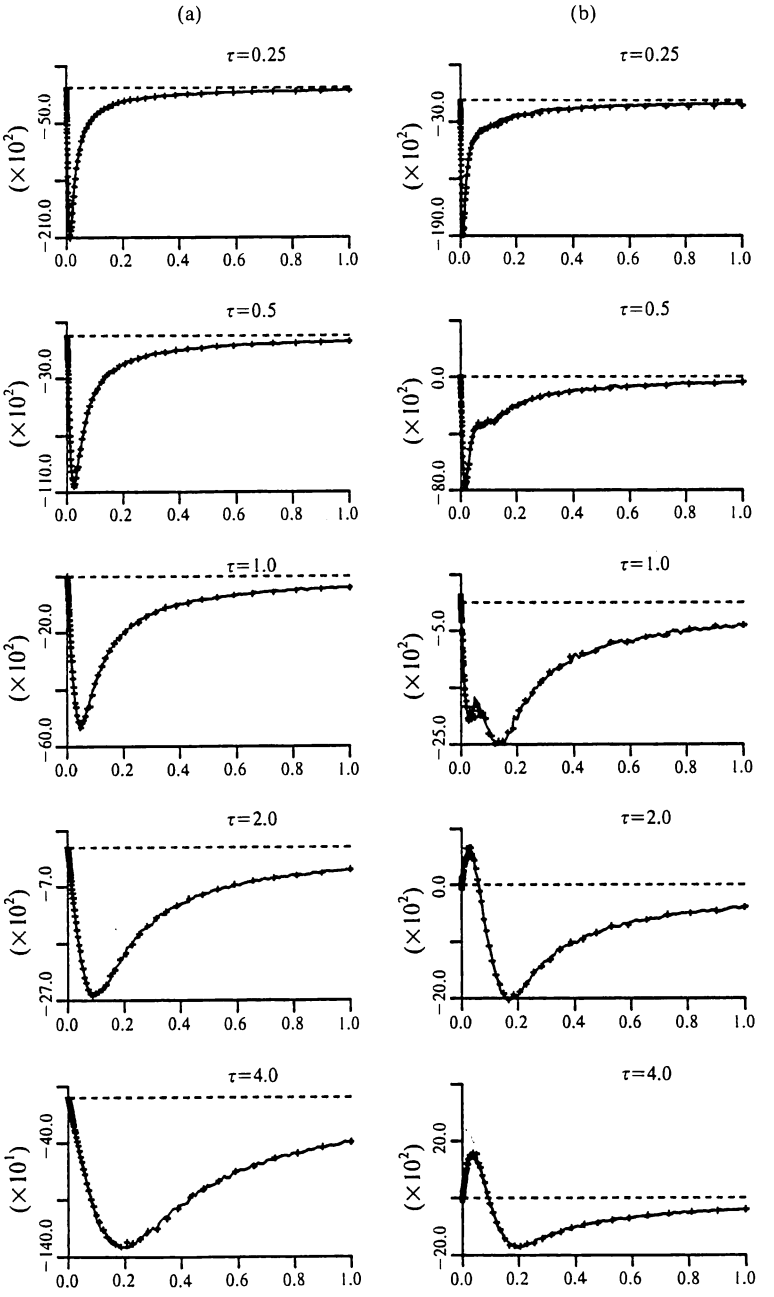


Fig. 2. ψ-functions by the way of (2.6) of the Gaussian smoothing model (4.3) for each τ: (a) $\psi^{(1)}$ for the prior and (b) $\psi^{(2)}$ for the posterior, respectively. Solid lines are obtained by connecting the ψ-values at the equidistantly sampled nodal points. + signs are ψ-values of distinct experiments at the geometrically sampled nodal points.

Table 1. Integral estimates of a Gaussian prior model: For A's and B's see the relations in (3.3) and (3.10), respectively. (a) and (c) are for the equidistant partition, and (b) and (d) for the geometrical partition. In the rows for $\log A$, their averages and corresponding standard errors are given.

	τ	0.25	0.5	1.0	2.0	4.0
(a)	A_1+A_2	−2216.0±2.5	−1859.7±2.1	−1573.8±1.9	−1334.7±1.2	−1190.9±1.1
		−2232.7±2.5	−1863.6±2.5	−1570.1±1.7	−1332.6±1.2	−1190.2±1.2
		−2212.2±2.8	−1855.7±2.3	−1572.1±1.6	−1333.7±1.2	−1188.6±1.2
		−2217.0±2.5	−1864.0±2.2	−1571.8±1.6	−1337.8±1.3	−1186.3±1.1
		−2222.3±2.3	−1863.9±1.9	−1572.2±1.6	−1332.7±1.2	−1188.4±1.2
	A_3	−1877.2±1.8	−1605.4±1.5	−1324.1±1.1	−1044.0±0.8	−768.9±0.5
		−1880.7±2.1	−1591.6±1.6	−1321.2±1.1	−1045.6±0.8	−769.0±0.5
		−1876.9±1.7	−1600.3±1.6	−1323.5±1.0	−1044.0±0.8	−769.4±0.5
		−1873.7±2.2	−1597.5±1.4	−1318.3±1.1	−1046.1±0.7	−767.9±0.6
		−1880.7±1.8	−1599.6±1.6	−1318.0±1.2	−1042.6±0.8	−767.7±0.5
	$\log A$	−342.2±3.2	−262.5±2.7	−250.9±2.0	−289.8±1.4	−420.3±1.3
(b)	A_1+A_2	−2431.0±3.9	−1866.1±3.4	−1570.2±3.2	−1334.9±2.3	−1187.5±1.9
		−2339.0±3.5	−1868.3±3.1	−1578.8±2.9	−1334.3±2.8	−1189.1±1.9
		−2394.4±4.0	−1886.8±3.6	−1565.2±2.6	−1338.2±2.2	−1188.2±1.7
		−2387.0±4.5	−1866.7±3.6	−1570.9±2.9	−1342.7±2.3	−1194.2±2.0
		−2364.0±3.8	−1885.5±3.7	−1578.3±2.8	−1329.8±1.8	−1192.3±2.0
	A_3	−1881.6±3.1	−1602.6±1.9	−1325.1±1.9	−1043.3±1.5	−770.3±1.0
		−1883.3±2.9	−1599.0±1.9	−1328.7±2.0	−1044.8±1.5	−765.5±1.0
		−1883.7±3.4	−1597.0±2.3	−1327.7±1.7	−1045.1±1.5	−767.4±1.0
		−1878.2±3.1	−1601.1±2.1	−1322.2±1.8	−1046.4±1.4	−765.5±1.1
		−1878.8±2.7	−1602.8±2.4	−1324.4±1.8	−1046.7±1.6	−766.5±1.0
	$\log A$	−502.0±5.0	−274.2±4.1	−247.1±3.4	−290.7±2.7	−423.2±2.2
(c)	B_1	−3686.8±13.	−2084.2±3.3	−1591.9±1.7	−1340.4±1.2	−1191.9±1.0
		−3671.9±11.	−2079.4±1.7	−1597.5±1.6	−1342.0±1.3	−1191.7±1.1
		−3640.6±14.	−2086.0±2.8	−1598.2±2.3	−1343.4±1.2	−1192.2±1.2
		−3701.1±5.7	−2092.2±3.0	−1593.1±1.8	−1335.9±1.2	−1195.3±1.2
		−3689.4±9.3	−2089.5±3.1	−1599.9±1.7	−1337.7±1.2	−1192.2±1.1
	B_2	−3346.8±7.5	−1800.9±3.0	−1336.9±1.4	−1048.5±0.8	−766.5±0.6
		−3306.8±7.5	−1805.0±2.5	−1340.3±1.7	−1048.3±0.8	−768.9±0.5
		−3337.0±5.9	−1814.9±2.5	−1339.6±1.5	−1048.8±0.9	−769.4±0.6
		−3351.8±6.5	−1810.7±3.6	−1332.6±1.4	−1047.6±1.1	−766.8±0.5
		−3338.6±11.	−1799.4±2.9	−1339.7±1.2	−1049.7±0.8	−768.3±0.5
	$\log A$	−341.8±14.	−280.8±4.1	−258.3±2.3	−291.3±1.5	−424.7±1.2
(d)	B_1	−2197.7±2.1	−1865.2±1.4	−1576.1±1.5	−1338.0±1.0	−1192.5±0.9
		−2195.5±2.2	−1863.2±1.9	−1573.3±1.3	−1337.1±1.1	−1194.2±1.2
		−2189.9±2.0	−1863.1±1.6	−1578.6±1.3	−1338.0±1.2	−1188.6±1.1
		−2190.9±2.0	−1860.4±2.0	−1572.8±1.5	−1340.0±1.1	−1192.0±1.1
		−2188.9±2.0	−1870.1±1.8	−1573.5±1.3	−1337.5±1.2	−1190.3±1.0
	B_2	−1876.9±2.1	−1595.3±1.8	−1323.0±1.5	−1044.2±1.0	−768.5±0.7
		−1883.9±2.0	−1602.1±1.8	−1325.2±1.4	−1050.9±1.0	−768.1±0.9
		−1879.6±2.3	−1604.0±1.5	−1318.6±1.2	−1048.1±1.1	−768.8±0.8
		−1880.3±2.3	−1597.5±1.6	−1321.4±1.3	−1047.1±1.0	−769.2±0.7
		−1877.2±2.0	−1601.9±1.7	−1321.0±1.4	−1048.8±1.1	−767.0±0.7
	$\log A$	−313.0±3.0	−264.2±2.4	−253.0±1.9	−290.3±1.5	−423.2±1.3

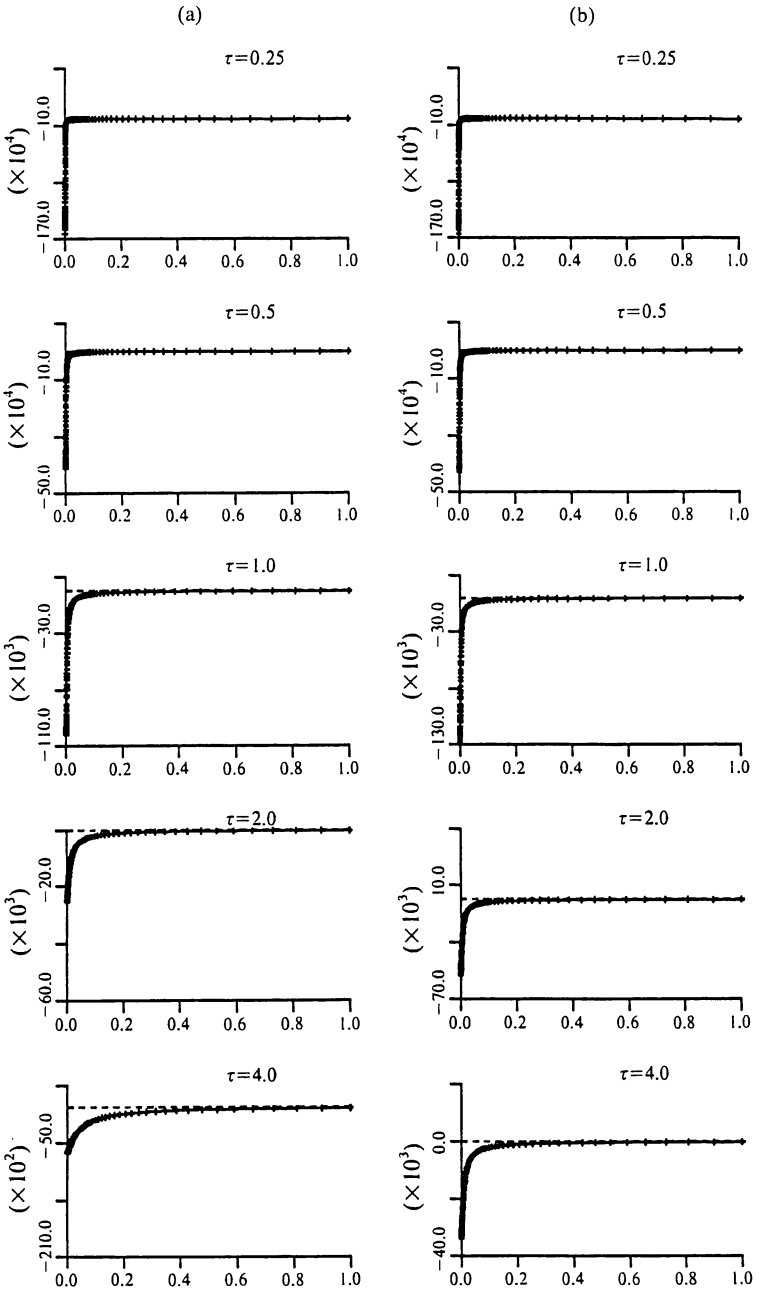


Fig. 3. ϕ-functions in (3.8) and (3.9) of the Gaussian smoothing model (4.3) for each τ: (a) $\phi^{(1)}$ for the prior and (b) $\phi^{(2)}$ for the posterior, respectively. Solid lines and + signs are the same as those described in Fig. 2.

B_1 and B_2, respectively, and the corresponding integral estimates are almost within a few times of the standard error to each other, except when the value of τ is very small. To investigate the significant difference of the integral estimates when τ is small, the relation (3.11) is used: since the values of A_3 and B_2 agree very well with one another when $\tau = 4.0$, the average of A_3's and B_2's for all experiments is regarded to be the true value. Then, the theoretical values for other τ's are calculated using (3.11). These are -1877.0, -1599.8, -1322.5, -1045.2 and -768.0 for $\tau = 0.25$, 0.5, 1.0, 2.0 and 4.0, respectively. Comparing these with the corresponding estimates in Tables 1(a)–(d), the equidistant partition for the method by way of (3.3) and the geometrical partition for (3.10) (that is, (a) and (d) in Table 1) are recommended for the computation of the current model with a limited number of nodal points. The maximum log Bayesian likelihood is attained at $\tau = 1.0$.

In order to find the parameter $\theta = (\theta_{ij})$ which (globally) maximizes the log posterior, or the penalized log-likelihood

$$\log L(\theta; Y) - Q(\theta; \tau) , \tag{4.4}$$

but when it is not easy to use the standard nonlinear optimization technique, a useful method will be Besag's ICM (iterated conditional mode; Besag (1986)). This is related to the Gibbs sampler (Geman, S. and Geman, D. (1984)) in the current model. That is to say, a set of samples can be simulated without any rejection using the Gaussian conditional transition probability in such a way that the sample at any coordinate (i, j) is given by the Gaussian random variable

$$N\left(\frac{Y_{ij}\tau^2 + 2\bar{\Theta}_{ij}\sigma^2}{\tau^2 + 2\sigma^2}, \frac{\sigma^2\tau^2}{2(\tau^2 + 2\sigma^2)} \right), \tag{4.5}$$

where $\bar{\Theta}_{ij} = (1/4) \sum_{(k,m) \in R_{ij}} \Theta_{km}$ is the average of the nearest neighbours, and Y_{ij} is the observed data at (i, j). In order to carry out the maximization of the posterior, take an arbitrary coordinate (i, j) by turns, and then the value Θ_{ij} is replaced by the mean $(Y_{ij}\tau^2 + 2\bar{\Theta}_{ij}\sigma^2)/(\tau^2 + 2\sigma^2)$ of the normal conditional distribution in (4.5) on the nearest neighbours. This is continued until the maximum of the penalized log likelihood in (4.4) is attained. Then the eventual states $\{\hat{\Theta}_{ij}\}$ are expected to realize the posterior mode.

Besides the ICM procedure for such a Gaussian model, we mainly adopt the posterior mean (Bayes estimates) in (3.12), or (3.14), especially when we consider a non-Gaussian model. The posterior mean is plotted in Fig. 4 for $\tau = 0.25$, 0.5, 1.0, 2.0 and 4.0. It appears that the selected estimate with $\tau = 1.0$ is reasonably smooth and suggests some discontinuous jumps across the mid interior, despite the low signal to noise ratio in the

data. Incidentally, the estimate of the pattern obtained by the ICM was quite similar to the posterior mean in the present case. The difference between this estimate and the true pattern seems to be almost inside the estimated standard error based on (3.15).

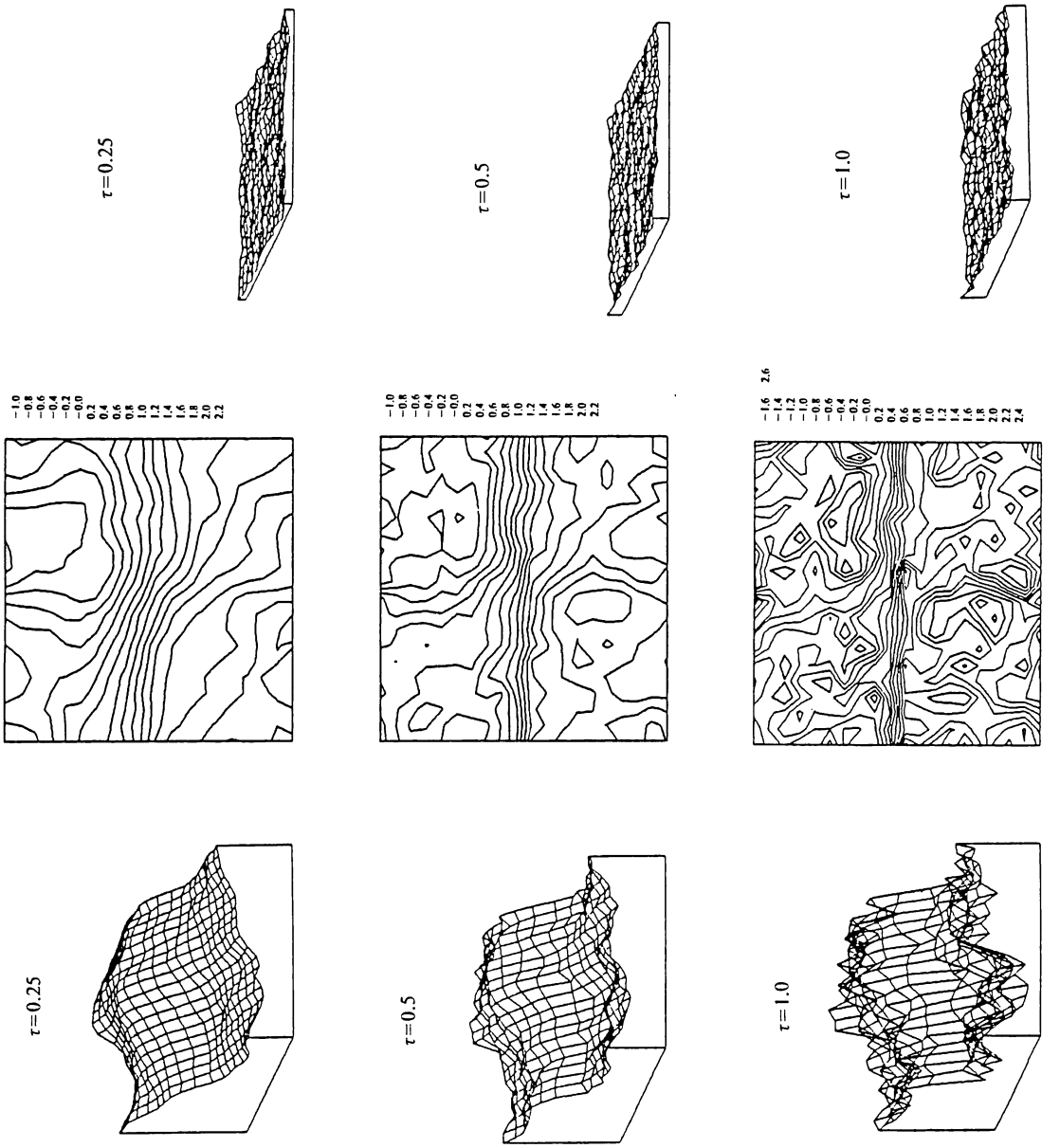
Next, we consider the potential of the penalty function

$$Q(\theta;\tau)=\frac{1}{2}\sum_{i=1}^{20}\sum_{j=1}^{20}\sum_{(k,m)\in R_{ij}}\log\left\{1+\left(\frac{\theta_{ij}-\theta_{km}}{\tau}\right)^2\right\}, \tag{4.6}$$

where R_{ij} is the same as defined in the above. This potential function aims at the situation where a portion of the probability is allotted to outliers for the discontinuity besides the smooth changes in the other portion, following the success in Kitagawa (1987) where the Cauchy distribution was used to recover a step function from a noisy data in a one-dimensional case. In a similar manner to that previously treated in a Gaussian prior model, the integral estimates are compared between the two types of potentials in (3.4) and (3.6) for the posterior as well as (3.5) and (3.7) for the prior. The figures related to the aboves are given in Figs. 5(a), (b), 6(a) and (b), respectively, corresponding to those in Figs. 2(a), (b), 3(a) and (b). Table 2 provides a list of the integral estimates and the Bayesian log likelihood, together with their standard errors. The corresponding values agree reasonably well with one another and they are also almost within a few times of the standard error to each other, except when the value of τ is small. The theoretical values for A_3 were calculated by the relation (3.11) in a similar manner to that in the above Gaussian prior model: they are -1929.9, -1652.6, -1375.4, -1098.1 and -820.9 for $\tau=0.25$, 0.5, 1.0, 2.0 and 4.0, respectively. Comparing these with the estimates in Table 2, the geometrical partition for (3.10) is recommended to be used here. The maximum log Bayesian likelihood is attained at $\tau=1.0$, but the values themselves are not improved when compared to those in the Gaussian prior model. Also, the estimated pattern of the posterior mean shown in Fig. 7 is not what I expected from our potential for a Cauchy type distribution. It may not be that 39 discontinuous jump among 400 random variables on a 20×20-lattice can be treated as outliers expressed by the heavy tailed distribution such as a Cauchy type.

Another type of modeling for the smoothing with some possible discontinuity for the current example in (4.1) leads to the use of line processes introduced by Geman, S. and Geman, D. (1984). Consider the potential for the prior,

$$Q(\theta;\tau)=\frac{1}{2}\sum_{i=1}^{20}\sum_{j=1}^{20}\frac{1}{2\tau^2}\sum_{(k,m)\in R_{ij}}v_{ij;km}(\theta_{ij}-\theta_{km})^2+\sum_{v\in C}R(v), \tag{4.7}$$


τ=0.25
τ=0.5
τ=1.0
τ=0.25
τ=0.5
τ=1.0

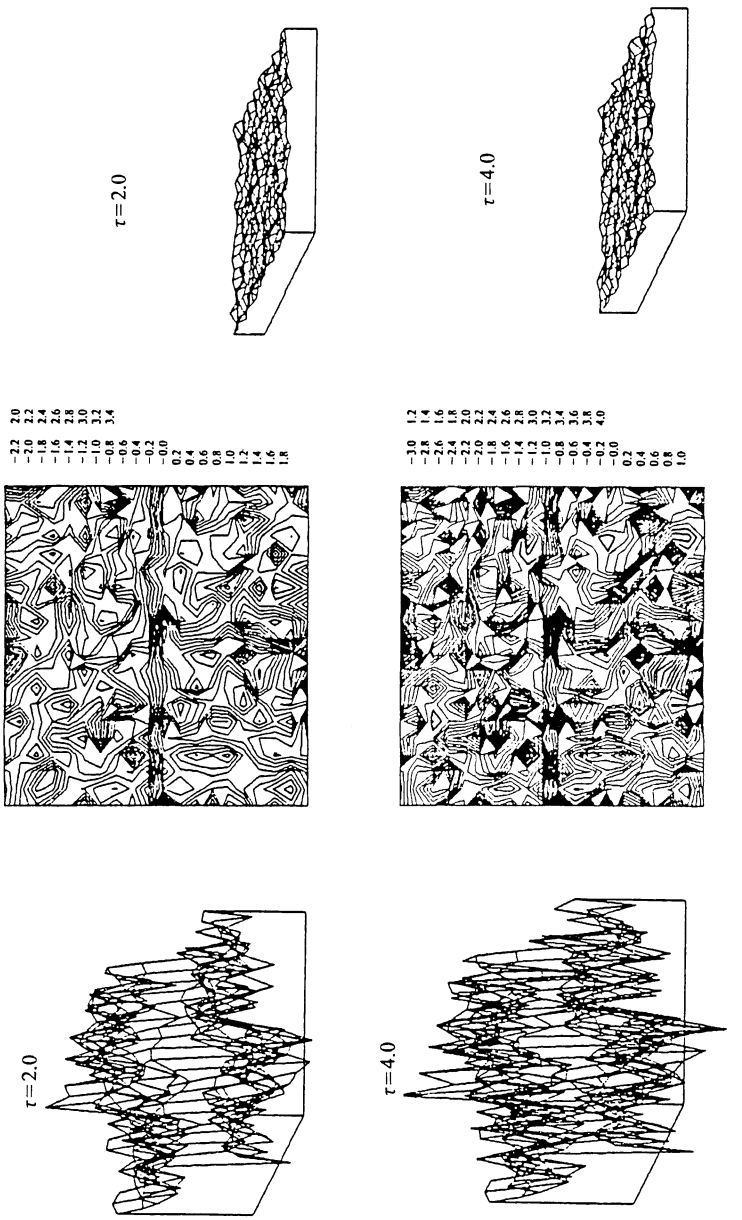


Fig. 4. Graphs of bird's-eye-view and contour lines of the posterior mean estimates of the Gaussian smoothing model (4.3) and bird's-eye-view of their standard errors at every lattice points for each $\tau = 0.25, 0.5, 1.0, 2.0$ and 4.0.

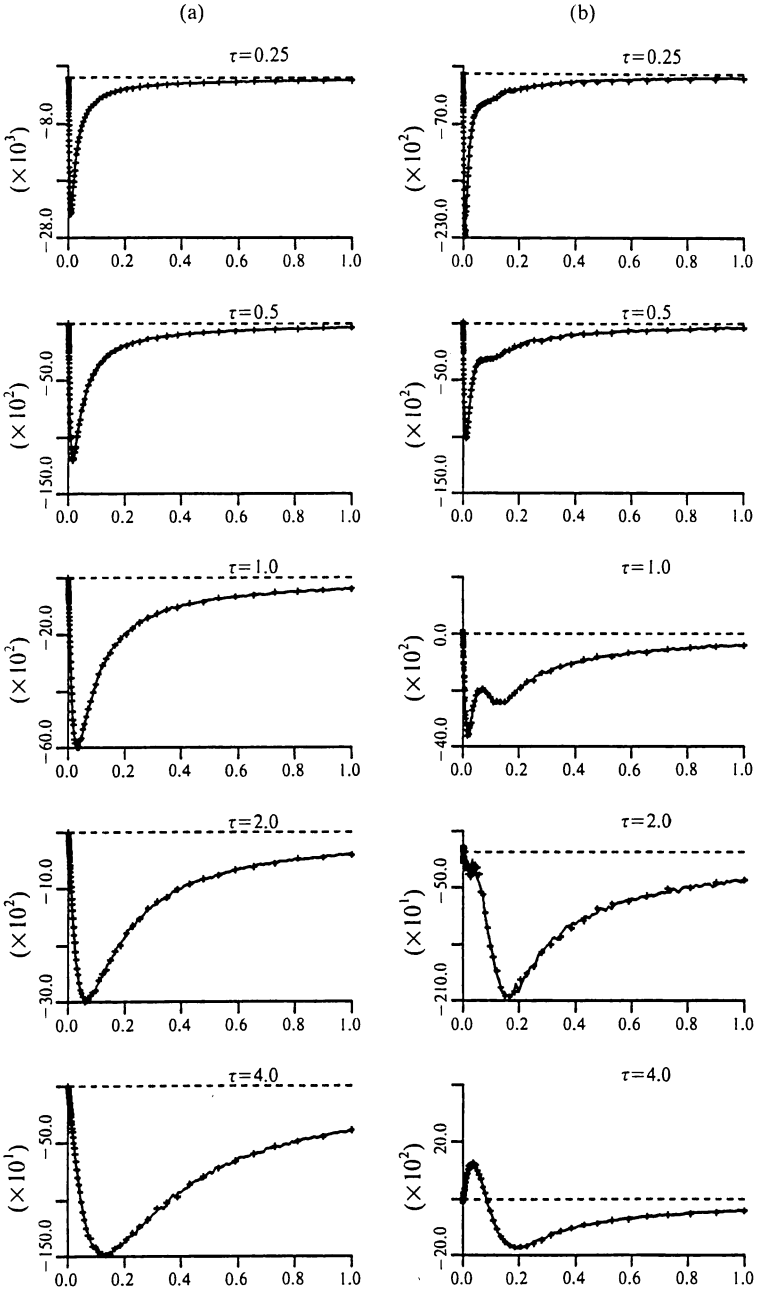


Fig. 5. ψ-functions by the way of (2.6) of the Cauchy type smoothing model (4.6) for each τ: (a) $\psi^{(1)}$ for the prior and (b) $\psi^{(2)}$ for the posterior, respectively. Solid lines and + signs are the same as those described in Fig. 2.

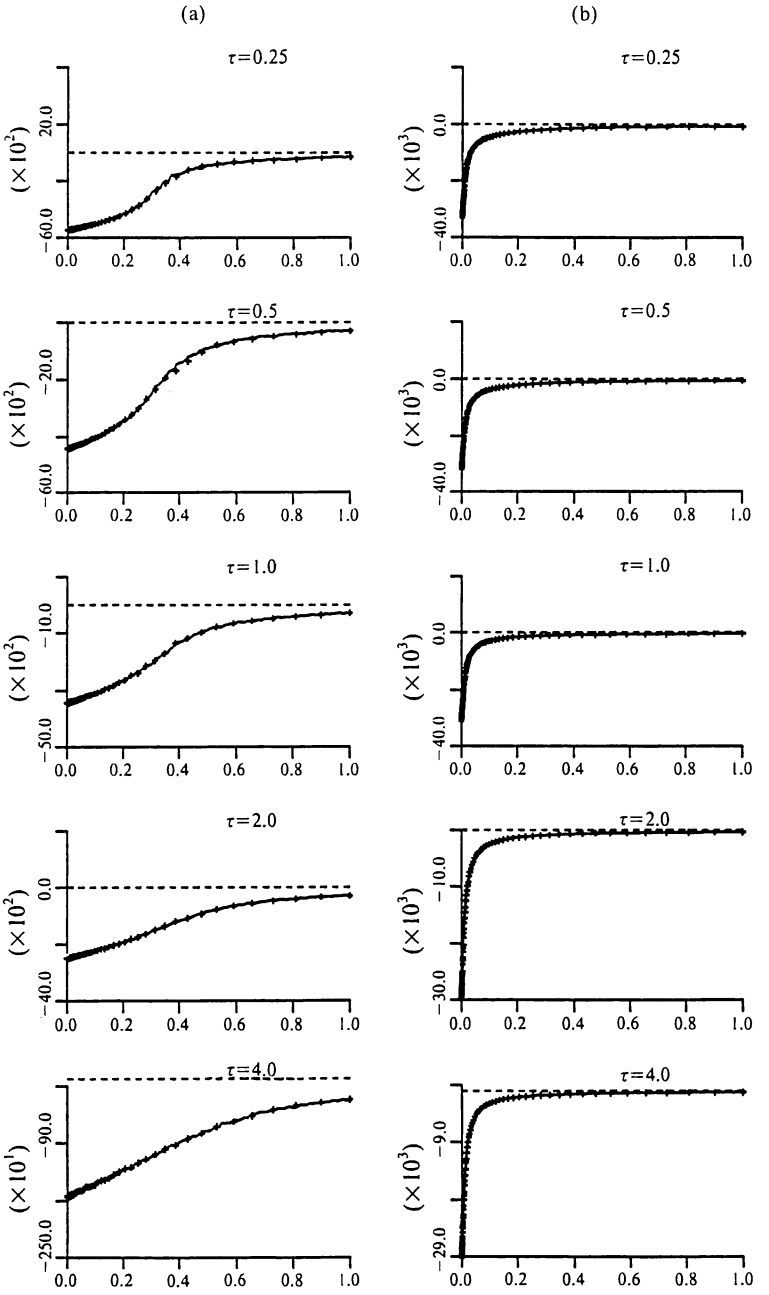


Fig. 6. ϕ-functions of the Cauchy type smoothing model (4.6) for each τ: (a) $\psi^{(1)}$ for the prior and (b) $\psi^{(2)}$ for the posterior, respectively. Solid lines and + signs are the same as those described in Fig. 2.

Table 2. Integral estimates of a Cauchy type prior model: See Table 1 for another description.

	τ	0.25	0.5	1.0	2.0	4.0
(a)	A_1+A_2	−2244.0±2.1	−1904.3±1.9	−1626.6±1.6	−1395.8±1.2	−1230.3±1.1
	A_3	−1925.2±1.6	−1644.5±1.2	−1374.3±1.0	−1098.0±0.7	−820.7±0.5
	$\log A$	−318.8±2.6	−259.8±2.2	−252.3±1.9	−297.8±1.4	−409.6±1.3
(b)	A_1+A_2	−2361.4±3.5	−1921.6±4.2	−1625.0±2.6	−1393.4±2.3	−1235.7±2.1
	A_3	−1941.1±2.6	−1656.3±1.9	−1372.4±1.7	−1099.0±1.2	−820.6±0.8
	$\log A$	−420.3±4.4	−265.3±4.6	−252.6±3.1	−294.4±6.7	−415.1±2.2
(c)	B_1	−2222.4±1.9	−1904.5±1.7	−1626.0±1.2	−1394.4±1.2	−1234.0±1.1
	B_2	−1942.7±2.5	−1677.0±2.8	−1378.5±1.3	−1100.0±1.3	−820.4±0.8
	$\log A$	−279.7±3.1	−227.5±3.3	−247.5±1.8	−294.4±1.8	−413.6±1.4
(d)	B_1	−2214.4±1.4	−1902.7±1.1	−1628.6±1.1	−1395.2±1.0	−1234.7±0.9
	B_2	−1934.7±1.4	−1650.7±1.2	−1377.4±0.9	−1095.4±0.7	−820.4±0.4
	$\log A$	−279.7±2.0	−252.0±1.6	−251.2±1.4	−299.8±1.2	−414.3±1.0

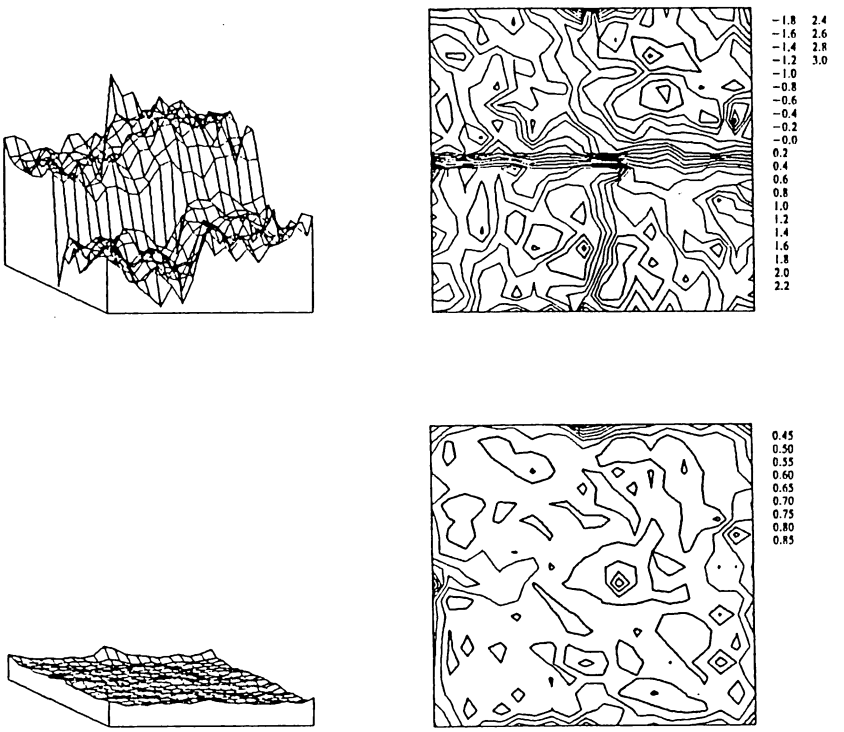


Fig. 7. Graphs of bird's-eye-view and contour lines of the posterior mean estimates of the Cauchy type smoothing model and those of their standard errors at every lattice points for $\tau = 1.0$, where the Bayesian likelihood is the maximum.

which is extended from the one in (4.3). Here, $\nu = (\nu_{ij;km} \in \{0, 1\})$ is defined on a set of cliques C, $\nu_{ij;km} = 0$ and 1 stands for the discontinuity or continuity, respectively, at the clique connecting (i, j) and its nearest neighbour (k, m), and $R(\nu)$ is the potential associated with the cliques. The construction of a good potential $R(\nu)$ is a considerable task in its own right and also outside the scope of the present paper. Rather, we simply assume here that the configuration of the discontinuous cliques in (4.7) is known and fixed, so that we can ignore the second term of the potential in (4.7) as a constant. Then we would like to see that the maximized log Bayesian likelihood is significantly improved here, compared to any of the previous analyzed models. A similar evaluation was considered by Akaike and Ishiguro (1983) where $\nu_{ij;km} = 0.1^2$ was set as the discontinuity on the requirement of the feasibility of the analytical calculation of ABIC for a certain time series.

In a similar manner to that shown for the previous models, the integral estimates are compared among the combination of two types of potentials and the two distinct configurations of the nodes of the partition. For $\tau = 0.125$, 0.25, 0.5, 1.0 and 2.0, the estimated figures of ϕ- and ψ-functions of the aboves are given in Figs. 8(a), (b), 9(a) and (b), respectively, corresponding to those in Figs. 2(a), (b), 3(a) and (b). Table 3 provides the list of integral estimates and the Bayesian log likelihood together with their standard errors. The theoretical values for A_3 were calculated by the relation (3.11): they are -2133.7, -1856.5, -1579.2, -1301.9 and -1024.7 for $\tau = 0.125$, 0.25, 0.5, 1.0 and 2.0, respectively. The significant discrepancies in the corresponding values for smaller τ's were seen due both to the unsatisfactory number of steps for the simulation and to the design of the partition for the trapezoidal formula: see the extremely large size of the troughs in ψ- and ϕ-functions in Figs. 8 and 9. Nevertheless, the maximum of log Bayesian likelihood is attained at $\tau = 0.5$, the value of which is significantly improved compared to that in the previously analyzed models. The patterns of the posterior mean and its marginal standard errors are shown in Fig. 10.

5. Concluding remark and discussion

The Monte Carlo method for the objective selection of the optimal prior distribution is provided, where the integration of very high dimensional functions is required to get the normalizing constants of the posterior and even of the prior distribution for the implementation. The logarithm of the high dimensional integral is reduced to a one-dimensional integration of ψ and ϕ-functions with respect to a scalar parameter over the range of the unit interval.

The theoretical error estimate for the integration is given, and two distinct and rather independent methods for integral estimation are sug-

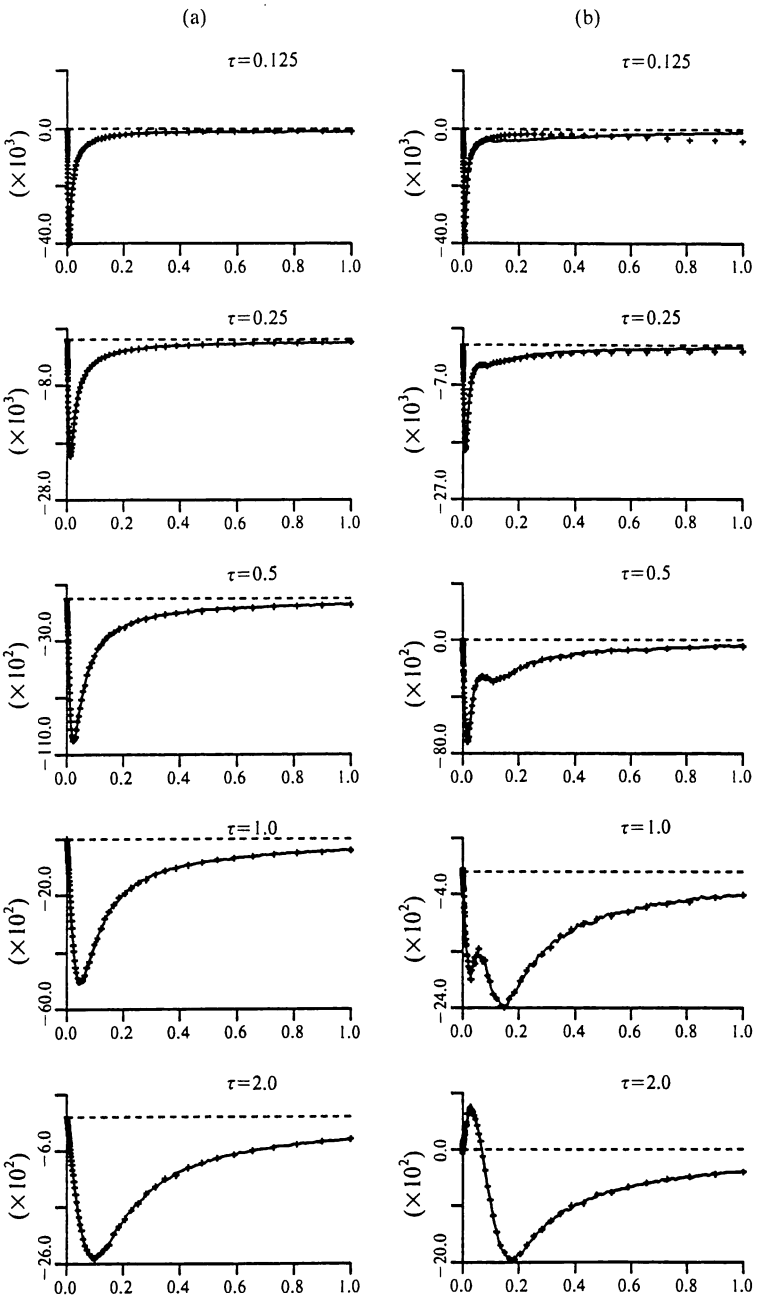


Fig. 8. ψ-functions by the way of (2.6) of the Gaussian smoothing model (4.7) with a known discontinuity: (a) $\psi^{(1)}$ for the posterior and (b) $\psi^{(2)}$ for the prior, respectively. Solid lines and + signs are the same as those described in Fig. 2.

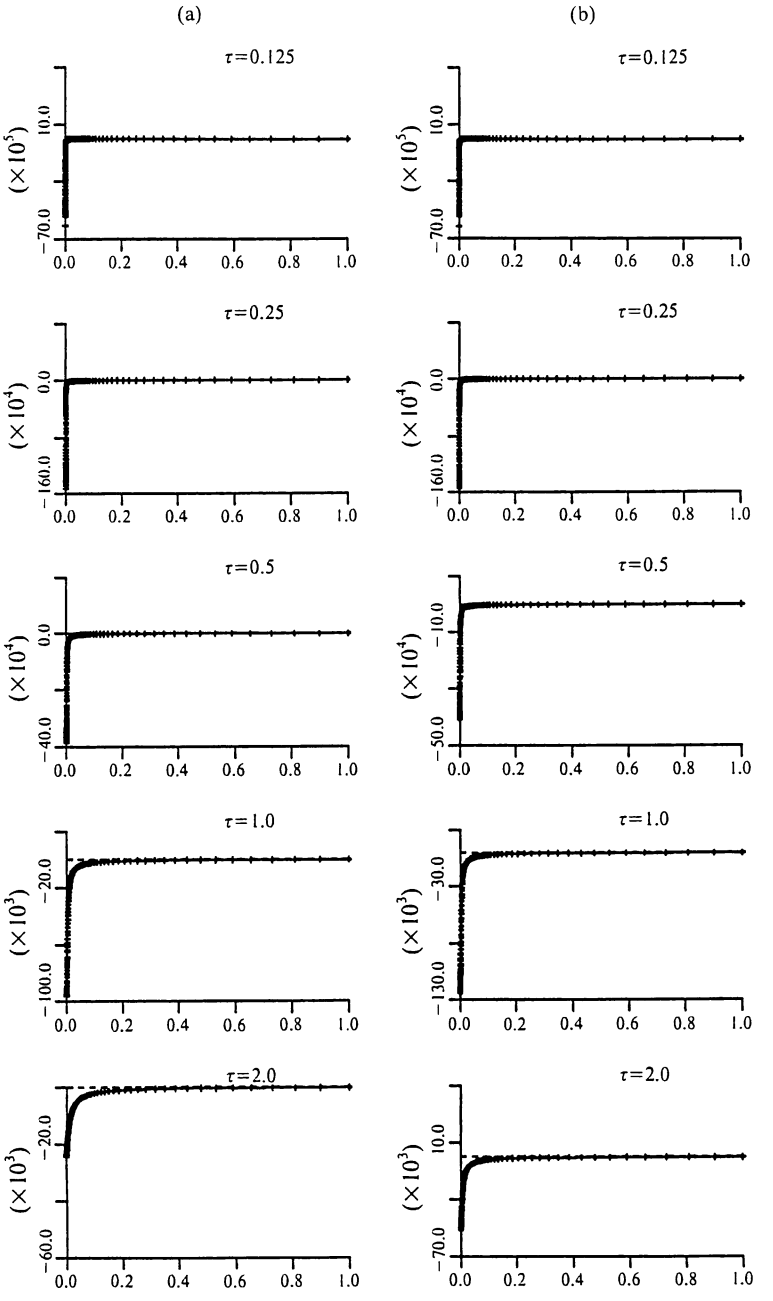


Fig. 9. ϕ-functions of the Gaussian smoothing model (4.7) with a known discontinuity: (a) $\phi^{(1)}$ for the posterior and (b) $\phi^{(2)}$ for the prior, respectively. Solid lines and + signs are the same as those described in Fig. 2.

Table 3. Integral estimates of a Gaussian prior model with discontinuity: See Table 1 for another description.

	τ	0.125	0.25	0.5	1.0	2.0
(a)	A_1+A_2	−3819.5±4.9	−2172.9±2.9	−1806.7±2.0	−1538.2±1.8	−1315.5±1.3
	A_3	−2343.8±2.5	−1887.1±2.0	−1586.4±1.6	−1303.5±1.1	−1024.0±0.8
	$\log \Lambda$	−1457.7±5.5	−285.8±3.5	−220.3±2.6	−234.7±2.1	−291.5±1.5
(b)	A_1+A_2	−3833.7±8.8	−2526.4±3.9	−1834.7±3.7	−1546.7±3.0	−1314.5±2.2
	A_3	−2309.8±3.2	−1906.7±2.3	−1604.3±2.4	−1311.0±1.6	−1023.2±1.2
	$\log \Lambda$	−1523.9±9.4	−619.7±4.5	−230.4±4.4	−235.7±3.4	−291.3±2.5
(c)	B_1	−4016.9±6.1	−2077.9±2.0	−1794.3±1.9	−1535.9±1.4	−1315.1±1.1
	B_2	−2168.6±2.3	−1998.7±2.0	−1584.7±1.9	−1303.0±1.1	−1026.5±1.0
	$\log \Lambda$	−1848.3±6.5	−389.4±3.0	−209.6±2.7	−232.9±1.8	−288.6±1.5
(d)	B_1	−10837.7±54.	−3779.3±3.5	−2005.5±3.7	−1555.4±1.5	−1316.5±1.0
	B_2	−9139.7±15.	−3377.1±8.3	−1777.1±3.4	−1325.2±1.4	−1025.0±1.0
	$\log \Lambda$	−1698.0±56.	−402.2±9.0	−228.4±5.0	−230.2±2.1	−291.5±1.4

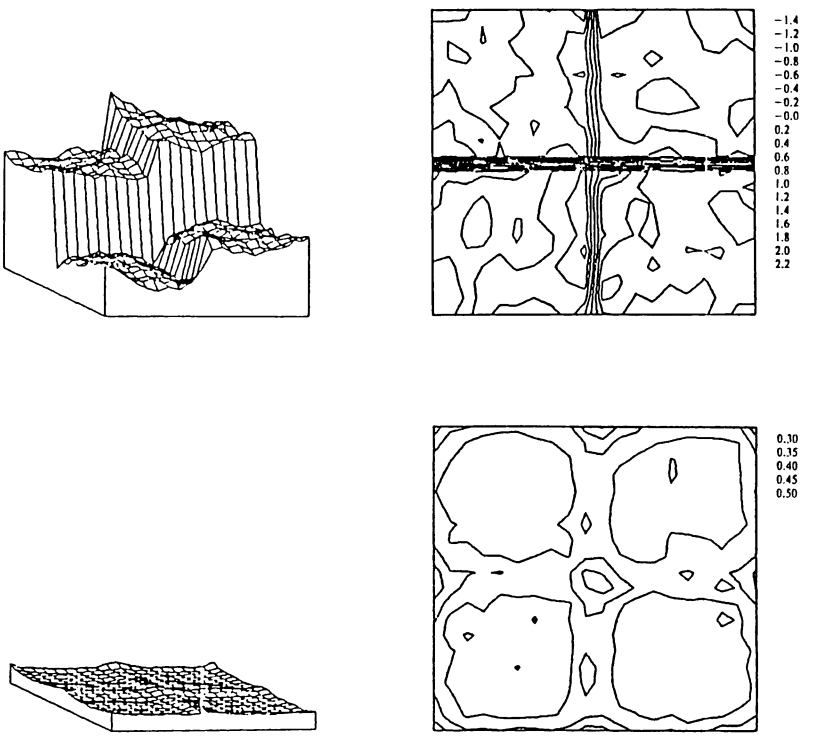


Fig. 10. Graphs of bird's-eye-view and contour lines of the posterior mean estimates of the model (4.7) with a known configuration of discontinuity and those of their standard errors at every lattice points for $\tau = 0.5$, where the Bayesian likelihood is the maximum.

gested to be used for the comparison of their outputs. To check the reliability of the proposed procedure, illustrative artificial data of the lattice system were analyzed by the suggested two integration methods for a few models under two conditions of the integration. The improvement of the integration's accuracy is substantial in comparison with the conventional crude Monte Carlo integration. Having decided the prior, the Bayes estimate or the posterior mean is mainly used instead of the posterior mode. All of these methods are based on the simulation of Gibbs distributions such as Metropolis' Monte Carlo algorithm. An advantage of the present method is that we have essentially no practical restrictions in modeling the prior and the likelihood.

It was found that extremely deep troughs in ψ- and ϕ-functions near $\sigma = 0$ cause the significant bias of the integrals under a limited number of partitions. Needless to say, the size of the integrand reflects the size of the integral, and consequently the size of the estimated error. For an objective Bayesian procedure, we are interested in the accuracy of the ABIC value up to an order less than 1.0. Therefore, on the one hand, for the optimal design of the nodal points in the numerical integration, I hope that the automatic adaptive routine (see Mori (1986), Chap. 12, for instance) will be suitable for our integration, since it is crucial that the configuration of nodal points should be properly designed, taking the variation of the functions into consideration. On the other hand, for the reduction of the size of ψ- and ϕ-functions, I would recommend to find a proper function $f_0(\boldsymbol{x})$, in Subsection 2.2, although I have not worked this out in the present paper. For example, for the extended model (4.7), the model (4.3) would be a good choice of $f_0(\boldsymbol{x})$, where the estimated integral is used in (2.13).

Acknowledgements

I have benefited greatly from useful discussions with Masaharu Tanemura who has been working with me on the estimation of spatial point patterns. Koichi Katsura generously helped with programming and computational implementations; the figures in this paper were also prepared by him. I would also like to thank Yukito Iba for his useful information on relevant references. Finally, thanks are due to the referees for their useful suggestions in the revision of this paper.

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. Bernard *et al.*), University Press, Valencia, Spain.

Akaike, H. and Ishiguro, M. (1983). Comparative study of the X-11 and BAYSEA procedure of seasonal adjustment, *Applied Time Series Analysis of Economic Data*, 17–53, U.S. Department of Commerce, Bureau of the Census.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussions), *J. Roy. Statist. Soc. Ser. B*, **36**, 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussions), *J. Roy. Statist. Soc. Ser. B*, **48**, 259–302.

Binder, K. (1986). Introduction: Theory and technical aspects of Monte Carlo simulations, *Monte Carlo Methods in Statistical Physics*, Topics in Current Physics, Vol. 7, (ed. K. Binder), Springer-Verlag, Berlin.

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

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

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities, *Biometrika*, **58**, 255–277.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*, Methuen, London.

Ibragimov, I. A. and Linnik, Yu. V. (1971). *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff, Groningen.

Ishiguro, M. and Sakamoto, Y. (1983). A Bayesian approach to binary response curve estimation, *Ann. Inst. Statist. Math.*, **35**, 115–137.

Kirkpatrick, S., Gellat, C. D., Jr. and Vecchi, M. P. (1983). Optimization by simulated annealing, *Science*, **220**, 671–680.

Kitagawa, G. (1987). Non-Gaussian state space modeling of nonstationary time series (with discussion), *J. Amer. Statist. Assoc.*, **82**, 1032–1063.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines, *J. Chem. Phys.*, **21**, 1087–1092.

Mori, M. (1986). *Programming of FORTRAN77 for Numerical Analyses* (in Japanese), The Iwanami Computer Science Series, Iwanami Publisher, Tokyo.

Ogata, Y. (1988). A Monte Carlo method for the objective Bayesian procedure, Research Memo. No. 347, The Institute of Statistical Mathematics, Tokyo.

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

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

Ogata, Y. and Tanemura, M. (1981*a*). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure, *Ann. Inst. Statist. Math.*, **33**, 315–338.

Ogata, Y. and Tanemura, M. (1981*b*). Approximation of likelihood function in estimating the interaction potentials from spatial point patterns, Research Memo. No. 216, The Institute of Statistical Mathematics, Tokyo.

Ogata, Y. and Tanemura, M. (1981*c*). A simple simulation method for quasi-equilibrium point patterns, Research Memo. No. 210, The Institute of Statistical Mathematics, Tokyo.

Ogata, Y. and Tanemura, M. (1984*a*). Likelihood analysis of spatial point patterns, Research Memo. No. 241, The Institute of Statistical Mathematics, Tokyo.

Ogata, Y. and Tanemura, M. (1984*b*). Likelihood analysis of spatial point patterns, *J. Roy. Statist. Soc. Ser. B*, **46**, No. 3, 496–518.

Ogata, Y. and Tanemura, M. (1989). Likelihood estimation of soft-core interaction potentials for Gibbsian point patterns, *Ann. Inst. Statist. Math.*, **41**, 583–600.

Ogata, Y., Imoto, M. and Katsura, K. (1989). Three-dimensional spatial variation of *b*-values of magnitude frequency distribution beneath the Kanto District, Japan, Research Memo. No. 369, The Institute of Statistical Mathematics, Tokyo.

Tanabe, K. and Tanaka, T. (1983). Fitting curves and surfaces by Bayesian method (in Japanese), *Chikyuu* (*Earth*), **5**, No. 3, 179–186.

Wood, W. W. (1968). Monte Carlo studies of simple liquid models, *Physics of Simple Liquids*, (eds. H. N. V. Temperley, J. S. Rowlinson and G. S. Rushbrooke), 115–230, Chap. 5, North-Holland, Amsterdam.