AN APPLICATION OF METROPOLIS-TYPE ALGORITHM TO A COMPLEX CLASSIFICATION PROBLEM

YUKITO IBA

Institute of Statistical Mathematics 106 Minami Azabu Tokyo

Email: iba@ism.ac.jp

ABSTRACT

'Classification of players' model is introduced and a Metropolis-type Monte Carlo algorithm for manipulating the model is presented. The model is an typical example of models for the 'complex' classification, in which the posterior means of statistics of discrete-valued parameters (labels, indicators) are prohibitively expensive to calculate with a naive method. The existence of the local maxima of the posterior distribution of this model can lead to the anomalous dynamics of the Markov chain and the long relaxation time of the Metropolis-type algorithm. The posterior marginals are, however, calculated in some examples in spite of the existence of a large number of local maxima. The Metropolis-type algorithm are shown to be useful in validating estimates as well as in calculating the optimal estimate under a given loss function. The estimation of the continuous-valued hyper parameters of the model is also discussed.

ver.20

1 INTRODUCTION

Analogy between Bayesian statistics and statistical mechanics draws much attention of statisticians recently. The algorithms which originate in statistical mechanics such as Metropolis-type Monte Carlo algorithm are successfully applied to difficult problems in Bayesian statistics (for example, Geman,D and Geman,S(1986), Hinton and Sejnowski(1986), Gelfand et al.(1990), see also reference in Section 3.1)

These algorithms are especially useful when we are working with statistical models with a large number of parameters. An important class of such models consists of statistical models for classification which contain a large number of discrete-valued parameters as labels or indices. These models can be interpreted as the extensions of the finite mixture models (Kay and Titterington(1986)). In this paper we call them 'complex classification models' because the posterior distributions of these models can have much more complex structure than that of ordinary finite mixture model as shown in Chapter 2.

Most of the previous studies of this type of models which has been conducted under the statistical mechanical analogy treat image processing problems (Geman,S. and Geman,D.(1984), Geman,D and Geman,S(1986), Marroquin et al.(1987), Gidas(1989), Besag et al.(1991), Qian and Titterington(1992), Iba(1991)) , where the segmentation or labelling of an image is considered. However, the analogy with statistical mechanics can be applied to the complex classification models which do not have any spatial structure.

In this paper we study an application of Metropolis-type Monte Carlo algorithm to an complex classification model without spatial structures. In Chapter 2 the model is introduced. In Chapter 3 a Metropolis-type Monte Carlo algorithm is applied to the model. In Chapter 4 the results of the numerical experiments are discussed.

2 AN EXAMPLE OF COMPLEX CLASSIFICATION — 'CLASSIFICATION OF PLAYERS' MODEL

Here we present an example of models for complex classification without any spatial structure. In this model, data are given as the *relations* among objects to be classified, in other words, data are assumed to be generated through the interaction among two or more objects to be classified.

We consider matches(pair comparisons) among K players. A player indexed by i plays N_{ij} matches against a player j and wins n_{ij} times to j.

For each pair of players i and j, the numbers N_{ij} and n_{ij} are recorded. The number N_{ij} can be equal to 0 for some pairs. By definition, $n_{ij} + n_{ji} = N_{ij}$. We assume there are several groups denoted by $1, 2, 3..G$ of players and each player belongs to one of these groups. Following conditions are assumed to characterize the groups.

- A player in the group A wins a player in the group B($A \neq B$, A=1,2.., B=1,2..) with probability p^{AB}.
- A players wins players in the same group with probability $\frac{1}{2}$. By definition $p^{BA} = 1 - p^{AB}$.

The likelihood of the model is

$$likelihood \sim \Pi_{i=1}^{K} \Pi_{j>i}^{K} L_{ij} \tag{1}$$

where

$$L_{ij} = \begin{cases} (p^{AB})^{n_{ij}} (p^{BA})^{n_{ji}} & \text{if i} \in \text{A and j} \in \text{B}(\neq A). \\ (\frac{1}{2})^{N_{ij}} & \text{if i and j are in the same group.} \end{cases} \tag{2}$$

The parameters $\{p^{AB}\}$ ($A = 1, 2..G$, $B = 1, 2..G$) of the binomial distributions determine dominance relations among the groups.

Under these assumptions, we want to infer, for example,

- the most plausible grouping of the players, or
- the probability that a player i wins a player$j$

based on the data $\{n_{ij}\}, \{N_{ij}\}$ of a set of matches. The estimation of 'hyper' parameters $\{p^{AB}\}$ is also an important problem.

We introduce discrete-valued parameters $\{S_i\}$, $S_i \in \{1, 2..G\}$, where each S_i indicates the group to which the player i belongs, i.e.

$$S_i = A \leftrightarrow S_i \in group\, A\,. \tag{3}$$

If we assign the same prior probability for any $\{S_i\}$, the posterior distribution of $\{S_i\}$ is

$$P_{pos}(\{S_i\}) = \frac{\exp(-E_{pos})}{\sum_{config.} \exp(-E_{pos})} \tag{4}$$

$$-E_{pos} = \sum_{i,j} n_{ij} f(S_i, S_j) \tag{5}$$

$$f(A,B) = \log p^{AB} \tag{6}$$

$$f(A,A) = \log \frac{1}{2} \tag{7}$$

where the summation $\sum_{config.}$ run through all the possible combination of $\{S_i\}$.

In our Bayesian framework, the most probable classification of the players is given as the mode of the prior distribution (4). The estimation of quantity X such as 'the probability that a player i wins a player$j$' is reduced to the calculation of the posterior mean of the corresponding statistics $X(\{S_i\})$

$$\langle X(\{S_i\})\rangle_{pos} = \frac{\sum_{config.} X(\{S_i\}) \exp(-E_{pos})}{\sum_{config.} \exp(-E_{pos})} \tag{8}$$

and the reliability of the estimate is measured by the tails of the marginal distribution.

In this paper we assume the uniform prior distribution of $\{S_i\}$. This means that we implicitly assume the sizes of each group are nearly equal. When groups of considerably different sizes are expected, we should assume a different prior distribution of $\{S_i\}$. For example the prior distribution

$$\pi(\{S_i\}) = \Pi_{A=1..G} (\alpha_A)^{K_A} \tag{9}$$

could be used, where α_A is the hyper parameter which determines the prior probability that a player belongs to the group A and K_A is the number of the members of the group A. When the number of data is small and the hyper parameters $\{\alpha_A\}$ ($\sum_{A=1..G} \alpha_A = 1$) are estimated from the data, the posterior distribution with this prior would differ significantly from the one with the uniform prior.

Our model can be regarded as a generalization of the finite mixture model. For the finite mixture model we can also introduce discrete-valued parameter S_i as an indicator of the origin of each data. For example, when 1:1 mixture of two distributions f_a and g_b is considered, the discrete-valued parameter S_i is defined as follows: If y_i comes from the distribution f_a then we set $S_i = 1$; if y_i comes from the distribution g_b, $S_i = -1$. When the parameters of the components a, b are fixed and treated as hyper parameters, the posterior distribution $P_{pos}(\{S_i\})$ of $\{S_i\}$ has a following form:

$$P_{pos}(\{S_i\}) = \frac{\exp(-E_{pos}(\{S_i\})}{\sum_{config.} \exp(-E_{pos}(\{S_i\})} \tag{10}$$

$$-E_{pos}(\{S_i\}) = \sum_i (\frac{1+S_i}{2} \log f_a(y_i) + \frac{1-S_i}{2} \log g_b(y_i)) \tag{11}$$

These expressions for the finite mixture model correspond to (4) and (5) for our model. There is, however, an essential difference between our model (5) and the finite mixture model (11). In the finite mixture model the expression $-E_{pos}(\{S_i\})$ is the sum of the terms each of which contains only one S_i and does not contain interaction terms among S_i. Thus posterior means of a statistics $X(\{S_i\})$ is readily calculated for the finite mixture model with fixed hyper parameters when the statistics $X(\{S_i\})$ is represented as the sum of the terms each of which contains small number of S_i. In our model $-E_{pos}(\{S_i\})$ contains the direct interaction terms among S_i's and the posterior distribution of discrete-valued parameters $\{S_i\}$ is much more complicated. Posterior means are not easy to calculate even for simple statistics.

For our models, classification to three (or more) groups is much more complex than classification to two groups. The fundamental distinction between them is that the absolute 'strength' of each player cannot de defined in the three (or more) group case while it can be defined in the two group case. The typical example of the situation where absolute strength cannot be defined is found in three-group-classification model with $p^{12} = p^{23} = p^{31}$. This model is schematically shown in Fig.1. In this case the dominance relations of groups are *cyclic* and similar to that of Japanese game 'Jang Keng', where 'stone' wins 'scissors', 'scissors' win 'paper' and 'paper' wins 'stone'.(In our model the dominance relations are stochastic while they are deterministic in 'Jang Ken'.)

——Fig.1——

The three-group-classification model with 'complete cyclic dominance'.

Note that the posterior probability that

- a player i belongs to group A

does not always have proper meaning in our framework. For example, the posterior probability that a player i belongs to group A is 1/3 for any i and any A for the model defined in Fig.1. This is an obvious consequence of the fact that the posterior probability of a state is invariant under the 'cyclic' change of the labels of the groups in the model in Fig.1. This suggests an interesting application of the concept 'spontaneous symmetry breaking' in statistical physics. In this paper, however, we do not discuss this point further. Related topics are discussed in the theory of the associative memory

(Amit et al.(1985), see also papers reprinted in Mezard et al(1987).). If we concentrate on the statistics which is invariant under the 'cyclic' change of the labels of the groups, this aspect of the model can be ignored.

3 METROPOLIS-TYPE ALGORITHM

3.1 COMPUTATIONAL DIFFICULTIES IN OUR PROBLEM

As mentioned in the previous chapter, the direct calculation of the summation $\sum_{config.}$ in (8) is prohibitively expensive in our model even for medium size problems. It involves summation over G^K terms, where K is the number of players and G is the number of the groups. This is a typical situation which is often encountered in Bayesian statistics.

Another difficulty in our problem is the existence of a large number of local maxima in the posterior distribution. The search for the posterior mode by the hill climbing algorithm (local improvement algorithm) is often trapped by a local maximum. The number of local maxima can be quite large in the three-group-classification models even when data are well described by the model.

Approximation methods developed in statistical physics can be powerful tools for treating these difficulties. Metropolis-type Monte Carlo algorithm (Metropolis et al.(1953), Binder and Stauffer(1984), Binder(1986)) is the most general one among them and has been successfully applied to difficult problems in Bayesian statistics. [1] Interesting examples of the application of Metropolis-type algorithm to Bayesian statistics are found in Geman,S. and Geman,D.(1984), Geman,D and Geman,S(1986), Geman,S. and McClure(1987), Hinton and Sejnowski(1986), Smolensky(1986), Rumelhart et al.(1986), Marroquin et al.(1987), Ogata(1990), Gelfand et al.(1990), Besag et al.(1991), Ohtsuki and Kawato(1991) , Qian and Titterington(1992), and Iba(1991,1992).

[1] Note that Metroplis-type algorithm have also been applied to statistical problems in non-Bayesian framework, see Tanemura and Ogata(1981,1984), Penttinen (1984) and Frank and Strauss(1986) ; in these works the normalization constant of the likelihood function is calculated with Metropolis-type algorithm.

3.2 APPLICATION OF METROPOLIS-TYPE ALGORITHM

In a Metropolis-type algorithm, we construct a Markov chain whose unique invariant distribution coincides with the desired distribution. For the Markov chain with these properties, each state $\{S_i\}_t$ in its sample path can be regarded as sample from the given distribution, where $t = 1.2...$ is 'time step' of the Markov chain. Thus the average of any statistics X over the given distribution equals to the average of X over the sample path of the Markov chain and can be calculated through the simulation of the Markov chain.

The Markov chain should have the finite relaxation time τ, which assures that the states $\{S_i\}_t$ and $\{S_i\}_{t+\triangle}$ are independent samples if $\triangle >> \tau$. The length of the relaxation time τ is crucial to the convergence of the average of statistics. Unfortunately it is difficult to estimate the relaxation time τ *a priori*, which depends on the details of the algorithm as well as on the properties of the posterior distribution.

There are many ways to construct a Markov chain with these properties, most of which are based upon the 'detailed balance condition' (Binder(1986)). The two of the simplest are the original form of Metropolis algorithm (Metropolis et al.(1953)) and Gibbs sampler(heat bath algorithm). Here we apply the original form of Metropolis algorithm to our model to generate samples from the posterior distribution $P_{pos}(\{S_i\})$.

In our algorithm, starting from an arbitrary initial state, we repeat following procedures to generate a sample path $\{\{S_i\}_t\}$:

1. Choose $i \in \{1..K\}$ and $S_i^{new} \in \{1..G\}$ randomly.

2. Calculate the odds P_{new}/P_{old}, where P_{old} is the posterior probability of the current state and P_{new} is the posterior probability when we change S_i to $S_i{}^{new}$.

3. Generate a uniform random number $rnd \in (0, 1)$.
   Change S_i to $S_i{}^{new}$, if and if only $rnd < P_{new}/P_{old}$.

4. $\{S_i\}_t :=\{S_i\}$ and $t := t + 1$.

The benefit of Metropolis-type algorithm is that the ratio P_{new}/P_{old} does not contain the normalization constant of the distribution which is extremely costly to calculate.

The number of the iteration of Metropolis-type algorithm is measured by the unit 'MCS' (Monte Carlo Steps). 1 MCS corresponds to 'system size' iterations. In our case, 1000 MCS means 30000 iterations of the procedures defined in this section when the number of the players K is 30.

3.3 TEMPERATURE

It is interesting to compare Metropolis-type algorithm with the hill climbing algorithm for the search of the posterior mode. A simple version of the hill climbing algorithms is obtained as a modification of the procedure 3 of the Metropolis algorithm as follows.

3. * Change S_i to S_i^{new}, if and if only $1 < P_{new}/P_{old}$.

It is convenient to introduce artificial 'temperature' T to interpolate Metropolis algorithm and the hill climbing algorithm(Kirkpatrick et al.(1983), Geman,S and Geman,D(1984), Smolensky(1986), Hinton and Sejnowski(1986)). Temperature T defines the following modification of the procedure 3:

3. ** Change S_i to S_i^{new}, if and if only $rnd < (P_{new}/P_{old})^{1/T}$.

This algorithm, which will be called 'Metropolis algorithm at temperature T', generates a sample from the following 'temperature T' distribution:

$$P_T(X) = \frac{P_{pos}(X)^{1/T}}{\sum_{config.} P_{pos}(X)^{1/T}} \tag{12}$$

When temperature $T \to 0$, the Metropolis algorithm at T becomes the hill climbing algorithm and the distribution P_T shrinks to (the sum of) delta distribution(s) placed on posterior mode(s). Hereafter the average of a statistics X over the 'temperature T' distribution is denoted by $\langle X \rangle_T$. In particular $\langle X \rangle_{T=1} = \langle X \rangle_{pos}$ and $\langle X \rangle_{T=0}$ coincides with the estimator of X based on posterior mode if the mode is unique.

In the Metropolis algorithm at temperature $T \neq 0$ the change of the state is possible even when $P_{new} < P_{old}$ whereas in the hill climbing algorithm the change of the state is permitted only when $P_{new} > P_{old}$. This means the sample path of the Markov chain at temperature $T \neq 0$ can 'escape' from local maxima at $T = 0$.

4 NUMERICAL EXPERIMENTS

4.1 CONDITIONS OF EXPERIMENTS

We concentrate on the three-group-classification model with complete cyclic dominance (Fig.1). Artificial data are generated under the following conditions: (1) The total number of players K is 30 and the sizes of groups 1,2,3 are 10,10,10 respectably. (2) The number N_{ij} of matches between player i

and j is 1 with probability r and 0 with probability $1-r$. (3) The 'true' probability p^{AB}_{true} that a member in group A wins a member in group B is set as $p^{12}_{true} = p^{23}_{true} = p^{31}_{true} = 0.8$ for the most of the experiments. In Section 4.7, the probabilities $p^{12}_{true} = p^{23}_{true} = p^{31}_{true} = 0.7$ and $p^{12}_{true} = p^{23}_{true} = p^{31}_{true} = 0.5$ are also used.

In the experiments in Sections 4.3,4.4,4.5 and 4.6, the parameters p^{12}, p^{23} and p^{31} are *not* estimated from data. They are fixed to the 'true value' $p^{12} = p^{23} = p^{31} = 0.8$. In Section 4.7 they are estimated from data.

The physical random number generator at the Institute of Statistical Mathematics is mainly used as the source of random numbers. A pseudo random number generator based on *M-sequence* is also used(Fushimi and Tezuka(1981)).

4.2 PERFORMANCE OF THE COMPUTER PROGRAM

The required computational time is less than 1 minute / 10000 MCS on the mainframe HITAC M-682H for the $K = 30$ problem described in the previous section. This CPU includes CPU for the calculation of some statistics.

4.3 DYNAMICS

An important problem about Metropolis-type algorithm is how it works in problems in which hill climbing search for the posterior mode is often interrupted by local maxima. Interesting examples of such systems are found in, for example, van Hemmen and Morgenstern(1987) and Mezard et al.(1987).

In this section we examine the effects of the local maxima at temperature $T = 0$ on the dynamics of the Markov chain generated by the Metropolis algorithm in our model. The average number r of matches between a pair of players is 0.6 for the artificial data used in the following experiments.

While the concept of 'local maximum' cannot be rigorously defined in Metropolis-type algorithm at $T \neq 0$ in the finite size system, it is convenient to define some analogue of local maximum at $T = 0$. Let us consider disjoint subsets $\{M_\gamma\}(\gamma = 1..m)$ of the space of all possible states. We define Π_γ as the probability that a sample path starting from a state in M_γ remains in M_γ for sufficiently long time steps. If the probability Π_γ is large at temperature T, we regard M_γ as a sort of local maximum and referred as a 'metastable state' at temperature T. If each M_γ is a metastable state, the system is said to have m metastable states. A simplest type of metastable states consists of a single local maximum at $T = 0$ and states 'around' it. They can also contain several local maxima. Another example is given by

putting $M_1 = \{$global maximum and small number of states around it$\}$ and $M_2 = \{$all of the rest states$\}$. If the transition rate between M_1 and M_2 is considerably lower, we can regard M_1 and M_2 as metastable states. Note that the contribution of a metastable state to the averages of statistics (the 'weight' of the metastable state) is not determined by the *average* of the probabilities of the states in it; it is determined by the *sum* of them. This point is important because the number of states in a metastable state can be quite large.

In Fig.2a-Fig.2d the behavior of the statistics over the sample paths of the Markov chain is shown for the same data at various temperatures. In the upper panel of each figure, the 'distance' between $\{S_k\}_t$ and 'true classification'$\{S_k\}^{true}$ is plotted against t. The distance is measured by the Kullback-Leibler divergence

$$J(\ \{p_{ij}(\{S_k\})^{true}\}, \{p_{ij}(\{S_k\}_t)\}\ ) = \sum_{i \neq j} p_{ij}(\{S_k\}^{true}) \log \frac{p_{ij}(\{S_k\}^{true})}{p_{ij}(\{S_k\}_t)} \quad (13)$$

Here the probability $p_{ij}(\{S_k\})$ is defined as 'the probability with which the player i wins the player j when the classification $\{S_k\}$ is assumed', i.e.

$$p_{ij}(\{S_k\}) = p^{S_i S_j} \quad (14)$$

In the lower panel, the statistics $-E_{pos}(\{S_k\}_t)$ (=log likelihood + constant) is plotted against the same sample path. In Fig.3a-Fig.3d the first 650 MCS of the corresponding sequences of the state $\{S_k\}_t$ are shown.

— Fig.2a, Fig.2b, Fig.2c, Fig.2d —

The plots of the statistics over the sample paths, where the first 1000 MCS from randomly chosen initial configurations are shown. In the upper and lower panel of each figure, the statistics $J(\ \{p_{ij}(\{S_k\})^{true}\}, \{p_{ij}(\{S_k\}_t)\}\ )$ and $-E_{pos}(\{S_k\}_t)$ (=log likelihood + constant) are plotted against the fictitious time of the Markov chain. Temperatures T are 4.0, 2.0, 1.0 and 0.4 in Fig.2a, Fig.2b, Fig.2c, Fig.2d respectably.

— Fig.3a, Fig.3b, Fig.3c, Fig.3d —

The sample paths $\{\{S_k\}_t\}$. Fig.3a-Fig.3d correspond to the first 650 MCS of Fig.2a-Fig.2d respectably. The fictitious time of the Markov chain t is increased from the top to the bottom of each figure. Each raw of the pattern corresponds to the state $\{S_k\}$ (k = 1..30) at intervals of 10 MCS. In Fig.3b, a remarkable change observed in the neighborhood of 450 MCS is indicated by an arrow.

At temperature $T = 4.0$, a large number of states are visited in the sample path(Fig.2a, Fig.3a). At temperature $T = 2.0$, the states with larger posterior probability dominate (Fig.2b, Fig.3b). In Fig.2b, however, a 'burst' in the statistics is observed in the neighborhood of 450 MCS, which corresponds to the appearance of the states with lower posterior probabilities. The sequence $\{\{S_k\}_t\}$ of the patterns corresponding to Fig.2b is shown in Fig.3b, in which a remarkable change is observed in the neighborhood of 450 MCS. This change corresponds to the cyclic change of the labels of the groups ($1 \to 3$, $2 \to 1$, $3 \to 2$) and does not alter the values of the statistics shown in Fig.2b. The 'burst' of the statistics shown in Fig.2b corresponds to the apparently disordered metastable state in Fig.3b between two 'equivalent' sets of states with large posterior probability. At temperature $T = 1.0$ the sample path stays around the global mode of the posterior in the most of our observational time(Fig.2c, Fig.3c). The change corresponds to the cyclic change of the labels of the groups is not observed in Fig.3c. Finally, at temperature $T = 0.4$, the sample path is easily 'trapped' by one of the metastable states with smaller weights and considerable number of steps are needed for escaping (Fig.2d, Fig.3d).

We could expect following properties of the Metropolis algorithm for our model with uniform prior for $\{S_k\}$; if the model fits given data well and the number of the data is enough to determine the plausible solutions, the sample path is expected to stays around the global maximum at $T = 1.0$ as in Fig.2c; if given data are not enough to determine any plausible solution, a large number of states is observed at $T = 1.0$. There is, however, no guarantee that the relaxation time of the Markov chain is within reasonable range of time at temperature $T = 1.0$. Even when data are well described by the model, peculiar behavior due to local maxima can be observed at $T = 1.0$. In Fig.4 and Fig.5 an interesting example of such behavior is shown.

—– Fig.4 —–

The plots of statistics over the sample path for another artificial data, where the first 5000 MCS from a randomly chosen initial configuration is shown. In the upper and lower panel of each figure, the statistics $J(\ \{p_{ij}(\{S_k\})^{true}\}, \{p_{ij}(\{S_k\}_t)\}\ )$ and $-E_{pos}(\{S_k\}_t)$ (=log likelihood + constant) are plotted against the fictitious time of the Markov chain. Temperature T is 1.0.

—– Fig.5 —–

The sample path $\{\{S_k\}_t\}$ corresponds to Fig.4. The fictitious time of the Markov chain t is increased from the top to the bottom of each figure. Each raw of the pattern corresponding to the state $\{S_k\}$ $(k = 1..30)$ at intervals of 100 MCS.

The behavior of the sample path shown in Fig.4 and Fig.5 suggests at least two metastable states with large weights exist, which are not in the same equivalent class under the cyclic change of the labels of the groups. One of them contains the global mode of the posterior. The relaxation of the Markov chain and the convergence of the average of statistics is severely affected by stochastic jumping between the metastable states.

In Fig.4, the weights of two metastable states are not so different each other. If the weight of one metastable state is much smaller than the other state but not negligible, the occasional appearance of the states in the metastable state with smaller weight is observed. In such cases, the distribution of statistics is highly non-Gaussian even when the average is taken over the sample path much longer than the relaxation time. In our experiments, the distribution of the average of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{p_{ij}(\{S_k\})\}\ )$ over such a large Monte Carlo steps as 10000 MCS has heavy tail for some artificial data.

4.4 VALIDATION OF ESTIMATES

As we mention in Chapter 2, the marginal distributions calculated with Metropolis-type algorithm contain valuable information. One of the interesting applications of this calculation with our model is measurement of goodness of the most probable classification. In our Bayesian framework, the goodness can be measured by the distribution of a suitably chosen distance between a sample from the posterior and the mode of the posterior distribution.

Examples of the calculation are shown in Fig.6a-Fig.6d. Again we choose the Kullback-Leibler divergence

$$J(\ \{p_{ij}(\{S_k\}^{mode})\}, \{p_{ij}(\{S_k\}^{sample})\}\ ) = \sum_{i \neq j} p_{ij}(\{S_k\}^{mode}) \log \frac{p_{ij}(\{S_k\}^{mode})}{p_{ij}(\{S_k\}^{sample})} \tag{15}$$

$$p_{ij}(\{S_k\}) = p^{S_i S_j} \tag{16}$$

as a measure of the distance between the mode $\{S_k\}^{mode}$ of the posterior and a sample $\{S_k\}^{sample}$ from the posterior.

— Fig.6a, Fig.6b, Fig.6c, Fig.6d —

Histograms of the divergence $J(\ \{p_{ij}(\{S_k\}^{mode})\}, \{p_{ij}(\{S_k\}^{sample})\}\ )$, where $\{S_k\}^{mode}$ is the mode of the posterior and $\{S_k\}^{sample}$ is a randomly chosen sample from the posterior. Fig.6a and Fig.6b correspond to data with the average number r of matches 0.6, and Fig.6c and Fig.6d correspond to data with $r = 0.4$. Each histogram consists of 1000 samples. They are sampled at intervals of 100 MCS after the first 30000 MCS are discarded.

In Fig.6a the histogram is well concentrated at the origin, while it has a long tail and/or multiple peaks in Fig.6b-Fig.6d. This means that $\{p_{ij}\}$ defined by the most probable classification is more reliable in Fig.6a than in Fig.6b, Fig.6c or Fig.6d.

The uncertainty of the hyper parameters and the assumed model structure is not considered in this section. The effects of the uncertainty of the hyper parameters will be studied in Section 4.7.

4.5 OPTIMAL ESTIMATES

Another interesting aspects of Metropolis-type Carlo algorithm is that the optimal Bayesian estimator can be calculated under any loss function (Marroquin et al.(1987), Ogata(1990), Iba(1989)).

In this section we consider the estimation of the probability p_{ij} with which the player i wins the player j. When a classification of players specified with $\{S_k\}$ is given, the probability p_{ij} is determined through $p_{ij}(\{S_k\}) = p^{S_i S_j}$. Thus a natural candidate of the estimator of p_{ij} is the one based on the most probable classification, i.e. the estimator $p_{ij}(\{S_k\}^{mode})$. When

the posterior distribution has multiple modes, we define a corresponding estimator by choosing one of the modes.

However the estimator based on the mode of the posterior is not always the optimal estimator of p_{ij}. For example, when we assume the loss function

$$J(\ \{p_{ij}(\{S_k\}^{true})\}, \{p_{ij}\}\ ) = \sum_{i \neq j} p_{ij}(\{S_k\}^{true}) \log \frac{p_{ij}(\{S_k\}^{true})}{p_{ij}} \tag{17}$$

the optimal estimator $\{p_{ij}^{opt}\}$ of p_{ij} is the one which minimizes the posterior mean of the loss function,

$$\langle\ J(\ \{p_{ij}(\{S_k\})\}, \{p_{ij}^{opt}\}\ )\ \rangle_{pos} \tag{18}$$

By using $p_{ji}^{opt} = 1 - p_{ij}^{opt}$, it is easy to show

$$\frac{\partial}{\partial p_{ij}^{opt}} \langle\ J(\ \{p_{ij}(\{S_k\})\}, \{p_{ij}^{opt}\}\ )\ \rangle_{pos} = 0 \ \Leftrightarrow\ p_{ij}^{opt} = \langle\ p_{ij}(\{S_k\})\ \rangle_{pos} \tag{19}$$

This means that under the loss function(17) the posterior mean $\langle p_{ij}(\{S_k\})\rangle_{pos}$ is better estimator of p_{ij} than the estimator $p_{ij}(\{S_k\}^{mode})$ based on the posterior mode. The posterior mean $\langle p_{ij}(\{S_k\})\rangle_{pos}$ can be calculated with Metropolis-type algorithm.

To show the merit of the optimal estimator and the utility of Metropolis-type algorithm for computing it, we calculate the estimate $\langle p_{ij}(\{S_k\})\rangle_T$ using the Metropolis algorithm at various temperature T. Hereafter we refer it as $\langle p_{ij}(\{S_k\})\rangle_T^{metro}$. The results of the experiments are shown in Fig.7a ($r = 0.4$) and Fig.7b ($r = 0.6$), where the sample-average of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_T^{metro}\}\ )$ over 50 artificial data is plotted against T. In these figures, the minimum of J is found near the temperature $T = 1.0$, which supports our prediction that the posterior mean is better estimator than the one based on the posterior mode under the loss function J. Note that $\langle p_{ij}(\{S_k\})\rangle_{pos} = \langle p_{ij}(\{S_k\})\rangle_{T=1}$ by definition and $\langle p_{ij}(\{S_k\}^{mode})\rangle = \langle p_{ij}(\{S_k\})\rangle_{T=0}$ if the effect of degeneracy of mode is negligible.

In these experiments we apply *simulated annealing procedure* (Kirkpatrick et al(1983), Geman,S and Geman,D(1984)). We start from higher temperature T_0 and gradually decrease T until we reach the desired temperature T^{goal}. Since trapping by the local maxima with small posterior probabilities can be avoided by the careful annealing procedure, we can expect the estimate $\langle p_{ij}(\{S_k\})\rangle_{T=0}^{metro}$ coincides with the estimate based on global mode. In order to check this, we try another method for the search of global mode; the hill climbing algorithm is repeated from many different initial states and

the local maximum which gives the largest posterior probability is selected as the global mode. As shown in Fig.7a and Fig.7b, this algorithm gives nearly the same sample-averages of J. For some artificial data, however, the simulated annealing algorithm fails to find the global maximum even with a large number of Monte Carlo steps. This suggests that in our model the repeated application of the hill climbing algorithm is more efficient than the simulated annealing algorithm for searching global mode.

At low temperature T, $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_T^{metro}\}\ )$ is not necessary a good approximation of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_T\}\ )$ even with a careful annealing procedure, because the simulated annealing procedure does not ensure the relaxation of the Markov chain at the desired temperature $T = T^{goal}$. In order to confirm the value of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_{T=1}^{metro}\}\ )$ calculated in our experiments is a reasonable approximation of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_{pos}\}\ )$, we test the relaxation of Metropolis-type algorithm at temperature $T = 1.0$ for the data with the same conditions. The results of the test is reported in Section 4.6, which suggest the reasonable approximation of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_{T=1}\}\ )$ is obtained for most of these data.

— Fig.7a, Fig.7b —

—– Fig.8 —–

The sample-average of $J(\{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_T^{metro}\}\ )$ over 50 artificial data is plotted against temperature T. The average numbers r of matches are 0.4 in Fig.7a and $r = 0.6$ in Fig.7b. The parameters of the algorithm $ANSTEP$, $AVSTEP$ and T_0 are defined in Fig.8, where the annealing and averaging schedules are shown.

At temperature $T \neq 0$:
For $r = 0.4$, the symbol `x` corresponds to the experiment with $AVSTEP$=3000 MCS, $ANSTEP$=2000 MCS and $T_0 = 2.0$;
the symbol `o` corresponds to the experiment with $AVSTEP$=18000 MCS, $ANSTEP$=20000 MCS and $T_0 = 2.0$.
For $r = 0.6$, the symbol `x` corresponds to the experiment with $AVSTEP$=3000 MCS, $ANSTEP$=2000 MCS and $T_0 = 2.0$;
the symbol `o` corresponds to the experiment with $AVSTEP$=18000 MCS, $ANSTEP$=20000 MCS and $T_0 = 8.0$.

At temperature $T = 0$:
The symbol o corresponds to the experiment with
$AVSTEP$=4000 MCS, $ANSTEP$=80000 MCS and $T_0 = 2.0$.
The results obtained by the repeated application of the hill climbing search are also shown with the symbols + , which correspond to a set of trials with 2000 repetition from different initial conditions. Another set of trials gives optimal states with different values of J for some data. The posterior probabilities of the optimal states are exactly the same even in these cases and the differences of J are due to non-uniqueness of the modes of the posterior.

For some of our data, the mode of the posterior distribution is not unique. In such cases, the averages of $p_{ij}(\{S_k\})$ over *the all of the modes of the posterior distribution* is another candidate of 'the estimator at $T = 0$'. Although it gives better estimates than the one based on one of the modes, the average of the loss function J corresponding to them is still considerably larger than the average corresponding to the posterior means.

4.6 TEST OF RELAXATION

When we apply Metropolis-type algorithm, we should carefully check the relaxation of the Markov chain. For example, the method for validating estimates presented in Section 4.4 gives incorrect results when the relaxation of the Markov chain is not achieved in the observational time.

In this section we give an example of the test of the relaxation, in which the averages over sample paths with different initial conditions are compared. We are interested in the statistics which are invariant under the cyclic transformations of the labels of the groups, hence, the relaxation of the Markov chain in the state space of the equivalent classes of this transformation.

Fig.9 and Fig.10 show the results of the test at temperature $T = 1.0$, where $\langle p_{ij}(\{S_k\})\rangle_T = \langle p_{ij}(\{S_k\})\rangle_{pos}$. For each of 10 artificial data, 10 sample paths are generated with different initial conditions and different random numbers. The average $\langle p_{ij}(\{S_k\})\rangle_{T=1}^{metro}$ of $p_{ij}(\{S_k\})$ on these sample paths are calculated and the statistics $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_{T=1}^{metro}\}\ )$ with different number of averaging steps are recorded. Samples from the first 2000 MCS of each path are discarded and not used for the average. The simulated annealing procedure is not used because it effectively reduces the variety of the initial conditions, which is crucial to our test.

—– Fig.9 —–

—– Fig.10 —–

The results of the test of the relaxation at temperature $T = 1.0$. The statistics $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_{T=1}^{metro}\}\ )$ are plotted with various sample paths. The panels correspond to the averages of $p_{ij}(\{S_k\})$ over 2000,4000,8000,16000 and 30000 MCS respectably. The symbols (o) in each column of the panels correspond to the results for the same artificial data with different initial conditions. The result corresponds to the average over a very long run(150000 MCS after 50000 MCS discarded) is also shown with a short bar(-) for each of artificial data in each panel. The average numbers r of matches are 0.4 in Fig.9 and 0.6 in Fig.10 respectably.

Although the convergence of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_{T=1}^{metro}\}\ )$ is considerably slow for a few data with $r = 0.6$, the results suggest that the sufficient relaxation of the Markov chain is achieved in the experiments. Reasonable approximations of $J(\ \{p_{ij}(\{S_k\}^{true})\}, \{\langle p_{ij}(\{S_k\})\rangle_{T=1}\}\ )$ are obtained in moderate computational time for most of these data.

The results of the same experiment for the same data at temperature $T = 0.2$ are shown in Fig.11 and Fig.12. The premature convergence of the averages is evident.

—– Fig.11 —–

—– Fig.12 —–

The results of the test of the relaxation at temperature $T = 0.2$. The average numbers r of matches are 0.4 in Fig.11 and 0.6 in Fig.12 respectably.

Although this test cannot detect metastable states which can be approached only from the small fractions of initial states, it is much better than the tests based on a single sample path where we can easily miss the contribution of the metastable states with large weights.

4.7 ESTIMATION OF HYPER PARAMETERS

In the experiments in previous sections, it is assumed that we know *a priori* the true value of the 'hyper' parameters $\{p^{AB}\}\,(A,B=1..G)$, which determine the dominance property among the groups of players. In this section we give preliminary results on the estimation of these hyper parameters.

Here we will work with the subspace $p^{12}=p^{23}=p^{31}\equiv p>0$ and try to estimate p. If we assume $p^{12}=p^{23}=p^{31}\equiv p$, the mode of posterior distribution $P_{pos}(\{S_i\})$ is not sensitive to the choice of p. The marginal distributions are, however, more sensitive to the value of p.

An approach for estimating the value of the hyper parameter p is the maximization of 'type II likelihood' $l(p)$ (Good(1965), Dempster et al.(1977), Akaike(1980)):

$$l(p)=\log\sum_{config.}P_{pos}(\{S_i\}) \tag{20}$$

The summation $\sum_{config.}$ means sum over all possible $\{S_i\}$'s. This method corresponds to 'the maximum likelihood estimation' in finite mixture model(see, for example, Everitt and Hand(1981)). Another interesting example of the inference with the likelihood marginalized over discrete parameters is given by Thorne et al.(1991) in their study of the DNA alignment problem.

The most general technique for calculating the type II likelihood with Metropolis-type algorithm is through the calculation of the differential coefficients of type II likelihood. This approach is introduced by Hinton and Sejnowski(1986) in their study of a neural network model (the 'Boltzmann machine') and applied to image processing problems by Geman and McClure(1987) and Ogata(1990,1991). In Ogata's work the maximum value of type II likelihood is calculated from the values of its differential coefficient with the trapezoidal formula. He used it for a model selection criterion as was proposed by Akaike(1980).

In our model, the differential coefficient of type II likelihood with respect to p is

$$\frac{\partial}{\partial p}l(p)=\frac{\langle n_++n_-\rangle_{pos}}{p(1-p)}\left(\frac{\langle n_+\rangle_{pos}}{\langle n_++n_-\rangle_{pos}}-p\right) \tag{21}$$

, where n_+,n_-,χ_+,χ_- is defined as follows:

$$\chi_+(1,2)=\chi_+(2,3)=\chi_+(3,1)=1 \tag{22}$$

$$\chi_+(i,j)=0\quad\text{for other}\,i\,,j \tag{23}$$

$$\chi_-(2,1)=\chi_-(3,2)=\chi_-(1,3)=1 \tag{24}$$

$$\chi_-(i,j)=0\quad\text{for other}\,i\,,j \tag{25}$$

$$n_+ = \sum_{ij} n_{ij}\chi_+(i,j) \tag{26}$$

$$n_- = \sum_{ij} n_{ij}\chi_-(i,j) \tag{27}$$

Then the value of type II likelihood as a function of p is expressed as

$$l(p) = l(p=0.5) + \int_{0.5}^{p} dp \frac{\langle n_+ + n_- \rangle_{pos}}{p(1-p)} \left(\frac{\langle n_+ \rangle_{pos}}{\langle n_+ + n_- \rangle_{pos}} - p \right) \tag{28}$$

With this expression, the calculation of the type II likelihood is reduced to the calculation of the posterior means $\langle n_+ \rangle_{pos}$ and $\langle n_- \rangle_{pos}$, which can be directly calculated with Metropolis-type algorithm.

Here we will give the results of the experiments for the sets of data with $p_{true} \equiv {p_1}^{true} = {p_2}^{true} = {p_3}^{true} = 0.8, 0.5$ ($r = 0.4$) and with $p_{true} = 0.8, 0.7, 0.5$ ($r = 0.6$). Each set of data consists of 15 artificial data. We note that the case with $p_{true} = 0.5$ correspond to a 'null hypothesis' , in which no group structure can be identified. In these experiments, the differential coefficient $\frac{\partial}{\partial p} l(p)$ is calculated for each of 18 different values of p ($p = 0.525, 0.550, .., 0.950$) by using the Metropolis algorithm and the value of $l(p) - l(0.5)$ are estimated with the trapezoidal formula at each p. The value of p which gives the largest value of $l(p)$ is selected as the estimate p_{est} of p ($p_{est} = 0.525, 0.550, .. ,0.950$ or 0.5). The average is taken over 10000 $\sim$ 20000 MCS at $T = 1.0$ after 2000 MCS of annealing from the starting temperature T_0=3.0 with the first 5000 $\sim$ 10000 MCS discarded. The occasional appearance of metastable states may cause considerable error in the averages. The long relaxation time of the Markov chain is also observed for the large value of p. This shows a limitation on the reliability of our results. The results are summerized in Fig.13a-Fig.13c, where p_{est} and $l(p_{est}) - l(0.5)$ is plotted for each of the data.

— Fig.13a, Fig.13b, Fig.13c —

The estimation of p through the maximum type II likelihood method. The x-axis and y-axis in each figure correspond to p_{est} and $l(p_{est}) - l(0.5)$ respectively.

In Fig.13a the results for the data with $p_{true} = 0.8$(symbol `o`) are compared to the results with $p_{true} = 0.5$ (symbol `x`) which correspond to the results with a 'null hypothesis'. The average number r of matches is 0.4 in Fig.13a. In Fig.13b and Fig.13c similar

comparisons are made for the data with $r = 0.6$; in Fig.13b the results with $p_{true} = 0.8$ (symbol `o`) and $p_{true} = 0.5$ (symbol `x`) are compared; in Fig.13c the results with $p_{true} = 0.7$ (symbol `o`) and $p_{true} = 0.5$ (symbol `x`) are compared.

Each set of data consists of 15 artificial data. Note that some of the symbols are overlapped with one another.

The results in Fig.13a-Fig.13c show that the variance of the estimator p_{est} is considerably large when $p_{true} = 0.5$. Besides, the estimates of p with smaller values of $l(p_{est}) - l(0.5)$ are easily affected by the numerical error of the Metropolis algorithm.

One method to remove the instability of the estimator is to introduce a significance test or model selection procedure, i.e. we set the value of estimator of p to 0.5 when $l(p_{est}) - l(0.5) < \epsilon$. Here ϵ is a constant which control the 'significance level'. To make proper choice of ϵ, the detailed knowledge about the distribution of the statistics $l(p_{est}) - l(0.5)$ is required. If we set $\epsilon = 1.0$ as a conventional choice, the results shown in the Table 1 are obtained.

—– Table 1 —–

$r = 0.4$

p_{true}	0.8	0.5
p (estimated)		
0.500	8	14
0.700		
0.725		
0.750		
0.775		1
0.800	1	
0.825	6	
0.850< p		

$r = 0.6$

p_{true}	0.8	0.7	0.5
p (estimated)			
0.500		11	14
0.700			
0.725	1		
0.750	2	2	1
0.775	2	2	
0.800	6		
0.825	4		
0.850< p			

The estimation of p through the maximum type II likelihood method.

Each number in the tables is the frequency that the corresponding value of p is selected as the estimate. The significance test or model selection procedure described in the text is used with the threshold $\epsilon = 1.0$.

As shown in Table 1, the hyper parameter p can be estimated from the data when $p_{true} = 0.8$ and $r = 0.6$, while the estimation is almost fails when $p_{true} = 0.7$ and $r = 0.6$. The situation is critical when $p_{true} = 0.8$ and $r = 0.4$, where $p = 0.5$ is selected as the estimate for about half of the data.

The threshold ϵ which corresponds to '5 % level' in traditional statistical significance test may be considerably larger than 1.0. More simulation is, however, required to determine it.

Another method to resolve the difficulty arising from the instability of the estimator is to consider the marginal distribution of p. In this approach we assume a prior distribution $\tilde{\pi}(p)$ of p and define the simultaneous posterior distribution $\tilde{P}_{pos}(\{S_i\}, p)$ of $\{S_i\}$ and p. Thus the marginal distribution $\tilde{P}(p)$ of p is easily calculated from $l(p)$ by using

$$\tilde{P}(p) = \frac{\exp(l(p))\tilde{\pi}(p)}{\int_{0.5}^{1} dp \exp(l(p))\tilde{\pi}(p)} \tag{29}$$

Examples of the marginal distribution $\tilde{P}(p)$ calculated with this method are shown in Fig.14a and Fig.14b. The uniform prior of p is assumed in these examples.

— Fig.14a, Fig.14b —

Marginal distributions $\tilde{P}(p)$ of p.

The results in Fig.14a and Fig.14b correspond to the 'significant' and 'not significant' cases respectably (with $\epsilon = 1.0$). The data with $r = 0.4$ and $p_{true} = 0.8$ are used for the both results. The uniform prior of p is assumed.

So far we calculate the type II likelihood of the model with the Metropolis algorithm and successfully used it for estimating the hyper parameter p. There are, however, other approaches for estimating the hyper parameter p. We will touch on some of the possible approaches in the rest of this section.

We note that in our model the simultaneous posterior distribution $\tilde{P}_{pos}(\{S_i\}, p)$ of $\{S_i\}$ and p can be directly sampled by a Metropolis-type algorithm. A similar method is used in an image processing problem by Besag et al.(1991). For our model this approach is especially attractive because the average of $\tilde{P}_{pos}(\{S_i\}, p)$ over the parameter p can be analytically carried out for fixed $\{S_i\}$ by using the expression

$$\int_0^1 p^{n_+}(1-p)^{n_-} dp = \frac{1}{n_+ + n_- + 1} \frac{1}{\left(\begin{array}{c} n_+ + n_- \\ n_+ \end{array} \right)} . \tag{30}$$

when the uniform prior(in general any conjugate prior) of p is assumed. Thus we can construct Metropolis-type algorithm which generate samples $\{S_i\}$ from the distribution $\int_0^1 dp\tilde{P}_{pos}(\{S_i\}, p)$. Using these samples, we can calculate required marginal distributions of $\tilde{P}_{pos}(\{S_i\}, p)$. The possible drawback of this approach is that the relaxation of the Metropolis-type algorithm might be slow in some situations.

One might consider another method for estimating the hype-parameter p, the maximization of the simultaneous posterior distribution $\tilde{P}_{pos}(\{S_i\}, p)$ with the parameters $\{S_i\}$ and p. This method is different from other methods in this section *in its principle* because the marginalization with discrete parameters $\{S_i\}$ is not considered. The method also gives good estimates of p when $p_{true} = 0.8$ and $r = 0.6$. It gives, however, over-estimates of p when $p_{true} = 0.5$. Although this difficulty can be circumvented with a suitable testing or model selection procedure based on the ratio of the posterior probability, the threshold corresponds to ϵ should be much larger than 1.0. This difference between the maximum simultaneous posterior method and the maximum type II likelihood method is mostly due to our assumption of uniform prior for $\{S_i\}$.

ACKNOWLEDGMENTS

The author would like to thank to Prof.K.Tanabe for many useful suggestions and discussions. The author also acknowledge Prof.T.Yanagimoto, Prof.Y.Ogata, Mr.T.Okamoto , Dr.M.Kawato, Dr.T.Inui, Dr.J.Suzuki and Prof.S.Amari for their continuous encouragements and kind advices.

REFERENCE

Akaike,H.(1980)
Likelihood and Bayes procedure
Bayesian statistics
Eds. Bernardo,J.M, DeGroot,M.H., Lindley,D.V., and Smith,A.F.M.
University press Valencia.

Amit,D.J., Gutfreund,H. and Somplinsky,H. (1985)
Spin-glass models of neural networks
Physical Review A 32 1007-1018.

Besag,J., York,J. and Mollié,A (1991)
Bayesian image restoration, with two applications in spatial statistics
Annals of Institute of Statistical Mathematics 43 1-59 (with discussions).

Binder,K. (1986)
Introduction: Theory and technical aspects of Monte Carlo simulations
Monte Carlo Methods in statistical physics (2nd ed.)
Topics in current physics vol.7
Ed. Binder,K.
Springer-Verlag Berlin.

Binder,K. and Stauffer,D. (1984)
A simple introduction to Monte Carlo simulation
and some specialized topics
Topics in current physics vol.36
Ed. Binder,K.
Springer-Verlag Berlin.

Dempster,A.P., Laird,N.M., Rubin,D.B. (1977)
Maximum likelihood from incomplete data via the EM algorithm
The Journal of the Royal Statistical Society B 39 1-38 (with discussions).

Everitt,B.S. and Hand,D.J.(1981)
Finite mixture distributions
Chapman and Hall London New York.

Frank,O. and Strauss,D.(1986)
Markov Graphs
Journal of the American Statistical Association 85 832-842.

Fushimi,M. and Tezuka,S. (1981)
Generation of the pseudo-random numbers which distribute uniformly in the high dimensional space.
Applied Statistics 10 151-163 (in Japanese).

Gelfand,A.E., Hills,S.E., Racine-poon,A. and Smith,A.F.M.(1990)
Illustration of Bayesian inference
in normal data models using Gibbs sampling
Journal of the American Statistical Association 85 972-985.

Geman,S. and Geman,D.(1984)
Stochastic relaxation, Gibbs distributions,
and the Bayesian restoration of images
IEEE Transactions on Pattern Analysis and Machine Intelligence 6 721-741.

Geman,D. and Geman,S.(1986)
Bayesian image analysis
NATO ASI series F20
Disordered Systems and Biological Organization
Ed. Bienenstock,E et al.
Springer-Verlag.

Geman,S. and McClure,D.E. (1987)
Statistical methods for tomographic image reconstruction
Proc. of the 46th Session of the ISI, Bulletin of the ISI, Vol.52.

Gidas,B (1989)
A renormalization group approach to image processing problems
IEEE Transactions on Pattern Analysis and Machine Intelligence 11 164-180.

Good,I.J.(1965)
The estimation of probabilities
MIT press Cambridge .

Hinton,G.E. and Sejnowski,T.J.(1986)
Learning and relearning in Boltzmann machines
in Parallel Distributed Processing Vol.1
Eds. Rumelhart,E. and McClelland,J.L.
MIT press Cambridge .

Iba,Y.(1989)
Bayesian statistics and statistical mechanics
In cooperative dynamics in complex physical systems
Ed. Takayama,H Springer-Verlag Berlin .

Iba,Y.(1991)
Macroscopic parameter estimation from incomplete data with Metropolis-type Monte Carlo algorithm
Proceedings of the institute of statistical mathematics Vol.39 No.1 1-21
(in Japanese with English summary).

Iba,Y.(1991)
Metropolis-type Monte Carlo algorithm and
quasi Bayesian estimation procedure
Proceedings of the institute of statistical mathematics Vol.39 No.2 225-244
(in Japanese with English summary).

Kay.J.W. and Titterington D.M. (1986)
Image labelling and statistical analysis of incomplete data
Proc. 2nd. Int. Conf. Image processing and Applications
Conf. Publ. No.265 London Inst. Elec. Engrs. 44-48.

Kirkpatrick,S, Gellatt,C.D. and Vecchi,M.P.(1983)
Optimization by simulated annealing
Science 220 671-680

Marroquin,J., Mitter,S. and Poggio,T. (1987)
Probabilistic solution of ill-posed problems in computational vision
Journal of the American Statistical Association Vol.82 76-89.

Metropolis.N, Rosenbluth,A.W., Rosenbluth,M.N.,
Teller,A.H. and Teller,E (1953)
Equation of state calculation by fast computing machines
Journal of the chemical physics 21 1087-1092.

Mezard,M., Parisi,G and Virasoro,M.A.(1987)
Spin glass theory and beyond
World Scientific Singapore .

Ogata,Y.(1990)
A Monte Carlo method for objective Bayesian procedure
Annals of Institute of Statistical Mathematics 42 403-433.

Ogata,Y.(1991)
Goodness-of-fit of Bayesian models by the Monte Carlo simulation
In discussions of the paper by Besag,J., York,J. and Mollié,A
Annals of Institute of Statistical Mathematics 43 25-32.

Ogata,Y. and Tanemura,M.(1981)
Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure
Annals of Institute of Statistical Mathematics 33 B 315-338.

Ogata,Y. and Tanemura,M.(1984)
Likelihood analysis of spatial point patterns
The Journal of the Royal Statistical Society B, Vol.46 496-518.

Ohtsuki,H. and Kawato,M. (1991)
Training a hierarchical image model :
Estimation of MRF line-process energy provided with image intensity data without a high-level teacher
Proc. International Joint Conference on Neural Networks Seattle.

Penttinen,A. (1984)
Modelling interactions in spatial point patterns:
parameter estimation by the maximum likelihood method
Jyväskylä studies in computer science economics and statistics vol.7
University of Jyväskylä, Finland (Ph.D Thesis).

Qian,W. and Titterington,D.M. (1992)
Stochastic relaxations and EM algorithms for Markov random fields
Journal of statistical computation and simulation vol.40 55-69.

Rumelhart,D.E, Smolensky,P., McClelland,J.L. and Hinton,G.E.(1986)
Schemata and sequential thought processes in PDP models
in Parallel Distributed Processing Vol.2
Eds. Rumelhart,E. and McClelland,J.L.
MIT press Cambridge .

Smolensky,P(1986)
Information processing in dynamical systems;
foundations of harmony theory
in Parallel Distributed Processing Vol.1
Eds. Rumelhart,E. and McClelland,J.L.
MIT press Cambridge .

Thorne,J.L., Kishino,H. and Felsenstein,J. (1991)
An evolutionary model for maximum likelihood alignment of DNA sequences
Journal of molecular evolution 33 114-124.

van Hemmen,J.L. and Morgenstern,I. (eds.) (1987)
Heidelberg colloquium on glassy dynamics
Lecture note in physics 275, Springer-Verlag Berlin.