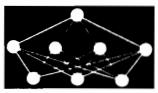

DATA AUGMENTATION ALGORITHM FOR GRAPHICAL MODELS WITH MISSING DATA

*Masahiro Kuroda**

Abstract: In this paper, we discuss an efficient Bayesian computational method when observed data are incomplete in discrete graphical models. The data augmentation (DA) algorithm of Tanner and Wong [8] is applied to finding the posterior distribution. Utilizing the idea of local computation, it is possible to improve the DA algorithm. We propose a local computation DA (LC-DA) algorithm and evaluate its computational efficiency.

Key words: *Data augmentation algorithm, local computation, graphical models, computational efficiency, missing data*

Received: August 1, 2004
Revised and accepted: June 1, 2005

1. Introduction

A graphical model is characterized by conditional independence relationships among variables of a statistical model. Graphical models are broadly used in various fields to describe complex statistical models and to specify the multivariate distributions, see Whittaker [9] and Edwards [4].

For a large graphical model, it is rare to obtain complete observed data. When observed data are incomplete, it is extremely difficult to obtain the exact posterior distribution for a graphical model and the calculation may take a long time when observed data are moderately large. To overcome this computational difficulty, various algorithms related to graph structures are proposed, see Cowell et al. [1]. In this paper, we apply the data augmentation (DA) algorithm of Tanner and Wong [8] to approximating the posterior distribution of a graphical model. Then, incorporating the idea of local computation into the DA algorithm, it is possible to reduce the computational effort. We propose a local computation DA (LC-DA) algorithm and evaluate its computational efficiency.

In Section 2, we show the exact Bayesian computation to find the posterior distribution for a discrete graphical model with missing data. In Section 3, instead

*M. Kuroda
Department of Socio-Information, Okayama University of Science, 1-1 Ridai-cho, Okayama 700-0005, Japan, E-mail: kuroda@soci.ous.ac.jp

of doing the infeasible exact computation, we give the DA algorithm to approximate the posterior distribution. In Section 4, we present the LC-DA algorithm. Section 5 discusses the efficiency of the LC-DA algorithm from the viewpoint of computational complexity.

2. Graphical model with missing data and exact Bayesian computation

Let V denote the set of vertices in a graph and $X_V = \{X_i \mid i \in V\}$ be the set of discrete random variables. Associated with each vertex $i \in V$, a random variable X_i takes values in a sample space Ω_i. For a subset $A \subseteq V$, we write X_A for $\{X_i \mid i \in A\}$ and $\Omega_A = \prod_{i \in A} \Omega_i$. Let the joint probability of X_V denote

$$p_V(x_V) = \Pr(X_V = x_V),$$

for every $x_V \in \Omega_V = \prod_{i \in V} \Omega_i$ and let $\theta_V = \{p_V(x_V) \mid x_V \in \Omega_V\}$. The marginal probability of X_A for $A \subset V$ can be written as

$$p_A(x_A) = \Pr(X_A = x_A) = \sum_{x_{V \setminus A}} p_V(x_V),$$

for every $x_A \in \Omega_A = \prod_{i \in A} \Omega_i$ and $\theta_A = \{p_A(x_A) \mid x_A \in \Omega_A\}$. The symbol "\" denotes the operator of a difference set. The conditional probability of X_A given $X_B = x_B$ is defined as

$$p_{A|B}(x_A|x_B) = \Pr(X_A = x_A \mid X_B = x_B) = p_{A \cup B}(x_{A \cup B})/p_B(x_B),$$

providing $A \cup B \subset V$ and $A \cap B = \emptyset$ where $\emptyset$ denotes the empty set, and also $\theta_{A|B} = \{p_{A|B}(x_A|x_B) \mid x_A \in \Omega_A, x_B \in \Omega_B\}$.

In this paper, we assume that the graph of X_V has the *global independence* that, for a triplet (A, B, C) of mutually disjoint subsets of V and $V = A \cup B \cup C$, each vertex of A is separated from each vertex of B given the subset C. Then, under the global independence structure, X_A is independent of X_B given X_C. Thus we have

$$p_V(x_V) = p_C(x_C) p_{A|C}(x_A|x_C) p_{B|C}(x_B|x_C),$$

so that $\{\theta_C, \theta_{A|C}, \theta_{B|C}\}$ are mutually independent. Suppose that observed data can be classified into three groups such that one is complete data and the others are incomplete data with X_B and X_A missing. The observed data patterns are indicated by $T = \{t_0, t_1, t_2\} = \{V, A \cup C, B \cup C\}$. In addition, we assume missingness at random in the sense of Rubin [7]. Each of observed data is denoted by $n^0 = \{n_{t_0}(x_{t_0}) \mid x_{t_0} \in \Omega_{t_0}\}$, $n^1 = \{n_{t_1}(x_{t_1}) \mid x_{t_1} \in \Omega_{t_1}\}$ and $n^2 = \{n_{t_2}(x_{t_2}) \mid x_{t_2} \in \Omega_{t_2}\}$. The sizes of the incomplete data n^1 and n^2 are considerably larger than the size of the complete data n^0. Assuming that observed data $n = (n^0, n^1, n^2)$ have a multinomial distribution with θ_V, the likelihood $L(n|\theta_V)$ is given by

$$\begin{aligned} L(n|\theta_V) &= f(n^0|\theta_{t_0}) f(n^1|\theta_{t_1}) f(n^2|\theta_{t_2}) \\ &\propto \prod_{0 \le i \le 2} \left\{ \prod_{x_{t_i} \in \Omega_{t_i}} p_{t_i}(x_{t_i})^{n_{t_i}(x_{t_i})} \right\}. \end{aligned} \tag{1}$$

For the multinomial model, we assume that the prior distribution of θ_V is a Dirichlet distribution which has the density function

$$\pi(\theta_V|\alpha_V) \propto \prod_{x_V \in \Omega_V} p_V(x_V)^{\alpha_V(x_V)-1}, \tag{2}$$

where $\alpha_V = \{\alpha_V(x_V) \mid x_V \in \Omega_V\}$ is a hyper-parameter. Then, according to the mutually independence relationships among $\{\theta_C, \theta_{A|C}, \theta_{B|C}\}$, it is possible to factorize $\pi(\theta_V|\alpha_V)$ into

$$\pi(\theta_V|\alpha_V) = \pi(\theta_C|\alpha_C)\pi(\theta_{A|C}|\alpha_{AC})\pi(\theta_{B|C}|\alpha_{BC}), \tag{3}$$

where $\alpha_C = \{\alpha_C(x_C) \mid x_C \in \Omega_C\}$, $\alpha_{AC} = \{\alpha_{AC}(x_{A\cup C}) \mid x_{A\cup C} \in \Omega_{A\cup C}\}$ and $\alpha_{BC} = \{\alpha_{BC}(x_{B\cup C}) \mid x_{B\cup C} \in \Omega_{B\cup C}\}$. The Dirichlet prior distribution (3) describes conditional independence of prior distributions and is called hyper Dirichlet prior distribution by Dawid and Lauritzen [3].

From the equations (1) and (2), we can obtain the mixture posterior distribution with the density

$$\begin{aligned} &\pi(\theta_V \mid n) \\ &\quad \propto \; L(n|\theta_V) \times \pi(\theta_V \mid \alpha_V) \\ &\quad \propto \sum_{\Omega(n^1)} \binom{n_{t_1}(x_{t_1})}{\{\tilde{n}_{t_1}(x_V)\}} \sum_{\Omega(n^2)} \binom{n_{t_2}(x_{t_2})}{\{\tilde{n}_{t_2}(x_V)\}} \prod_{x_V \in \Omega_V} p_V(x_V)^{\tilde{\alpha}_V(x_V)-1}, \end{aligned} \tag{4}$$

where, for $i = 1, 2$,

$$\sum_{\Omega(n^i)} \binom{n_{t_i}(x_{t_i})}{\{\tilde{n}_{t_i}(x_V)\}} = \prod_{x_{t_i} \in \Omega_{t_i}} \sum_{\Omega(n_{t_i}(x_{t_i}))} \binom{n_{t_i}(x_{t_i})}{\{\tilde{n}_{t_i}(x_V)\}}$$

and $\sum_{\Omega(n_{t_i}(x_{t_i}))}$ denotes the sum over all possible $\tilde{n}_{t_i}(x_V)$ for all $x_V \in \Omega_V$ under the conditions $\tilde{n}_{t_i}(x_V) \geq 0$ and $\sum_{x_{V\setminus t_i}} \tilde{n}_{t_i}(x_V) = n_{t_i}(x_{t_i})$, and

$$\tilde{\alpha}_V(x_V) = \alpha_V(x_V) + n_{t_0}(x_V) + \tilde{n}_{t_1}(x_V) + \tilde{n}_{t_2}(x_V).$$

Because of combinational explosion, the posterior density (4) has a very complicated function. Therefore, it is extremely difficult to calculate exactly $\pi(\theta_V|n)$ and these computation may take a long time when the observed data are moderately large.

Instead of performing the infeasible Bayesian computation, we use the DA algorithm which imputes incomplete data and finds $\pi(\theta_V|n)$ using the Monte Carlo method.

3. DA algorithm to graphical models

The DA algorithm is a type of Markov chain Monte Carlo. In the case that the incomplete-data posterior density is complicated as the posterior distribution (4)

and the complete-data posterior is relative easy to handle and draw from, the DA algorithm is very suitable. Each iteration of the DA algorithm consists of an *Imputation*-step and a *Posterior*-step.

For this case, the exact posterior distribution $\pi(\theta_C|n)$ can be obtained without any iterations, since the complete marginal data for X_C are calculated from n. Then the DA algorithm for the graphical model is given by the following iterative scheme:

Initialization: Set an initial distribution $\pi^{(0)}(\theta_V|n) = \pi(\theta_V|\alpha_V)$.

Imputation-step: Repeat the following steps for $l = 1, \ldots, L$ to obtain the augmented data $\tilde{n} = (n^0, \tilde{n}^1, \tilde{n}^2)$, where

$$\begin{aligned}\tilde{n}^1 &= \{\tilde{n}_{t_1}(x_V) \mid x_V \in \Omega_V, \sum_{x_B \in \Omega_B} \tilde{n}_{t_1}(x_V) = n_{t_1}(x_{t_1}), \tilde{n}_{t_1}(x_V) \geq 0\},\\ \tilde{n}^2 &= \{\tilde{n}_{t_2}(x_V) \mid x_V \in \Omega_V, \sum_{x_A \in \Omega_A} \tilde{n}_{t_2}(x_V) = n_{t_2}(x_{t_2}), \tilde{n}_{t_2}(x_V) \geq 0\}.\end{aligned}$$

1. Generate θ_V^* from the current approximation $\pi^{(t-1)}(\theta_V|n)$.
2. Generate $\tilde{n}^{1(l)}$ and $\tilde{n}^{2(l)}$ from the predictive multinomial distributions $f(\tilde{n}^1|\theta^*_{B|t_1}, n^1)$ and $f(\tilde{n}^2|\theta^*_{A|t_2}, n^2)$, where

$$\begin{aligned}\theta^*_{B|t_1} &= \{p^*_{B|t_1}(x_B|x_{t_1}) \mid x_B \in \Omega_B, x_{t_1} \in \Omega_{t_1}\},\\ \theta^*_{A|t_2} &= \{p^*_{A|t_2}(x_A|x_{t_2}) \mid x_A \in \Omega_A, x_{t_2} \in \Omega_{t_2}\}.\end{aligned}$$

Posterior-step: Update $\pi^{(t)}(\theta_V|n)$ given $\{\tilde{n}^{(l)} \mid 1 \leq l \leq L\}$ using the Monte Carlo method:

$$\pi^{(t)}(\theta_V|n) = \frac{1}{L}\sum_{l=1}^{L} \pi(\theta_C|n)\pi(\theta_{A|C}|\tilde{n}^{(l)})\pi(\theta_{B|C}|\tilde{n}^{(l)}).$$

Then

$$\begin{aligned}\pi(\theta_{A|C}|\tilde{n}^{(l)}) &\propto \prod_{x_C \in \Omega_C}\prod_{x_A \in \Omega_A} p_{A|C}(x_A|x_C)^{\tilde{\alpha}^{(l)}_{AC}(x_{A\cup C})-1},\\ \pi(\theta_{B|C}|\tilde{n}^{(l)}) &\propto \prod_{x_C \in \Omega_C}\prod_{x_B \in \Omega_B} p_{B|C}(x_B|x_C)^{\tilde{\alpha}^{(l)}_{BC}(x_{B\cup C})-1},\end{aligned}$$

where

$$\begin{aligned}\tilde{\alpha}^{(l)}_{AC}(x_{A\cup C}) &= \alpha_{AC}(x_{A\cup C}) + n_{t_0}(x_{A\cup C}) + n_{t_1}(x_{t_1}) + \sum_{x_B \in \Omega_B} \tilde{n}^{(l)}_{t_2}(x_V),\\ \tilde{\alpha}^{(l)}_{BC}(x_{B\cup C}) &= \alpha_{BC}(x_{B\cup C}) + n_{t_0}(x_{B\cup C}) + n_{t_2}(x_{t_2}) + \sum_{x_A \in \Omega_A} \tilde{n}^{(l)}_{t_1}(x_V).\end{aligned}$$

Until the approximations $\pi^{(t)}(\theta_{A|C}|n)$ and $\pi^{(t)}(\theta_{B|C}|n)$ converge to stationary distributions, the *Imputation*- and *Posterior*-steps are alternated repeatedly. Achieving convergence of the DA algorithm, the true posterior distribution $\pi(\theta_V|n)$ can be found.

In the practical implementation of the DA algorithm, the selection of the number of imputation (L) is perhaps more crucial. When the proportion of missing data in incomplete data is high and the size of the incomplete data is large, L must be considerably large. However, it is difficult to determine L on theoretical bases. In order to assess the convergence, diagnostic techniques are applied to output from the DA iteration. Cowles and Carlin [2] provide the comparative reviews of many convergence diagnostic techniques.

4. Application of LC-DA algorithm

In this section, we present the LC-DA algorithm. The important property of the LC-DA algorithm is that the DA algorithm is applied to each of factorized posterior distributions according to a graph structure and each posterior distribution is computed independently. Then it is possible to reduce the computational efforts from the viewpoint of computational complexity.

We denote the marginal data for $X_{A\cup C}$ and $X_{B\cup C}$ as $n_{AC} = (n^0_{AC}, n^1, n^2_C)$ and $n_{BC} = (n^0_{BC}, n^1_C, n^2)$, where

$$\begin{aligned} n^0_{AC} &= \{n_{t_0}(x_{A\cup C}) \mid x_{A\cup C} \in \Omega_{A\cup C}\},\ n^0_{BC} = \{n_{t_0}(x_{B\cup C}) \mid x_{B\cup C} \in \Omega_{B\cup C}\}, \\ n^1_C &= \{n_{t_1}(x_C) \mid x_C \in \Omega_C\},\ \ n^2_C = \{n_{t_2}(x_C) \mid x_C \in \Omega_C\}. \end{aligned}$$

With local computation to find $\pi(\theta_V|n)$, we can obtain the following theorem.

Theorem 4.1 *Suppose that C separates A and B in V. If $C \subseteq t$ for all $t \in T$, then the calculation of the posterior distributions of $\theta_{A|C}$ and $\theta_{B|C}$ can be done independently.*

Theorem 4.1 guarantees that the DA algorithm can execute separately to obtain the posterior distributions of $\theta_{A|C}$ and $\theta_{B|C}$ given n_{AC} and n_{BC}. The condition of $C \subseteq t$ for all $t \in T$ is called "lossless decomposition" by Geng and Li [5]. Then the LC-DA algorithm realizes the computation according to the following iterative scheme:

The DA iteration of $\pi(\theta_{A|C}|n_{AC})$

Initialization: Set an initial distribution $\pi^{(0)}(\theta_{A|C}|n_{AC}) = \pi(\theta_{A|C}|\alpha_{AC})$.

Imputation-step: Repeat the following steps for $l = 1, \ldots, L$ to impute $\tilde{n}^2_C$, where

$$\begin{aligned} \tilde{n}^2_C &= \{\tilde{n}_{t_2}(x_{A\cup C}) \mid x_{A\cup C} \in \Omega_{A\cup C}, \sum_{x_A} \tilde{n}_{t_2}(x_{A\cup C}) = n_{t_2}(x_C), \\ &\qquad \tilde{n}_{t_2}(x_{A\cup C}) \geq 0\}. \end{aligned}$$

1. Generate $\theta^*_{A|C}$ from the current approximation $\pi^{(t-1)}(\theta_{A|C}|n_{AC})$.
2. Generate the imputed data $\tilde{n}^{2(l)}_C$ from the predictive multinomial distribution $f(\tilde{n}^2_C|\theta^*_{A|C}, n^2_C)$.

Posterior-step: Update $\pi^{(t-1)}(\theta_{A|C}|n_{AC})$ by the Monte Carlo method:

$$\pi^{(t)}(\theta_{A|C}|n_{AC}) = \frac{1}{L}\sum_{l=1}^{L}\pi(\theta_{A|C} \mid \tilde{\alpha}_{AC}^{(l)}),$$

where

$$\tilde{\alpha}_{AC}^{(l)}(x_{A\cup C}) = \alpha_{AC}(x_{A\cup C}) + n_{t_0}(x_{A\cup C}) + n_{t_1}(x_{A\cup C}) + \tilde{n}_{t_2}^{(l)}(x_{A\cup C}).$$

The DA iteration of $\pi(\theta_{B|C}|n_{BC})$

The DA algorithm to obtain $\pi(\theta_{B|C}|n_{BC})$ are similar to the DA iteration of $\pi(\theta_{A|C}|n_{AC})$: The *Imputation*-step generates $\{\tilde{n}_C^{1(l)} \mid 1 \leq l \leq L\}$, where $\tilde{n}_C^1 = \{\tilde{n}_{t_1}(x_{B\cup C}) \mid x_{B\cup C} \in \Omega_{B\cup C}, \sum_{x_B}\tilde{n}_{t_1}(x_{B\cup C}) = n_{t_1}(x_C), \tilde{n}_{t_1}(x_{B\cup C}) \geq 0\}$. The *Posterior*-step finds $\pi^{(t)}(\theta_{B|C}|n_{BC})$ using n_{BC}^0, n^2 and $\{\tilde{n}_C^{1(l)} \mid 1 \leq l \leq L\}$.

When each of the approximations $\pi^{(t)}(\theta_{A|C}|n_{AC})$ and $\pi^{(t)}(\theta_{B|C}|n_{BC})$ converges to a stationary distribution, the true posterior distribution $\pi(\theta_V|n)$ can be calculated.

5. Computational efficiency of LC-DA algorithm

We now evaluate the computational efficiency of the LC-DA algorithm from the viewpoint of computational complexity. Here we introduce two quantities:

- $||\Omega_V||$ = the number of all possible values in Ω_V
- $||\Omega_A||$ = the number of all possible values in Ω_A where $A \subset V$

As for the space complexity, the amount of the storage space required by the DA algorithm is $||\Omega_V||$. Alternatively, in the LC-DA algorithm, it can not exceed $\max(||\Omega_{A\cup C}||, ||\Omega_{B\cup C}||)$. Next consider the time complexity under the worst-case assumption. The time complexity of the DA algorithm can be expressed by $O(||\Omega_V||)$. The implementation of the LC-DA algorithm can be done in $O(\max(||\Omega_{A\cup C}||, ||\Omega_{B\cup C}||))$.

The LC-DA algorithm is more efficient than the DA algorithm from both aspects of the space and time complexities and then can reduce the computational efforts.

Finally, we briefly describe the convergence speed of the LC-DA algorithm. The LC-DA algorithm is regarded as the collapsed Gibbs sampler of Liu [6]. Then, according to Liu's [6] result, the convergence speed of the LC-DA algorithm is faster than the speed of the DA algorithm. We shall investigate its convergence speed in detail.

References

[1] Cowell R. G., Dawid A. P., Lauritzen S. L., Spiegelhalter D. J.: Probabilistic networks and expert systems. Springer-Verlag, New York, 1999.

[2] Cowles M. K., Carlin D. B.: Markov chain Monte Carlo convergence diagnostics: A comparative review. Journal of the American Statistical Association, **91**, 1996, pp. 883-904.

[3] Dawid A. P., Lauritzen S. L.: Hyper Dirichlet laws in the statistical analysis of decomposable graphical models. The Annals of Statistics, **21**, 1993, pp. 1272-1317.

[4] Edwards D.: Introduction to graphical modeling. Second edition. Springer-Verlag, New York, 2000.

[5] Geng Z., Li K.: Factorization of posteriors and partial imputation algorithm for graphical models with missing data. Statistics & Probability Letters, **64**, 2003, pp. 369-379.

[6] Liu J.: The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. Journal of the American Statistical Association, **89**, 1994, pp. 958-966.

[7] Rubin D. B.: Inference and missing data. Biometrika, **63**, 1976, pp. 581-592.

[8] Tanner M. A., Wong W. H.: The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association, **82**, 1987, pp. 528-540.

[9] Whittaker J.: Graphical models in applied multivariate statistics. John Wiley & Sons, Chichester, 1990.

Appendix

Proof of Theorem 1

Since $C \subseteq t$ for all $t \in T$, we have $t \cap C = C$ and

$$
\begin{aligned}
&\prod_{x_t \in \Omega_t} p_t(x_t)^{n_t(x_t)} \\
&= \left\{ \prod_{x_C \in \Omega_C} p_C(x_C)^{n_t(x_t)} \right\} \\
&\quad \times \left\{ \prod_{x_{t\cap(A\cup C)} \in \Omega_{t\cap(A\cup C)}} \sum_{\Omega(n_t(x_{t\cap(A\cup C)}))} \binom{n_t(x_{t\cap(A\cup C)})}{\{\tilde{n}_t(x_{A\cup C})\}} \right. \\
&\quad \left. \prod_{x_C \in \Omega_C} \prod_{x_A \in \Omega_A} p_{A|C}(x_A|x_C)^{\tilde{n}_t(x_{A\cup C})} \right\} \\
&\quad \times \left\{ \prod_{x_{t\cap(B\cup C)} \in \Omega_{t\cap(B\cup C)}} \sum_{\Omega(n_t(x_{t\cap(B\cup C)}))} \binom{n_t(x_{t\cap(B\cup C)})}{\{\tilde{n}_t(x_{B\cup C})\}} \right. \\
&\quad \left. \prod_{x_C \in \Omega_C} \prod_{x_B \in \Omega_B} p_{B|C}(x_B|x_C)^{\tilde{n}_t(x_{B\cup C})} \right\} \\
&= \left\{ \prod_{x_C \in \Omega_C} p_C(x_C)^{n_t(x_t)} \right\} \\
&\quad \times \left\{ \sum_{\Omega_{t\cap(A\cup C)}(n_t)} \binom{n_t(x_{t\cap(A\cup C)})}{\{\tilde{n}_t(x_{A\cup C})\}} \prod_{x_C \in \Omega_C} \prod_{x_A \in \Omega_A} p_{A|C}(x_A|x_C)^{\tilde{n}_t(x_{A\cup C})} \right\} \\
&\quad \times \left\{ \sum_{\Omega_{t\cap(B\cup C)}(n_t)} \binom{n_t(x_{t\cap(B\cup C)})}{\{\tilde{n}_t(x_{B\cup C})\}} \prod_{x_C \in \Omega_C} \prod_{x_B \in \Omega_B} p_{B|C}(x_B|x_C)^{\tilde{n}_t(x_{B\cup C})} \right\}.
\end{aligned}
$$

Then for any $s \subset V$, $t \cap s = s$, it holds $n_t(x_s) = \tilde{n}_t(x_s)$ and

$$\sum_{\Omega_{t\cap s}(n_t)} \binom{n_t(x_{t\cap s})}{\{\tilde{n}_t(x_s)\}} = 1.$$

Therefore it is possible to factorize the likelihood (1) as follows:

$$L(\theta_V|n) = L(\theta_C|n_C)L(\theta_{A|C}|n_{AC})L(\theta_{B|C}|n_{BC}). \tag{5}$$

From the prior distribution (3) and the likelihood (5), we can obtain

$$\begin{aligned} \pi(\theta_V|n) &\propto \{L(\theta_C|n_C)\pi(\theta_C|\alpha_C)\} \times \{L(\theta_{A|C}|n_{AC})\pi(\theta_{A|C}|\alpha_{AC})\} \\ &\quad \times \{L(\theta_{B|C}|n_{BC})\pi(\theta_{B|C}|\alpha_{BC})\} \\ &= \pi(\theta_C|n_C)\pi(\theta_{A|C}|n_{AC})\pi(\theta_{B|C}|n_{BC}). \end{aligned}$$

Since it also holds the mutual independence among the posterior distributions, we can compute each posterior distribution independently. ∎