The 24th Symposium on Information Theory
and Its Applications (SITA2001)
Kobe, Hyogo, Japan, Dec. 4 - 7, 2001

不完全データを含む分割表におけるベイズ予測
Bayesian Prediction Algorithm in Contingency Tables with Incomplete Data

本田真理* 後藤正幸† 平澤茂一*
Mari HONDA Masayuki GOTO Shigeichi HIRASAWA

Abstract— The EM algorithm is popularly known as a method for maximum likelihood estimation in incomplete data settings. In this paper, we propose two algorithms for prediction in contingency table with incomplete data and compare the precision of proposed methods with those of conventional methods. From the simulation experiments, we show effectiveness of the proposed algorithms.

Keywords— contingency table, incomplete-data, prediction, EM algorithm, Bayesian statistics

1 はじめに

アンケート調査など，結果がカテゴリカルデータとして得られる場合，分割表の形で表すことがある．データに欠測がない完全データの場合は各セルの出現確率の最尤推定およびベイズ最適な推定やデータの予測を行うのは容易であるが，データの一部分が欠測している不完全データの場合は，問題は難しくなってくる．

不完全データを含む統計モデルから最尤推定値を求めるには，EM アルゴリズムを適用できることが知られている [1],[2]．不完全データを含む分割表の場合，特にデータ数が少ないときには不完全データを EM アルゴリズムの E-step を用いて各セルに割り振るということは逆に推定・予測精度を悪くすることが実験的に確かめられる．これは必ずしも最尤推定量を用いた予測が精度の面で優れている訳ではないことによる．不完全データによる予測精度の劣化を防ぐためには，最尤を追い求めるのを諦め，完全データとして得られたデータのみを用いる方法が最も簡単である．

一方，得られた分割表の真のモデルを 1 つに限定するのではなく，分割表を構成する要因ごとにセルをまとめることによって得られるパラメータ数の少ない，より単純なモデルも考え，モデル選択やベイズ最適な混合モデルを用いることにより推定・予測精度を向上させる方法が提案されている [3]．モデル選択による方法は，データに含まれる情報に適切なパラメータ数のモデルを選ぶことにより推定・予測精度を向上させる方法であり，一方，混合モデルによる方法はベイズ最適な予測はすべてのモデルの混合によって与えられるというベイズ最適性に基づく方法である．

本研究では，分割表を構成する要因ごとにセルをまとめることによって得られるモデル族において，複雑なモデルでは不完全データとみなされるようなデータも，単純なモデルでは情報の損失がない完全データとみなして推定を行うことが可能となる，という分割表特有の性質に着目する．本稿では推定・予測精度の向上を目的とし，不完全データを含む分割表における各セルの出現確率を推定する方法として，この性質をうまく利用した 2 つの方法を提案する．具体的には，各モデルごとに完全データとみなせるすべてのデータのみを用いることにより，①事後確率を計算してモデルの混合をとるベイズ統計的な方法 (提案法 1) と，②最大対数尤度を計算し AIC を用いてモデル選択を行う方法 (提案法 2) を提案し，不完全データをいっさい用いないで推定する方法 (従来法) よりも，真のデータ出現確率に対する推定精度という点で優れていることを示す．この結果は，モデルクラスによってはモデル族の性質を利用したヒューリスティックな方法で推定・予測精度を向上させることが可能であることを示している．

2 問題設定

2.1 完全データのみから構成される分割表

要因 A と要因 B で構成される分割表において，要因 A の水準を $\{a_1, a_2, \cdots, a_\eta\}$，要因 B の水準を $\{b_1, b_2, \cdots, b_\xi\}$ とする．分割表のセルは二つの水準の組み合わせによって表され，セル $a_i b_j$ に入るデータ数を x_{ij}，データの出現確率を $\theta_{ij}(\sum_{i=1}^{\eta}\sum_{j=1}^{\xi}\theta_{ij}=1)$ とする．以下で，要因 A，B の水準がそれぞれ 2 つのとき $(\eta=\xi=2)$ に，欠測データを含まない完全データ X としてデータが得られた場合の例を示す．

表 1:(a, b) の二元分割表 (完全データ)

	b_1	b_2
a_1	x_{11}	x_{12}
a_2	x_{21}	x_{22}

2.2 不完全データを含む分割表

どちらか片方の要因についてはどの水準に属するのか分かっているが，もう一方は分からない 1 次欠測データおよび，いずれの要因も欠測している 2 次欠測データを含む不完全データ Y について考える．ここで，セル $a_i b_j$ に入るデータの個数を y_{ij} とする．

以下で，要因 A，B の水準がそれぞれ 2 つの場合 $(\eta=\xi=2)$ の例を示す．このとき，各要因において水準 3 を水準 1，2 のどちらであるかが不明であるデータが入る水準とする．例えば要因 A については水準 1 であることがわかっているが，要因 B についてはわからないデータはセル $a_1 b_3$ に入り，どちらの要因についても不明な場合はセル $a_3 b_3$ に入る．したがって

欠測のないデータ：$y_{11}, y_{12}, y_{21}, y_{22}$

1 次欠測データ：$y_{13}, y_{23}, y_{31}, y_{32}$

* 早稲田大学 理工学部 経営システム工学科，〒169-8555 東京都新宿区大久保 3-4-1. School of Science and Engineering, Waseda University, 3-4-1 Ohkubo, Shinjuku-ku, Tokyo, 169-8555, Japan.
† 東京大学大学院 工学系研究科 環境海洋工学専攻，〒113-8656 東京都文京区本郷 7-3-1, Graduate School of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.

2 次欠測データ：y_{33}
である.

表 2:(a,b) の二元分割表 (不完全データ)

	b_1	b_2	b_3
a_1	y_{11}	y_{12}	y_{13}
a_2	y_{21}	y_{22}	y_{23}
a_3	y_{31}	y_{32}	y_{33}

例えば，y_{13} はセル a_1b_1 かセル a_1b_2 から生起していることを示す数である. よって，y_{13} のうち d_{11} がセル a_1b_1 から，d_{12} がセル a_1b_2 から生起しているものとすると，$y_{13}=d_{11}+d_{12}$ である. 同様にして 1 次欠測および 2 次欠測データの構造は以下のようになっている.

$$y_{13}=d_{11}+d_{12} \quad , \quad y_{23}=d_{21}+d_{22} \quad ,$$
$$y_{31}=e_{11}+e_{21} \quad , \quad y_{32}=e_{12}+e_{22} \quad ,$$
$$y_{33}=f_{11}+f_{12}+f_{21}+f_{22}$$

ただし，d_{ij} は y_{i3} のうちセル a_ib_j に入る個数，e_{ij} は y_{3j} のうちセル a_ib_j に入る個数，f_{ij} は y_{33} のうちセル a_ib_j に入る個数を表す.

以上を表 3 にまとめる.

表 3:欠測データの各セルへの割り振り

	a_1b_1	a_1b_2	a_2b_1	a_2b_2
y_{13}	d_{11}	d_{12}		
y_{23}			d_{21}	d_{22}
y_{31}	e_{11}		e_{21}	
y_{32}		e_{12}		e_{22}
y_{33}	f_{11}	f_{12}	f_{21}	f_{22}

このとき，観測できない完全データ X との関係は,

$$x_{ij}=y_{ij}+d_{ij}+e_{ij}+f_{ij},\ i,j\in\{1,2\}$$

となり，表 2 の不完全データの問題は，上式の y_{ij} が観測でき，d_{ij},e_{ij},f_{ij} は観測できない変数と考えることと等価である.

3 従来研究

3.1 EM アルゴリズム [2]

Dempster らにより提案された EM アルゴリズム [2] は，E-step と M-step と呼ばれる 2 つの手続きの反復からなっている. まず，不完全データが得られたもとで，完全データの対数尤度の条件付き期待値を計算することにより，擬似的な完全データを生成する (E-step). この擬似完全データより最尤推定値を求める (M-step). さらに得られたパラメータ推定値から再び擬似完全データを生成し，それからパラメータの推定値を求め直すという手続きを繰り返し行う.

E-step: p 時点で得られているパラメータ $\theta^{(p)}$ を用いて，対数尤度の期待値 $Q(\theta|\theta^{(p)})$ を求める.

$$Q(\theta|\theta^{(p)})=E[\log f(x|\theta)|y,\theta^{(p)}] \qquad (1)$$

M-step: $Q(\theta|\theta^{(p)})$ を最大化する θ を $\theta^{(p+1)}$ とおく.

$$\theta^{(p+1)}=\arg\max_{\theta\in\Theta}Q(\theta|\theta^{(p)}) \qquad (2)$$

3.2 ベイズ決定理論に基づく推定 [3]

ここでは，分割表が完全データとして得られた場合のベイズ推定について述べる. 本研究では事前分布として自然共役事前分布であるディリクレ分布を仮定する.

$$f(\theta)=\frac{\Gamma(\alpha_{11}+\alpha_{12}+\cdots+\alpha_{\eta\xi})}{\Gamma(\alpha_{11})\Gamma(\alpha_{12})\cdots\Gamma(\alpha_{\eta\xi})}\theta_{11}^{\alpha_{11}-1}\theta_{12}^{\alpha_{12}-1}\cdots\theta_{\eta\xi}^{\alpha_{\eta\xi}-1}$$

ただし，α_{ij} は事前分布のパラメータである. このとき，事後分布は

$$f(\theta|x)=\frac{p(x|\theta)f(\theta)}{p(x)}$$
$$=\frac{\Gamma(x_{11}+x_{12}+\cdots+x_{\eta\xi}+\alpha_{11}+\alpha_{12}+\cdots+\alpha_{\eta\xi})}{\Gamma(x_{11}+\alpha_{11})\Gamma(x_{12}+\alpha_{12})\cdots\Gamma(x_{\eta\xi}+\alpha_{\eta\xi})}$$
$$\cdot\,\theta_{11}^{x_{11}+\alpha_{11}-1}\theta_{12}^{x_{12}+\alpha_{12}-1}\cdots\theta_{\eta\xi}^{x_{\eta\xi}+\alpha_{\eta\xi}-1} \qquad (3)$$

で表される.

3.3 E-step とベイズ推定を用いたアルゴリズム [4]

事前の知識を事前分布として仮定できれば，EM アルゴリズムにおける M-step(最尤推定) の代わりに，ベイズ推定を用いることができる. 以下で $\eta=\xi=2$ の二元分割表の場合のアルゴリズムを与える.

E-step：

1) 欠測データを次式を用いて，完全に得られているデータのセルに割り振る.

$$x_{ij}^{(p)}=y_{ij}+\sum_{d_{ij}=0}^{y_{i3}}z_{d_{ij}}^{(p)}\cdot d_{ij}+\sum_{e_{ij}=0}^{y_{3j}}z_{e_{ij}}^{(p)}\cdot e_{ij}$$
$$+\sum_{f_{ij}=0}^{y_{33}}z_{f_{ij}}^{(p)}\cdot f_{ij},\ i,j\in\{1,2\} \qquad (4)$$

ただし，$z^{(p)}$ を以下のように定める.

$$z_{d_{i1}}^{(p)}=\frac{\left(\frac{\theta_{i1}^{(p)}}{\theta_{i2}^{(p)}}\right)^{d_{i1}}}{\sum_{d_{i1}=0}^{y_{i3}}\left(\frac{\theta_{i1}^{(p)}}{\theta_{i2}^{(p)}}\right)^{d_{i1}}},\ z_{e_{1j}}^{(p)}=\frac{\left(\frac{\theta_{1j}^{(p)}}{\theta_{2j}^{(p)}}\right)^{e_{1j}}}{\sum_{e_{1j}=0}^{y_{3j}}\left(\frac{\theta_{1j}^{(p)}}{\theta_{2j}^{(p)}}\right)^{e_{1j}}},$$

$$z_{f_{ij}}^{(p)}=\frac{\left(\frac{\theta_{ij}^{(p)}}{\theta_{22}^{(p)}}\right)^{f_{ij}}}{\sum_{f_{ij}=0}^{y_{33}}\left(\frac{\theta_{ij}^{(p)}}{\theta_{22}^{(p)}}\right)^{f_{ij}}},\ i,j\in\{1,2\}$$

2) 得られた疑似完全データの対数尤度 $\log f(x^{(p)}|\theta)$ の，データ y とパラメータ $\theta^{(p)}$ に関する条件付き平均を求める.

$$Q(\theta|\theta^{(p)}) = x_{11}^{(p)}\log\theta_{11}+x_{12}^{(p)}\log\theta_{12}$$
$$+x_{21}^{(p)}\log\theta_{21}+x_{22}^{(p)}\log\theta_{22} \qquad (5)$$

ベイズ推定： 擬似完全データから事後分布の期待値を計算する.

$$E[\theta_{ij}|\boldsymbol{x}^{(p)}] = \int \theta_{ij} f(\theta|\boldsymbol{x}^{(p)}) d\theta$$
$$= \frac{x_{ij}^{(p)} + \alpha_{ij}}{\sum_{s=1}^{\eta}\sum_{t=1}^{\xi} x_{st}^{(p)} + \sum_{s=1}^{\eta}\sum_{t=1}^{\xi} \alpha_{st}} \quad (6)$$

4　提案法

データが不完全データとして得られた場合，モデルによっては不完全データも有効なデータとして用いることができる．以下では簡単のため各水準が 2 つの 2×2 分割表について考える．各水準間で出現確率が同じとみなせるセルをまとめて新たなセルを作るとすると，次のような 3 つのモデルが考えられる．

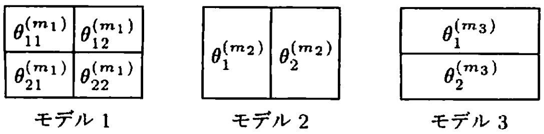

図 1: 3 つのモデル

要因 A について各要素間で差が見られないようなモデル 2 の場合は，モデル 1 におけるセル a_1b_1 とセル a_2b_1，セル a_1b_2 とセル a_2b_2 をまとめて新たな拡大セル B_1，B_2 とする．このとき，要因 A に関する 1 次欠測データ y_{31},y_{32} は損失のない完全データとみなしてよい．同様に，モデル 3 では要因 B について各要素間で差が見られないと考えることにより，モデル 1 におけるセル a_1b_1 とセル a_1b_2，セル a_2b_1 とセル a_2b_2 をまとめて新たな拡大セル A_1,A_2 とし，要因 B に関する 1 次欠測データ y_{13},y_{23} も完全データとみなす．

4.1　提案法 1 (混合モデルを用いた推定)

データに欠測がない完全データの場合には，次に出現するデータの予測は，1 つのモデルを選択するのではなく，考え得るすべてのモデルの混合をとって予測することがベイズ最適であることが知られている．提案法 1 では各モデル $m \in \{m_1, m_2, m_3\}$ ごとにセル a_ib_j の出現確率 $\theta_{ij}^{(m)}$ を求め，それらを各モデルの事後確率で重み付けし，出現確率の推定値を求める．

4.1.1　各セルの出現確率の推定

[モデル 1 の場合]

セルが 4 分割されたモデル 1 を考えた場合のデータ予測では，不完全データは用いずに完全に得られたデータのみ用いる．各セルのデータ出現確率の推定値 $\theta_{ij}^{(m_1)}$ は次式のようなラプラス型推定量によって与えられる．

$$\theta_{ij}^{(m_1)} = \frac{y_{ij} + \alpha_{ij}}{\sum_{s=1}^{2}\sum_{t=1}^{2} y_{st} + \sum_{s=1}^{2}\sum_{t=1}^{2} \alpha_{st}} \quad (7)$$

[モデル 2 の場合]

モデル 2 は二項分布に従うモデルであるので，事前分布はベータ分布 (パラメータ：α,β) を仮定し，拡大セル B_1,B_2 の出現確率 $\theta_1^{(m_2)}$,$\theta_2^{(m_2)}$ を推定する．

$$\theta_1^{(m_2)} = \frac{\sum_{s=1}^{3} y_{s1} + \alpha}{\sum_{s=1}^{3}\sum_{t=1}^{2} y_{st} + \alpha + \beta} \quad (8)$$

$$\theta_2^{(m_2)} = \frac{\sum_{s=1}^{3} y_{s2} + \beta}{\sum_{s=1}^{3}\sum_{t=1}^{2} y_{st} + \alpha + \beta} \quad (9)$$

さらに，求めた出現確率を拡大セルを構成するセルの個数 (ここでは 2) で割り，セル a_1b_1 からセル a_2b_2 までの出現確率 $\theta_{ij}^{(m_2)}$ を計算する．

$$\theta_{11}^{(m_2)} = \theta_{21}^{(m_2)} = \theta_1^{(m_2)}/2 \quad (10)$$
$$\theta_{12}^{(m_2)} = \theta_{22}^{(m_2)} = \theta_2^{(m_2)}/2 \quad (11)$$

[モデル 3 の場合]

モデル 2 と同様に，事前分布にベータ分布を仮定し，拡大セル A_1,A_2 の出現確率 $\theta_1^{(m_3)}$,$\theta_2^{(m_3)}$ を推定する．

$$\theta_1^{(m_3)} = \frac{\sum_{t=1}^{3} y_{1t} + \alpha}{\sum_{s=1}^{3}\sum_{t=1}^{2} y_{st} + \alpha + \beta} \quad (12)$$

$$\theta_2^{(m_3)} = \frac{\sum_{t=1}^{3} y_{2t} + \beta}{\sum_{s=1}^{3}\sum_{t=1}^{2} y_{st} + \alpha + \beta} \quad (13)$$

$$\theta_{11}^{(m_3)} = \theta_{12}^{(m_3)} = \theta_1^{(m_3)}/2 \quad (14)$$
$$\theta_{21}^{(m_3)} = \theta_{22}^{(m_3)} = \theta_2^{(m_3)}/2 \quad (15)$$

4.1.2　モデルの混合

まず各モデルの事後確率を計算する．

$$p(m|x) = \frac{p(x|m)p(m)}{p(x)} \quad (16)$$

ここでは例としてモデル 1 の事後確率を求める．モデル 1 におけるデータの尤度 $p(x|m)$ は次式で表される．

$$\begin{aligned} p(x|m) &= \int p(x|\theta, m) f(\theta|m) d\theta \\ &= \frac{\Gamma(\alpha_{11} + \cdots + \alpha_{22})}{\Gamma(\alpha_{11}) \cdots \Gamma(\alpha_{22})} \prod_{s=1}^{2}\prod_{t=1}^{2} \theta_{st}^{\alpha_{st}+y_{st}-1} \\ &= \frac{\prod_{s=1}^{2}\prod_{t=1}^{2}\prod_{u=0}^{y_{st}} (\alpha_{st} + u)}{(\prod_{v=0}^{n-1} \sum\sum \alpha_{st} + v)} \end{aligned} \quad (17)$$

ただし，

$$n = \sum_{s=1}^{2}\sum_{t=1}^{2} y_{st}$$

各モデルが選ばれる確率 $p(m)$ が等確率であると仮定すると，ここではモデルが 3 つ考えられるので

$$p(m_1) = p(m_2) = p(m_3) = 1/3 \quad (18)$$

となる．
以下同様にしてモデル 2，モデル 3 についても事後確率を計算し，それらを用いて各セルの出現確率の推定値を求める．

$$\hat{\theta}_{ij} = \sum_m \theta_{ij}^{(m)} \cdot p(m|x) \quad (19)$$

4.2 提案法 2 (AIC を用いたモデル選択)

各モデルごとに完全データとみなせるデータから対数尤度を計算し，AIC によって選択されたモデルの出現確率を推定値とする

$$MLL(m) = (\text{モデル } m \text{ の下での最大対数尤度})$$
$$k = (\text{モデル } m \text{ に含まれる未知パラメータの数})$$
$$AIC(m) = -2 \times MLL(m) + 2 \times k$$

［モデル 1 の場合］

$$AIC(m_1) = -2\left(\sum_{s=1}^{2}\sum_{t=1}^{2}\left(y_{st}\cdot\log\frac{y_{st}}{\sum_{s=1}^{2}\sum_{t=1}^{2}y_{st}}\right)\right)+2\cdot 3 \qquad (20)$$

［モデル 2 の場合］

$$AIC(m_2) = -2\left(\left(\sum_{s=1}^{3}y_{s1}\right)\log\frac{\sum_{s=1}^{3}y_{s1}}{\sum_{s=1}^{3}\sum_{t=1}^{2}y_{st}} + \left(\sum_{s=1}^{3}y_{s2}\right)\log\frac{\sum_{s=1}^{3}y_{s2}}{\sum_{s=1}^{3}\sum_{t=1}^{2}y_{st}}\right)+2\cdot 1 \qquad (21)$$

［モデル 3 の場合］

$$AIC(m_3) = -2\left(\left(\sum_{t=1}^{3}y_{1t}\right)\log\frac{\sum_{t=1}^{3}y_{1t}}{\sum_{s=1}^{2}\sum_{t=1}^{3}y_{st}} + \left(\sum_{t=1}^{3}y_{2t}\right)\log\frac{\sum_{t=1}^{3}y_{2t}}{\sum_{s=1}^{2}\sum_{t=1}^{3}y_{st}}\right)+2\cdot 1 \qquad (22)$$

5 シミュレーション

提案法の有効性を検証するために，従来法（完全データのみ用いる方法）と，提案法 1，2 により推定したデータ出現確率と真の出現確率との平均二乗誤差を求める．

5.1 シミュレーション条件

以下の条件により A，B の水準が共に 2 つの場合 ($\eta = \xi = 2$) のシミュレーションを行う．

(1) 事前分布：一様分布を仮定
(2) データの欠測率：要因 A，B 共に 0.2
(3) データ数：10,20,30,$\cdots$,170
(4) 同一データ数での試行繰り返し数：1000 回
(5) 真の出現確率：各モデルごとに 100 パターンずつ，計 300 パターンを乱数によってランダムに設定

5.2 シミュレーションの結果と考察

シミュレーション結果を図 1 に示す．

① 提案法による推定は，完全データのみ用いた場合に，モデルの混合をとる方法やモデル選択を行う場合と同様に，データ数の増加に伴って平均二乗誤差が減少する挙動が見られる．

② 平均二乗誤差は混合，モデル選択のいずれの方法についても，提案法の方が従来法よりも小さい値となっている．

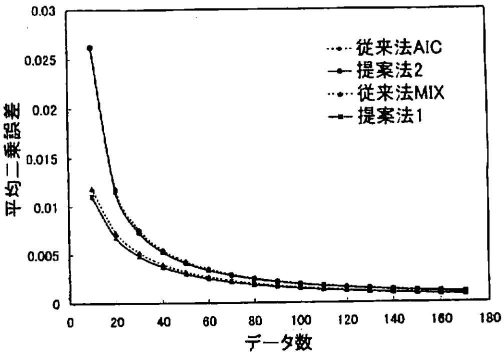

図 1：データ数に対する平均二乗誤差

③ 特にデータ数が少ないときには，混合をとる方法の方がモデル選択をする場合よりも平均二乗誤差が小さい値となっている．

①，③は完全データに対するベイズ解の挙動 [3] と類似しており，提案法 1 で事前分布を適切に利用できていることがわかる．

6 まとめと今後の課題

本研究では，不完全データを含む分割表において各セルの出現確率を推定する方法について議論し，不完全データを用いずに完全に得られたデータのみから推定を行う従来法に比べて推定精度という面で優れている 2 つの推定方法を提案した．

提案法では用いるデータ数が増えるため出現確率の推定精度が高くなる．事後確率については従来法と同様のものを用いているため，結果的に混合をとった推定値が従来法で求めた推定値よりも推定精度が高くなるということがシミュレーションによって示された．今回のシミュレーションでは各要因の欠測率を 0.2 に設定したが，欠測率を 0.5 に設定した場合に，従来法と提案法の差がより大きくなるという結果が得られた．よって，提案法はデータの欠測率が高いほど有効であると考えられる．

今後は分割表を構成する要因の数を増やし，提案法の有効性をさらに検討する必要がある．また，提案法の理論的解析も今後の課題である．

謝辞: 著者の一人本田は，本研究に対して多くのご討論とご助言を頂きました小林学氏，および石田崇氏に深く感謝致します．また，日頃よりお世話になっている平澤研究室各氏に心から感謝致します．なお，本研究の一部は 2001 年度早稲田大学特定課題研究助成費 (課題番号 2001A-566) の助成による．

参考文献

[1] 赤穂昭太郎，"EM アルゴリズムの幾何学，" 情報処理，Vol.37, No1, pp43-51, 1996.
[2] G.J.McLachlan, T.Krishnan, The EM Algorithm and Extensions, Wiley-Interscience, 1997.
[3] 菊池淑子，後藤正幸，俵信彦，" 順序カテゴリカルデータ解析における母数推定に関する研究，" 日本経営工学会論文誌，Vol.50，No.3, pp163-170, 1999.
[4] 本田真理，後藤正幸，平澤茂一，"不完全データからのベイズ推定に関する一考察，" 電子情報通信学会技術研究報告，IT2000-15, pp.7-12, 2000.