2004 年情報論的学習理論ワークショップ
2004 Workshop on Information-Based Induction Sciences (IBIS2004)
Tokyo, Japan, November 8 - 10, 2004.

隠れマルコフモデルの EM 法における学習誤差と汎化誤差のふるまい

Numerical behavior of training and generalization errors in HMM

星野 力* 渡辺 澄夫†
Tikara Hosino Sumio Watanabe

Abstract: Hidden Markov Models(HMMs) are used in many fields such as speech processing and bioinfomatics. However HMMs are non-identifiable statistical models and their mathematical property has not yet been clarifyed. Recently, in the Bayesian case, the asymptotic order of thier stochastic complexity is derived. In this paper we evaluate numerical behaviour of training and generalization errors optimized by EM and VB methods. The result shows that generalization error of VB method is much less than EM method and their values are approximately equal to the theoretical bayes upper bound.
キーワード：隠れマルコフモデル，特異モデル，EM 法， VB 法，汎化誤差

1 まえがき

隠れマルコフモデルは音声認識やバイオインフォマティクス等，非線型の時系列を表現するモデルとして広く使われている [1]， [2].

しかし，隠れマルコフモデルは，モデルが真の分布を含む場合，最適なパラメータが一点ではなく，解析的集合となる識別不能なモデルで，かつその解析的集合は特異点を持つ．そのため，従来の枠組みで汎化誤差の評価やモデル選択を行うことには問題がある [3].

近年，この問題に対して，特異な場合も含むモデルにおいて，ベイズ法を行った場合の，確率的複雑さと汎化誤差に関する漸近論が，代数解析的手法を用いて，非常に一般的な形で展開された [7]．また，その応用として，代数幾何の手法 (ブローアップ) により，隠れマルコフモデルでの自由エネルギーの上界が得られている [8]

隠れマルコフモデルのパラメータ推定については，EM 法を用いて尤度を最適化する方法が広くおこなわれており，さらに，近年そのベイズ法的な拡張として，変分自由エネルギーを最適化する変分ベイズ法 (以下 VB 法と略) が提案された [4].

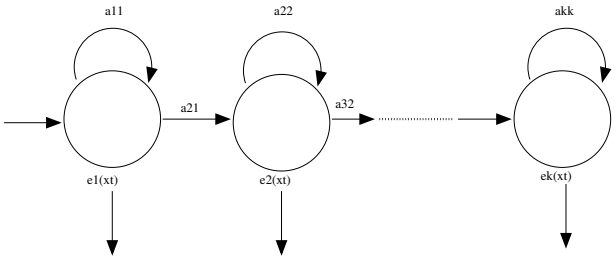

図 1: 隠れマルコフモデル (Left-to-Right)

それぞれの最適化法について，実験的な比較や精度評価はおこなわれているが [5]，[6]，モデルの基本的性質である学習誤差と汎化誤差の振舞いはあまり知られていない．本論では，特異モデルの視点から，隠れマルコフモデルが真の分布を含む場合の，汎化誤差と学習誤差を数値実験により評価し，理論値と比較することにより，隠れマルコフモデルにおける EM 法と VB 法の性質を明らかにする．

2 隠れマルコフモデルと EM 法

隠れマルコフモデルは，マルコフ過程をもつ離散状態と，状態ごとに定義される出力分布を持つモデルである．

同一の出力をする状態が複数あり，実際に観測される出力列からは，内部の状態列を特定することができない (隠れている) ため，”隠れ”マルコフモデルと呼ばれる．

内部状態の次元が K，観測される出力が C 次元で離

*東京工業大学大学院 総合理工学研究科，226-8503 横浜市緑区長津田 4259, tel. 045-924-5018, e-mail thosino@cs.pi.titech.ac.jp, Tokyo Inistitute of Technology, 4259 Nagatsuda, Midori-ku, Yokohama, 226-8503 Japan
日本ユニシス株式会社，135-8560 東京都江東区 1-1-1
Nihon Unisys, Ltd. 1-1-1 Toyosu, Koutou-ku, Tokyo 135-8560 Japan
†東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology

散 ($B = \{b_1, \ldots, b_C\}$)，系列長が T の場合を定式化する．

観測系列を $X = \{x_1, \ldots, x_T\} \in B^T$, 内部状態の系列を $S = \{s_1, \ldots, s_T\} \in \{0,1\}^{K \times T}$ として，

$$P(s_1) = \prod_{i=1}^{K} \pi_i^{s_{1,i}} \quad (1)$$

$$P(s_t|s_{t-1}) = \prod_{i=1}^{K}\prod_{j=1}^{K} A_{ij}^{s_{t,i}s_{t-1,j}} \quad (2)$$

$$P(x_t|s_t) = \prod_{i=1}^{K}\{\sum_{c=1}^{C} \delta(x_t - b_c)E_{ic}\}^{s_{t,i}} \quad (3)$$

を用いて結合確率密度関数は，

$$P(S,X) = P(s_1)\prod_{t=1}^{T} P(s_t|s_{t-1})\prod_{t=1}^{T} P(x_t|s_t) \quad (4)$$

である．パラメータの制約は，

$$\sum_{i=1}^{K}\pi_i = 1, \sum_{j=1}^{K} A_{ij} = 1, \sum_{c=1}^{C} E_{ic} = 1 \quad (5)$$

となる．

ただし，$s_{t,i} \in \{0,1\}$ は時刻 t での内部状態をあらわす指示変数であり，学習するパラメータは，初期状態のパラメータ $\pi \in R^K$，状態遷移確率 $A \in R^K \times R^K$，および出力確率 $E \in R^K \times R^C$ である．

また，遷移確率行列に制約を加えることにより，状態遷移の制約を表現することができる．例えば，$A_{i,j}(i \neq j$ もしくは $i+1 \neq j) = 0$ の制約は，状態遷移が，時間的に逆行することがなく，自分自身もしくは，次の状態への遷移へ制限されていることを表していて Left-to-Right モデルと呼ばれている (図 1)．

隠れマルコフモデルは，隠れ変数をもつ指数分布族に入るため，パラメータの尤度を最適化するのに EM 法を適用することが可能で，実際にも広く使われている．特に EM 法の E-step である，観測データが与えられた時の，隠れ状態の事後分布 $P(s_t|X)$ を計算するための効率的な方法が知られており，前向き後ろ向きアルゴリズムと呼ばれている．

前向き確率を時刻 t までに，データ $x_1, \ldots, x_t$ を観測し，時刻 t で，状態 k である確率とする．つまり，

$$f_k(t) = P(x_1, \ldots, x_t, s_{t,k}) \quad (6)$$

とすると，時刻 $t+1$ で状態 l である前向き確率は，漸化式

$$f_l(t+1) = e_l(x_{t+1})\sum_k f_k(t)A_{lk} \quad (7)$$

で書ける．ただし，状態 l で出力 x_t を観測する確率を，

$$e_k(x_t) \equiv \sum_{c=1}^{C}\delta(x_t - b_c)E_{kc} \quad (8)$$

とおいた．このとき，データに対するモデルの尤度は，前向き確率を用いて，

$$P(X) = \sum_S P(X,S) = \sum_{k=1}^{K} f_k(T) \quad (9)$$

と表わすことができる．

次に，後ろ向き確率を，時刻 t での状態が，k であり，時刻 $t+1$ から時刻 T までに $x_{t+1}, \ldots, x_T$ を出力する確率で定義する．

$$b_k(t) = P(x_{t+1}, \ldots, x_T|s_{t,k}) \quad (10)$$

とすると，時刻 $t-1$ で状態 l である後ろ向き確率は，漸化式

$$b_l(t-1) = \sum_{k=1}^{K} A_{kl}e_k(x_t)b_k(t) \quad (11)$$

で書ける．

前向き確率，後ろ向きの確率を上のように定義すると，時刻 t で状態 k であり，$x_1, \ldots, x_T$ を観測する確率が，

$$\begin{aligned} P(X, s_{t,k}) &= P(x_1, \ldots, x_t, s_{t,k})P(x_{t+1}, \ldots, x_T|s_{t,k}) \\ &= f_k(t)b_k(t) \end{aligned} \quad (12)$$

と書き下せる．

よって，E Step は，

$$P(s_{t,k}|X) = \frac{f_k(t)b_k(t)}{P(X)} \equiv \gamma_k(t) \quad (13)$$

で与えられる．

M step は $x_1, \ldots x_T$ が時刻 t で状態 k であり，時刻 $t+1$ で，状態 l である確率を，

$$\gamma_{kl}(t) \equiv P(s_{t,k}, s_{t+1,l}|X) = \frac{f_k(t)a_{lk}e_l(x_{t+1})b_l(t+1)}{P(X)} \quad (14)$$

で定義すると，パラメータの更新式は，各時系列に対するインデックスを X^j で表わすと，遷移確率については，

$$A_{kl}^{new} = \frac{\sum_j \sum_{t=1}^{T-1}\gamma_{lk}^j(t)}{\sum_{k=1}^{K}\sum_j \sum_{t=1}^{T-1}\gamma_{lk}^j(t)} \quad (15)$$

となり，出力確率については，

$$E_{kc}^{new} = \frac{\sum_j \sum_{\{t|x_k^j = b_c\}}\gamma_k^j(t)}{\sum_{c=1}^{C}\sum_j \sum_{\{t|x_k^j = b_c\}}\gamma_k^j(t)} \quad (16)$$

と書き表せる．

適当な初期値から，E Step と M Step を繰り返すことで，モデルの対数尤度が単調に増加し，局所最大値に収束することは示されているが，大域的な最適解である保証はなく，収束先はパラメータの初期値や，データセットに依存する．

3 VB 法

近年ベイズ法の近似アルゴリズムとして提案された，VB 法について簡単に述べる [4],[6]．隠れ変数を持つ指数分布族における VB 法は，パラメータ w の事前分布 $\phi(w)$ を導入し，データ X が与えられたもとで，事後分布を隠れ変数 S とパラメータ w の事後分布が独立なもの，つまり

$$q(S)q(w) \tag{17}$$

と書けるクラスの中でカルバック距離で測って最も真の事後分布 $P(S,w|X)$ に近いものを探索する方法である．この近似は，次式で定義される変分自由エネルギーの最大化によって実現されることが示されている．

$$F(q) \equiv < \log \frac{p(X,S|w)\phi(w)}{q(S,w)} >_q \tag{18}$$

(ただし，$<>_p$ は分布 p での平均をとることを意味する．)

VB 法は，EM 法の自然な拡張になっており，アルゴリズム的にも EM 法から，遷移確率のパラメータ更新式が，

$$\begin{aligned} A_{kl}^{\hat{new}} = \quad & \exp \quad \{\Psi(\sum_j \sum_{t=1}^{T-1} \gamma_{lk}^j(t) + \phi_0) \\ & - \quad \Psi(\sum_{k=1}^{K}(\sum_j \sum_{t=1}^{T-1} \gamma_{lk}^j(t) + \phi_0))\} \end{aligned} \tag{19}$$

に，出力確率のパラメータ更新式が，

$$\begin{aligned} E_{kc}^{\hat{new}} = \quad & \exp \quad \{\Psi(\sum_j \sum_{\{t|x_k^j=b_c\}} \gamma_k^j(t) + \xi_0) \\ & - \quad \Psi(\sum_{c=1}^{C}(\sum_j \sum_{\{t|x_k^j=b_c\}} \gamma_k^j(t) + \xi_0))\} \end{aligned} \tag{20}$$

へと変わるだけである．(ただし，ϕ_0, ξ_0 は，事前分布のパラメータで，Ψ は digamma 関数．) そのため，繰り返し毎の計算量も EM 法と同じになる．

4 ベイズ法における汎化誤差の理論解析

汎化誤差の理論は，最尤法ではまだ解決されていないが，ベイズ法を用いた場合の解析はおこなわれている．後に EM 法と VB 法における汎化誤差の振舞と比較するために，ベイズ法の汎化誤差について簡単に述べておく．

データ $X^n = \{X_1, \ldots, X_n\}$ が与えられた時，真の分布 $q(x)$ を，モデル $p(x|w)$ で学習する場合を考える．パラメータの事前分布を $\phi(w)$ とすると，パラメータの事後分布は，

$$p(w|X^n) = \frac{1}{Z_0(X^n)}\phi(w)\prod_{i=1}^{n} p(X_i|w) \tag{21}$$

で定義される．ただし，$Z_0(X^n)$ は規格化定数である．予測分布は，事後分布を用いて

$$p(x|X^n) = \int p(x|w)p(w|X^n)dw \tag{22}$$

で書き表せる．

そのとき，ベイズ法における汎化誤差，

$$G(n) \quad = \quad E_{\mathbf{x}^n}[\int q(x)\log\frac{q(x)}{p(x|X^n)}dx] \tag{23}$$

は，カルバック情報量，

$$H(w) = \int q(x)\log\frac{q(x)}{p(x|w)}dx \tag{24}$$

から誘導される，複素数 z の関数，

$$J(z) = \int H(w)^z\phi(w)dw \tag{25}$$

の最大極 $-\lambda$ を用いて，漸近的に

$$G(n) \quad = \quad \frac{\lambda}{n} + O(\frac{1}{n}) \tag{26}$$

で書けることが証明されている [7]．特に，識別可能な正則モデルにおいては，モデルのパラメータ数を d として，$\lambda = \frac{d}{2}$ となり，

$$G(n) = \frac{d}{2n} + O(\frac{1}{n}) \tag{27}$$

である．

隠れマルコフモデルにおいては，真の分布の状態数が H，学習モデルの状態数が K，出力分布の次元が C の場合，カルバック情報量の分割と，ブローアップによって，$J(z)$ の極の一つが求まり，汎化誤差 $G(n)$ が，

$$\frac{H(2K-H+C-1)}{2n} \tag{28}$$

で，上から押えることができることが証明されている [8]．

特に Left-to-Right モデルの場合は，状態遷移確率が疎なことを利用して，

$$\frac{HC+H+1}{2n} \tag{29}$$

と，学習モデルの状態数 K によらない形で上から押えることができる [8]．

5　数値実験

出力の分布は，$C = 2$ 値とした ($x_{ijk} = \{0, 1\}$)．真のモデル $q(x)$ は状態数 $H = 2$ の Left-to-Right モデルとし，時系列の長さを $T = 20$ で固定した ($X_{mi} = \{x_{mi1}, \ldots, x_{miT}\}$)．真の分布から $N = 100$ 個の系列をサンプリングし，それを 1 セットとする ($X_m = \{X_{m1}, \ldots, X_{mN}\}$)．さらに，データの出方に対するばらつきを表すために，上のデータを $M = 1000$ セット用意する ($X = \{X_1, \ldots, X_M\}$)．学習は，各データセットについて，真の分布を含む，状態数 $K = \{2, 4, 6, 8\}$ のモデル $p(x|w)$ を用いた．各モデルごとに，ランダムな初期値から始めて EM 法を適用し，収束は，各反復での対数尤度の変化が $1.0E - 6$ 以下で判定した．極端に悪い局所解を避けたい理由から，同じデータに対し異なる初期値で 10 回 EM 法を適用し，最も尤度の高いパラメータを推定結果とした．

VB 法では，事前分布のパラメータが，$\phi_0 = \{0.1, 1.0\}$ の場合について，尤度のかわりに変分自由エネルギーを用いて，同様の推定を行なった．

評価は，m セット目のデータで推定されたパラメータを $w_m^\star$ として，学習誤差

$$\frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N}\log\frac{q(X_{mn})}{p(X_{mn}|w_m^\star)} \tag{30}$$

および，汎化誤差

$$E_{X^n}[E_{w^\star}[\int q(x)\log\frac{q(x)}{p(x|w^\star)}dx]] \tag{31}$$

を大数の法則で近似した，

$$\frac{1}{M}\frac{1}{N}\sum_{m=1}^{M}\sum_{n=1}^{N'}\log\frac{q(\hat{X}_{mn})}{p(\hat{X}_{mn}|w_m^\star)} \tag{32}$$

を用いた．(ただし，$\hat{X}$ は学習に使用していないサンプルで，$N' = 10000$ 個の平均をとる．)

各状態数における，学習誤差，汎化誤差の結果を，図 3,4 に示す．

6　考察

6.1　誤差の振る舞い

数値実験によって得た汎化誤差と，

- 同じパラメータ数の識別可能なモデル $\frac{d}{2N}$
- 4 節で示したベイズ法における上界 $\frac{HC+H+1}{2N}$

を比較したものを図 2 に示す．

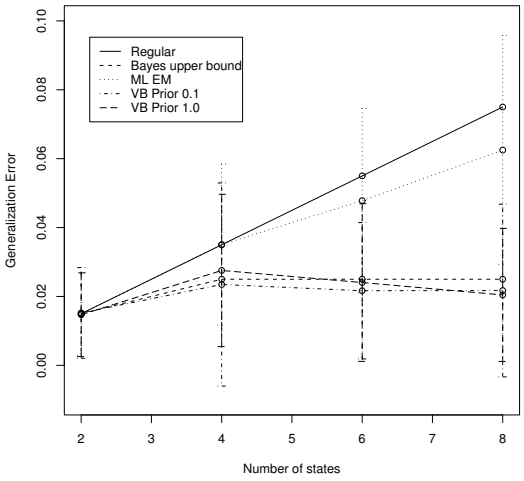

図 2: 汎化誤差の比較

図 2 から，汎化誤差については，EM 法ではベイズ法の上界よりは大きく，同じパラメータ数の正則モデルよりは小さいことがわかった．VB 法の汎化誤差は，EM 法よりずっと小さいベイズ法の理論的上界に近い値であり，学習モデルの大きさにも依存しないことが観察された．事前分布の影響は，今回選択した 2 値では大きな違いは見られなかったが，詳細は今後の課題である．

学習誤差に関しては，図 3,4 および表 2 をもとに，絶対値に対する比率で見ると，EM 法の場合には，汎化誤差とほぼ対称であり，VB 法では学習誤差が若干小さくなることが観察された．

6.2　繰り返しと汎化誤差

EM 法と，VB 法の目的関数である尤度および変分自由エネルギーと汎化誤差の関係を調べるため，異なる 200 組のサンプルセットについて EM，VB ステップの繰り返しごとの汎化誤差を評価した．

汎化誤差の繰り返し中での最小値と 1000 回繰り返し後の値を比較した図 5 より，EM 法，VB 法ともに過学習が観察されるが，VB 法の方が汎化誤差の絶対値，その差ともに小さいことがわかる．

また，図 6,7 に EM 法，VB 法の繰り返しごとの汎化誤差の典型的な振る舞いを示す. 繰り返しに対して，EM 法では尤度，VB 法では変分自由エネルギーが単調に動くことから，厳密な意味でそれらが汎化誤差の最小化にはなっていないことがわかる．グラフより，これら 2 つの振る舞いは顕著に異なることが示唆され，共に汎化誤差は一度下がり，EM 法ではその後ほぼ単調に増加すること，VB 法では上昇した後にまた下がることがしばしば観察された．

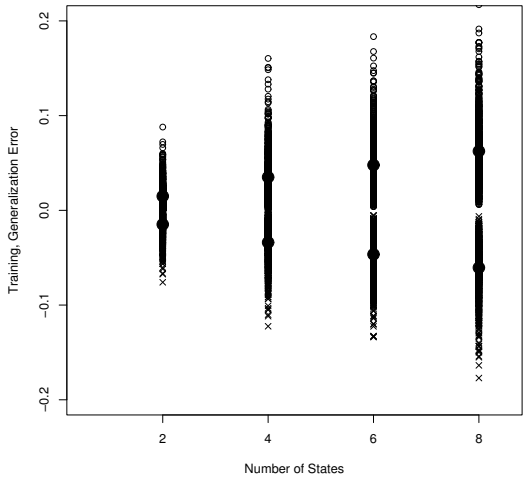

図 3: EM 法の学習誤差 (< 0) と汎化誤差 (> 0) (大円は平均)

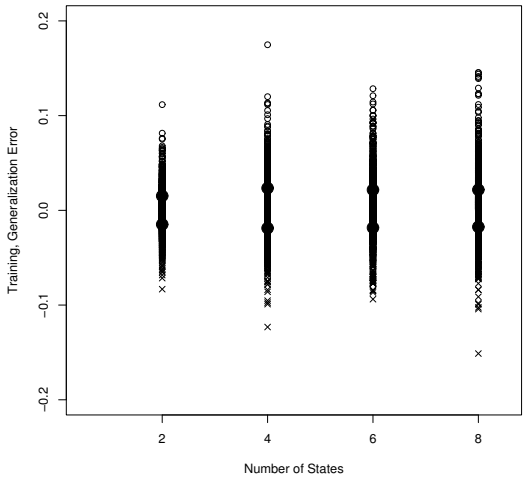

図 4: VB 法 ($\phi_0 = 0.1$) の学習誤差 (< 0) と汎化誤差 (> 0) (大円は平均)

繰返し	汎化誤差最小	1000 回繰り返し後
EM 学習誤差	-0.0192 ± 0.0289	-0.0390 ± 0.0338
VB0.1 学習誤差	-0.0159 ± 0.0151	-0.0192 ± 0.0153
EM 汎化誤差	0.04 ± 0.031	0.0577 ± 0.0326
VB0.1 汎化誤差	0.0190 ± 0.0175	0.0224 ± 0.0185

表 1: 最小な汎化誤差と 1000 回繰り返し後

7 おわりに

隠れマルコフモデルについて，モデルが真の分布を含み特異な場合に EM 法と VB 法でパラメータを推定した結果，汎化誤差は，VB 法のほうが EM 法より小さく，VB 法ではほぼベイズ法の理論値に近い値になることが分かった．また，EM 法，VB 法ともに過学習が見られるが，VB 法の方が過学習の程度が小さいこともわかった．

今後の課題としては，今回観察された振る舞いが Left-to-Right モデルだけでなく，一般的な構造をもつモデルにおいても観察されるか調べることがあげられる．

参考文献

[1] Lawrence Rabiner, Biing-Hwang Juang , “Fundamentals of Speech Recognition”, Pearson Education, 1993.

[2] Pierre Baldi, Soren Brunak, “Bioinformatics”, The MIT Press, 1998.

[3] 福水 健次, 栗木 哲, 竹内 啓, 赤平 昌文,“特異モデルの統計学”, 岩波書店, 2004

[4] H.Attias, “A Variational Bayesian Framework for Graphical Models”, NIPS12, MIT Press, 2000.

[5] Shinji Watanabe, Yasuhiro Minami, Atsushi Nakamura and Naonori Ueda, “Application of Variational Bayesian Approach to Speech Recognition”, NIPS15, MIT Press, 2003.

[6] Matthew J. Beal, “Variational Algorithms for Approximate Bayesian Inference”, PhD.Thesis, Gatsby Computaional Neuroscience Unit, University College London, 2003.

[7] Sumio Watanabe, “Algebraic analysis for nonidentifiable learning machines”, Neural Computation, 13(4), 899-933, 2001.

[8] Keisuke Yamazaki, Sumio Watanabe, “Stochastic Complexities of Hidden Markov Models”, IEEE International Workshop On Neural Networks For Signal Processing, 2003.

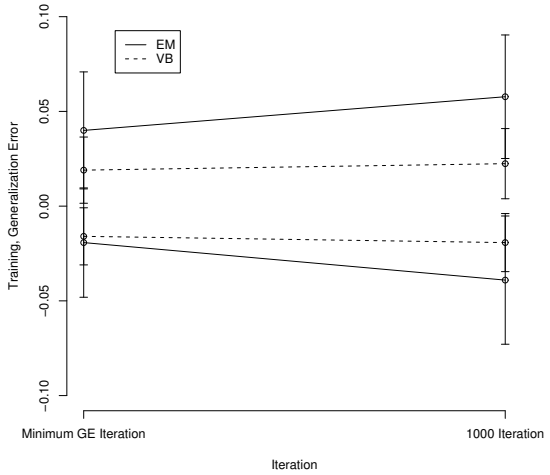

図 5: 最小の汎化誤差と 1000 回繰り返し後

状態数	2	4
EM 学習誤差	-0.0148 ± 0.0116	-0.038 ± 0.0196
VB0.1 学習誤差	-0.0148 ± 0.0121	-0.0188 ± 0.0162
VB1.0 学習誤差	-0.0144 ± 0.0117	-0.0272 ± 0.0195
EM 汎化誤差	0.0149 ± 0.012	0.035 ± 0.0234
VB0.1 汎化誤差	0.0152 ± 0.013	0.0234 ± 0.0294
VB1.0 汎化誤差	0.0147 ± 0.0122	0.0275 ± 0.0221
VB0.1 FE	-25.80 ± 1.22	-28.72 ± 4.54
VB1.0 FE	-8.27 ± 1.18	-15.77 ± 1.67
状態数	6	8
EM 学習誤差	-0.0464 ± 0.022	-0.0605 ± 0.0272
VB0.1 学習誤差	-0.0183 ± 0.0171	-0.0175 ± 0.0177
VB1.0 学習誤差	-0.0223 ± 0.02	-0.0178 ± 0.0173
EM 汎化誤差	0.0478 ± 0.0268	0.0625 ± 0.0333
VB0.1 汎化誤差	0.0217 ± 0.0198	0.0217 ± 0.0333
VB1.0 汎化誤差	0.0240 ± 0.0228	0.0204 ± 0.0193
VB0.1 FE	-28.79 ± 4.78	-28.59 ± 4.44
VB1.0 FE	-17.90 ± 3.52	-16.95 ± 3.28

表 2: 学習誤差と汎化誤差の平均と標準偏差 (FE:変分自由エネルギー)

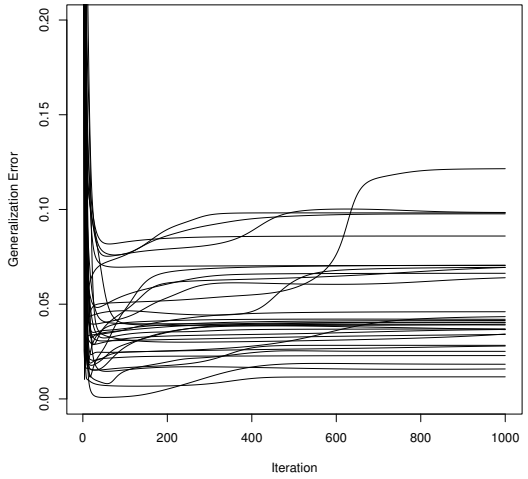

図 6: EM 法における繰り返しと汎化誤差

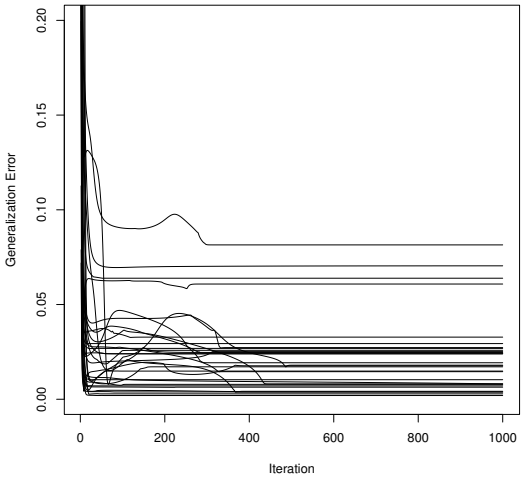

図 7: VB 法における繰り返しと汎化誤差