2005 年情報論的学習理論ワークショップ
2005 Workshop on Information-Based Induction Sciences (IBIS2005)
Tokyo, Japan, November 9 - 11, 2005.

線形神経回路網における変分ベイズ法の汎化特性
Generalization Properties of Variational Bayes Approach in Linear Neural Networks

中島 伸一* 渡辺 澄夫†
Shinichi Nakajima Sumio Watanabe

Abstract: It is known that, in unidentifiable models, the Bayes estimation provides much better generalization performance than the maximum likelihood estimation, however, its accurate approximation by Markov chain Monte Carlo methods requires huge computational costs. As an alternative, a tractable approximation method, called the variational Bayes (VB) approach, has recently been proposed and been showing good generalization performance in many applications, nevertheless, little of its generalization properties has been theoretically clarified yet. In this paper, we prove that, in three-layer linear neural networks, the VB approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation, and clarify its free energy and generalization error, which are less simply related to each other, unlike in the Bayes estimation.

Keywords: variational Bayes, unidentifiable, singular, James-Stein, generalization

1 はじめに

神経回路網，ベイジアンネット，混合分布モデル，隠れマルコフモデル等，機械学習の分野で用いられる多くのモデルは識別不能であり，そのパラメータ空間に特異点を持つ．そのため，最尤推定量やベイズ事後分布の漸近正規性に基づいた，正則モデルにおいて成立する学習理論は，これらのモデル（以下特異モデルと呼ぶ）においては成立しない [19]．近年，いくつかの特異モデルにおいてその汎化性能が解明され，以下のことが知られるようになった．「一般に特異モデルにおいて，最尤法を用いる場合，正則モデルよりも汎化誤差が大きく [7]，ベイズ法を用いる場合，汎化誤差が小さい [2,18,22].」したがって特異モデルにおいては，ベイズ法を用いて学習することが望ましいといえる．しかしながら，ベイズ法を実現するために用いられるマルコフ鎖モンテカルロ（MCMC）法は，多くの計算コストを要する．この問題を解決するため，変分ベイズ（variational Bayes；VB）法と呼ばれる繰り返しアルゴリズムが提案され，実験的に良い汎化性能を示している [3,8,9,12,20]．しかし，その汎化性能は理論的には解明されていない．

ところで，正則モデルにおいても最尤推定量が必ずしも漸近最適な汎化性能を持たないことが証明されている．James-Stein（JS）縮小推定量は，最尤推定量を優越する推定量として提案された [10,23]．いくつかの論文において，ベイズ学習法と JS 推定との関係が議論されている．[6] では，JS 推定量が，事前分布の分散をハイパーパラメータとする経験ベイズ法によって導出されることが示された．[17] では，特異モデルにおけるベイズ法と，正則モデルにおける JS 推定との間の，汎化誤差の振る舞いの類似性が指摘されている．また，[21] では，本論で解析するのと同じモデルである 3 層線形神経回路網において，部分空間ベイズ法（パラメータの一部をハイパーパラメータとみなし，経験ベイズ法を適用する方法）が，JS 推定の拡張である JS 型打ち切り推定（positive-part JS type estimation）と等価であることが示された．実は，本論で導出する変分ベイズ解は，部分空間ベイズ法の解と漸近的に一致する．

本論では，最も簡単な特異モデルである 3 層線形神経回路網（three-layer linear neural networks；LNN）の解析を通して，変分ベイズ法という，機械学習において近年注目されている方法と，縮小推定という，統計学に

*東京工業大学〒 226-8503 神奈川県横浜市緑区長津田 4259, R2-5
Tokyo Institute of Technology, Mailbox R2-5, 226-8503 Japan
tel. 045-924-5018, e-mail nakajima.s@cs.pi.titech.ac.jp
（株）ニコン, 〒 360-8559 埼玉県熊谷市大字御稜威ヶ原 201-9
Nikon Corporation, 201-9 360-8559 Japan
†東京工業大学, Tokyo Institute of Technology, e-mail swatanab@pi.titech.ac.jp

おける古典的な概念との関係をあきらかにし，変分ベイズ法の汎化誤差を理論的に解明する．本論の主要な結果を以下に述べる．

1. 入力が直交する LNN において，変分ベイズ法は打ち切り型縮小推定と漸近的に等価である．
2. 変分ベイズ法は多くの場合，ベイズ法に匹敵する汎化性能を示すが，最尤法に似た性質も合わせ持つ．
3. ベイズ法と異なり，自由エネルギーと汎化誤差との間には簡単な関係が成立しない．

2 線形神経回路網（LNN）

入力（列）ベクトルを $x \in \mathbb{R}^M$，出力ベクトルを $y \in \mathbb{R}^N$，パラメータベクトルを w とする．出力ベクトルには $\mathcal{N}_N(0, \Sigma)$ に従うノイズが付加されると仮定する．(ここで $\mathcal{N}_d(\mu, \Sigma)$ は，平均 μ，共分散行列 Σ の d 次元正規分布を表す．また，以下でその密度関数を $\mathcal{N}_d(\cdot; \mu, \Sigma)$ と表す.）すると，H 個の中間素子を持つ LNN の条件付き密度関数は，

$$p(y|x, w) = \mathcal{N}_N(y; f(x; w), \Sigma), \tag{1}$$

$$\text{where} \quad f(x; w) = BAx = \textstyle\sum_{h=1}^{H} b_h a_h^t x \tag{2}$$

で与えられる．ここで，$A = (a_1, \ldots, a_H)^t$ は $H \times M$ 入力パラメータ行列，$B = (b_1, \ldots, b_H)$ は $N \times H$ 出力パラメータ行列であり，上付き添え字 t は行列の転置を表す．LNN は縮小ランク回帰とも呼ばれ，多変数線形回帰問題において，出力を支配する要因の次元が，入力次元および出力次元よりも小さいことが予想されるような場合に使われるモデルである．任意の $H \times H$ 正則行列 T に対し，変換 $(A, B) \mapsto (TA, BT^{-1})$ は写像を変えないため，縮小ランク回帰モデルのパラメータ次元は

$$K = H(M + N) - H^2. \tag{3}$$

である．本論では $H \leq N \leq M$ を仮定する．

3 ベイズ学習法

3.1 ベイズ法

真の分布 $q(x, y) = q(x)q(y|x)$ から得られる n 個の iid サンプルを $(X^n, Y^n) = (\{x_1, \ldots, x_n\}, \{y_1, \ldots, y_n\})$ とする．このときモデル $p(y|x, w)$ の周辺尤度は

$$Z(Y^n|X^n) = \int \phi(w) \textstyle\prod_{i=1}^{n} p(y_i|x_i, w) dw \tag{4}$$

で与えられる．ここで $\phi(w)$ はパラメータの事前分布である．ベイズ事後分布は

$$p(w|X^n, Y^n) = \frac{\phi(w) \prod_{i=1}^{n} p(y_i|x_i, w)}{Z(Y^n|X^n)} \tag{5}$$

であり，予測分布はモデルの事後分布による平均

$$p(y|x, X^n, Y^n) = \int p(y|x, w) p(w|X^n, Y^n) dw \tag{6}$$

で与えられえる．

自由エネルギーは $F(Y^n|X^n) = -\log Z(Y^n|X^n)$ で定義されるが，[1] 真の分布のエントロピーを減算したものを規格化自由エネルギーと呼ぶことにする．規格化自由エネルギーの，サンプルのでかたによる平均 $F(n) = \langle F(Y^n|X^n) + \log q(Y^n|X^n) \rangle_{q(X^n, Y^n)}$（ここで $\langle \cdot \rangle_{q(X^n, Y^n)}$ は n 個の学習サンプルのでかたについての平均を表す.）は一般に，

$$F(n) = \lambda' \log n + o(\log n) \tag{7}$$

と漸近展開される．本論では汎化誤差を，真の分布から見た予測分布のカルバック擬距離の，サンプルのでかたによる平均

$$G(n) = \left\langle \int q(x) q(y|x) \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} dx dy \right\rangle_{q(X^n, Y^n)} \tag{8}$$

で定義する．汎化誤差も一般に漸近展開され，

$$G(n) = \lambda n^{-1} + o(n^{-1}) \tag{9}$$

と表される．本論では，主要項の係数 λ' および λ をそれぞれ，自由エネルギー係数および汎化係数と呼ぶことにする．ベイズ法においては，

$$G(n) = F(n + 1) - F(n) \tag{10}$$

なる関係 [11] から，$\lambda = \lambda'$ が導かれる．

3.2 変分ベイズ法

任意の試行分布 $r(w)$ に対し，Jensen の不等式より

$$F(Y^n|X^n) \leq \left\langle \log \frac{r(w)}{p(Y^n|X^n, w)\phi(w)} \right\rangle_{r(w)} \tag{11}$$

$$= \bar{F}(Y^n|X^n)$$

が成立する．ここで，$\langle \cdot \rangle_p$ は分布 p に関する期待値を示し，$\bar{F}(Y^n|X^n)$ は変分自由エネルギーと呼ばれる．[2] 変分ベイズ法では，制限された関数クラスの中で，変分自由エネルギーを最小にする $r(w)$（これを変分事後分布と呼ぶ.）で事後分布を近似する．変分ベイズ法はもともと神経回路網において，事後分布の関数クラスを正規分布に限定することによって導入されたが [9]，近年，混合

[1] 自由エネルギーは，エビデンスあるいは確率的複雑さとも呼ばれ，モデル選択やハイパーパラメータ最適化等に使用される [6].

[2] 以下，特に混同の恐れのない場合，変分自由エネルギーの最小値も変分自由エネルギーと呼び，同じ記号 $\bar{F}(Y^n|X^n)$ を使って記す.

分布モデルやベイジアンネット等において，隠れ変数とパラメータの事後分布の独立性を仮定し，[3] 適切な事前分布を用いることによって，Expectation-Maximization (EM) アルゴリズム [5] に似た繰り返しアルゴリズムが導出されることが見出された [3,20].

LNN においても，入力パラメータ行列 A と出力パラメータ行列 B の独立性を事後分布に仮定することにより，同様のアルゴリズムが，以下のように導出される．

$$r(w) = r(A,B) = r(A)r(B) \tag{12}$$

を仮定すると，変分自由エネルギーは

$$\bar{F}(Y^n|X^n) = \int r(A)r(B)\log\frac{r(A)r(B)}{p(Y^n|X^n,A,B)\phi(A,B)}dw \tag{13}$$

と書ける．これを変分法を用いて解くことにより，

$$r(A) \propto \phi(A)\exp\langle\log p(Y^n|X^n,A,B)\rangle_{r(B)} \tag{14}$$

$$r(B) \propto \phi(B)\exp\langle\log p(Y^n|X^n,A,B)\rangle_{r(A)} \tag{15}$$

を得る．ただし，事前分布の独立性 $\phi(A,B) = \phi(A)\phi(B)$ も仮定した．対数尤度関数 $\log p(Y^n|X^n,A,B)$ が A，B それぞれについて 2 次形式であることから，式 (14) および (15) より，事前分布が正規分布であれば事後分布も正規分布となることがわかる．

変分自由エネルギー $\bar{F}$ と同様に，変分ベイズ法（すなわち式 (6) において，$p(w|X^n,Y^n)$ を変分事後分布で代用する学習法）で得られる変数を，上線を付けて表す．例えば自由エネルギー係数は $\bar{\lambda}'$，汎化係数は $\bar{\lambda}$，などである．変分ベイズ法では，式 (10) に相当する関係が成立しないため，一般に $\bar{\lambda} \neq \bar{\lambda}'$ である．すなわち，変分ベイズ法における自由エネルギーと汎化誤差とは，ベイズ法の場合のような単純な関係にはない．

4 理論解析

4.1 変分条件

出力ノイズの共分散行列が単位行列 I_N に等しいと仮定すると，(以下，$d \times d$ 単位行列を I_d と書く.）モデルの密度関数は

$$p(y|x,A,B) = \mathcal{N}_N(y; BAx, I_N) \tag{16}$$

と書ける．事前分布として，

$$\phi(A,B) = \mathcal{N}_{HM}(\mathrm{vec}(A^t); 0, c_a^2 I_{HM}) \cdot \mathcal{N}_{NH}(\mathrm{vec}(B); 0, c_b^2 I_{NH}) \tag{17}$$

を用いる．ここで，$0 < c_a^2, c_b^2 < \infty$ は分散に対応する定数であり，[4] $\mathrm{vec}(\cdot)$ は行列の列ベクトルを縦 1 列に並べて作った列ベクトルを示す．本論では，パラメータの真の値には $*$，推定値には $\hat{}$ を付けて表す．真の写像 B^*A^* のランクが $H^* \leq H$ であると仮定する．簡単のため，入力ベクトルの規格直交性 $\int xx^t q(x)dx = I_M$ を仮定すると，[5] 中心極限定理により以下が得られる．

$$Q(X^n) = n^{-1}\textstyle\sum_{i=1}^n x_i x_i^t = I_M + O_p(n^{-1/2}), \tag{18}$$

$$R(X^n,Y^n) = n^{-1}\textstyle\sum_{i=1}^n y_i x_i^t = B^*A^* + O_p(n^{-1/2}). \tag{19}$$

ここで $\{Q(X^n), R(X^n,Y^n)\}$ は十分統計量であり，それぞれ $M \times M$ 対称行列，$N \times M$ 行列である．以後，学習サンプル依存性を省略して $\{Q,R\}$ と書く．

行列 $RQ^{-1/2}$ の h 番目に大きい特異値を γ_h とし，対応する右および左特異ベクトルをそれぞれ ω_{a_h} および ω_{b_h} とする.（ここで，$1 \leq h \leq H$ である.）漸近極限において，特異値の大きい方から H^* 個は，真の分布を表現するのに必要な成分に確率 1 で対応する．従って式 (19) より，$H^* < h \leq H$ に対して γ_h はオーダー $O_p(n^{-1/2})$ であることがわかり，式 (18) を用いると，

$$\omega_{b_h} RQ^\rho = \omega_{b_h} R + O_p(n^{-1}) \quad \text{for } H^* < h \leq H \tag{20}$$

が成立する．ただし $-\infty < \rho < \infty$ は任意の定数である．写像 BA の特異値分解の存在により，写像（すなわち尤度）を変えずにすべての行ベクトルが互いに直交している行列 A および，すべての列ベクトルが互いに直交している行列 B の組を取れる．事前分布 (17) は各成分が直交するとき最大値をとるので，一般性を失わずに変分事後分布を

$$r(A,B) = \mathcal{N}_M(a_h; \mu_{a_h}, \Sigma_{a_h})\mathcal{N}_N(b_h; \mu_{b_h}, \Sigma_{b_h}) \tag{21}$$

として良い．ただし，$\{\mu_{a_h}; h = 1,\ldots,H\}$ および $\{\mu_{b_h}; h = 1,\ldots,H\}$ はそれぞれ互いに直交する H 個のベクトルの集合である．式 (21) を式 (14)，(15) に代入することにより，変分事後分布に関する以下の変分条件を得る．

$$\mu_{a_h} = n\Sigma_{a_h} R^t \mu_{b_h}, \tag{22}$$

$$\Sigma_{a_h} = (n(\|\mu_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h}))Q + c_a^{-2} I_M)^{-1}, \tag{23}$$

$$\mu_{b_h} = n\Sigma_{b_h} R\mu_{a_h}, \tag{24}$$

$$\Sigma_{b_h} = (n(\mu_{a_h}^t Q\mu_{a_h} + \mathrm{tr}(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2})) + c_b^{-2})^{-1} I_N. \tag{25}$$

[3] この仮定は，以下で LNN に対して課す，異なる層のパラメータ間の独立性（これは，変分ベイズ法により計算を容易にする上で，本質的な役割を果たす.）よりも強い仮定である．

[4] 分散 c_a^2 および c_b^2 は 4.2 章で示すように，正で有界である限り，漸近結果に影響しない．

[5] 規格直交性の影響の解析は今後の課題である．ここでは，サンプルデータにより入力を直交化し（$O_p(n^{-1})$ の精度で可能），新しく取り直した入力ベクトルの基底に対して事前分布 (17) を定義するような学習方法を想定している．

ここで，tr(·) は行列のトレースを示す．同様に，式 (21) を式 (13) に代入することにより，変分自由エネルギーは以下のように書ける．

$$2\bar{F}(Y^n|X^n) = \sum_{h=1}^{H} \Big\{ -\log \sigma_{a_h}^{2M}\sigma_{b_h}^{2N} - 2n\mu_{b_h}^t R\mu_{a_h} + n\Big(\mu_{a_h}^t Q\mu_{a_h} + \mathrm{tr}(\Sigma_{a_h}^{1/2} Q \Sigma_{a_h}^{1/2})\Big)\big(\|\mu_{b_h}\|^2 + \mathrm{tr}(\Sigma_{b_h})\big)\Big\} + \sum_{i=1}^{n}\|y_i\|^2 + O_p(1) + \mathrm{const.} \quad (26)$$

ここで，$\sigma_{a_h}^2$ および $\sigma_{b_h}^2$ はそれぞれ以下で定義される．

$$\Sigma_{a_h} = \sigma_{a_h}^2 (I_M + O_p(n^{-1/2})), \quad (27)$$
$$\Sigma_{b_h} = \sigma_{b_h}^2 (1 + O_p(n^{-1/2})) I_N. \quad (28)$$

4.2 変分ベイズ解

変分条件 (22)–(25) は解析的に解くことができて，以下の定理を得る．

定理 1 $L = \max(M, N) = M$，$L'_h = \max(L, n\gamma_h^2)$ とする．*LNN* の写像の変分ベイズ推定量は以下で与えられる．

$$\hat{B}\hat{A} = \sum_{h=1}^{H} \big(1 - LL_h'^{-1}\big)\, \omega_{b_h}\omega_{b_h}^t RQ^{-1} + O_p(n^{-1}). \quad (29)$$

（証明）事前分布 (17)，および変分自由エネルギー (26) の形を考慮することにより，μ_{a_h}，μ_{b_h}，Σ_{a_h} および Σ_{b_h} の最適値はかならず変分条件 (22)–(25) を満たさねばならないことがわかる．式 (22)，(24) より，

$$\hat{\mu}_{a_h} = n^2 \hat{\Sigma}_{a_h} R^t \hat{\Sigma}_{b_h} R \hat{\mu}_{a_h}, \quad (30)$$
$$\hat{\mu}_{b_h} = n^2 \hat{\Sigma}_{b_h} R \hat{\Sigma}_{a_h} R^t \hat{\mu}_{b_h}, \quad (31)$$

であるので，$\hat{\mu}_{a_h}$ と $\hat{\mu}_{b_h}$ はそれぞれ R^tR と $R\hat{\Sigma}_{a_h}R^t$ の固有ベクトルであるか，あるいは $\|\hat{\mu}_{a_h}\| = \|\hat{\mu}_{b_h}\| = 0$ である．真の分布を表現するのに必要な成分（$h \le H^*$）に関して，観測される R の特異値はオーダー $O_p(1)$ である．従って自由エネルギー (26) が最小化されるのは，$\hat{\mu}_{a_h}$ および $\hat{\mu}_{b_h}$ が共にオーダー 1 の場合に限られる．よって，変分条件 (22)–(25) は以下のように書きなおせる．

$$\hat{\mu}_{a_h} = \|\hat{\mu}_{b_h}\|^{-2} Q^{-1} R^t \hat{\mu}_{b_h} + O_p(n^{-1}), \quad (32)$$
$$\hat{\Sigma}_{a_h} = n^{-1}\|\hat{\mu}_{b_h}\|^{-2} Q^{-1} + O_p(n^{-2}), \quad (33)$$
$$\hat{\mu}_{b_h} = (\hat{\mu}_{a_h}^t Q \hat{\mu}_{a_h})^{-1} R \hat{\mu}_{a_h} + O_p(n^{-1}), \quad (34)$$
$$\hat{\Sigma}_{b_h} = n^{-1}(\hat{\mu}_{a_h}^t Q \hat{\mu}_{a_h})^{-1} I_N + O_p(n^{-2}). \quad (35)$$

以上により，以下の変分ベイズ推定量を得る．

$$\hat{\mu}_{b_h}\hat{\mu}_{a_h}^t = \omega_{b_h}\omega_{b_h}^t RQ^{-1} + O_p(n^{-1}). \quad (36)$$

一方，冗長成分（$h > H^*$）に関して，式 (20)，(27) および (28) を用いると，変分条件 (22)–(25) は以下のように書ける．

$$\hat{\mu}_{a_h} = n\hat{\sigma}_{a_h}^2 R^t \hat{\mu}_{b_h}(1 + O_p(n^{-1/2})), \quad (37)$$
$$\hat{\sigma}_{a_h}^2 = (n(\|\hat{\mu}_{b_h}\|^2 + N\hat{\sigma}_{b_h}^2) + c_a^{-2})^{-1}, \quad (38)$$
$$\hat{\mu}_{b_h} = n\hat{\sigma}_{b_h}^2 R \hat{\mu}_{a_h}(1 + O_p(n^{-1/2})), \quad (39)$$
$$\hat{\sigma}_{b_h}^2 = (n(\|\hat{\mu}_{a_h}\|^2 + M\hat{\sigma}_{a_h}^2) + c_b^{-2})^{-1}. \quad (40)$$

導出は省略するが，以下が条件 (37)–(40) の解である．

1. When $M > N$,

$$\hat{\mu}_{a_h} = \left(\left(1 - \frac{L}{L'_h}\right)\frac{L-l}{c_a^{-2}}\right)^{1/2}\omega_{a_h} + O_p(n^{-1}), \quad (41)$$
$$\hat{\sigma}_{a_h}^2 = \frac{L-l}{c_a^{-2}L'_h} + O_p(n^{-1}), \quad (42)$$
$$\hat{\mu}_{b_h} = \left(\left(1 - \frac{L}{L'_h}\right)\frac{c_a^{-2}}{L-l}\right)^{1/2}\gamma_h\omega_{b_h} + O_p(n^{-1}), \quad (43)$$
$$\hat{\sigma}_{b_h}^2 = \frac{c_a^{-2}}{n(L-l)} + O_p(n^{-2}). \quad (44)$$

2. When $M = N$,

$$\hat{\mu}_{a_h} = \left(\frac{c_a}{c_b}\left(1 - \frac{L}{L'_h}\right)\gamma_h\right)^{1/2}\omega_{a_h} + O_p(n^{-1}), \quad (45)$$
$$\hat{\sigma}_{a_h}^2 = \frac{c_a}{c_b\sqrt{nL'_h}} + O_p(n^{-1}), \quad (46)$$
$$\hat{\mu}_{b_h} = \left(\frac{c_b}{c_a}\left(1 - \frac{L}{L'_h}\right)\gamma_h\right)^{1/2}\omega_{b_h} + O_p(n^{-1}), \quad (47)$$
$$\hat{\sigma}_{b_h}^2 = \frac{c_b}{c_a\sqrt{nL'_h}} + O_p(n^{-1}). \quad (48)$$

ただし，$l = \min(M, N) = N$ である．特異値を大きい方から H 個選んだとき，自由エネルギー (26) は最小になる．式 (36) と，必要成分（$1 \le h \le H^*$）において $LL_h'^{-1} = O_p(n^{-1})$ が成り立つという事実，また，上の解と式 (20) とを合わせることにより，定理 1 の変分ベイズ推定量を得る．(Q.E.D.)

予測分布を真の分布で割ったもの $\bar{p}(y|x, X^n, Y^n)/q(y|x)$ を展開することにより，以下の補題が得られる．

補題 1 $\hat{V} = I_N + O_p(n^{-1})$ とする．*LNN* の変分ベイズ法における予測分布は以下のように書ける．

$$\bar{p}(y|x, X^n, Y^n) = \mathcal{N}_N(y; \hat{V}\hat{B}\hat{A}x, \hat{V}) + O_p(n^{-3/2}). \quad (49)$$

式 (29) と，最尤推定量

$$\hat{B}\hat{A}_{MLE} = \sum_{h=1}^{H}\omega_{b_h}\omega_{b_h}^t RQ^{-1} \quad (50)$$

[4] とを比較してみると，変分ベイズ推定量の各成分はそれぞれ，JS 型打ち切り推定量となっていることがわかる．[6] さらに補題 1 より，予測分布として変分ベイズ

[6] パラメータ w の JS 型打ち切り推定量 [6, 10, 23] は，
$$\hat{w}_{PJS} = \theta(\|\hat{w}_{MLE}\|^2 > \chi/n)(1 - \chi/n\|\hat{w}_{MLE}\|^2)\hat{w}_{MLE}$$
$$= \big(1 - \chi\chi'^{-1}\big)\,\hat{w}_{MLE}$$
で定義される．ただし，$\hat{w}_{MLE}$ は最尤推定量，$\theta(\cdot)$ は定義関数，$\chi > 0$ は定数，$\chi' = \max(\chi, n\|\hat{\mu}_{MLE}\|^2)$ である．

推定量に対応するモデル（すなわちパラメータ空間上の 1 点）を用いても，漸近的に汎化誤差には影響しない．よって，LNN における変分ベイズ法と JS 型打ち切り推定との漸近等価性が示された．

定理 1 の証明において，変分事後分布が求まったので，それを式 (26) に代入することにより，変分自由エネルギー（の最小値）を求めることができる．

補題 2 *LNN* の変分規格化自由エネルギーは $\bar{F}(n) = \bar{\lambda}' \log n + O(1)$ と漸近展開される．ただし，自由エネルギー係数は以下である．

$$2\bar{\lambda}' = H^*(M+N) + (H-H^*)l. \qquad (51)$$

(証明) 必要成分（$h \leq H^*$）に対して，式 (36) より $\hat{\mu}_{a_h}$ と $\hat{\mu}_{b_h}$ とがともにオーダー $O_p(1)$ であることがわかる．よって式 (33) および (35) より，Σ_{a_h} および Σ_{b_h} はともにオーダー $O_p(n^{-1})$ である．これらの事実を式 (26) に代入すると，式 (51) の第 1 項が得られる．一方，冗長成分（$h > H^*$）に対して，変分ベイズ解 (41)–(48) を式 (26) に代入することにより，式 (51) の第 2 項を得る．(Q.E.D.)

必要成分からの寄与が，2 章で述べた自明な冗長性を含んでいることに注意する．(式 (51) の第一項と式 (3) とを比較せよ．) これは A と B との独立性が，変分事後分布が自明な縮退方向に伸びるのを妨げることに起因し，自明な冗長性をもつモデルでのみ起こる現象である．

4.3 汎化誤差

変分ベイズ解は，部分空間ベイズ法の漸近解 [21] と同じ形をしているため，汎化誤差は同様の方法で求めることができる．本論では，証明なしで変分ベイズ法の汎化誤差に関する定理を述べる．自由度 m，尺度行列 Σ の d 次元ウィシャート分布を $\mathcal{W}_d(m, \Sigma)$ と書く．

定理 2 LNNの変分ベイズ法における汎化誤差は $\bar{G}(n) = \bar{\lambda} n^{-1} + O(n^{-3/2})$ と漸近展開される．ただし汎化係数は

$$2\bar{\lambda} = (H^*(M+N) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > L) \left(1 - \frac{L}{\gamma_h'^2}\right)^2 \gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})} \qquad (52)$$

である．ここで，$\theta(\cdot)$ は定義関数，$\gamma_h'^2$ は $\mathcal{W}_{N-H^*}(M-H^*, I_{N-H^*})$ に従うランダム行列の h 番目に大きい固有値，$\langle\cdot\rangle_{q(\{\gamma_h'^2\})}$ はウィシャート分布に関する期待値を示す．

式 (52) の第 2 項は初等関数ではないが，ウィシャート分布に従うサンプルを発生させることにより，比較的容易に計算できる．また，M，N，H および H^* を同じ比率で無限大に発散させた極限で収束する関数も得られる [21]．

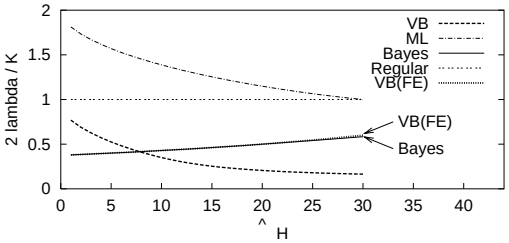

図 1: 汎化係数 （$M = 50, N = 30, H^* = 0$）．

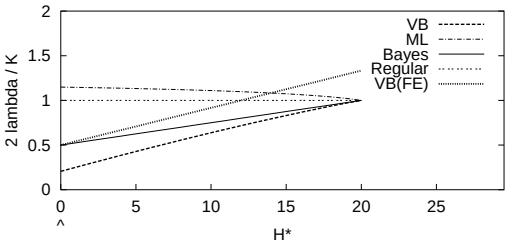

図 2: 汎化係数（$M = 50, N = 30, H = 20$）．

5 考察

図 1 は，真のランクが $H^* = 0$ である場合の，入力素子数 $M = 50$，出力素子数 $N = 30$，中間素子数 $H = 1, \ldots, 30$ の LNN の汎化係数の理論値を示す．縦軸は汎化係数を示すが，式 (3) で与えられるパラメータ次元 K で規格化されている．横軸は H を示す．4 本のはっきり区別できる曲線はそれぞれ，本論で得られた変分ベイズ法，[7] で得られた最尤法，[2] で得られたベイズ法，および正則モデルの汎化係数を示す．[7] この図から，変分ベイズ法の汎化性能が良いことがわかるが，一方で，H への依存のしかたが，ベイズ法ではなくむしろ最尤法に似ていることも見てとれる．このことは，定理 1 および 2 からわかるように，変分ベイズ法が「ランダム行列の特異値のなかから最大のものを（すなわち最もノイズに適合する基底を）選ぶことによる過学習の促進」という，最尤法の性質を備えていることに起因する．実際，変分ベイズ法の汎化係数は，$M = N = 80$, $H = 1$, $H^* = 0$ の場合 $2\bar{\lambda}/K \sim 1.04$ となり，正則モデルの汎化係数を超えるが，このようなことはベイズ法では起こらないことが証明されている [16]．

実は図 1 には，VB(FE) とラベル付けされた曲線も描かれている．これは変分自由エネルギー係数を示すので

[7] 正則モデルでは，最尤法，ベイズ法のいずれの場合でも，汎化係数は K に等しく，このことが赤池情報量基準（AIC）の基礎を与えている [1]．なお，ベイズ法における自由エネルギー係数もやはり K に等しく，これがベイズ情報量基準（BIC）[15] および最小記述長基準（MDL）[14] の基礎を与えている．

あるが，ベイズ汎化係数（これはベイズ自由エネルギー係数に一致する）とほぼ重なっている．他の条件についても検討した結果，$H^* = 0$ でありかつ，出力次元と入力次元との比が $N/M \leq 2/3$ の場合には，変分自由エネルギーがベイズ自由エネルギーをよく近似するといえる．それに対し，上に見たように，変分ベイズ汎化誤差とベイズ汎化誤差の振る舞いは大きく異なる．

図 2 は，$M = 50$，$N = 30$，$H = 20$ の LNN において，横軸に真のランク $H^* = 1, \ldots, 20$ を取ったときの係数の振る舞いを示している．[8] この LNN においては，真のランク H^* にかかわらず，常に変分ベイズ法の汎化係数はベイズ法のそれより小さい．しかしこのことは，変分ベイズ法がベイズ法を優越する（dominate）ことを必ずしも意味しないことに注意する．漸近論における優越性の判定には，特異点から見た真の分布のカルバック擬距離がオーダー n^{-1} であるような，より「微妙な」状況を検討する必要がある [17, 21]．[9] 図 2 ではまた，変分自由エネルギー係数の振る舞いが，ベイズ自由エネルギー係数のそれと大きく異なることも見てとれるが，これは LNN 特有の自明な冗長性に起因したものであり（4.2 章の最後の段落参照），変分ベイズ法の一般的な特徴ではないことを付け加えておく．

6 結論

本論では，3 層線形神経回路網において，変分ベイズ法が James-Stein 型打ち切り推定と漸近的に等価であることを示し，その汎化特性を理論的に解明した．今後の課題は，(非線形）神経回路網や，隠れ変数を持つモデル（混合分布モデル，ベイジアンネットなど）の解析である．我々は，非線形性は汎化性能を悪化させるが，本論で得られたいくつかの特徴は，他のモデルにおいても成立するのではないかと予想している．

参考文献

[1] H. Akaike. A New Look at Statistical Model. *IEEE Trans. on Automatic Control*, Vol. 19, pp. 716–723, 1974.

[2] M. Aoyagi and S. Watanabe. The Generalization Error of Reduced Rank Regression in Bayesian Estimation. In *Proc. of ISITA*, pp. 1068–1073, Parma, Italy, 2004.

[3] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proc. of UAI*, 1999.

[4] P. F. Baldi and K. Hornik. Learning in Linear Neural Networks: a Survey. *IEEE Trans. on Neural Networks*, Vol. 6, pp. 837–858, 1995.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood for Incomplete Data Via the EM Algorithm. *J. R. Statistical Society*, Vol. 39-B, pp. 1–38, 1977.

[6] B. Efron and C. Morris. Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach. *J. of Am. Stat. Assoc.*, Vol. 68, pp. 117–130, 1973.

[7] K. Fukumizu. Generalization Error of Linear Neural Networks in Unidentifiable Cases. In *Proc. of ALT*, pp. 51–62. Springer, 1999.

[8] Z. Ghahramani and M. J. Beal. Graphical Models and Variational Methods. In *Advanced Mean Field Methods*. MIT Press, 2000.

[9] G. E. Hinton and D. van Camp. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. In *Proc. of COLT*, 1993.

[10] W. James and C. Stein. Estimation with Quadratic Loss. In *Proc. of the 4th Berkeley Symp. on Math. Stat. and Prob.*, pp. 361–379, 1961.

[11] E. Levin, N. Tishby, and S. A. Solla. A Statistical Approaches to Learning and Generalization in Layered Neural Networks. In *Proc. of IEEE*, Vol. 78, pp. 1568–1674, 1990.

[12] D. J. C. MacKay. Developments in Probabilistic Modeling with Neural Networks—Ensemble Learning. In *Proc. of the 3rd Ann. Symp. on Neural Networks*, pp. 191–198, 1995.

[13] S. Nakajima and S. Watanabe. Generalization Error and Free Energy of Variational Bayes Approach of Linear Neural Networks. *To appear in Proc. of ICONIP*, 2005.

[14] J. Rissanen. Stochastic Complexity and Modeling. *Annals of Statistics*, Vol. 14, pp. 1080–1100, 1986.

[15] G. Schwarz. Estimating the Dimension of a Model. *Annals of Statistics*, Vol. 6, No. 2, pp. 461–464, 1978.

[16] S. Watanabe. Algebraic Information Geometry for Learning Machines with Singularities. In *Advances in NIPS*, Vol. 13, pp. 329–336, 2001.

[17] S. Watanabe and S. Amari. Learning Coefficients of Layered Models When the True Distribution Mismatches the Singularities. *Neural Computation*, Vol. 15, pp. 1013–1033, 2003.

[18] K. Yamazaki and S. Watanabe. Singularities in Mixture Models and Upper Bounds of Stochastic Complexity. *Neural Networks*, Vol. 16, No. 7, pp. 1029–1038, 2003.

[19] 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, 東京, 2004.

[20] 上田修功. ベイズ学習. 電子情報通信学会誌, Vol. 85, No. 4,6,7,8, April–August 2002.

[21] 中島伸一, 渡辺澄夫. 線形神経回路網における部分空間ベイズ法の解析. 信学技報（NC 研究会）, October 2005.

[22] 渡辺澄夫. データ学習アルゴリズム. 共立出版, 東京, 2001.

[23] 久保川達也. モデル選択（第 3 部：スタインのパラドクスと縮小推定の世界）. 岩波書店, 東京, 2004.

[8] 図 1 における $H = 20$ の場合と図 2 における $H^* = 0$ の場合とは（いずれも横軸数値の下に ˆ 印を付けた），同じ場合を示していることに注意する．

[9] [21] では，変分ベイズ法と漸近等価な縮小推定の，「微妙な」状況における汎化係数に関する定理が導出されており，[17] で得られたベイズ法のそれと比較した上で，ベイズ法に対する優越性を否定している [21]．