社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

線形神経回路網における部分空間ベイズ法の解析

— ベイズ的推定法と縮小推定との関係 —

中島　伸一†,††　　渡辺　澄夫†

† 東京工業大学総合理工学研究科　〒 226–8503 神奈川県横浜市緑区長津田 4259，R2-5
†† 株式会社ニコン　〒 360–8559 埼玉県熊谷市大字御稜威ヶ原 201–9
E-mail: †nakajima.s@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

あらまし　神経回路網や混合分布モデルなどの特異モデルは，正則モデルとは異なる汎化性能を示す．近年の研究によると，一般に正則モデルと比較して，最尤法では過学習しやすく，ベイズ法では過学習しにくいことが知られている．従って特異モデルの学習においては，ベイズ法を用いることが望ましいのであるが，事後分布を実現するための方法であるマルコフ連鎖モンテカルロ法は，膨大な計算量を必要とする．このため，変分ベイズ法などの近似手法が提案されている．本論では，別の近似方法である，部分空間ベイズ法についてその汎化性能を解析し，縮小推定，変分ベイズ法との関係を議論する．得られた結論は以下である．3 層線形神経回路網において，部分空間ベイズ法は Jame-Stein 型縮小推定と漸近等価であり，多くの場合，ベイズ法に匹敵する汎化性能を示す．また，変分ベイズ法とも強く関連している．
キーワード　部分空間ベイズ，経験ベイズ，変分ベイズ，縮小推定，神経回路網，特異モデル

Analysis of Subspace Bayes Approach in Linear Neural Networks

— Relation between Baysian Approach and Shrinkage Estimation —

Shinichi NAKAJIMA†,†† and Sumio WATANABE†

† Tokyo Institute of Technology, Mailbox R2-5, 4259 Nagatsuda, Yokohama-shi, Kanagawa, 226-8503 Japan
†† Nikon Corporation, 201-9 Oaza-Miizugahara, Kumagaya-shi, Saitama, 360-8559 Japan
E-mail: †nakajima.s@cs.pi.titech.ac.jp, ††swatanab@pi.titech.ac.jp

Abstract　It is well known that the generalization performance of unidentifiable models differs from that of the regular models. According to recent works, it is known that the Bayes estimation has the advantage over the maximum likelihood estimation. However, accurate approximation of the posterior distribution requires huge computational costs. In this paper, we consider an alternative approximation method, which we call a subspace Bayes approach, and discuss the relation to the shrinkage estimation and the variational Bayes approach. We show that, in three-layer linear neural networks, the subspace Bayes approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation, that it provides as good generalization performance as the Bayes estimation in typical cases, and that it is strongly related to the variational Bayes approach.
Key words　subspace Bayes, empirical Bayes, variational Bayes, shrinkage, neural networks, singular model

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などは，そのパラメータと分布とが一対一対応しておらず，識別不能である．識別不能なモデルはパラメータ空間に，フィッシャー情報行列が縮退する特異点をもち，それゆえに正則モデルを対象とした古典的な学習理論が適用できない [2].（以後，このようなモデルを特異モデルと呼ぶ.）特異モデルについては，以下のことが知られている．学習に最尤（maximum likelihood；ML）法を用いた場合，正則モデルよ

りも過学習しやすいが [3]，ベイズ法を用いた場合には，過学習が抑制される [4]〜[6]．従って，特異モデルの学習においてはベイズ法を用いることが望ましいのであるが，ベイズ事後分布を実現するために用いられるマルコフ鎖モンテカルロ（Markov chain Monte Carlo；MCMC）法は，膨大な計算量を必要とする．この問題を解決するために，近似的解法として変分ベイズ（variational Bayes；VB）法などが提案された [7]〜[10]．(注1)

本論では，別の近似法である部分空間ベイズ法について，その汎化性能を解析する．部分空間ベイズ法とは，パラメータの一部をハイパーパラメータとみなして経験ベイズ法を適用する方法である．いくつかの 3 層モデルにおいては，周辺化を解析的に実行できる場合があり，そのような場合にはハイパーパラメータの最適化を行うだけで学習ができるため，MCMC 法を用いて分布の期待値を得るよりも，はるかに少ない計算量で学習を行うことができる．

2. では線形神経回路網について簡単に述べ，3. ではベイズ法，経験ベイズ法および部分空間ベイズ法について述べる．その後，4. で部分空間ベイズ法と縮小推定 [12], [13] との漸近等価性を証明し，その汎化性能を解明する．5. および 6. で，それぞれ考察および結論を述べる．

2. 線形神経回路網

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．H 個の中間素子を持つ 3 層神経回路網は，写像

$$f(x;w) = \sum_{h=1}^{H} b_h \psi\left(a_h^t x\right) \tag{1}$$

によって定義される．ここで，$w = \{(a_h, b_h) \in \mathbb{R}^M \times \mathbb{R}^N; h = 1, \ldots, H\}$ はすべてのパラメータをひとまとめに表したものである．$\psi(\cdot)$ は活性化関数であり，通常 $\tanh(\cdot)$ のような，単調，有界，反対称な非線形関数が用いられる．上付き添え字 t は行列の転置を示す．平均ベクトル μ，共分散行列 Σ の d 次元正規分布を $\mathcal{N}_d(\mu, \Sigma)$ と書き，その密度関数を $\mathcal{N}_d(\cdot; \mu, \Sigma)$ で表す．出力ベクトルの持つノイズが $\mathcal{N}_N(0, \Sigma)$ に従うと仮定すると，3 層神経回路網の密度関数は

$$p(y|x,w) = \mathcal{N}_N(y; f(x;w), \Sigma) \tag{2}$$

で与えられる．本論では，活性化関数が線形である場合，すなわち線形神経回路網（Linear Neural Network；LNN）について議論する．LNN は縮小ランク回帰モデルとも呼ばれ，多変量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用いられる統計モデルである．LNN の写像は，$H \times M$ 入力パラメータ行列 $A = (a_1, \ldots, a_H)^t$ および $N \times H$ 出力パラメータ行列 $B = (b_1, \ldots, b_H)$ を用いて，

$$f(x; A, B) = BAx \tag{3}$$

と表される．任意の $H \times H$ 正則行列 T に対し，変換 $(A, B) \mapsto (TA, BT^{-1})$ が写像を変えないことから，LNN は H^2 個の自明な冗長性を持っていることがわかる．従ってパラメータ次元は

$$K = H(M + N) - H^2 \tag{4}$$

とみなされる．本論では $H \leqq N \leqq M$ 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 i.i.d. 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

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

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

$$p(w|X^n, Y^n) = \frac{\phi(w) \prod_{i=1}^{n} p(y_i|x_i, w)}{Z(Y^n|X^n)} \tag{6}$$

であり，予測分布はモデルの事後分布による平均，すなわち

$$p(y|x, X^n, Y^n) = \int p(y|x,w) p(w|X^n, Y^n) dw \tag{7}$$

で与えられえる [14], [15].

本論では，汎化誤差を

$$G(n) = \langle G(X^n, Y^n) \rangle_{q(X^n, Y^n)} \tag{8}$$

で定義する．ただし

$$G(X^n, Y^n) = \int q(x)q(y|x) \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} dxdy \tag{9}$$

は真の分布から見た予測分布のカルバック擬距離であり，$\langle \cdot \rangle_{q(X^n, Y^n)}$ は n 個の学習サンプルのでかたについての平均を表す．

3.2 経験ベイズ法と部分空間ベイズ法

事前分布に関する情報を持っていない学習者が，ベイズ法を用いて学習を行う場合について考える．情報を殆ど持たないとみなせるような事前分布を用いるという方法もあるが，別の方法として，事前分布にハイパーパラメータを持たせ，これを周辺尤度を規準として決定する方法があり，経験ベイズ（empirical Bayes；EB）法と呼ばれる．例として，ハイパーパラメータ τ_1 を持つ，以下の事前分布を用いる場合を考える．(注2)

$$\phi(w\|\tau_1) = \frac{1}{(2\pi\tau_1^2)^{K/2}} \exp\left(-\frac{\|w\|^2}{2\tau_1^2}\right). \tag{10}$$

このとき，周辺尤度 (5) も τ_1 の関数となる．経験ベイズ法では，τ_1 を周辺尤度を最大化することによって決定する [14], [16]．この考えを拡張し，モデルにハイパーパラメータを導入することもできる．本論で解析する部分空間ベイズ（subspace Bayes；SB）法では，パラメータの一部をハイパーパラメータとみな

(注1)：線形神経回路網における変分ベイズ法の汎化誤差は最近解明された [11].

(注2)：本論では，$\|$ によってパラメータとハイパーパラメータとを区別する．

す．以下で，2 種の部分空間ベイズ法，すなわち出力パラメータ（式 (3) における B）をハイパーパラメータとみなし，入力パラメータ（A）空間を周辺化する方法（marginalizing input parameter space；MIP）および，入力パラメータをハイパーパラメータとみなし，出力パラメータ空間を周辺化する方法（marginalizing output parameter space；MOP）について解析する．

なお，以下では，部分空間ベイズ法で得られる変数には上線を付ける．例えば汎化誤差は $\bar{G}$，などである．

4. 理 論 解 析

4.1 部分空間ベイズ解

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

$$p(y|x, A\|B) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\|y - BAx\|^2}{2}\right) \qquad (11)$$

と書ける．事前分布として，

$$\phi(A) = \frac{1}{(2\pi)^{HM/2}} \exp\left(-\frac{\mathrm{tr}\left(A^t A\right)}{2}\right) \qquad (12)$$

を用いる．(MOP 法における，$p(y|x, B\|A)$ および $\phi(B)$ も同様に定義できる．) 真の写像 B^*A^* のランクを $H^* \leqq H$ であると仮定する．(注3) 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，中心極限定理により以下が得られる．

$$Q(X^n) = n^{-1}\textstyle\sum_{i=1}^n x_i x_i^t = I_M + O_p(n^{-1/2}), \qquad (13)$$

$$R(X^n, Y^n) = n^{-1}\textstyle\sum_{i=1}^n y_i x_i^t = B^*A^* + O_p(n^{-1/2}). \qquad (14)$$

ここで $\{Q(X^n), R(X^n, Y^n)\}$ は十分統計量であり，それぞれ $M \times M$ 対称行列および $N \times M$ 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．(ただし，$1 \leqq h \leqq H$ 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 (14) より，$H^* < h \leqq H$ に対して γ_h は $O_p(n^{-1/2})$ のオーダーであることがわかり，式 (13) を用いると，

$$\omega_{b_h} RQ^\rho = \omega_{b_h} R + O_p(n^{-1}) \qquad \text{for } H^* < h \leqq H \qquad (15)$$

が成立する．ただし $-\infty < \rho < \infty$ は任意の定数である．

部分空間ベイズ推定量（事後分布に関する期待値）は，次の定理で与えられる．

［定理 1］ MIP 法において $L = M$，MOP 法において $L = N$ とし，また，$L'_h = \max(L, n\gamma_h^2)$ とする．LNN の写像の部分空間ベイズ推定量は以下で与えられる．

$$\hat{B}\hat{A} = \textstyle\sum_{h=1}^H (1 - LL'^{-1}_h)\omega_{b_h}\omega_{b_h}^t RQ^{-1} + O_p(n^{-1}). \qquad (16)$$

(注3)：以下，パラメータの真の値を $*$ をつけて表し，また，パラメータの推定値には $\hat{}$ をつけて表す．

(証明は付録 1. で与える．)

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］ LNN の部分空間ベイズ予測分布は，以下のように書ける．

$$\bar{p}(y|x, X^n, Y^n) = \left((2\pi)^N |\hat{V}|\right)^{-1/2}$$
$$\cdot \exp\left(-(y - \hat{V}\hat{B}\hat{A}x)^t \frac{\hat{V}^{-1}}{2}(y - \hat{V}\hat{B}\hat{A}x)\right) + O_p(n^{-3/2}). \qquad (17)$$

ただし，$\hat{V} = I_N + O_p(n^{-1})$ であり，$|\cdot|$ は行列式を表す．

式 (16) と，最尤推定量

$$\hat{B}\hat{A}_{MLE} = \textstyle\sum_{h=1}^H \omega_{b_h}\omega_{b_h}^t RQ^{-1} \qquad (18)$$

[17] とを比較してみると，部分空間ベイズ推定量の各成分はそれぞれ，James-Stein（JS）型打ち切り推定量となっていることがわかる．(注4) さらに補題 1 より，予測分布として部分空間ベイズ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ことがわかる．よって LNN における，部分空間ベイズ法と JS 型打ち切り推定との漸近等価性が示された．

4.2 汎 化 誤 差

任意の真の写像 B^*A^* に対して，その特異値分解が存在することにより，すべての行ベクトルが互いに直交する行列 A^* および，すべての列ベクトルが互いに直交する行列 B^* が存在する．従って，一般性を失うことなく上の直交性を仮定してよい．よって補題 1 より，n 個の学習サンプル系列を 1 セット得たときのカルバック擬距離 (9) は

$$G(X^n, Y^n) = \left\langle \frac{\|(B^*A^* - \hat{B}\hat{A})x\|^2}{2} \right\rangle_{q(x)} + O_p(n^{-3/2})$$
$$= \textstyle\sum_{h=1}^H G_h(X^n, Y^n) + O_p(n^{-3/2}) \qquad (19)$$

と書ける．ただし，

$$G_h(X^n, Y^n) = \frac{1}{2}\mathrm{tr}\left((b_h^* a_h^{*t} - \hat{b}_h\hat{a}_h^t)^t(b_h^* a_h^{*t} - \hat{b}_h\hat{a}_h^t)\right) \qquad (20)$$

は第 h 成分からの寄与である．ここで，$\mathrm{tr}(\cdot)$ は行列のトレースを表す．自由度 m，スケール行列 Σ，非心度行列 Λ である d 次元ウィシャート分布を $\mathcal{W}_d(m, \Sigma, \Lambda)$ と書く．また，非心度行列が零行列のとき，これを省略して $\mathcal{W}_d(m, \Sigma)$ と書く．

［定理 2］ 部分空間ベイズ法における，LNN の汎化誤差は

$$G(n) = \lambda n^{-1} + O(n^{-3/2})$$

と漸近展開される．ただし，主要項の係数（本論では，これを汎化係数と呼ぶ）は

(注4)：パラメータ w の JS 型打ち切り推定量 [12], [13], [16] は，

$$\hat{w}_{PJS} = \theta(\|\hat{w}_{MLE}\|^2 > \chi/n)(1 - \chi/n\|\hat{w}_{MLE}\|^2)\hat{w}_{MLE}$$
$$= \left(1 - \chi\chi'^{-1}\right)\hat{w}_{MLE}$$

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，$\chi > 0$ は定数であり，$\chi' = \max(\chi, n\|\hat{\mu}_{MLE}\|^2)$ である．

$$2\lambda = (H^*(M+N) - H^{*2}) + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > L) \left(1 - \frac{L}{\gamma_h'^2}\right)^2 \gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})} \quad (21)$$

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，$\langle\cdot\rangle_{q(\{\gamma_h'^2\})}$ はこのウィシャート分布に関する期待値を示す．

(略証) 真の分布を実現するために必要な成分は識別可能であるため，これらの成分の寄与は正則な場合と同じとなり，式 (21) の第 1 項を得る．冗長な成分は，$n^{-1/2}R'$ の大きい方から $(H-H^*)$ 個の特異値成分に倣う．ただし，R' は各成分が $\mathcal{N}_1(0,1)$ に従う $(N-H^*)\times(M-H^*)$ 確率行列である．(よって $R'R'^t$ は $\mathcal{W}_{N-H^*}(M-H^*, I_{N-H^*})$ に従う．) 定理 1 を用いることにより，式 (21) の第 2 項を得る．(証明終)

4.3 高次元極限近似

[3] における最尤法の解析と同様の方法で，M，N，H および H^* が同じオーダーで無限大となる極限（高次元極限）において，式 (21) の第 2 項は解析的に計算される．$\alpha = N'/M' = (N-H^*)/(M-H^*)$，$\beta = H'/N' = (H-H^*)/(N-H^*)$ および $\kappa = L/M' = L/(M-H^*)$ とする．W を $\mathcal{W}_{N'}(M', I_{N'})$ に従う確率行列とし，$\{u_1, \ldots, u_{N'}\}$ を行列 $M'^{-1}W$ の固有値とする．固有値の経験分布の測度を

$$\delta P = N'^{-1} \{\delta(u_1) + \delta(u_2) + \cdots + \delta(u_{N'})\} \quad (22)$$

で定義する．ただし，$\delta(u)$ は u におけるディラック測度である．測度 (22) は，測度の集合の上に値を取る確率変数であるが，$N' \to \infty$ のとき，確率変数として殆どいたるところ，測度としての位相で

$$p(u)du = \frac{\sqrt{(u-u_m)(u_M-u)}}{2\pi\alpha u}\theta(u_m < u < u_M)du \quad (23)$$

に収束する．ただし，$u_m = (\sqrt{\alpha}-1)^2$ および $u_M = (\sqrt{\alpha}+1)^2$ である [18]．分布 (23) のモーメントを計算することにより，以下の定理を得る．

［定理 3］ 高次元極限における LNN の汎化係数は，

$$2\lambda \sim (H^*(M+N) - H^{*2}) + \frac{(M-H^*)(N-H^*)}{2\pi\alpha} \left\{J(s_t;1) - 2\kappa J(s_t;0) + \kappa^2 J(s_t;-1)\right\} \quad (24)$$

で与えられる．ただし，

$$J(s;1) = 2\alpha(-s\sqrt{1-s^2} + \cos^{-1} s),$$

$$J(s;0) = -2\sqrt{\alpha}\sqrt{1-s^2} + (1+\alpha)\cos^{-1} s - (1-\alpha)\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s + 2\alpha}{2\alpha s + \sqrt{\alpha}(1+\alpha)},$$

$$J(s;-1) = \begin{cases} 2\sqrt{\alpha}\frac{\sqrt{1-s^2}}{2\sqrt{\alpha}s+1+\alpha} - \cos^{-1} s + \frac{1+\alpha}{1-\alpha}\cos^{-1}\frac{\sqrt{\alpha}(1+\alpha)s+2\alpha}{2\alpha s+\sqrt{\alpha}(1+\alpha)} & (0<\alpha<1) \\ 2\sqrt{\frac{1-s}{1+s}} - \cos^{-1} s & (\alpha = 1) \end{cases},$$

および $s_t = \max\left((\kappa - (1+\alpha))/2\sqrt{\alpha}, J^{-1}(2\pi\alpha\beta; 0)\right)$ である．ここで，$J^{-1}(\cdot;k)$ は $J(s;k)$ の逆関数を表す．

4.4 「デリケートな」状況

通常の漸近解析においては，真のモデルのすべての成分の振幅はそれぞれ，零であるか，あるいははっきりと非零である（定数オーダー）かのどちらかであると仮定される．定理 2 も，そのような状況においてのみ成立する．しかし，真の写像 B^*A^* が，微小だが無視できないオーダーの（すなわち $\gamma_h^* \sim O(n^{-1/2})$ なる）特異値を持つ場合を考えることは，モデル選択や検定問題を考える上で非常に重要である．特に，学習法の優越性を議論する場合には必須である．

定理 1 は，式 (16) の第 2 項を $o_p(n^{-1/2})$ とすれば，この場合にも成立する．H^* を，はっきりと非零である（すなわち $\gamma_h^{*-1} = o(\sqrt{n})$ なる）特異値の数であると定義しなおす．一般性を失わずに，B^*A^* は非負の一般対角行列であり，その対角成分は左上から右下へ大きい順に並んでいると仮定する．R''^* を，B^*A^* の右下 $(N-H^*)\times(M-H^*)$ 行列とし，R'' を $\mathcal{W}_{N-H^*}(M-H^*, I_{N-H^*}, nR''^*R''^*)$ に従う確率行列とする．$H^* < h \leqq H$ なる成分は，行列 $n^{-1/2}R''$ の大きい方から $(H-H^*)$ 個の特異値成分に倣うことから，以下の定理が得られる．

［定理 4］ B^*A^* が $0 < \sqrt{n}\gamma_h^* < \infty$ なる特異値を持つ場合，LNN の汎化係数は

$$2\lambda = (H^*(M+N) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L) \left\{\left(1 - \frac{L}{\gamma_h''^2}\right)^2 \gamma_h''^2 - 2\left(1 - \frac{L}{\gamma_h''^2}\right)\gamma_h'' \omega_{b_h}''^t \sqrt{n} R'^* \omega_{a_h}''\right\}\right\rangle_{q(R'')} \quad (25)$$

で与えられる．ただし，γ_h''，ω_{a_h}'' および ω_{b_h}'' はそれぞれ，R'' の h 番目に大きい特異値，その右および左特異ベクトルである．また，$\langle\cdot\rangle_{q(R'')}$ は非心ウィシャート分布に関する期待値を表す．

5. 考　　察

5.1 最尤法およびベイズ法との比較

図 1 は，入力素子数 $M = 50$，出力素子数 $N = 30$，中間素子数 $H = 20$ の LNN の汎化係数を示している．横軸は真の中間素子数（ランク）H^* であり，縦軸は汎化係数を式 (4) で与えられるパラメータ次元 K で規格化したものである．最尤（ML）法の汎化係数は [3]，ベイズ（Bayes）法の汎化係数は [6] で，それぞれ解明されたものを図示しており，また，正則（Regular）な場合の値も示している．(注5) 図 1 の結果は，高次元極限近似（定理 3）によって得られたものであるが，ウィシャート分布に従う確率行列を生成し，定理 2 を用いて数値的にも計算を行ってみたが，この場合，ほとんど見分けが付かないくらい両者の結果は一致した．図 1 から，部分空間ベイズ（SB）法がベイズ法と同等の汎化性能を示すことがわかる．特に MIP 法は，

(注5)：正則モデルにおいては，最尤法およびベイズ法のいずれを用いても，汎化係数は常に K に等しい．このことは，赤池情報量規準（AIC）の理論的根拠のひとつである [19].

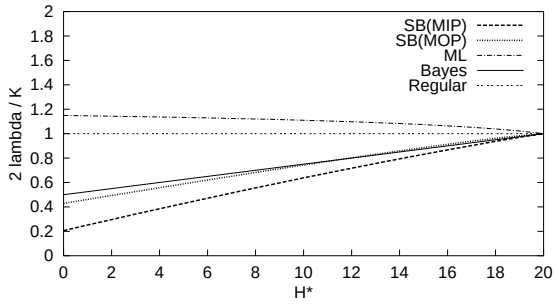

図 1　汎 化 誤 差

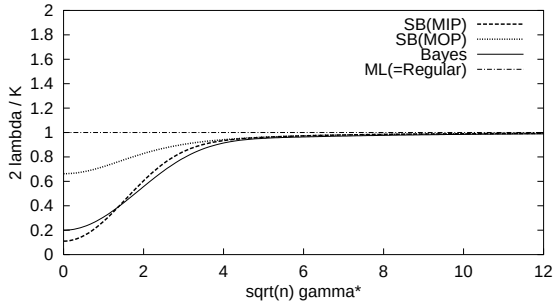

図 2 「デリケートな」状況

すべての H^* の値において，ベイズ法より汎化誤差が小さい．ただし，このことはベイズ法に対する MIP 法の優越性を意味するものではないことに注意する．優越性を議論するには，4. 4 で述べた「デリケートな」状況を考慮せねばならない．

定理 4 を用いれば,「デリケートな」状況における汎化係数の計算が可能となる．しかし,「デリケートな」状況におけるベイズ法の汎化係数は，単出力（single-output；SO）LNN の場合にのみ結果が得られているので [20]，ここではこの場合について比較する．(すなわち $N=H=1$ の場合.)(注6) 図 2 は，$M=5$ の場合の SOLNN の汎化係数を示す．真の特異値の $\sqrt{n}$ 倍が横軸に示されている．図 2 より，部分空間ベイズ法が，特異点のエントロピー（状態密度）が過学習を抑制するという，ベイズ法に似た性質を持つことがわかる．また，真の特異値の値によっては，MIP 法の汎化係数がベイズ法のそれを超える場合があるため，MIP 法がベイズ法を優越しないことがわかる．部分空間ベイズ法は，ベイズ法と同等か，場合によってはより強い特異点の過学習抑制効果を持つと結論することができる．

部分空間ベイズ法は，MCMC 法などと比較してはるかに少ない計算量で実行できるので，この方法によってベイズ法と同等な汎化性能が常に得られるのであれば，非常に都合が良い．しかし残念ながら，部分空間ベイズ法は，最尤法に似た性質も併せ持っている．この性質とは，定理 2 からもわかるように，ノイズからなるランダムな確率行列の特異値のうち，大きい方を選択することにより，過学習が促進されるという性質である．最尤推定量が，サポートがコンパクトでない分布の極値統計量となるため，例えば $(M-H^*)>(N-H^*)\gg(H-H^*)$ である場合，$\mathcal{W}_{N-H^*}(M-H^*, I_{N-H^*})$ に従う確率行列の，大きい方から $(H-H^*)$ 個の特異値は，L と比較して非常に大きい値をとる．従って，定理 2 に現れる固有値 $\{\gamma_h^2\}$ は,「縮小」が有効となる範囲から十分離れてしまい，その結果，殆ど最尤法と同じ結果が得られることになる．実際，ベイズ法においては，特異モデルの汎化誤差が正則な場合を超えることはないということが証明されているのであるが [5]，部分空間ベイズ法においては，例えば $M=N=80, H=1, H^*=0$ の場合には汎化係数は $2\lambda/K\sim1.04$ となり，正則な場合より大きくなりうる．

(注6)：SOLNN は，最尤法の観点では正則モデルとして取り扱われる．なぜなら，$b_1a_1 \mapsto w \in \mathbb{R}^M$ により，普通の線形モデルに変換され，その汎化係数は正則な場合と同じになるためである．しかし，ベイズ法の観点ではこれも特異モデルとしての特徴を持ち，図 2 に見られるように，正則な場合とは異なる汎化係数を持つ．

5. 2　縮小推定量，経験ベイズ法および変分ベイズ法の関係

線形（正則）モデルにおいて，式 (10) で示されるような事前分布を用い，τ_1^{-2} を経験ベイズ法に基づいて決定すると，JS 推定量が得られることが知られている [16]．SOLNN においては，$b_1a_1 \mapsto w \in \mathbb{R}^M$ によってモデルを線形に，かつ，事前分布を式 (10) の形に変換することができるので，MIP 法において b_1 が τ_1 と同じ役割を果たすといえる．より一般に，3 層の特異モデルにおいて，ある階層のパラメータがハイパーパラメータとみなされたとき，それは，対をなすパラメータの事前分布の標準偏差ハイパーパラメータと似た働きをすると考えられるため，特異モデルにおける部分空間ベイズ法と，縮小推定との類似は自然であると言える．

LNN における，変分ベイズ法の解析も行われた [11]．実はこれによると，入力パラメータ空間と出力パラメータ空間のうち，次元の小さい方をハイパーパラメータとみなす部分空間ベイズ法と，変分ベイズ法とから，漸近的に同じ解（予測分布）が得られることがわかった．これらの一致も，両者の事後分布を比較してみると，自然であることがわかる．例えば $M>N$ である場合，MIP 法では付録 1. の式 (A·2) および (A·9) より，冗長な成分に対応する事後分布が，入力パラメータ空間にオーダー 1 で広がることがわかるが（出力パラメータ空間に関しては 1 点に決めるので，デルタ関数となる），変分ベイズ法においても，冗長な成分に対応する事後分布は，入力パラメータ空間にオーダー 1 で広がる（変分ベイズ法では，出力パラメータ空間にはオーダー n^{-1} の分散で収束する）．

5. 3　今後の課題

非線形な神経回路網における，部分空間ベイズ法および変分ベイズ法の汎化性能を調べることが，今後の課題である．活性化関数の非線形性が基底選択の自由度を高め，それによって汎化誤差が増大することが予想される．

6. 結　　論

モデルのパラメータの一部をハイパーパラメータとみなして経験ベイズ法を適用するという，部分空間ベイズ法を紹介し，その解を線形神経回路網において導出した．汎化性能を解明した結果，多くの場合，部分空間ベイズ法がベイズ法に匹敵する汎化性能を示すことが明らかになった．

付　　録

1.　定理 1 の証明

はじめに MIP 法の場合を考える．特異値分解により，任意の写像

BA を，すべての行ベクトルが互いに直交する行列 A と，すべての列ベクトルが互いに直交する行列 B とに分解することができる．さらに，この場合に事前分布 (12) が最大になることも容易にわかる．従って一般性を失うことなく，B の（周辺尤度を最大化する）最適値は，互いに直交する列ベクトルからなると仮定してよい．すると，周辺尤度と事後分布は以下のように分解できる．

$$Z(Y^n|X^n\|B) = \prod_{h=1}^{H} Z(Y^n|X^n\|b_h),$$
$$p(A|X^n, Y^n\|B) = \prod_{h=1}^{H} p(a_h|X^n, Y^n\|b_h).$$

式 (11) および (12) を，式 (5) および (6) に代入することにより，成分ごとの周辺尤度と事後分布が以下のように得られる．

$$Z(Y^n|X^n\|b_h) \propto |S_h|^{-1/2} \exp\left(\frac{nb_h^t R S_h^{-1} R^t b_h}{2}\right), \qquad \text{(A·1)}$$

$$p(a_h|X^n, Y^n\|b_h) \propto \exp\left(-\left(a_h - S_h^{-1}R^t b_h\right)^t \frac{nS_h}{2}\left(a_h - S_h^{-1}R^t b_h\right)\right). \qquad \text{(A·2)}$$

ただし，$S_h = (\|b_h\|^2 Q + n^{-1} I_M)$ である．$F(Y^n|X^n\|b_h)$ を，確率的複雑さ（対数周辺尤度の符号反転）の第 h 成分からの寄与であるとし，また，$F'(Y^n|X^n\|b_h) = F(Y^n|X^n\|b_h) + \text{const.}$ とすると，

$$\begin{aligned} 2F'(Y^n|X^n\|b_h) &= -2\log Z(Y^n|X^n\|b_h) + \text{const.} \\ &= \log|S_h| - nb_h^t R S_h^{-1} R^t b_h \end{aligned} \qquad \text{(A·3)}$$

と書ける．以後 $F'(Y^n|X^n\|b_h)$ を $F'(b_h)$ と書く．以下で，真の分布を実現するために必要な成分と，冗長な成分とに分けて，式 (A·3) を最小化する．

必要成分（$h \leqq H^*$）に対応する $RQ^{-1/2}$ の特異値は，$O_p(1)$ のオーダーである．よって式 (A·3) は

$$\begin{aligned} 2F'(b_h) = M\log\|b_h\|^2 &- n\|b_h\|^{-2} b_h^t R Q^{-1} R^t b_h \\ &+ \|b_h\|^{-4} b_h^t R Q^{-2} R^t b_h + O_p(n^{-1}) \end{aligned} \qquad \text{(A·4)}$$

と書ける．式 (A·4) を最小化する際，b_h の方向余弦は主要項である第 2 項が決定し，その値は $\hat{b}_h = \|\hat{b}_h\|(\omega_{b_h} + O_p(n^{-1}))$ となる．第 2 項は b_h のノルムに依存しないので，ノルムは第 1 項と第 3 項によって決められる．よって，ハイパーパラメータの最適値は以下となる．

$$\hat{b}_h = \sqrt{\frac{\omega_{b_h}^t R Q^{-2} R^t \omega_{b_h}}{M}}\omega_{b_h} + O_p(n^{-1}). \qquad \text{(A·5)}$$

a_h の事後分布 (A·2) に関する期待値は $\hat{a}_h = S_h^{-1} R^t b_h$ であるので，（写像の）必要成分の部分空間ベイズ推定量が以下のように得られる．

$$\hat{b}_h \hat{a}_h^t = \omega_{b_h}\omega_{b_h}^t R Q^{-1} + O_p(n^{-1}). \qquad \text{(A·6)}$$

一方，冗長成分（$h > H^*$）に関しては，式 (15) により式 (A·3) を以下のように近似できる．

$$2F'(b_h) = M\log\left(\|b_h\|^2 + n^{-1}\right) - \frac{nb_h^t R R^t b_h}{\|b_h\|^2 + n^{-1}} + O_p(n^{-1/2}). \qquad \text{(A·7)}$$

すると，b_h の方向余弦は式 (A·7) の第 2 項によって決定され，その値は $O_p(n^{-1/2})$ の精度で ω_{b_h} に等しいことがわかる．従って，$b_h^t R R^t b_h = \gamma_h^2 \|b_h\|^2 (1 + O_p(n^{-1/2}))$ が成り立つ．これを用いた上で，式 (A·7) を b_h のノルムで偏微分すると

$$\begin{aligned} 0 = 2\frac{\partial F'(b_h)}{\partial \|b_h\|^2} = \frac{M}{(\|b_h\|^2 + n^{-1})^2}&\left(\|b_h\|^2 - \frac{n\gamma_h^2 - M}{nM}\right) \\ &+ O_p(\|b_h\|^{-2} n^{-1/2}) \end{aligned} \qquad \text{(A·8)}$$

が得られるが，これにより，$\gamma_h < \sqrt{M/n}$ の場合には，式 (A·7) が $\|b_h\|$ に関して単調増加であることがわかる．よってハイパーパラメータの最適値は

$$\hat{b}_h = \sqrt{\frac{L'_h - M}{nM}}\omega_{b_h} + O_p(n^{-1}) \qquad \text{(A·9)}$$

となり，冗長成分の部分空間ベイズ推定量が以下のように得られる．

$$\hat{b}_h \hat{a}_h^t = (1 - ML_h'^{-1})\omega_{b_h}\omega_{b_h}^t R + O_p(n^{-1}) \qquad \text{(A·10)}$$

式 (A·3) は，特異値を大きい方から順に選んだとき最小となるので，式 (A·6)，(A·10)，(15) および，必要成分に対して $ML_h'^{-1} = O_p(n^{-1})$ が成り立つことを用いて，定理 1 の部分空間ベイズ推定量を得る．MOP 法の場合も同様にして証明できる．(証明終)

文　献

[1] S. Nakajima and S. Watanabe: "Generalization Error of Linear Neural Networks in an Empirical Bayes Approach", Proc. of IJCAI, Edinburgh, U.K., pp. 804–810 (2005).

[2] 福水, 栗木, 竹内, 赤平："特異モデルの統計学", 岩波書店, 東京 (2004).

[3] K. Fukumizu: "Generalization Error of Linear Neural Networks in Unidentifiable Cases", Proc. of ALT, Springer, pp. 51–62 (1999).

[4] 渡辺："データ学習アルゴリズム", 共立出版, 東京 (2001).

[5] S. Watanabe: "Algebraic Information Geometry for Learning Machines with Singularities", Advances in NIPS, Vol. 13, pp. 329–336 (2001).

[6] M. Aoyagi and S. Watanabe: "The Generalization Error of Reduced Rank Regression in Bayesian Estimation", Proc. of ISITA, Parma, Italy, pp. 1068–1073 (2004).

[7] G. E. Hinton and D. van Camp: "Keeping Neural Networks Simple by Minimizing the Description Length of the Weights", Proc. of COLT (1993).

[8] D. J. C. MacKay: "Developments in Probabilistic Modeling with Neural Networks—Ensemble Learning", Proc. of the 3rd Ann. Symp. on Neural Networks, pp. 191–198 (1995).

[9] H. Attias: "Inferring Parameters and Structure of Latent Variable Models by Variational Bayes", Proc. of UAI (1999).

[10] Z. Ghahramani and M. J. Beal: "Graphical Models and Variational Methods", Advanced Mean Field Methods, MIT Press (2000).

[11] 中島, 渡辺："線形神経回路網における変分ベイズ法の汎化特性", 情報論的学習理論ワークショップ (IBIS2005) に投稿中 (2005).

[12] W. James and C. Stein: "Estimation with Quadratic Loss", Proc. of the 4th Berkeley Symp. on Math. Stat. and Prob., pp. 361–379 (1961).

[13] 久保川："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).

[14] H. Akaike: "Likelihood and Bayes Procedure", J. M. Bernald, Bayesian statistics, University Press, pp. 143–166 (1980).

[15] D. J. C. MacKay: "Bayesian Interpolation", Neural Computation, **4**, 2, pp. 415–447 (1992).

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

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

[18] K. W. Watcher: "The Strong Limits of Random Matrix Spectra for Sample Matrices of Independent Elements", Ann. Prob., **6**, pp. 1–18 (1978).

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

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