IEICE TRANS. INF. & SYST., VOL.E89–D, NO.3 MARCH 2006

PAPER

Generalization Performance of Subspace Bayes Approach in Linear Neural Networks

Shinichi NAKAJIMA[†,††a)], ***Student Member and*** **Sumio WATANABE**[†b)], ***Member***

SUMMARY In unidentifiable models, the Bayes estimation has the advantage of generalization performance 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. A subspace Bayes approach is an empirical Bayes approach where a part of the parameters are regarded as hyperparameters. Consequently, in some three-layer models, this approach requires much less computational costs than Markov chain Monte Carlo methods. We show that, in three-layer linear neural networks, a subspace Bayes approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation, and theoretically clarify its generalization error and training error. We also discuss the domination over the maximum likelihood estimation and the relation to the variational Bayes approach.

key words: *empirical Bayes, variational Bayes, neural networks, reduced-rank regression, James-Stein, unidentifiable*

1. Introduction

Unidentifiable parametric models, such as neural networks, mixture models, hidden Markov models, Bayesian networks, and so on, have a wide range of applications. These models have singularities in the parameter space, on which the Fisher information matrix degenerates hence the log-likelihood cannot be approximated by any quadratic form of the parameter. Therefore, neither the distribution of the maximum likelihood estimator nor the Bayes posterior distribution asymptotically converges to the normal distribution, which prevents the conventional learning theory of the regular statistical models to hold [2]–[5]. Accordingly, statistical model selection methods such as Akaike's information criterion (AIC) [6], Bayesian information criterion (BIC) [7], and the minimum description length criterion (MDL) [8] have no theoretical foundation in unidentifiable models.

Some properties of learning in unidentifiable models have been theoretically clarified. In the maximum likelihood (ML) estimation, which is asymptotically equivalent to the maximum a posterior (MAP) estimation, the asymptotic behavior of the log-likelihood ratio in some unidentifiable models was analyzed [2], [9]–[12], facilitated by the idea of the locally conic parameterization [13]. It has, thus, been known that the ML estimation, in general, provides poor generalization performance, and in the worst cases, the ML estimator diverges. In linear neural networks, on which we focus in this paper, the generalization error was clarified and proved to be greater than that of the regular models whose dimension of the parameter space is the same when the model is redundant to learn the true distribution [14], although the ML estimator is in a finite region [15].

On the other hand, for analysis of generalization performance of the Bayes estimation in unidentifiable models, an algebraic geometrical method was developed, by which the asymptotic behavior of the generalization error or its upper bound in some unidentifiable models was clarified and proved to be less than that of the regular models [16]–[21], and moreover, the generalization error of any model having singularities was proved to be less than that of the regular models when we use a prior distribution having positive values on the singularities [22].

According to the previous works above, it can be said that, in unidentifiable models, the Bayes estimation provides better generalization performance than the ML estimation. However, the Bayes posterior distribution can seldom be exactly realized. Furthermore, Markov chain Monte Carlo (MCMC) methods, often used for approximation of the posterior distribution, require huge computational costs. As an alternative, the variational Bayes approach, where the correlation between parameters and the other parameters, or the correlation between the parameters and the hidden variables is neglected, was proposed [23]–[26]. We have just derived the variational Bayes solution of linear neural networks and clarified its generalization error and training error [27].

In this paper, we consider another alternative, which we call a subspace Bayes (SB) approach. An SB approach is an empirical Bayes (EB) approach where a part of the parameters of a model are regarded as hyperparameters. If we regard the parameters of one layer as hyperparameters, we can analytically calculate the marginal likelihood in some three-layer models. Consequently, what we have to do is only to find the hyperparameter value maximizing the marginal likelihood. The computational costs of the SB approach is thus much less than that of posterior distribution approximation by MCMC methods.

At first in this paper, we prove that, in three-layer linear neural networks, an SB approach is equivalent to a positive-part James-Stein type shrinkage estimation [28]. Then, we clarify its generalization error and training error, also con-

Manuscript received May 10, 2005.
Manuscript revised August 24, 2005.

[†]The authors are with Tokyo Institute of Technology, Yokohama-shi, 226–8503 Japan.
[††]The author is with Nikon Corporation, Kumagaya-shi, 360–8559 Japan.
a) E-mail: nakajima.s@cs.pi.titech.ac.jp
b) E-mail: swatanab@pi.titech.ac.jp

DOI: 10.1093/ietisy/e89–d.3.1128

sidering *delicate* situations, the most important situations in model selection problems and in statistical tests, when the Kullback-Leibler divergence of the true distribution from the singularities is comparable to the inverse of the number of training samples [29]. We thus conclude that the SB approach provides as good performance as the Bayes estimation in typical cases.

In Sect. 2, neural networks and linear neural networks are briefly introduced. The framework of the Bayes estimation, that of the EB approach, and that of the SB approach are described in Sect. 3. The significance of singularities for generalization performance and the importance of analysis of *delicate* situations are explained in Sect. 4. The SB solution and its generalization error, as well as training error, are derived in Sect. 5. Discussion and conclusions follow in Sect. 6 and in Sect. 7, respectively.

2. Linear Neural Networks

Let $x \in \mathbb{R}^M$ be an input (column) vector, $y \in \mathbb{R}^N$ an output vector, and w a parameter vector. A neural network model can be described as a parametric family of maps $\{f(\cdot; w) : \mathbb{R}^M \mapsto \mathbb{R}^N\}$. A three-layer neural network with H hidden units is defined by

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

where $w = \{(a_h, b_h) \in \mathbb{R}^M \times \mathbb{R}^N; h = 1, \ldots, H\}$ summarizes all the parameters, $\psi(\cdot)$ is an activation function, which is usually a bounded, non-decreasing, antisymmetric, nonlinear function like $\tanh(\cdot)$, and t denotes the transpose of a matrix or vector. Assume that the output is observed with a noise subject to $\mathcal{N}_N(0, \sigma^2 I_N)$, where $\mathcal{N}_d(\mu, \Sigma)$ denotes the d-dimensional normal distribution with average vector μ and covariance matrix Σ, and I_d denotes the $d \times d$ identity matrix. Then, the conditional distribution is given by

$$p(y|x, w) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{\|y - f(x; w)\|^2}{2\sigma^2}\right). \tag{2}$$

In this paper, we focus on linear neural networks, whose activation function is linear, as the simplest multilayer models.[†] A linear neural network model (LNN) is defined by

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

where $A = (a_1, \ldots, a_H)^t$ is an $H \times M$ input parameter matrix and $B = (b_1, \ldots, b_H)$ is an $N \times H$ output parameter matrix. Because the transform $(A, B) \mapsto (TA, BT^{-1})$ does not change the map for any non-singular $H \times H$ matrix T, the parameterization in Eq. (3) has trivial redundancy. Accordingly, the essential dimension of the parameter space is

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

We assume that $H \leq N \leq M$ throughout this paper.

3. Framework of Learning Methods

3.1 Bayes Estimation

Let $X^n = \{x_1, \ldots, x_n\}$ and $Y^n = \{y_1, \ldots, y_n\}$ be arbitrary n training samples independently and identically taken from the true distribution $q(x, y) = q(x)q(y|x)$. The marginal conditional likelihood of a model $p(y|x, w)$ is given by

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

where $\phi(w)$ is the prior distribution. The posterior distribution is given by

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

and the predictive distribution is defined as the average of the model over the posterior distribution as follows:

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

The generalization error, a criterion of generalization performance, and the training error are defined by

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

$$T(n) = \langle T(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{9}$$

respectively, where

$$G(X^n, Y^n) = \int q(x) q(y|x) \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} dxdy \tag{10}$$

is the Kullback-Leibler (KL) divergence of the predictive distribution from the true distribution,

$$T(X^n, Y^n) = n^{-1} \sum_{i=1}^{n} \log \frac{q(y|x)}{p(y|x, X^n, Y^n)} \tag{11}$$

is the empirical KL divergence, and $\langle \cdot \rangle_{q(X^n, Y^n)}$ denotes the expectation value over all sets of n training samples.

3.2 Empirical Bayes Approach and Subspace Bayes Approach

We often have little information about the prior distribution, with which an EB approach was originally proposed to cope. We can introduce hyperparameters in the prior distribution; for example, when we use a prior distribution that depends on a hyperparameter τ_1 such as

$$\phi(w\|\tau_1) = \frac{1}{(2\pi\tau_1^2)^{K/2}} \exp\left(-\frac{\|w\|^2}{2\tau_1^2}\right), \tag{12}$$

[†]A linear neural network model, also known as a reduced-rank regression model, is not a toy but an useful model in many applications [30].

the marginal likelihood, Eq.(5), also depends on τ_1.[†] In an EB approach, τ_1 is estimated by maximizing the marginal likelihood or by a slightly different way [31]–[33].

Extending the idea above, we can introduce hyperparameters also in a model distribution. What we call an SB approach is an EB approach where a part of the parameters of a model are regarded as hyperparameters. In the SB approach, we first separate the *whole* parameter w of an original model $p(y|x, w)$ into the parameter $\bar{w}$ and the hyperparameter τ, i.e., $w = \{\bar{w}, \tau\}$. Then we have the model distribution $p(y|x, \bar{w}\|\tau)$. Using the prior distribution $\phi(\bar{w})$, we get the marginal likelihood as follows:

$$Z(Y^n|X^n\|\tau) = \int \phi(\bar{w}) \prod_{i=1}^n p(y_i|x_i, \bar{w}\|\tau) d\bar{w}. \tag{13}$$

We estimate the hyperparameter value by maximizing Eq. (13), i.e.,

$$\hat{\tau} = \operatorname*{argmax}_{\tau} Z(Y^n|X^n\|\tau). \tag{14}$$

Then, the SB posterior distribution is given by

$$p(\bar{w}|X^n, Y^n\|\hat{\tau}) = \frac{\phi(\bar{w}) \prod_{i=1}^n p(y_i|x_i, \bar{w}\|\hat{\tau})}{Z(Y^n|X^n\|\hat{\tau})}. \tag{15}$$

We denote by a *hat* an estimator of a parameter or hyperparameter, and define the SB estimator of a hyperparameter as the optimal value maximizing the marginal likelihood, as in Eq. (14), and the SB estimator of a parameter as the expectation value over the SB posterior distribution, Eq.(15). The SB predictive distribution, the SB generalization error, and the SB training error are respectively given by:

$$\bar{p}(y|x, X^n, Y^n) = \int p(y|x, \bar{w}\|\hat{\tau}) p(\bar{w}|X^n, Y^n\|\hat{\tau}) d\bar{w}, \tag{16}$$

$$\bar{G}(n) = \langle \bar{G}(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{17}$$

$$\bar{T}(n) = \langle \bar{T}(X^n, Y^n) \rangle_{q(X^n, Y^n)}, \tag{18}$$

where

$$\bar{G}(X^n, Y^n) = \int q(x) q(y|x) \log \frac{q(y|x)}{\bar{p}(y|x, X^n, Y^n)} dxdy, \tag{19}$$

$$\bar{T}(X^n, Y^n) = n^{-1} \sum_{i=1}^n \log \frac{q(y|x)}{\bar{p}(y|x, X^n, Y^n)}. \tag{20}$$

In the following sections, we analyze two versions of SB approach: in the first one, we regard the output parameter matrix B of the map, Eq.(3), as a hyperparameter and then marginalize the likelihood in the input parameter space (MIP); and in the other one, we regard the input parameter matrix A, instead of B, as a hyperparameter and then marginalize in the output parameter space (MOP).

4. Unidentifiability and Singularities

We say that a parametric model is unidentifiable if the map from the parameter to the probability distribution is not one-to-one. A neural network model, Eq.(1), is unidentifiable because the model is independent of a_h when $b_h = 0$, or vice versa. The continuous points denoting the same distribution are called the singularities, because the Fisher information matrix on them degenerates. The shadowed locations in Fig. 1 indicate the singularities. We can see in Fig. 1 that the model denoted by the singularities has more neighborhoods and state density than any other model denoted by only one point each. When the true model is not on the singularities, they asymptotically do not affect prediction, and therefore, the conventional learning theory of the regular models holds. On the other hand, when the true model is on the singularities, they significantly affect generalization performance as follows: in the ML estimation, the increase of the neighborhoods of the true distribution leads to the increase of the flexibility of imitating noises, and therefore, accelerates overfitting; while in the Bayes estimation, the large state density of the true distribution increases its weight, and therefore, suppresses overfitting. In LNNs, the former property appears as acceleration of overfitting by selection of the largest singular value components of a random matrix, and in the SB approaches of LNNs, the latter property appears as James-Stein type shrinkage, as shown in the following sections.

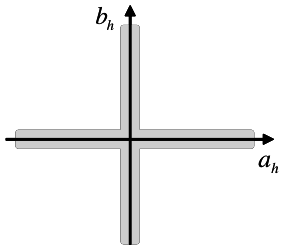


Fig. 1 Singularities of a neural network model.

Suppression of overfitting accompanies insensitivity to the true components with small amplitude. There is a trade-off, which would, however, be ignored in asymptotic analysis if we would consider only situations when the true model is *distinctly* on the singularities or not. Therefore, in this paper, we also consider *delicate* situations when the KL divergence of the true distribution from the singularities is comparable to the inverse of the number of training samples, n^{-1}, which are important situations in model selection problems and in statistical tests with finite number of samples for the following reasons: first, that there naturally exist a few true components with amplitude comparable to $n^{-1/2}$ when neither the smallest nor the largest model is selected; and secondly, that whether the selected model involves such components essentially affects generalization performance.

5. Theoretical Analysis

5.1 Subspace Bayes Solution

Assume that the variance of a noise is known and equal to

[†]By $\|$ we distinguish the hyperparameter from the parameter in this paper.

unity. Then the conditional distribution of an LNN in the MIP version of SB approach is given by

$$p(y|x, A\|B) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{\|y - BAx\|^2}{2}\right). \quad (21)$$

We use the following prior distribution:

$$\phi(A) = \frac{1}{(2\pi)^{HM/2}} \exp\left(-\frac{\mathrm{tr}\,(A^t A)}{2}\right). \quad (22)$$

Note that we can similarly prepare $p(y|x, B\|A)$ and $\phi(B)$ for the MOP version. We denote by $*$ the true value of a parameter, and assume that the true conditional distribution is $p(y|x, A^*\|B^*)$, where B^*A^* is the true map with rank $H^* \leq H$. For simplicity, we assume that the input vector is orthonormalized so that $\int xx^t q(x)dx = I_M$. Consequently, the central limit theorem leads to the following two equations:

$$Q(X^n) = n^{-1}\sum_{i=1}^n x_i x_i^t = I_M + O_p(n^{-1/2}), \quad (23)$$

$$R(X^n, Y^n) = n^{-1}\sum_{i=1}^n y_i x_i^t = B^*A^* + O_p(n^{-1/2}), \quad (24)$$

where $Q(X^n)$ is an $M \times M$ symmetric matrix and $R(X^n, Y^n)$ is an $N \times M$ matrix. Hereafter, we abbreviate $Q(X^n)$ as Q, and $R(X^n, Y^n)$ as R.

Let γ_h be the h-th largest singular value of the matrix $RQ^{-1/2}$, ω_{a_h} the corresponding right singular vector, and ω_{b_h} the corresponding left singular vector, where $1 \leq h \leq H$. We find from Eq. (24) that, in the asymptotic limit, the singular values corresponding to the necessary components to realize the true distribution converge to finite values, while the other ones corresponding to the redundant components converge to zero. Therefore, with probability 1, the largest H^* singular values correspond to the necessary components, and the others correspond to the redundant components. Combining Eqs. (23) and (24), we have

$$\omega_{b_h} RQ^\rho = \omega_{b_h} R + O_p(n^{-1}) \qquad \text{for } H^* < h \leq H, \quad (25)$$

where $-\infty < \rho < \infty$ is an arbitrary constant. The SB estimator is given by the following theorem:

Theorem 1: Let $L = M$ in the MIP version or $L = N$ in the MOP version, and $L'_h = \max(L, n\gamma_h^2)$. The SB estimator of the map of an LNN is given by

$$\hat{B}\hat{A} = \sum_{h=1}^H (1 - LL'^{-1}_h)\omega_{b_h}\omega_{b_h}^t RQ^{-1} + O_p(n^{-1}). \quad (26)$$

(The proof is given in Appendix.)

The following lemma, which states the localization of the SB posterior distribution of the map BA, also holds:

Lemma 1: The predictive distribution in the SB approaches can be written as follows:

$$\bar{p}(y|x, X^n, Y^n) = \left((2\pi)^N |\hat{V}|\right)^{-1/2} \cdot \exp\left(-(y - \hat{V}\hat{B}\hat{A}x)^t \frac{\hat{V}^{-1}}{2}(y - \hat{V}\hat{B}\hat{A}x)\right) + O_p(n^{-3/2}), \quad (27)$$

where $\hat{V} = I_N + O_p(n^{-1})$, and $|\cdot|$ denotes the determinant of a matrix.

(Proof) We will prove only in the MIP, as we can do also in the MOP in exactly the same way. The predictive distribution is written as follows:

$$\begin{aligned}\bar{p}(y|x, X^n, Y^n) &= \left\langle p(y|x, A\|\hat{B})\right\rangle_{p(A|X^n, Y^n\|\hat{B})} \\ &= q(y|x)\left\langle \frac{p(y|x, A\|\hat{B})}{q(y|x)}\right\rangle_{p(A|X^n, Y^n\|\hat{B})} \\ &\propto q(y|x)\left\langle \exp\left(y^t(\hat{B}A - B^*A^*)x\right)\right\rangle_{p(A|X^n, Y^n\|\hat{B})} \end{aligned} \quad (28)$$

where $\langle\cdot\rangle_p$ denotes the expectation value over a distribution p. We find from Eqs.(A· 5), (A· 8), and (A· 12) in Appendix that the random variable $(\hat{B}A - B^*A^*)$ is of order $O_p(n^{-1/2})$ when A is subject to $p(A|X^n, Y^n\|\hat{B})$. Hence we can expand Eq. (28) as follows:

$$\bar{p}(y|x, X^n, Y^n) \propto q(y|x)\left\langle 1 + y^t(\hat{B}A - B^*A^*)x + \frac{y^t v v^t y}{2n}\right\rangle_{p(A|X^n, Y^n\|\hat{B})} + O_p(n^{-3/2}), \quad (29)$$

where $v = \sqrt{n}(\hat{B}A - B^*A^*)x$ is an N-dimensional vector of order $O_p(1)$. Calculating the expectation value and expanding the logarithm of Eq. (29), we immediately arrive at Lemma 1. (Q.E.D.)

Comparing Eq. (26) with the ML estimator

$$\hat{B}\hat{A}_{MLE} = \sum_{h=1}^H \omega_{b_h}\omega_{b_h}^t RQ^{-1} \quad (30)$$

[15], we find that the SB estimator of each component is asymptotically equivalent to a positive-part James-Stein type shrinkage estimator [28] (See Sect. 6.2).

Moreover, by virtue of the localization of the SB posterior distribution, stated by Lemma 1, we can substitute the model at the SB estimator for the predictive distribution with asymptotically insignificant impact on generalization performance.[†] Therefore, we conclude that the SB approach is asymptotically equivalent to the shrinkage estimation. Note that the variance of the prior distribution, Eq. (22), asymptotically has no effect upon prediction and hence upon generalization performance, as far as it is a positive, finite constant. Remember that we can modify all the theorems in this paper for the ML estimation only by letting $L = 0$.

5.2 Generalization Error

Using the singular value decomposition of the true map B^*A^*, we can transform arbitrary A^* and B^* without change of the map into a matrix with its orthogonal row vectors and

[†]The SB approach, where the predictive distribution is the average of the models over the SB posterior distribution, can significantly differ from the shrinkage estimation, where the predictive distribution is the model denoted by the shrinkage estimator, even if the SB estimator, the average over the SB posterior distribution, of the map is equal to the shrinkage estimator. Lemma 1 enables us to identify the SB approach and the method where the predictive distribution is the model denoted by the SB estimator, equal to the shrinkage estimator, in the asymptotic limit.

another matrix with its orthogonal column vectors, respectively. Accordingly, we assume the above orthogonalities without loss of generality. Then, Lemma 1 implies that the KL divergence, Eq. (19), with a set of n training samples is given by

$$\begin{aligned}\bar{G}(X^n, Y^n) &= \frac{1}{2}\left\langle -\|y - B^*A^*x\|^2 - \log|\hat{V}^{-1}| \right. \\ &\left. +(y - \hat{V}\hat{B}\hat{A}x)^t\hat{V}^{-1}(y - \hat{V}\hat{B}\hat{A}x)\right\rangle_{q(x)q(y|x)} + O_p(n^{-3/2}) \\ &= \left\langle \frac{\|(B^*A^* - \hat{B}\hat{A})x\|^2}{2}\right\rangle_{q(x)} + O_p(n^{-3/2}) \\ &= \textstyle\sum_{h=1}^{H} \bar{G}_h(X^n, Y^n) + O_p(n^{-3/2}), \end{aligned} \tag{31}$$

where

$$\bar{G}_h(X^n, Y^n) = \frac{1}{2}\mathrm{tr}\left((b_h^*a_h^{*t} - \hat{b}_h\hat{a}_h^t)^t(b_h^*a_h^{*t} - \hat{b}_h\hat{a}_h^t)\right) \tag{32}$$

is the contribution of the h-th component. Here $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. We denote by $\mathcal{W}_d(m, \Sigma, \Lambda)$ the d-dimensional Wishart distribution with m degrees of freedom, scale matrix Σ, and noncentrality matrix Λ, and abbreviate as $\mathcal{W}_d(m, \Sigma)$ the central Wishart distribution.

Theorem 2: The generalization error of an LNN in the SB approaches can be asymptotically expanded as

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

where the coefficient of the leading term, called the generalization coefficient in this paper, is given by

$$\begin{aligned}2\lambda &= (H^*(M + N) - H^{*2}) \\ &+ \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > L)\left(1 - \frac{L}{\gamma_h'^2}\right)^2\gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \end{aligned} \tag{33}$$

Here $\theta(\cdot)$ is the indicator function of an event, i.e., which is equal to one if the event is true and to zero otherwise, $\gamma_h'^2$ is the h-th largest eigenvalue of a random matrix subject to $\mathcal{W}_{N-H^*}(M - H^*, I_{N-H^*})$, over which $\langle\cdot\rangle_{q(\{\gamma_h'^2\})}$ denotes the expectation value.

(Proof) According to Theorem 1, the difference between the SB and the ML estimators of a true component with a positive singular value is of order $O_p(n^{-1})$. Furthermore, the generalization error of the ML estimator of the component is the same as that of the regular models because of its identifiability. Hence, from Eq. (4), we obtain the first term of Eq. (33) as the contribution of the first H^* components. On the other hand, we find from Eq. (25) and Theorem 1 that, for a redundant component, identifying $RQ^{-1/2}$ with R affects the SB estimator only of order $O_p(n^{-1})$, which, hence, does not affect the generalization coefficient. We say that U is the general diagonalized matrix of an $N \times M$ matrix T if T is singular value decomposed as $T = \Omega_b U \Omega_a$, where Ω_a and Ω_b are an $M \times M$ and an $N \times N$ orthogonal matrices, respectively. Let D be the general diagonalized matrix of R, and D' the $(N - H^*) \times (M - H^*)$ matrix created by removing the first H^* columns and rows from D. Then, the first H^* diagonal elements of D correspond to the positive true singular value components and D' consists only of noises. Therefore, D' is the general diagonalized matrix of $n^{-1/2}R'$, where R' is an $(N - H^*) \times (M - H^*)$ random matrix whose elements are independently subject to $\mathcal{N}_1(0, 1)$, so that $R'R'^t$ is subject to $\mathcal{W}_{N-H^*}(M - H^*, I_{N-H^*})$. The redundant components imitate $n^{-1/2}R'$. Hence, using Theorem 1 and Eq. (32), we obtain the second term of Eq. (33) as the contribution of the last $(H - H^*)$ components. Thus, we complete the proof of Theorem 2. (Q.E.D.)

5.3 Large Scale Approximation

In a similar fashion to the analysis of the ML estimation [14], the second term of Eq. (33) can be analytically calculated in the large scale limit when M, N, H, and H^* go to infinity in the same order. We define the following scalars:

$$\alpha = N'/M' = (N - H^*)/(M - H^*), \tag{34}$$

$$\beta = H'/N' = (H - H^*)/(N - H^*), \tag{35}$$

$$\kappa = L/M' = L/(M - H^*). \tag{36}$$

Let W be a random matrix subject to $\mathcal{W}_{N'}(M', I_{N'})$, and $\{u_1, \ldots, u_{N'}\}$ the eigenvalues of $M'^{-1}W$. The measure of the empirical distribution of the eigenvalues is defined by

$$\delta P = N'^{-1}\left\{\delta(u_1) + \delta(u_2) + \cdots + \delta(u_{N'})\right\}, \tag{37}$$

where $\delta(u)$ denotes the Dirac measure at u. In the large scale limit, the measure, Eq. (37), converges almost everywhere to

$$p(u)du = \frac{\sqrt{(u-u_m)(u_M-u)}}{2\pi\alpha u}\theta(u_m < u < u_M)du, \tag{38}$$

where $u_m = (\sqrt{\alpha} - 1)^2$ and $u_M = (\sqrt{\alpha} + 1)^2$ [34]. Let

$$(2\pi\alpha)^{-1}J(u_t; k) = \int_{u_t}^{\infty} u^k p(u)du \tag{39}$$

be the k-th order moment of the distribution, Eq. (38), where u_t is the lower bound of the integration range. The second term of Eq. (33) consists of the terms proportional to the minus first, the zero, and the first order moments of the eigenvalues. Because only the eigenvalues greater than L among the largest H' eigenvalues contribute the generalization error, the moments with the lower bound $u_t = \max(\kappa, u_\beta)$ should be calculated, where u_β is the β-percentile point of $p(u)$, i.e.,

$$\beta = \int_{u_\beta}^{\infty} p(u)du = (2\pi\alpha)^{-1}J(u_\beta; 0).$$

Using the transform $s = (u - (u_m + u_M)/2)/(2\sqrt{\alpha})$, we can calculate the moments and thus obtain the following theorem:

Theorem 3: The generalization coefficient of an LNN in the large scale limit is given by

$$2\lambda \sim (H^*(M + N) - H^{*2}) + \frac{(M - H^*)(N - H^*)}{2\pi\alpha}$$

$$\Big\{J(s_t;1) - 2\kappa J(s_t;0) + \kappa^2 J(s_t;-1)\Big\}, \tag{40}$$

where

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

and $s_t = \max\Big((\kappa - (1+\alpha))/2\sqrt{\alpha}, J^{-1}(2\pi\alpha\beta;0)\Big)$. Here $J^{-1}(\cdot;k)$ denotes the inverse function of $J(s;k)$.

5.4 *Delicate* Situations

In ordinary asymptotic analysis, one considers only situations when the amplitude of each component of the true model is zero or *distinctly-positive*. Also Theorem 2 holds only in such situations. However, as mentioned in the last paragraph of Sect. 4, it is important to consider *delicate* situations when the true map B^*A^* has tiny but non-negligible singular values such that $\gamma_h^* \sim O(n^{-1/2})$, given sufficiently large but finite n. Theorem 1 still holds in such situations by replacing the second term of Eq. (26) with $o_p(n^{-1/2})$. We regard H^* as the number of *distinctly-positive* true singular values such that $\gamma_h^{*-1} = o(\sqrt{n})$. Without loss of generality, we assume that B^*A^* is a non-negative, general diagonal matrix with its diagonal elements arranged in non-increasing order. Let R''^* be the true submatrix created by removing the first H^* columns and rows from B^*A^*. Then, D', defined in the proof of Theorem 2, is the general diagonalized matrix of $n^{-1/2}R''$, where R'' is a random matrix such that $R''R''^t$ is subject to $\mathcal{W}_{N-H^*}(M-H^*, I_{N-H^*}, nR''^*R''^*)$. Therefore, we obtain the following theorem:

Theorem 4: The generalization coefficient of an LNN in the general situations when the true map B^*A^* may have *delicate* singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$ is given by

$$2\lambda = (H^*(M+N) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} + \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L) \left\{ \left(1 - \frac{L}{\gamma_h''^2}\right)^2 \gamma_h''^2 - 2\left(1 - \frac{L}{\gamma_h''^2}\right)\gamma_h'' \omega_{b_h}''^t \sqrt{n} R''^* \omega_{a_h}'' \right\} \right\rangle_{q(R'')}, \tag{41}$$

where γ_h'', ω_{a_h}'', and ω_{b_h}'' are the h-th largest singular value of R'', the corresponding right singular vector, and the corresponding left singular vector, respectively, of which $\langle\cdot\rangle_{q(R'')}$ denotes the expectation value over the distribution.

5.5 Training Error

Lemma 1 implies that the empirical KL divergence, Eq. (20), with a set of n training samples is given by

$$\bar{T}(X^n, Y^n) = -\frac{1}{2}\Big\{\mathrm{tr}\Big((B^*A^* - \hat{B}\hat{A}_{\mathrm{MLE}})^t(B^*A^* - \hat{B}\hat{A}_{\mathrm{MLE}}) - (\hat{B}\hat{A} - \hat{B}\hat{A}_{\mathrm{MLE}})^t(\hat{B}\hat{A} - \hat{B}\hat{A}_{\mathrm{MLE}})\Big)\Big\} + O_p(n^{-3/2}). \tag{42}$$

In the same way as the analysis of the generalization error, we obtain the following theorems.

Theorem 5: The training error of an LNN in the SB approaches can be asymptotically expanded as

$$\bar{T}(n) = \nu n^{-1} + O(n^{-3/2}),$$

where the coefficient of the leading term, called the training coefficient in this paper, is given by

$$2\nu = -(H^*(M+N) - H^{*2}) - \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h'^2 > L)\left(1 - \frac{L}{\gamma_h'^2}\right)\left(1 + \frac{L}{\gamma_h'^2}\right)\gamma_h'^2 \right\rangle_{q(\{\gamma_h'^2\})}. \tag{43}$$

Theorem 6: The training coefficient of an LNN in the large scale limit is given by

$$2\nu \sim -(H^*(M+N) - H^{*2}) - \frac{(M-H^*)(N-H^*)}{2\pi\alpha}\Big\{J(s_t;1) - \kappa^2 J(s_t;-1)\Big\}. \tag{44}$$

Theorem 7: The training coefficient of an LNN in the general situations when the true map B^*A^* may have *delicate* singular values such that $0 < \sqrt{n}\gamma_h^* < \infty$ is given by

$$2\nu = -(H^*(M+N) - H^{*2}) + \sum_{h=H^*+1}^{H} n\gamma_h^{*2} - \left\langle \sum_{h=1}^{H-H^*} \theta(\gamma_h''^2 > L)\left(1 - \frac{L}{\gamma_h''^2}\right)\left(1 + \frac{L}{\gamma_h''^2}\right)\gamma_h''^2 \right\rangle_{q(R'')}. \tag{45}$$

6. Discussion

6.1 Comparison with the ML Estimation and the Bayes Estimation

Figure 2 shows the theoretical results of the generalization and the training coefficients of an LNN with $M = 50$ input, $N = 30$ output, and $H = 20$ hidden units. The horizontal axis indicates the true rank H^*. The vertical axis indicates the coefficients normalized by the parameter dimension K, given by Eq. (4). The lines in the positive region correspond to the generalization coefficients of the SB approaches, clarified in this paper, that of the ML estimation, previously clarified [14], that of the Bayes estimation, also previously clarified [21], and that of the regular models, respectively; while

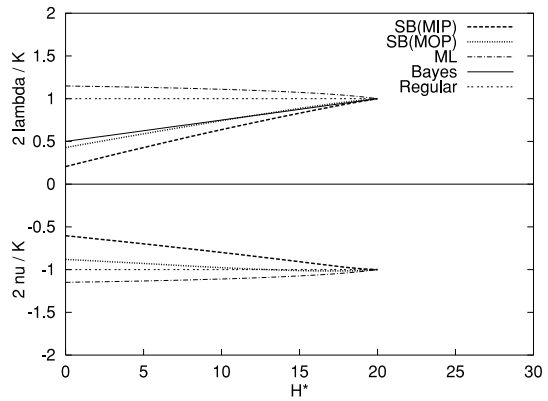


Fig. 2 Generalization error (in the positive region) and training error (in the negative region).

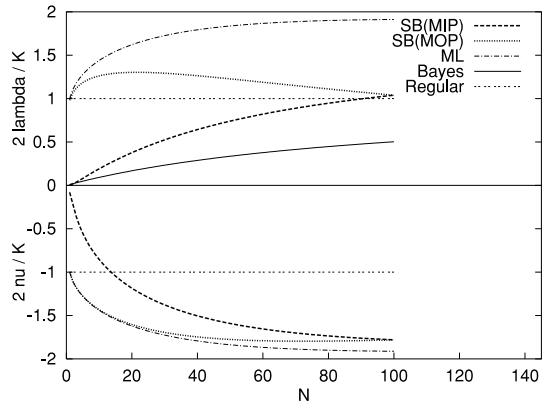


Fig. 3 N dependence when $M = 100, H = 1, H^* = 0$.

the lines in the negative region correspond to the training coefficients of the SB approaches, that of the ML estimation, and that of the regular models, respectively.[†] Unfortunately the Bayes training error has not been clarified yet.[††] Figure 3 similarly shows the coefficients of LNNs with $M = 100$ input units, $N = 1, \ldots, 100$ output units, indicated by the horizontal axis, and $H = 1$ hidden unit on the assumption that $H^* = 0$. The results in Fig. 2 and in Fig. 3 have been calculated in the large scale approximation, i.e., by using Theorems 3 and 6. We have also numerically calculated them by creating samples subject to the Wishart distribution and then using Theorems 2 and 5, and thus found that the both results almost coincide with each other so that we can hardly distinguish. We see in Fig. 2 that the SB approaches provide as good performance as the Bayes estimation, and that the MIP, moreover, has no greater generalization coefficient than the Bayes estimation for arbitrary H^*, which seems to be the asymptotic domination of the MIP over the Bayes estimation in this LNN.[†††] However, discussion of domination needs consideration of *delicate* situations.

Using Theorems 4 and 7, we can numerically calculate the SB, as well as the ML, generalization error and training error in *delicate* situations when the true distribution is near the singularities. Figure 4 shows the coefficients of an LNN with $M = 50$ input, $N = 30$ output, and $H = 5$ hidden units on the assumption that the true map consists of

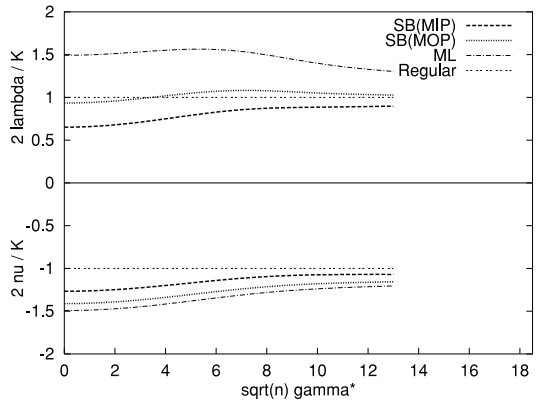


Fig. 4 With *delicate* true components.

$H^* = 1$ *distinctly-positive* component, three *delicate* components whose singular values are identical to each other, and the other one null component. The horizontal axis indicates $\sqrt{n}\gamma^*$, where $\gamma_h^* = \gamma^*$ for $h = 2, \ldots, 4$. Note that even in the ML estimation, the generalization error and the training error are asymmetrical with each other in *delicate* situations. The Bayes generalization error in *delicate* situations was previously clarified [29], but unfortunately, only in single-output (SO) LNNs, i.e., $N = H = 1$.[††††] Figure 5 shows the coefficients of an SOLNN with $M = 5$ input units on the assumption that $H^* = 0$ and the true singular value of the one component, indicated by the horizontal axis, is *delicate*. We see in Fig. 5 that the SB approaches have a property similar to the Bayes estimation, suppression of overfitting by the large state density of the singularities. We also see that, in some *delicate* situations, the MIP provides worse generalization performance than the Bayes estimation, though the MIP seems to dominate the Bayes estimation also in this SOLNN without consideration of *delicate* situations. We conjecture that, in general LNNs, the MIP could not dominate the Bayes estimation even in the case that it seems to dominate without consideration of *delicate* situations. We conclude that, in typical cases, the suppression by the singularities in the MIP is comparable to, or sometimes stronger than, that in the Bayes estimation.

[†] In the regular models, the normalized generalization and the training coefficients are always equal to one and to minus one, respectively, which leads to the penalty term of Akaike's information criterion [6].

[††] In unidentifiable models, it has not been known whether the Bayes generalization and training coefficients are symmetrical with each other, although we find from Theorems 2 and 5 that the ML generalization and training coefficients are symmetrical with each other.

[†††] We say a learning method α dominates another method β if the generalization error of α is no greater than that of β for any true distribution and that of α is smaller than that of β for a certain true distribution.

[††††] An SOLNN is regarded as a regular model at a view point of the ML estimation because the transform $b_1 a_1 \mapsto w \in \mathbb{R}^M$ makes the model linear and hence identifiable, and therefore, the ML generalization error is identical to that of the regular models. Nevertheless, an SOLNN has a property of unidentifiable models at a view point of the Bayesian learning methods, as shown in Fig. 5.

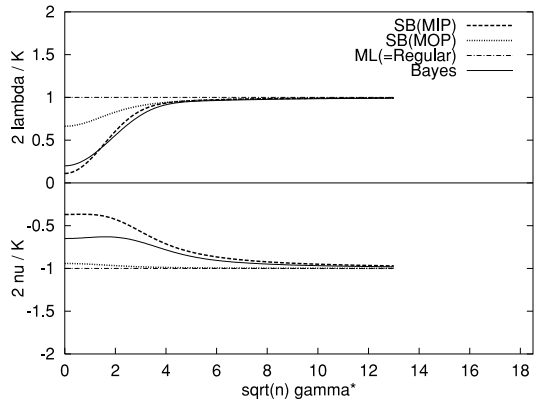


Fig. 5 Single-output LNN.

It would be more fortunate if any of the SB approaches, which require much less computational costs than MCMC methods, would always provide comparable generalization performance to the Bayes estimation. However, the SB approaches have also a property similar to the ML estimation, acceleration of overfitting by selection of the largest singular values of a random matrix. Because of selection from a large number of random variables subject to non-compact support distribution, the $(H - H^*)$ largest eigenvalues of a random matrix subject to $\mathcal{W}_{N-H^*}(M - H^*, I_{N-H^*})$ are much greater than L when $(M - H^*) > (N - H^*) \gg (H - H^*)$. Therefore, the eigenvalues $\{\gamma_h'^2\}$ in Theorem 2 go out of the effective range of shrinkage, and consequently, the SB approaches approximate the ML estimation in such atypical cases. Actually, we see in Fig. 3 that, when $N \sim 100$, the generalization error of the MIP exceeds that of the regular models, which never happens in the Bayes estimation [22].

6.2 Relation to Shrinkage Estimation

The SB estimator, given in Theorem 1, in an SOLNN is changed into the James-Stein (JS) estimator by letting $L = (M - 2)$ [28]. The relation between an EB approach and the JS estimator was discussed in a linear, hence identifiable, model as follows: based on the EB approach, the JS estimator can be derived as the solution of an equation with respect to an unbiased estimator of the hyperparameter τ_1^{-2}, introduced in Sect. 3.2 [31]. In an SOLNN, the transform $b_1 a_1 \mapsto w \in \mathbb{R}^M$ makes not only the model linear but also the prior distribution as the same form as Eq. (12). Therefore, b_1 plays the same role as τ_1. More generally, the parameters of one layer of an unidentifiable model those are regarded as hyperparameters in the SB approach can be considered to play a similar role as the deviation hyperparameters of the prior distribution in the EB approach. So, the similarity between the JS and the SB estimators is natural.

In the rest of this subsection, we focus on SOLNNs, which have the parameter $w = \{a_1 \in \mathbb{R}^M, b_1 \in \mathbb{R}\}$. In Fig. 5, the SB approaches and the Bayes estimation seem to dominate the ML estimation. The following asymptotic expansion of the generalization coefficient with respect to $\sqrt{n}\gamma_1^*$ provides a clue when it occurs:

$$2\lambda = M - \xi(\sqrt{n}\gamma_1^*)^{-2} + o\left((\sqrt{n}\gamma_1^*)^{-2}\right), \tag{46}$$

where ξ is the coefficient of the leading term when γ_1^* increases to be *distinctly-positive*. The sign of ξ indicates the direction of approach to the line $2\lambda = M$, which corresponds to the generalization coefficient of the regular models. It was found that $\xi = (M - 1)(M - 3)$ in the Bayes estimation, which leads to the conjecture that the Bayes estimation would dominate the ML estimation when $M \geq 4$ [29]. Now we consider the SB approaches. Let $a'^* = \sqrt{n}b_1^* a_1^*$ be an M-dimensional vector, so that $\|a'^*\|^2 = n\gamma_1^{*2}$. Then, Eq. (41) can be asymptotically expanded when $\sqrt{n}\gamma_1^*$ goes to infinity as follows:

$$\begin{aligned} 2\lambda &= \|a'^*\|^2 + \left\langle \left(1 - \frac{L}{\|a'^* + g\|^2}\right)^2 \|a'^* + g\|^2 \right. \\ &\quad \left. -2\left(1 - \frac{L}{\|a'^* + g\|^2}\right) a'^{*t}(a'^* + g) \right\rangle_{q(g)} + o\left(\frac{1}{\|a'^*\|^2}\right) \\ &= \left\langle \|g\|^2 + \frac{1}{\|a'^*\|^2}\left\{L^2 - 2L\|g\|^2 + 8L\frac{(a'^{*t}g)^2}{\|a'^*\|^2} \right.\right. \\ &\quad \left.\left. -4L\frac{(a'^{*t}g)^2}{\|a'^*\|^2}\right\}\right\rangle_{q(g)} + o\left(\frac{1}{\|a'^*\|^2}\right), \end{aligned} \tag{47}$$

where g is a random vector subject to $\mathcal{N}_M(0, I_M)$, over which $\langle\cdot\rangle_{q(g)}$ denotes the expectation value. Since $\langle\|g\|^2\rangle_{q(g)} = M$ and $\langle(a'^{*t}g)^2\rangle_{q(g)} = \|a'\|^2$, we have

$$2\lambda = M - \frac{L(2M - L - 4)}{\|a'^*\|^2} + o\left(\frac{1}{\|a'^*\|^2}\right). \tag{48}$$

Comparing Eqs. (46) and (48), we have

$$\xi = L(2M - L - 4), \tag{49}$$

and find that $\xi = M(M-4)$ in the MIP and that $\xi = (2M-5)$ in the MOP, which lead to the conjecture that the MIP when $M \geq 5$, as well as the MOP when $M \geq 3$, would dominate the ML estimation.

In the same format as Fig. 5, Figs. 6 and 7 show the input dimension, M, dependence of the generalization, as well as the training, coefficient in SOLNNs, numerically calculated by using Theorems 4 and 7. We see that the figures support the conjecture above. We also find from Eq. (49) that $\xi = (M - 2)^2$ in the JS estimation, which is consistent with its proved domination over the ML estimation when $M \geq 3$. The training coefficient is also asymptotically expanded as

$$2\nu = -M + \iota(\sqrt{n}\gamma_1^*)^{-2} + o\left((\sqrt{n}\gamma_1^*)^{-2}\right), \tag{50}$$

where ι is the leading coefficient. It was found that $\iota = (M - 1)^2$ in the Bayes estimation [29]. Expanding Eq. (45), we have

$$\iota = L^2 \tag{51}$$

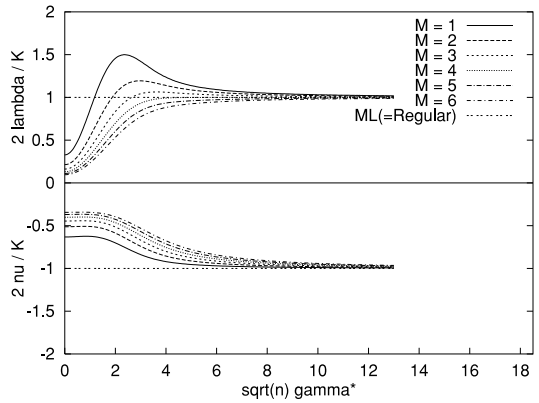


Fig. 6 M dependence in MIP.

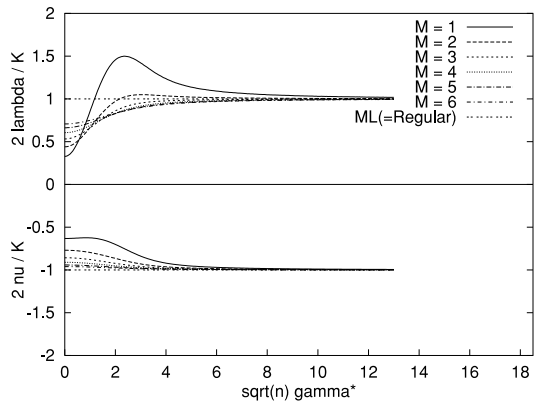


Fig. 7 M dependence in MOP.

in the SB approaches and in the JS type shrinkage estimations. We find that ξ and ι in the Bayes estimation are given by letting $L = (M-3)$ in Eq. (49) and by letting $L = (M-1)$ in Eq. (51), respectively. It does not seem to be trivial that the coefficients even in the Bayes estimation can be expressed by the forms of Eqs. (49) and (51), respectively, of which consideration can be a future work.

6.3 Relation to Variational Bayes Approach

The generalization error of the variational Bayes (VB) approach in LNNs has just been clarified [27]. In the parameter subspace corresponding to the redundant components, the VB posterior distribution extends with its variance of order 1 in the larger dimension parameter subspace either the input one or the output one; while the SB posterior distribution extends with its variance of order 1 in the parameter space $\bar{w}$, not in the hyperparameter space τ, as we find from Eqs. (A· 5) and (A· 12) in Appendix. Consequently, in LNNs, the VB approach is asymptotically equivalent to the MIP version of SB approach.

6.4 Future Work

As a future work, we would like to consider the effect of non-linearity of the activation function, $\psi(\cdot)$ in Eq. (1). We expect that the non-linearity would extend the range of basis selection and hence increase the generalization error.

7. Conclusions

We have introduced a subspace Bayes (SB) approach, an empirical Bayes approach where a part of the parameters are regarded as hyperparameters, and derived the solution of two versions of SB approach in three-layer linear neural networks (LNNs). As a result, we have discovered the asymptotic equivalence between the SB approach and a positive-part James-Stein type shrinkage estimation, and clarified its generalization error and training error. We also discuss the domination over the maximum likelihood estimation and the asymptotic equivalence to the variational Bayes approach in LNNs. We have concluded that the SB approaches have a property similar to the Bayes estimation and provide as good performance as the Bayes estimation in typical cases.

Acknowledgments

The authors would like to thank Kazuo Ushida, Masahiro Nei, and Nobutaka Magome of Nikon Corporation for encouragement to work on this subject.

References

[1] S. Nakajima and S. Watanabe, "Generalization error of linear neural networks in an empirical Bayes approach," Proc. IJCAI, pp.804–810, Edinburgh, U.K., 2005.
[2] J.A. Hartigan, "A Failure of likelihood ratio asymptotics for normal mixtures," Proc. Berkeley Conference in Honor of J. Neyman and J. Kiefer, pp.807–810, 1985.
[3] S. Watanabe, "A generalized Bayesian framework for neural networks with singular fisher information matrices," Proc. NOLTA, pp.207–210, 1995.
[4] S. Amari, H. Park, and T. Ozeki, "Geometrical singularities in the neuromanifold of multilayer perceptrons," Advances in NIPS, vol.14, pp.343–350, 2002.
[5] K. Hagiwara, "On the problem in model selection of neural network regression in overrealizable scenario," Neural Comput., vol.14, pp.1979–2002, 2002.
[6] H. Akaike, "A new look at statistical model," IEEE Trans. Autom. Control, vol.19, no.6, pp.716–723, 1974.
[7] G. Schwarz, "Estimating the dimension of a model," Annals of Statistics, vol.6, no.2, pp.461–464, 1978.
[8] J. Rissanen, "Stochastic complexity and modeling," Annals of Statistics, vol.14, no.3, pp.1080–1100, 1986.
[9] P. Bickel and H. Chernoff, Asymptotic Distribution of the Likelihood Ratio Statistic in a Prototypical Non Regular Problem, pp.83–96, Wiley Eastern Limited, 1993.
[10] A. Takemura and S. Kuriki, "Weights of chi-bar-square distribution for smooth or piecewise smooth cone alternatives," Annals of Statistics, vol.25, no.6, pp.2368–2387, 1997.
[11] S. Kuriki and A. Takemura, "Tail probabilities of the maxima of multilinear forms and their applications," Annals of Statistics, vol.29, no.2, pp.328–371, 2001.
[12] K. Fukumizu, "Likelihood ratio of unidentifiable models and multilayer neural networks," Annals of Statistics, vol.31, no.3, pp.833–851, 2003.
[13] D. Dacunha-Castelle and E. Gassiat, "Testing in locally conic models, and application to mixture models," Probability and Statistics, vol.1, pp.285–317, 1997.

[14] K. Fukumizu, "Generalization error of linear neural networks in unidentifiable cases," Proc. ALT, pp.51–62, Springer, 1999.
[15] P.F. Baldi and K. Hornik, "Learning in linear neural networks: A survey," IEEE Trans. Neural Netw., vol.6, no.4, pp.837–858, 1995.
[16] S. Watanabe, "Algebraic analysis for nonidentifiable learning machines," Neural Comput., vol.13, no.4, pp.899–933, 2001.
[17] K. Yamazaki and S. Watanabe, "Resolution of singularities in mixture models and its stochastic complexity," Proc. ICONIP, pp.1355–1359, Singapore, 2002.
[18] D. Rusakov and D. Geiger, "Asymptotic model selection for naive Bayesian networks," Proc. UAI, pp.438–445, Alberta, Canada, 2002.
[19] K. Yamazaki and S. Watanabe, "Stochastic complexities of hidden Markov models," Proc. Neural Networks for Signal Processing XIII (NNSP), pp.179–188, Toulouse, France, 2003.
[20] K. Yamazaki and S. Watanabe, "Stochastic complexity of Bayesian networks," Proc. UAI, pp.592–599, Acapulco, Mexico, 2003.
[21] M. Aoyagi and S. Watanabe, "The generalization error of reduced rank regression in Bayesian estimation," Proc. ISITA, pp.1068–1073, Parma, Italy, 2004.
[22] S. Watanabe, "Algebraic information geometry for learning machines with singularities," Advances in NIPS, vol.13, pp.329–336, 2001.
[23] G.E. Hinton and D. van Camp, "Keeping neural networks simple by minimizing the description length of the weights," Proc. COLT, pp.5–13, 1993.
[24] D.J.C. MacKay, "Developments in probabilistic modeling with neural networks—Ensemble learning," Proc. 3rd Ann. Symp. on Neural Networks, pp.191–198, 1995.
[25] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," Proc. UAI, 1999.
[26] Z. Ghahramani and M.J. Beal, "Graphical models and variational methods," in Advanced Mean Field Methods, pp.161–177, MIT Press, 2001.
[27] S. Nakajima and S. Watanabe, "Generalization error and free energy of variational Bayes approach of linear neural networks," Proc. ICONIP, pp.55–60, Taipei, Taiwan, 2005.
[28] W. James and C. Stein, "Estimation with quadratic loss," Proc. 4th Berkeley Symp. on Math. Stat. and Prob., pp.361–379, 1961.
[29] S. Watanabe and S. Amari, "Learning coefficients of layered models when the true distribution mismatches the singularities," Neural Comput., vol.15, pp.1013–1033, 2003.
[30] G.C. Reinsel and R.P. Velu, Multivariate Reduced-Rank Regression, Springer, 1998.
[31] B. Efron and C. Morris, "Stein's estimation rule and its competitors—An empirical Bayes approach," J. Am. Stat. Assoc., vol.68, pp.117–130, 1973.
[32] H. Akaike, "Likelihood and Bayes Procedure," in Bayesian Statistics, ed. J.M. Bernald, pp.143–166, University Press, 1980.
[33] R.E. Kass and D. Steffey, "Approximate Bayesian inference in conditionally independent hierarchical models (Parametric empirical Bayes models)," J. Am. Stat. Assoc., vol.84, pp.717–726, 1989.
[34] K.W. Watcher, "The strong limits of random matrix spectra for sample matrices of independent elements," Ann. Prob., vol.6, pp.1–18, 1978.

Appendix: Proof of Theorem 1

First, we will prove in the MIP version, where the (conditional) marginal likelihood is given by

$$Z(Y^n|X^n||B) = \int \phi(A) \prod_{i=1}^n p(y_i|x_i, A||B) dA \propto \int \exp\left(-\frac{\sum_{i=1}^n ||y_i - BAx_i||^2 + \mathrm{tr}(A^t A)}{2}\right) dA, \quad (\mathrm{A}\cdot 1)$$

where $\int dA$ denotes the integral with respect to all the elements of the matrix A. We denote by $\otimes$ the Kronecker product and by $\mathrm{vec}(\cdot)$ the vector created from a matrix by stacking the column vectors below one another, for example, $\mathrm{vec}(V) = (v_1^t, \ldots, v_H^t)^t$ is the NH-dimensional column vector, where $V = (v_1, \ldots, v_H)$ is an $N \times H$ matrix. By using the Gaussian integral, we have the following form of the marginal likelihood:

$$Z(Y^n|X^n||B) \propto \left(n|\tilde{S}|^{-1}\right)^{-1/2} \exp\left(\frac{n\tilde{b}^t \tilde{R} \tilde{S}^{-1} \tilde{R}^t \tilde{b}}{2}\right), \quad (\mathrm{A}\cdot 2)$$

where $\tilde{a} = \mathrm{vec}(A^t)$, $\tilde{b} = \mathrm{vec}(B)$, $\tilde{R} = I_M \otimes R$, and $\tilde{S} = (B^t B \otimes Q) + n^{-1} I_{HM}$. Similarly, we also have the following form of the posterior distribution:

$$p(A|X^n, Y^n||B) = \frac{\phi(A) \prod_{i=1}^n p(y_i|x_i, A||B)}{Z(Y^n|X^n||B)} \propto \exp\left(-\left(\tilde{a} - \tilde{S}^{-1}\tilde{R}^t\tilde{b}\right)^t \frac{n\tilde{S}}{2} \left(\tilde{a} - \tilde{S}^{-1}\tilde{R}^t\tilde{b}\right)\right). \quad (\mathrm{A}\cdot 3)$$

Given an arbitrary map BA, we can have A with its orthogonal row vectors and B with its orthogonal column vectors by using the singular value decomposition. Just in that case, the prior probability, Eq. (22), is maximized. Accordingly, we assume without loss of generality that the optimal value of B consists of its orthogonal column vectors. Consequently, Eq. (A· 2), as well as Eq. (A· 3), factorizes as

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

where

$$Z(Y^n|X^n||b_h) \propto |S_h|^{-1/2} \exp\left(\frac{n b_h^t R S_h^{-1} R^t b_h}{2}\right), \quad (\mathrm{A}\cdot 4)$$

$$p(a_h|X^n, Y^n||b_h) \propto \exp\left(-\left(a_h - S_h^{-1}R^t b_h\right)^t \frac{nS_h}{2} \left(a_h - S_h^{-1}R^t b_h\right)\right). \quad (\mathrm{A}\cdot 5)$$

Here $S_h = (||b_h||^2 Q + n^{-1} I_M)$. Let $F'(Y^n|X^n||b_h) = F(Y^n|X^n||b_h) + \text{const.}$, where $F(Y^n|X^n||b_h)$ is the stochastic complexity, i.e., the negative log marginal likelihood, of the h-th component. Then, we get

$$2F'(Y^n|X^n||b_h) = -2\log Z(Y^n|X^n||b_h) + \text{const.} = \log|S_h| - n b_h^t R S_h^{-1} R^t b_h. \quad (\mathrm{A}\cdot 6)$$

Hereafter, separately considering the components imitating the positive true ones and the redundant components, we will find the optimal hyperparameter value $\hat{b}_h$ that minimizes Eq. (A· 6). We abbreviate $F'(Y^n|X^n||b_h)$ as $F'(b_h)$.

For a positive true component, $h \le H^*$, the corresponding observed singular value γ_h of $RQ^{-1/2}$ is of order 1 with probability 1. Then, from Eq. (A· 6), we get

$$2F'(b_h) = M\log||b_h||^2 - n||b_h||^{-2} b_h^t R Q^{-1} R^t b_h + ||b_h||^{-4} b_h^t R Q^{-2} R^t b_h + O_p(n^{-1}). \quad (\mathrm{A}\cdot 7)$$

To minimize Eq. (A· 7), the leading, second term dominates the determination of the direction cosine of $\hat{b}_h$ and leads to $\hat{b}_h = \|\hat{b}_h\|(\omega_{b_h} + O_p(n^{-1}))$. The first and the third terms determine the norm of $\hat{b}_h$ because the second term is independent of it. Thus, we get the optimal hyperparameter value as follows:

$$\hat{b}_h = \sqrt{\frac{\omega_{b_h}^t RQ^{-2}R^t\omega_{b_h}}{M}}\omega_{b_h} + O_p(n^{-1}). \quad (A\cdot 8)$$

Because the average of a_h over the posterior distribution, Eq. (A· 5), is $\hat{a}_h = S_h^{-1}R^t\hat{b}_h$, we obtain the SB estimator for the positive true component of the map BA as follows:

$$\hat{b}_h\hat{a}_h^t = \omega_{b_h}\omega_{b_h}^t RQ^{-1} + O_p(n^{-1}). \quad (A\cdot 9)$$

On the other hand, for a redundant component, $h > H^*$, Eq. (25) allows us to approximate Eq. (A· 6) as follows:

$$2F'(b_h) = M\log\left(\|b_h\|^2 + n^{-1}\right) - \frac{nb_h^t RR^t b_h}{\|b_h\|^2 + n^{-1}} + O_p(n^{-1/2}). \quad (A\cdot 10)$$

Then, we find that the direction cosine of $\hat{b}_h$, determined by the second term of Eq. (A· 10), is approximated by ω_{b_h} with accuracy $O_p(n^{-1/2})$. After substituting $\gamma_h^2\|b_h\|^2(1 + O_p(n^{-1/2}))$ for $b_h^t RR^t b_h$, we get the following extreme condition by partial differentiation of Eq. (A· 10) with respect to the norm of b_h:

$$0 = 2\frac{\partial F'(b_h)}{\partial\|b_h\|^2} = \frac{M}{(\|b_h\|^2 + n^{-1})^2}\left(\|b_h\|^2 - \frac{n\gamma_h^2 - M}{nM}\right) + O_p(\|b_h\|^{-2}n^{-1/2}). \quad (A\cdot 11)$$

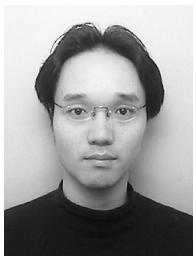
We find from Eq. (A· 11) that Eq. (A· 10) is an increasing function of $\|b_h\|$ if γ_h is less than $\sqrt{M/n}$. Therefore, we get the optimal hyperparameter value as follows:

$$\hat{b}_h = \sqrt{\frac{L'_h - M}{nM}}\omega_{b_h} + O_p(n^{-1}). \quad (A\cdot 12)$$

Thus, we obtain the SB estimator of the redundant component as follows:

$$\hat{b}_h\hat{a}_h^t = \frac{\hat{b}_h\hat{b}_h^t R}{\|\hat{b}_h\|^2 + n^{-1}} + O_p(n^{-1}) = (1 - ML_h'^{-1})\omega_{b_h}\omega_{b_h}^t R + O_p(n^{-1}). \quad (A\cdot 13)$$

Selecting the largest singular value components minimizes Eq. (A· 6). Hence, combining Eq. (A· 9) with the fact that $ML_h'^{-1} = O_p(n^{-1})$ for the positive true components, and Eq. (A· 13) with Eq. (25), we obtain the SB estimator in Theorem 1. We can also derive the SB estimator in the MOP version in exactly the same way. (Q.E.D.)



Shinichi Nakajima was born in Kobe, Japan. He received the master degree in 1995 from Kobe university. He has been working on research and development of semiconductor lithography tools at Nikon corporation from 1995 to the present. He has also been a doctoral course student of the department of computational intelligence and systems science, Tokyo Institute of Technology since 2003. His research interests are in learning theory and its application.

Sumio Watanabe received Ph.D. degree in applied electronics from Tokyo Institute of Technology, Japan, in 1993. He is currently a professor at Precision and Intelligence Laboratory, Tokyo Institute of Technology, Japan. His research interests include probability theory, algebraic geometry, and learning theory.