Analytic Solution of Hierarchical Variational Bayes in Linear Inverse Problem

Shinichi Nakajima[1] and Sumio Watanabe[2]

[1] Nikon Corporation, 201-9 Miizugahara, Kumagaya, 360-8559 Japan
nakajima.s@nikon.co.jp, swatanab@pi.titech.ac.jp
http://watanabe-www.pi.titech.ac.jp/~nkj23/index.html
[2] Tokyo Institute of Technology, Mailbox R2-5, 4259 Nagatsuda, Yokohama, 226-8503 Japan

Abstract. In singular models, the Bayes estimation, commonly, has the advantage of the generalization performance over the maximum likelihood estimation, however, its accurate approximation using Markov chain Monte Carlo methods requires huge computational costs. The variational Bayes (VB) approach, a tractable alternative, has recently shown good performance in the automatic relevance determination model (ARD), a kind of hierarchical Bayesian learning, in brain current estimation from magnetoencephalography (MEG) data, an ill-posed linear inverse problem. On the other hand, it has been proved that, in three-layer linear neural networks (LNNs), the VB approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation. In this paper, noting the similarity between the ARD in a linear problem and an LNN, we analyze a simplified version of the VB approach in the ARD. We discuss its relation to the shrinkage estimation and how ill-posedness affects learning. We also propose the algorithm that requires simpler computation than, and will provide similar performance to, the VB approach.

1 Introduction

It is known that the Bayes estimation provides better generalization performance than the maximum likelihood (ML) estimation when we use a model having singularities, on which the Fisher information matrix is singular, in the parameter space. However, Markov chain Monte Carlo (MCMC) methods, used for approximation of the Bayes posterior distribution, require huge computational costs. The variational Bayes (VB) approach was proposed as a tractable alternative [1, 2], and is often applied to singular models, for example, mixture models and hidden Markov models. Recently, the VB approach has been applied also to the automatic relevance determination model (ARD) [3] in a linear inverse problem, i.e., brain current estimation from magnetoencephalography (MEG) data [4]. Although the advantage of the VB approach has been experimentally shown in many applications, its generalization performance had been theoretically clarified in no singular model until quite recently. Last year, proving the asymptotic equivalence between the VB approach and a positive-part James-Stein (JS) type shrinkage estimation [5], we have clarified the generalization error of the VB approach in three-layer linear neural networks (LNNs), the simplest singular models [6].

In this paper, noting the similarity between the ARD in a linear problem and an LNN, we clarify some properties of the VB approach in the ARD and then propose the alternative that requires less computational costs. In Section 2, we shortly describe the framework of the VB approach. In Section 3, we explain the brain current estimation and the ARD, and then, discuss the similarity between the ARD and an LNN. In Section 4, we, in detail, describe the setting assumed in our theoretical analysis. After that, in Section 5, we analyze the VB approach in the ARD, and show its relation to the JS type shrinkage estimation. Discussion including our proposal is in Section 6, and finally, conclusions and future work are in Section 7.

2 Variational Bayes Approach

Let $Y^n = \{y_1, \ldots, y_n\}$ be arbitrary n training samples independently and identically taken from the true distribution. In the framework of the Bayes estimation, the posterior distribution of the parameter w of a model $p(y|w)$ is given by

$$p(w|Y^n) = \frac{\phi(w) \prod_{i=1}^{n} p(y_i|w)}{Z(Y^n)}, \quad \text{where} \quad Z(Y^n) = \int \phi(w) \prod_{i=1}^{n} p(y_i|w) dw \quad (1)$$

is the marginal likelihood, and $\phi(w)$ is the prior distribution. The predictive distribution is defined as the average of the model over the posterior distribution.

In the variational Bayes (VB) approach [2], called the mean field approximation in statistical physics, we first define the following functional, called the generalized free energy in this paper, of an arbitrary trial posterior distribution $r(w|Y^n)$:[3]

$$\bar{F}(r) = -S(r) + nE(r), \qquad \text{where} \qquad (2)$$

$$S(r) = -\left\langle \log r(w) \right\rangle_{r(w)} \quad \text{and} \quad E(r) = -n^{-1} \left\langle \log \left(\phi(w) \prod_{i=1}^{n} p(y_i|w) \right) \right\rangle_{r(w)}$$

are the entropy and the energy, respectively. Here, $\langle \cdot \rangle_p$ denotes the expectation value over a distribution p. Note that $n^{-1}\bar{F}(r)$ corresponds to the Helmholtz free energy if we consider n to be the inverse temperature, hence, the Bayes posterior distribution, Eq.(1), which minimizes the Helmholtz free energy, corresponds to the equilibrium distribution [7]. Then, in the VB approach, restricting the space of allowed $r(w)$, we minimize the generalized free energy, Eq.(2). The optimum of $r(w)$ is called the VB posterior distribution, over which the expectation value of w is called the VB estimator.

3 Model

3.1 MEG inverse problem

Magnetoencephalography (MEG) is one of the major recording means of brain activity, in which we estimate the electric current distribution in a brain from the magnetic fields that are induced by the current and observed on the head [8]. For simplicity, we assume that the current and the field are scalar. Let $a' \in \mathbb{R}^M$ be the current vector of which each element corresponds to the current value at each site in a brain, and $y \in \mathbb{R}^N$ the field

[3] We will hereafter abbreviate $r(w|Y^n)$ by $r(w)$ or by r.

vector of which each element corresponds to the field value at each site on the head. We utilize the following linear regression model:

$$y = Va' + \varepsilon, \tag{3}$$

where V is the $N \times M$ constant matrix, called the lead field matrix, that represents the field induced by the current, and $\varepsilon \in \mathbb{R}^N$ is an observation noise [4,8]. By $\mathcal{N}_d(\mu, \Sigma)$ we denote the d-dimensional normal distribution with average μ and covariance matrix Σ, and by $\mathcal{N}_d(\cdot; \mu, \Sigma)$ its probability density. Assume that the noise, ε in Eq.(3), is subject to $\mathcal{N}_N(0, \sigma_y^2 I_N)$, where $0 < \sigma_y^2 < \infty$ and I_d is the $d \times d$ identity matrix. Then, the probability density of the field is given by

$$p(y|a') = \mathcal{N}_N(y; Va', \sigma_y^2 I_N). \tag{4}$$

In typical MEG estimation problems, the number of sites at which the fields are observed is smaller than the number of sites at which we want to know the brain currents, i.e., $N < M$, hence, the MEG estimation is an ill-posed problem. Therefore, in the a' space, the region in which any point gives the maximum likelihood is not a point, given an observed field. So, a prior assumption is needed to select one point in that region. One of the most popular methods is the minimum norm method, in which the point giving the minimum norm is selected from the points giving the maximum likelihood [8]. We can easily find that the maximum a posterior (MAP) estimation with the following prior distribution provides the minimum norm solution as well: $\phi(a') = \mathcal{N}_M(a'; 0, I_M)$.

3.2 Automatic Relevance Determination

The automatic relevance determination model (ARD), a kind of hierarchical Bayesian learning, was proposed to eliminate irrelevant connections from neural networks [3]. In the ARD, we first introduce a prior distribution of the parameters, i.e., the weight vectors, with the hyperparameters corresponding to the variances. Then, we introduce a prior distribution of the hyperparameters, called a hyperprior. If we apply the Bayes estimation to this model, many weight vectors tend to be eliminated as irrelevant connections, because of the singularities caused by the hierarchy. (See [9] for detail.)

Now, we focus on its application to the MEG inverse problem, whose distribution is given by Eq.(4). We use the following prior distribution of a':

$$\phi(a' \| B^{-2}) = \mathcal{N}_M(a'; 0, B^2), \tag{5}$$

where B^{-2} is the hyperparameter. We consider in this paper the simplest ARD, where B^{-2} is diagonal.[4] Then, increasing the (m, m)-th element of B^{-2} eliminates the m-th element, a'_m, of the current as an irrelevant one. We can estimate the value of B^{-2} based on the empirical Bayes (EB) approach [10], where the hyperparameter is estimated by maximizing the marginal likelihood, $Z(Y^n)$ in Eq.(1), or can estimate the posterior distribution of B^{-2} by introducing the hyperprior, such as

$$\phi(B^{-2}) = \prod_{m=1}^{M} \Gamma\left(B_{mm}^{-2}; \bar{\kappa}_{0m}, \bar{\nu}_{0m}\right), \tag{6}$$

[4] The differences between the setting assumed in this paper and that in [4] are summarized at the end of Section 4.2.

where we denote by $\mathit{\Gamma}(\kappa, \nu)$ the Gamma distribution with shape parameter κ and scale parameter ν, and by $\mathit{\Gamma}(\cdot; \kappa, \nu)$ its probability density. To the latter method, we can apply the VB approach and obtain the iterative algorithm, restricting the posterior distribution such that a' and B^{-2} are independent of each other [4].

The following point is important: MEG data are time series, and we want to know the current at each point of time; on the other hand, the hyperparameter B^{-2} is considered to be invariant during some time period in [4], which essentially affects learning and enhances elimination of irrelevant elements, as will be shown in the following sections.

3.3 Similarity to Linear Neural Networks

Let $x' \in \mathbb{R}^M$ be the formal input vector of which all the elements are equal to one. Then, the transform $a' \to Ba$, where $a \in \mathbb{R}^M$, makes the model distribution, Eq.(4), and the prior distribution, Eq.(5), as

$$p(y|x', A, B) = \mathcal{N}_N(y; VBAx', \sigma_y^2 I_N), \tag{7}$$

$$\phi(a) = \mathcal{N}_M(a; 0, I_M), \tag{8}$$

where A is the $M \times M$ diagonal matrix whose (m, m)-th element is equal to the m-th element of a. We thus find that the model, Eq.(7), is similar to a linear neural network model (LNN),[5] in which the VB approach has been analyzed in [6]. However, there is an important difference, i.e., the existence of the lead field matrix, V in Eq.(7), although the ARD is equivalent to an LNN when V is general diagonal. We do not like to transform the basis of the current vector, a', space, so that V is general diagonal, since the purpose of that application is to find the few sites where synapses fire.

4 Setting

4.1 Restriction on Posterior Distribution

As discussed in Section 3.3, the ARD in the linear inverse problem, Eq.(4), with the prior distribution, Eq.(5), is equivalent to the model, Eq.(7), with the prior distribution, Eq.(8). By b we denote the M-dimensional vector whose m-th element is equal to the (m, m)-th element of B. We introduce the following prior distribution of b, which is substituted for the hyperprior of B^{-2} in the ARD:

$$\phi(b) = \mathcal{N}_M(b; 0, c_b^2 I_M), \tag{9}$$

where $0 < c_b^2 < \infty$ is a *constant* hyperparameter. Note that B_{mm}^2 / c_b^2 is then subject to $\mathit{\Gamma}(1/2, 2)$. For symmetry, we also introduce a *constant* hyperparameter $0 < c_a^2 < \infty$ in the prior distribution of a as follows:

$$\phi(a) = \mathcal{N}_M(a; 0, c_a^2 I_M). \tag{10}$$

[5] The definition of the LNN is described in Appendix A.

Actually, it will be shown in the following sections that the values of the *constant* hyperparameters, c_a^2 and c_b^2, do not asymptotically affect learning as far as they are positive and finite; while whether the hyperparameter B^{-2} is constant or estimated from observation strongly affects learning even in the asymptotic limit.

Now we apply the VB approach to the transformed ARD model, Eq.(7), with the prior distributions, Eqs.(10) and (9). We restrict the trial posterior distribution such that a and b are independent of each other:

$$r(a,b) = r(a)r(b). \tag{11}$$

Then, the generalized free energy, Eq.(2), can be written as follows:

$$\bar{F}(Y^n) = \int r(a)r(b) \log \frac{r(a)r(b)}{p(Y^n|a,b)\phi(a)\phi(b)} dadb. \tag{12}$$

Using the variational method [2], we obtain the following condition:

$$r(a) \propto \phi(a) \exp\langle \log p(Y^n|a,b)\rangle_{r(b)}, \quad r(b) \propto \phi(b) \exp\langle \log p(Y^n|a,b)\rangle_{r(a)}. \tag{13}$$

We find from Eq.(13) that the VB posterior distribution is the normal, because the log-likelihood, $\log p(Y^n|a,b)$, is a biquadratic function of a and b, and we use the normal prior distributions, Eqs.(10) and (9). In this paper, we furthermore restrict $r(a,b)$ such that all the elements are independent of each other for simplicity, which results that

$$r(a_m) \propto \phi(a_m) \exp\langle \log p(Y^n|a,b)\rangle_{r(a)r(b)/r(a_m)}, \tag{14}$$
$$r(b_m) \propto \phi(b_m) \exp\langle \log p(Y^n|a,b)\rangle_{r(a)r(b)/r(b_m)}. \tag{15}$$

4.2 Summary of Setting

We summarize our setting in the following. Let $A^{(u)}$ be an $M \times M$ diagonal parameter matrix at the time u, B another $M \times M$ diagonal parameter matrix, which is assumed to be invariant during the time period $u = 1, \ldots, U$, and $y^{(u)}$ an N-dimensional observed vector. By $a^{(u)}$ we denote the M-dimensional parameter vector representing the diagonal elements of $A^{(u)}$, i.e., $a_m^{(u)} = A_{mm}^{(u)}$, and by b the M-dimensional parameter vector representing the diagonal elements of B, i.e., $b_m = B_{mm}$. Suppose that we have n training samples, i.e., n sets of U time series data, denoted by Y^n.

In this paper, restricting the trial posterior distribution $r(a,b)$ such that all the elements are independent of each other, we analyze the VB approach in the model

$$p(\{y^{(u)}\}|\{a^{(u)}\}, b) = \textstyle\prod_{u=1}^{U} \mathcal{N}_N(y^{(u)}; V \sum_{m=1}^{M} b_m a_m^{(u)} 1_m, \sigma_y^2 I_N) \tag{16}$$

with the prior distributions

$$\phi(\{a^{(u)}\}) = \textstyle\prod_{u=1}^{U} \mathcal{N}_M(a^{(u)}; 0, c_a^2 I_M), \qquad \phi(b) = \mathcal{N}_M(b; 0, c_b^2 I_M), \tag{17}$$

where $V = (v_1, \ldots, v_M)$ is an $N \times M$ constant matrix, and 1_m denotes the M-dimensional vector whose m-th element is equal to unity and all the other elements are equal to zero. The noise variance, σ_y^2, is assumed to be known.

Finally, we summarize the major differences of our setting from that in [4]:

1. The spatial correlation of the brain current distribution is considered by introducing the smoothness prior, where the hyperparameter B^{-2} in Eq.(5) is not assumed to be diagonal, in [4]; while B is assumed to be diagonal in this paper.
2. The restriction on the VB posterior distribution is only the independence between a' and B^{-2} in [4]; while the independence among the elements of a, as well as those of b, is also assumed in this paper.
3. The prior distribution of b_m^{-2} is $\Gamma\left(b_m^{-2}; \bar{\kappa}_{0m}, \bar{\nu}_{0m}\right)$ in [4]; while that of its inverse, b_m^2, is $\Gamma(b_m^2/c_b^2; 1/2, 2)$ in this paper.
4. The number of samples for estimation of each site and each point of time is only one, i.e., $n = 1$, in [4]; while we consider the case that we have sufficiently large n training samples in this paper. However, we will derive also the non-asymptotic solution in the case that $U = 1$, at the end of Section 5.2.

5 Theoretical Analysis

5.1 Variational Condition

Define the following M-dimensional vector:

$$j^{(u)}(Y^n) = n^{-1} \textstyle\sum_{i=1}^n V^t y_i^{(u)}, \quad \text{i.e.,} \quad j_m^{(u)}(Y^n) = n^{-1} \textstyle\sum_{i=1}^n v_m^t y_i^{(u)}, \tag{18}$$

where t denotes the transpose of a matrix or vector. We hereafter abbreviate $j^{(u)}(Y^n)$ as $j^{(u)}$. We find from Eqs.(14) and (15) that the VB posterior distribution factorizes as

$$r(\{a^{(u)}\}, b) = \textstyle\prod_{m=1}^M \left\{ \left(\textstyle\prod_{u=1}^U \mathcal{N}_1(a_m^{(u)}; \mu_{a_m}^{(u)}, \sigma_{a_m}^{2\,(u)} I_M) \right) \mathcal{N}_1(b_m; \mu_{b_m}, \sigma_{b_m}^2 I_M) \right\}, \tag{19}$$

where $\mu_{a_m}^{(u)}$, μ_{b_m}, $\sigma_{a_m}^{2\,(u)}$, and $\sigma_{b_m}^2$ are scalar for $m = 1, \ldots, M$ and $u = 1, \ldots, U$. Note that the VB estimator of the m-th element of the current at the time u is given by $(\hat{a}_m'^{(u)})_{\text{VB}} = (\hat{b}_m \hat{a}_m^{(u)})_{\text{VB}} = \mu_{b_m} \mu_{a_m}^{(u)}$. By $\tilde{\cdot}$ we denote the U-dimensional time series vector, for example, $\tilde{a}_m = (a_m^{(1)}, \ldots, a_m^{(U)})^t$. Then, we obtain the following variational condition by substituting Eq.(19) into Eqs.(14) and (15):

$$\tilde{\mu}_{a_m} = n\sigma_y^{-2} \|v_m\|^2 \sigma_{a_m}^2 \tilde{z}_m \mu_{b_m}, \tag{20}$$

$$\sigma_{a_m}^{2\,(u)} = n^{-1} \sigma_y^2 \left(\|v_m\|^2 (\mu_{b_m}^2 + \sigma_{b_m}^2) + n^{-1} \sigma_y^2 c_a^{-2} \right)^{-1}, \tag{21}$$

$$\mu_{b_m} = n\sigma_y^{-2} \|v_m\|^2 \sigma_{b_m}^2 \tilde{z}_m^t \tilde{\mu}_{a_m}, \tag{22}$$

$$\sigma_{b_m}^2 = n^{-1} \sigma_y^2 \left(\|v_m\|^2 (\|\tilde{\mu}_{a_m}\|^2 + U\sigma_{a_m}^2) + n^{-1} \sigma_y^2 c_b^{-2} \right)^{-1}, \tag{23}$$

$$\text{where} \qquad \tilde{z}_m = \|v_m\|^{-2} (\tilde{j}_m - \textstyle\sum_{m' \neq m} \mu_{b_{m'}} \tilde{\mu}_{a_{m'}} v_m^t v_{m'}). \tag{24}$$

We denote $\sigma_{a_m}^{2\,(u)}$ by $\sigma_{a_m}^2$ in Eqs.(20) and (23), since Eq.(21) implies that it is invariant for u. Similarly, substituting Eq.(19) into Eq.(12), we also have the following form of the generalized free energy:

$$\begin{aligned} 2\bar{F}(Y^n) = \text{const.} + \textstyle\sum_{m=1}^M \Big\{ & -\log \sigma_{a_m}^{2U} \sigma_{b_m}^2 + \tfrac{\|\tilde{\mu}_{a_m}\|^2 + U\sigma_{a_m}^2}{c_a^2} + \tfrac{\mu_{b_m}^2 + \sigma_{b_m}^2}{c_b^2} \\ & -2n\sigma_y^{-2} \|v_m\|^2 \left(\tilde{z}_m^t \tilde{\mu}_{a_m} \mu_{b_m} \right) + n\sigma_y^{-2} \|v_m\|^2 (\|\tilde{\mu}_{a_m}\|^2 + U\sigma_{a_m}^2)(\mu_{b_m}^2 + \sigma_{b_m}^2) \Big\}. \end{aligned} \tag{25}$$

5.2 Variational Bayes Solution

The variational condition, Eqs.(20)–(23), can be analytically solved, which leads to the following theorem:

Theorem 1. *The VB estimator of the m-th element of the current is given by*

$$(\hat{b}_m \hat{\hat{a}}_m)_{VB} = \begin{cases} 0 & \textit{for } m \textit{ such that } v_m = 0 \\ \mathcal{S}(\tilde{z}_m; \sigma_y^2 U/\|v_m\|^2) + O_p(n^{-1}) & \textit{for } m \textit{ such that } v_m \neq 0 \end{cases}, \tag{26}$$

where $$\mathcal{S}(z;\chi) = \theta(n\|z\|^2 > \chi)\left(1 - \chi/n\|z\|^2\right) z \tag{27}$$

is the positive-part James-Stein (JS) type shrinkage operator with the degree of shrinkage $\chi > 0$.[6] *Here $\theta(\cdot)$ denotes the indicator function of an event.*

(Outline of the proof) We will find the solution of the variational condition, Eqs.(20)–(23). We easily have the solution for the elements such that $v_m = 0$. For the other elements such that $v_m \neq 0$, we have the following variances by solving Eqs.(21) and (23):

$$\hat{\sigma}_{a_m}^2 = \frac{-(\hat{\eta}_m^2 - n^{-1}\sigma_y^2\|v_m\|^2(U-1)) + \sqrt{(\hat{\eta}_m^2 + n^{-1}\sigma_y^2\|v_m\|^2(U+1))^2 - 4n^{-2}\sigma_y^4 U\|v_m\|^4}}{2U\|v_m\|^2(\|v_m\|^2\hat{\mu}_{b_m}^2 + n^{-1}\sigma_y^2 c_a^{-2})}, \tag{28}$$

$$\hat{\sigma}_{b_m}^2 = \frac{-(\hat{\eta}_m^2 + n^{-1}\sigma_y^2\|v_m\|^2(U-1)) + \sqrt{(\hat{\eta}_m^2 + n^{-1}\sigma_y^2\|v_m\|^2(U+1))^2 - 4n^{-2}\sigma_y^4 U\|v_m\|^4}}{2\|v_m\|^2(\|v_m\|^2\|\hat{\hat{\mu}}_{a_m}\|^2 + n^{-1}\sigma_y^2 c_b^{-2})}, \tag{29}$$

where $$\hat{\eta}_m^2 = (\|v_m\|^2\|\hat{\hat{\mu}}_{a_m}\|^2 + n^{-1}\sigma_y^2 c_b^{-2})(\|v_m\|^2\hat{\mu}_{b_m}^2 + n^{-1}\sigma_y^2 c_a^{-2}). \tag{30}$$

By using Eqs.(20), (22), (28), and (29), we have

$$\hat{\eta}_m^2 = \left(1 - \frac{\sigma_y^2}{n\|v_m\|^2\|\tilde{z}_m\|^2}\right)\left(1 - \frac{\sigma_y^2 U}{n\|v_m\|^2\|\tilde{z}_m\|^2}\right)\|\tilde{z}_m\|^2, \tag{31}$$

$$\sigma_y^2(Uc_a^{-2}\hat{\delta}_m - c_b^{-2}\hat{\delta}_m^{-1}) = n(U-1)\|v_m\|^2(\|\tilde{z}_m\| - \hat{\gamma}_m), \tag{32}$$

where $$\hat{\gamma}_m = \|\hat{\hat{\mu}}_{a_m}\|\hat{\mu}_{b_m} \quad \text{and} \quad \hat{\delta}_m = \|\hat{\hat{\mu}}_{a_m}\|/\hat{\mu}_{b_m}. \tag{33}$$

Solving Eqs.(30)–(32), we obtain the VB estimator in Theorem 1. (Q.E.D.)

Moreover, we obtain the following non-asymptotic expression of the VB estimator when $U = 1$:

Theorem 2. *The VB estimator of the m-th element of the current when $U = 1$ and $v_m \neq 0$ is given by*

$$(\hat{b}_m \hat{a}_m)_{VB} = \mathrm{sign}(z_m) \cdot \max\left(0, \|\mathcal{S}(z_m; \sigma_y^2/\|v_m\|^2)\| - \sigma_y^2(nc_a c_b\|v_m\|^2)^{-1}\right), \tag{34}$$

where sign$(\cdot)$ *denotes the sign of a scalar.*

(Outline of the proof) We find from Eq.(32) that $\|\hat{\delta}_m\| = c_a/c_b$, which makes Eqs.(30) and (31) rigorously solvable and leads to Theorem 2. (Q.E.D.)

Note that neither Theorem 1 nor Theorem 2 provides any explicit expression of the VB estimator, since $\tilde{z}_m$, given by Eq.(24), depends on the other elements of the VB estimator, i.e., $(\hat{b}_{m'}\hat{\hat{a}}_{m'})_{\mathrm{VB}}$ for $m' \neq m$. So, further consideration is needed.

[6] The positive-part JS type shrinkage estimator, as well as operator, is explained in Appendix B.

5.3 Comparison with Shrinkage Estimation

By $(\cdot)^-$ we denote the Moore-Penrose generalized inverse of a matrix. Consider the following positive-part JS type shrinkage estimator based on the minimum norm maximum likelihood (MNML) estimator:

$$(\hat{b}_m\hat{\hat{a}}_m)_{\text{PJS}} = \mathcal{S}\left((\hat{b}_m\hat{\hat{a}}_m)_{\text{MN}}; \sigma_y^2 U/\|v_m\|^2\right), \text{ where } (\hat{B}\hat{a}^{(u)})_{\text{MN}} = (V^tV)^- j^{(u)} \quad (35)$$

is the MNML estimator. Hereafter, we compare the VB estimator, Eq.(26), and the shrinkage estimator, Eq.(35). From the definition of $\tilde{z}_m$, given by Eq.(24), we find that $\tilde{z}_m$ is the *unique* ML estimator and hence the VB and the shrinkage estimators of the m-th element are asymptotically equivalent to each other, if $v_m^t v_{m'} = 0$ for $^\forall m' \neq m$. However, nonorthogonality and linear dependence, which causes ill-posedness, of the set of the lead field column vectors, i.e., $\{v_m\}$, makes a difference between them.

Consider the simplest ill-posed case where all the lead field vectors, $\{v_m\}$ for $m = 1, \ldots, M$, are parallel to each other. Then, the MNML estimator at u is given by

$$(\hat{B}\hat{a}^{(u)})_{\text{MN}} = (\textstyle\sum_{m=1}^M \|v_m\|^2)^{-1} j^{(u)}, \quad (36)$$

from which we find that all the elements of the MNML estimator, naturally, have the same sign. Hence, we find from Eq.(24) that the fact that $\|(\hat{b}_m\hat{a}_m^{(u)})_{\text{VB}}\| < \|(\hat{b}_m\hat{a}_m^{(u)})_{\text{MN}}\|$ leads to the fact that $\|(\hat{b}_m\hat{a}_m^{(u)})_{\text{MN}}\| < \|z_m^{(u)}\|$. Consequently, we conclude that, in this case, the amplitude of the positive-part JS type shrinkage estimator gives the lower bound of the amplitude of the VB estimator, i.e.,

$$\|(\hat{b}_m\hat{a}_m^{(u)})_{\text{PJS}}\| < \|(\hat{b}_m\hat{a}_m^{(u)})_{\text{VB}}\|, \quad (37)$$

because $\|\mathcal{S}(z;\chi)\|$ is an increasing function of $\|z\|$. However, if there is any pair of v_m and $v_{m'}$ that are neither orthogonal nor parallel to each other, neither the asymptotic equivalence between the VB solution and the shrinkage estimator nor Inequality (37) necessarily hold. Further consideration is future work.

6 Discussion

6.1 Features

The authors of [4] compared their approach with a previous work, the MAP estimation or the Wiener filter method with inaccurate prior information, where the hyperparameter, B^{-2} in Eq.(5), is regarded as a constant. Consider the situation, with which all the model selection and the regularization methods have been proposed to cope, when we do not accurately know the true prior and may use a model with irrelevant elements or redundant degree of freedom. Because the MAP estimation causes no singularity, it provides the generalization performance asymptotically equivalent to that of the regular models. On the other hand, because an LNN is singular, it provides different generalization performance even in the asymptotic limit [6]. The case in this paper corresponds to the case of a single-output (SO) LNN, i.e., an LNN with one output unit and one hidden unit, with U input units. (See Appendix A.) Because it has been shown that the VB approach asymptotically dominates the ML, as well as the MAP, estimation in SOLNNs when U is sufficiently large [9],[7] we expect that the ARD will provide better performance than the MAP estimation. Moreover, in SOLNNs, the suppression of overfitting caused by the singularities is stronger in the VB approach than in the Bayes estimation, which means that the elimination of irrelevant elements is enhanced in the VB approach. In addition, note that the time period U significantly affects performance because the degree of shrinkage, χ, is proportional to U, as we find from Eq.(26).

6.2 Proposition

We propose to simply use the positive-part JS type shrinkage estimator, Eq.(35), based on the MNML estimator. It only requires the calculation of the Moore-Penrose generalized inverse like the MAP estimation; while it is expected to eliminate irrelevant elements to suppress overfitting like the VB approach, which has been shown to provide better performance than the MAP estimation [4] and requires relatively costly iterative calculation. If the noise variance, σ_y^2 in Eq.(35), is unknown, its ML estimator should be substituted for it.

Note that the shrinkage estimation, as well as the VB approach, is not coordinate-invariant unlike the ML estimator, and there is a difference between the shrinkage estimation and the VB solution in nonorthogonal cases, as shown in Section 5.3. Although Inequality (37) states that, in a special case, ill-posedness makes the elimination effect of the shrinkage estimation stronger than that of the VB approach, the discussion in Section 5.3 also seems to imply that the VB approach can be less affected by the nonorthogonality, and hence more desirable than the shrinkage estimation. Further analysis is future work.

7 Conclusions and Future Work

In this paper, noting the similarity between the automatic relevance determination model (ARD) in a linear problem and a linear neural network model, we have found the relation between the variational Bayes (VB) approach in the ARD and a positive-part James-Stein (JS) type shrinkage estimation. It has let us propose to use the shrinkage estimation as an alternative, which requires less costs and behaves like the VB approach.

The relation between the empirical Bayes (EB) approach in a linear model and the JS estimation was previously discussed in [10], where the JS estimator was derived as an EB estimator. We have recently pointed out the equivalence between the EB approach in a linear model and a subspace Bayes (SB) approach in a single-output LNN [9], and found the asymptotic equivalence between the VB and the SB approaches in LNNs [6]. The previous works above and this paper have revealed the similarity between the VB approach and the shrinkage estimation. But in this paper, it has also been found that the nonorthogonality of the basis makes a difference, on which we will focus from now. Consideration of what our simplification, i.e., the differences in setting between in [4] and in this paper, itemized at the end of Section 4.2, causes is also future work.

[7] It was conjectured that, in SOLNNs, the VB approach asymptotically dominates the ML estimation when $U \geq 5$ [9].

Acknowledgments

The authors would like to thank Dr. Okito Yamashita of ATR, Japan for the discussion on [4], which produced the motivation of this work.

A Definition of Linear Neural Networks

Let $x \in \mathbb{R}^M$ be an input vector, $y \in \mathbb{R}^N$ an output vector, and w a parameter vector. Assume that the output is observed with a noise subject to $\mathcal{N}_N(0, \Sigma)$. Then, the probability density of a three-layer linear neural network model (LNN) with H hidden units, also known as the reduced rank regression model with rank H, is given by

$$p(y|x, A, B) = \mathcal{N}_N(BAx, \Sigma), \tag{38}$$

where A and B are an $H \times M$ and an $N \times H$ parameter matrices, respectively. It has been proved that, in LNNs, the VB approach is asymptotically equivalent to a positive-part James-Stein type shrinkage estimation, and its generalization error has been clarified [6].

B James-Stein type Shrinkage Estimator

A positive-part James-Stein type shrinkage estimator [5, 10], which can dominate the maximum likelihood (ML) estimator, of the parameter w is defined by

$$\hat{w}_{PJS} = \theta(n\|\hat{w}_{MLE}\|^2 > \chi)\left(1 - \chi/n\|\hat{w}_{MLE}\|^2\right)\hat{w}_{MLE} \equiv \mathcal{S}(\hat{w}_{MLE}; \chi), \tag{39}$$

where $\hat{w}_{MLE}$ is the ML estimator, $\theta(\cdot)$ is the indicator function of an event, and $\chi > 0$ is a constant, called the degree of shrinkage in this paper.

References

1. Hinton, G.E., van Camp, D.: Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. In: Proc. of COLT. (1993) 5–13
2. Attias, H.: Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In: Proc. of UAI. (1999)
3. Neal, R.M.: Bayesian Learning for Neural Networks. Springer (1996)
4. Sato, M., Yoshioka, T., Kajihara, S., Toyama, K., Goda, N., Doya, K., Kawato, M.: Hierarchical Bayesian Estimation for MEG inverse problem. Neuro Image **23** (2004) 806–826
5. James, W., Stein, C.: Estimation with Quadratic Loss. In: Proc. of the 4th Berkeley Symp. on Math. Stat. and Prob. (1961) 361–379
6. Nakajima, S., Watanabe, S.: Generalization Error and Free Energy of Variational Bayes Approach of Linear Neural Networks. In: Proc. of ICONIP, Taipei, Taiwan (2005) 55–60
7. Callen, H.B.: Thermodynamics. Wiley (1960)
8. Hamalainen, M., Hari, R., Ilmoniemi, R.J., Knuutila, J., Lounasmaa, O.V.: Magnetoencephalography — Theory, Instrumentation, and Applications to Noninvasive Studies of the Working Human Brain. Rev. Modern Phys. **65** (1993) 413–497
9. Nakajima, S., Watanabe, S.: Generalization Performance of Subspace Bayes Approach in Linear Neural Networks. IEICE Trans. **E89-D** (2006) 1128–1138
10. Efron, B., Morris, C.: Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach. J. of Am. Stat. Assoc. **68** (1973) 117–130