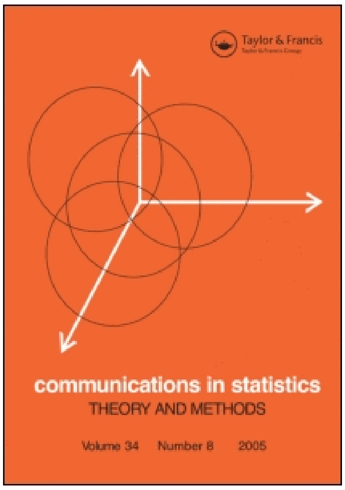

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

A Bayesian Learning Coefficient of Generalization Error and Vandermonde Matrix-Type Singularities

Miki Aoyagi[a]
[a] Advanced Research Institute for the Sciences and Humanities, Nihon University, Tokyo, Japan

Online publication date: 28 July 2010

To cite this Article Aoyagi, Miki(2010) 'A Bayesian Learning Coefficient of Generalization Error and Vandermonde Matrix-Type Singularities', Communications in Statistics - Theory and Methods, 39: 15, 2667 — 2687

To link to this Article: DOI: 10.1080/03610920903094899

URL: http://dx.doi.org/10.1080/03610920903094899

Communications in Statistics—Theory and Methods, 39: 2667–2687, 2010
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920903094899

A Bayesian Learning Coefficient of Generalization Error and Vandermonde Matrix-Type Singularities

MIKI AOYAGI

Advanced Research Institute for the Sciences and Humanities,
Nihon University, Tokyo, Japan

The coefficient of the main term of the generalization error in Bayesian estimation is called a Bayesian learning coefficient. In this article, we first introduce Vandermonde matrix type singularities and show certain orthogonality conditions of them. Recently, it has been recognized that Vandermonde matrix type singularities are related to Bayesian learning coefficients for several hierarchical learning models. By applying the orthogonality conditions of them, we show that their log canonical threshold also corresponds to the Bayesian learning coefficient for normal mixture models, and we obtain the explicit computational results in dimension one.

Keywords Generalization error; Hierarchical learning models; Normal mixture models; Resolution of singularities; Zeta function.

Mathematics Subject Classification 62D05; 62M20; 32S10; 14Q15.

1. Introduction

The theoretical study of hierarchical learning models has been rapidly developed in recent years. The data analyzed by such learning models are associated with image or speech recognition, artificial intelligence, the control of a robot, genetic analysis, data mining, time series prediction, and so on. They are very complicated and not usually generated by a simple normal distribution, as they are influenced by many factors. Hierarchical learning models such as the normal mixture model, the Boltzmann machine, layered neural network, and reduced rank regression may be known as effective learning models. They, however, likewise have complicated, i.e., non regular statistical structures, which cannot be analyzed using the classic theories of regular statistical models (Fukumizu, 1996; Hartigan, 1985; Hagiwara et al., 1993; Sussmann, 1992). The theoretical study has therefore been started to construct a mathematical foundation for non regular statistical models.

Watanabe (2001a,b) proved that the largest pole of a zeta function for a non regular statistical model gives the main term of the generalization error of

Received December 11, 2008; Accepted June 1, 2009

Address correspondence to Miki Aoyagi, Department of Mathematics, College of Science and Technology, Nihon University, Tokyo, Japan; E-mail: aoyagi.miki@nihon-u.ac.jp

hierarchical learning models in Bayesian estimation. The generalization error of a learning model is a difference between a true density function and a predictive density function obtained using distributed training samples. It is one of the most important topic in learning theory. The largest pole of a zeta function for a learning model, which is called a Bayesian learning coefficient, corresponds to the log canonical threshold in algebraic geometry. The log canonical threshold $\lambda_Z(Y, f)$ over the real field is analytically defined by

$$\lambda_Z(Y, f) = \sup\{c : |f|^{-c} \text{ is locally } L^1 \text{ near } Z\},$$

for a non zero regular f on a smooth variety Y, where $Z \subset Y$ is a closed subscheme.

In spite of these mathematical foundations, obtaining their largest pole, i.e., their log canonical threshold is still difficult for several reasons such that degeneration with respect to their Newton polyhedrons and non-isolation of their singularities (Fulton, 1993). Moreover, in algebraic geometry and algebraic analysis, these studies are usually done over an algebraically closed field (Kollár, 1997; Mustata, 2002). We have many differences between the real field and the complex field, for example, log canonical thresholds over the complex field are less than 1, while those over the real field are not necessarily less than 1. We cannot therefore apply results over an algebraically closed field to our cases, directly.

In this article, we first introduce Vandermonde matrix type singularities (Definition 3.3) and next show certain orthogonality conditions of their log canonical threshold (Theorem 3.1). We then show that the theorem enables us to connect Bayesian learning coefficients of normal mixture models with Vandermonde matrix-type singularities. By applying such results, we obtain explicitly the coefficients of normal mixture models with unit matrix variances in dimension one. Yamazaki and Watanabe obtained only upper bounds of these values (Yamazaki and Watanabe, 2003).

In the past few years, we have also obtained Bayesian learning coefficients for the three layered neural network (Aoyagi and Watanabe, 2005a; Aoyagi, 2006) and for the reduced rank regression (Aoyagi and Watanabe, 2005b). Rusakov and Geiger (2005) obtained them for Naive Bayesian networks.

This article consists of four sections. In Sec. 2, we summarize the framework of Bayesian learning theory. In Sec. 3, we state our main results. Our conclusion is given in Sec. 4.

2. Bayesian Learning Theory

In this article, we overview Bayesian learning theory, especially the stochastic complexity and the generalization error.

It is well known that Bayesian estimation is more appropriate than the maximum likelihood method when a learning machine is non-regular (Akaike, 1980; Mackay, 1992).

Let $q(x)$ be a true probability density function and $(x)^n := \{x_i\}_{i=1}^n$ be n training independent and identical samples from $q(x)$. Consider a learning model which is written by a probability form $p(x \mid w)$, where w is a parameter. The purpose of the learning system is to estimate $q(x)$ from $(x)^n$ by using $p(x \mid w)$.

Let $p(w \mid (x)^n)$ be the *a posteriori* probability density function:

$$p(w \mid (x)^n) = \frac{1}{Z_n}\psi(w)\prod_{i=1}^{n} p(x_i \mid w),$$

where $\psi(w)$ is an *a priori* probability density function on the parameter set W and

$$Z_n = \int_W \psi(w)\prod_{i=1}^{n} p(x_i \mid w)\mathrm{d}w.$$

So the average inference $p(x \mid (x)^n)$ of the Bayesian density function is given by

$$p(x \mid (x)^n) = \int p(x \mid w)p(w \mid (x)^n)\mathrm{d}w,$$

which is the predictive density function.

Set

$$K(q \,||\, p) = \int q(x)\log\frac{q(x)}{p(x \mid (x)^n)}\mathrm{d}x.$$

This is always a positive value and satisfies $K(q \,||\, p) = 0$ if and only if $q(x) = p(x \mid (x)^n)$.

The generalization error $G(n)$ is its expectation value E_n over n training samples:

$$G(n) = E_n\left\{\int q(x)\log\frac{q(x)}{p(x \mid (x)^n)}\mathrm{d}x\right\}.$$

Let

$$K_n(w) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{q(x_i)}{p(x_i \mid w)}.$$

The average stochastic complexity or the free energy is defined by

$$F(n) = -E_n\left\{\log\int \exp(-nK_n(w))\psi(w)\mathrm{d}w\right\}.$$

Then we have $G(n) = F(n+1) - F(n)$ for an arbitrary natural number n (Amari et al., 1992; Amari and Murata, 1993; Levin et al., 1990). $F(n)$ is known as the Bayesian criterion in Bayesian model selection (Schwarz, 1978), stochastic complexity in universal coding (Rissanen, 1986; Yamanishi, 1998), Akaike's Bayesian criterion in optimization of hyperparameters (Akaike, 1980), and evidence in neural network learning (Mackay, 1992). In addition, $F(n)$ is an important function for analyzing the generalization error.

It has recently been proved that the largest pole of a zeta function gives the generalization error of hierarchical learning models asymptotically (Watanabe, 2001a,b). We assume that the true density distribution $q(x)$ is included in the learning model, i.e., $q(x) = p(x \mid w_t^*)$ for $w_t^* \in W$, where W is the parameter space.

Define the zeta function $J(z)$ of a complex variable z for the learning model by

$$J(z) = \int K(w)^z \psi(w) \mathrm{d}w,$$

where $K(w)$ is the Kullback function:

$$K(w) = \int p(x \mid w_t^*) \log \frac{p(x \mid w_t^*)}{p(x \mid w)} \mathrm{d}x.$$

Then, for the largest pole $-\lambda$ of $J(z)$ and its order θ, we have

$$F(n) = \lambda \log n - (\theta - 1) \log \log n + O(1), \tag{1}$$

where $O(1)$ is a bounded function of n, and if $G(n)$ has an asymptotic expansion,

$$G(n) \cong \frac{\lambda}{n} - \frac{\theta - 1}{n \log n} \quad \text{as } n \to \infty. \tag{2}$$

Therefore, our aim is to obtain λ and θ.

Note that for $Z = \{w : K(w) = 0\}$, $\lambda = \lambda_Z(W, K(w)) = \sup\{c : |K|^{-c}$ is locally L^1 near $Z\}$, which is the log canonical threshold of $K(w)$.

To assist in achieving this aim, we introduce Hironaka's Theorem.

Theorem 2.1 (Desingularization, Hironaka, 1964). *Let f be a real analytic function in a neighborhood of $w = (w_1, \ldots, w_d) \in \mathbb{R}^d$ with $f(w) = 0$. There exist an open set $V \ni w$, a real analytic manifold U, and a proper analytic map μ from U to V such that*:

(1) $\mu : U - \mathscr{E} \to V - f^{-1}(0)$ *is an isomorphism, where* $\mathscr{E} = \mu^{-1}(f^{-1}(0))$;
(2) *For each $u \in U$, there is a local analytic coordinate system $(u_1, \ldots, u_d)$ such that $f(\mu(u)) = \pm u_1^{s_1} u_2^{s_2} \ldots u_d^{s_d}$, where $s_1, \ldots, s_d$ are non negative integers.*

Applying Hironaka's theorem to the Kullback function $K(w)$, for each $w \in K^{-1}(0) \cap W$, we have a proper analytic map μ_w from an analytic manifold U_w to a neighborhood V_w of w satisfying Hironaka's Theorems (1) and (2). Then the local integration on V_w of the zeta function $J(z)$ of the learning model is

$$\begin{aligned} J_w(z) &= \int_{V_w} K(w)^z \psi(w) \mathrm{d}w \\ &= \int_{U_w} \sum_u (u_1^{2s_1} u_2^{2s_2} \ldots u_d^{2s_d})^z \psi(\mu_w(u)) |\mu_w'(u)| \mathrm{d}u. \end{aligned}$$

Therefore, the poles of $J_w(z)$ can be obtained. For each $w \in W \setminus K^{-1}(0)$, there exists a neighborhood V_w such that $K(w') \neq 0$, for all $w' \in V_w$. So $J_w(z) = \int_{V_w} K(w)^z \psi(w) dw$ has no poles. It is known that μ of an arbitrary polynomial in Hironaka's Theorem can be obtained by using a blowing up process.

3. Main Result

In this article, we denote by a^*, b^* constants and denote by a^* if the variable a is in a sufficiently small neighborhood of a^*.

Define the norm of a matrix $C = (c_{ij})$ by $\|C\| = \sqrt{\sum_{i,j} |c_{ij}|^2}$. Denote by $\langle C \rangle$ the ideal generated by $\{c_{ij}\}$. Set $\mathbb{N}_{+0} = \mathbb{N} \cup \{0\}$.

Definition 3.1. Let f be a real analytic function defined near w^*.
Set $\lambda_{w^*}(f) = \sup\{c : |f|^{-c}$ is locally L^1 near $w^*\}$.

3.1. *Vandermonde Matrix Type Singularities*

Definition 3.2. Fix $Q \in \mathbb{N}$. Define $[b_1^*, b_2^*, \ldots, b_N^*]_Q = \gamma_i(0, \ldots, 0, b_i^*, \ldots, b_N^*)$ if $b_1^* = \cdots = b_{i-1}^* = 0$, $b_i^* \neq 0$, and $\gamma_i = \begin{cases} 1 & \text{if } Q \text{ is odd} \\ |b_i^*|/b_i^* & \text{if } Q \text{ is even.} \end{cases}$

Definition 3.3. Fix $Q \in \mathbb{N}$ and $m \in \mathbb{N}_{+0}$.
Let $MH + HN$ variables

$$w = \left\{ \begin{pmatrix} a_{11} & \cdots & a_{1H} \\ a_{21} & \cdots & a_{2H} \\ & \vdots & \\ a_{M1} & \cdots & a_{MH} \end{pmatrix}, \begin{pmatrix} b_{11} & \cdots & b_{1N} \\ b_{21} & \cdots & b_{2N} \\ & \vdots & \\ b_{H1} & \cdots & b_{HN} \end{pmatrix} \right\}$$

and $rM + rN$ constants

$$w_t^* = \left\{ \begin{pmatrix} a^*_{1,H+1} & \cdots & a^*_{1,H+r} \\ a^*_{2,H+1} & \cdots & a^*_{2,H+r} \\ & \vdots & \\ a^*_{M,H+1} & \cdots & a^*_{M,H+r} \end{pmatrix}, \begin{pmatrix} b^*_{H+1,1} & \cdots & b^*_{H+1,N} \\ b^*_{H+2,1} & \cdots & b^*_{H+2,N} \\ & \vdots & \\ b^*_{H+r,1} & \cdots & b^*_{H+r,N} \end{pmatrix} \right\}.$$

Let

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1H} & a^*_{1,H+1} & \cdots & a^*_{1,H+r} \\ a_{21} & \cdots & a_{2H} & a^*_{2,H+1} & \cdots & a^*_{2,H+r} \\ & \vdots & & & \vdots & \\ a_{M1} & \cdots & a_{MH} & a^*_{M,H+1} & \cdots & a^*_{M,H+r} \end{pmatrix}, \quad I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N,$$

$$B_I = \left(\prod_{j=1}^N b_{1j}^{\ell_j}, \prod_{j=1}^N b_{2j}^{\ell_j}, \ldots, \prod_{j=1}^N b_{Hj}^{\ell_j}, \prod_{j=1}^N {b^*_{H+1,j}}^{\ell_j}, \ldots, \prod_{j=1}^N {b^*_{H+r,j}}^{\ell_j} \right)^t$$

and $B = (B_I)_{\ell_1+\cdots+\ell_N = Qn+m, 0 \leq n \leq H+r-1}$ (t denotes the transpose) , where A is an $M \times (H+r)$ dimensional matrix and B is an $(H+r) \times \sum_{n=0}^{H+r-1} \frac{(Qn+m+N-1)!(Qn+m)!}{(N-1)!}$ dimensional matrix.

We call singularities of $\|AB\|^2 = 0$ Vandermonde matrix type singularities.

To simplify, we usually assume that

$$(a^*_{1,H+j}, a^*_{2,H+j}, \ldots, a^*_{M,H+j})^t \neq 0, \quad (b^*_{H+j,1}, b^*_{H+j,2}, \ldots, b^*_{H+j,N}) \neq 0$$

for $1 \leq j \leq r$ and

$$[b^*_{H+j,1}, b^*_{H+j,2}, \ldots, b^*_{H+j,N}]_Q \neq [b^*_{H+j',1}, b^*_{H+j',2}, \ldots, b^*_{H+j',N}]_Q$$

for $j \neq j'$.

Let w, w^*_t, A, and B be as in Definition 3.3. Let w be in a sufficiently small neighborhood of

$$w^* = \left\{ \begin{pmatrix} a^*_{11} & \cdots & a^*_{1H} \\ a^*_{21} & \cdots & a^*_{2H} \\ & \vdots & \\ a^*_{M1} & \cdots & a^*_{MH} \end{pmatrix}, \begin{pmatrix} b^*_{11} & \cdots & b^*_{1N} \\ b^*_{21} & \cdots & b^*_{2N} \\ & \vdots & \\ b^*_{H1} & \cdots & b^*_{HN} \end{pmatrix} \right\}.$$

Set $(b^{**}_{01}, b^{**}_{02}, \ldots, b^{**}_{0N}) = (0, \ldots, 0)$.

Let each $(b^{**}_{11}, b^{**}_{12}, \ldots, b^{**}_{1N})$, …, $(b^{**}_{r'1}, b^{**}_{r'2}, \ldots, b^{**}_{r'N})$ be a different real vector in

$$\begin{aligned} &[b^*_{i1}, b^*_{i2}, \ldots, b^*_{iN}]_Q \neq 0, \quad \text{for } i = 1, \ldots, H + r : \\ &\quad \times \{(b^{**}_{11}, \ldots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, \ldots, b^{**}_{r'N}); [b^*_{i1}, \ldots, b^*_{iN}]_Q \neq 0, i = 1, \ldots, H + r\}. \end{aligned}$$

Then $r' \geq r$ and set $(b^{**}_{i1}, \ldots, b^{**}_{iN}) = [b^*_{H+i,1}, \ldots, b^*_{H+i,N}]_Q$, for $1 \leq i \leq r$.

It is natural to assume that

$$\left.\begin{array}{c} [b^*_{11}, \ldots, b^*_{1N}]_Q \\ \vdots \\ [b^*_{H_0 1}, \ldots, b^*_{H_0 N}]_Q \end{array}\right\} = 0,$$

$$\left.\begin{array}{c} [b^*_{H_0+1,1}, \ldots, b^*_{H_0+1,N}]_Q \\ \vdots \\ [b^*_{H_0+H_1,1}, \ldots, b^*_{H_0+H_1,N}]_Q \end{array}\right\} = (b^{**}_{11}, \ldots, b^{**}_{1N}),$$

$$\left.\begin{array}{c} [b^*_{H_0+H_1+1,1}, \ldots, b^*_{H_0+H_1+1,N}]_Q \\ \vdots \\ [b^*_{H_0+H_1+H_2,1}, \ldots, b^*_{H_0+H_1+H_2,N}]_Q \end{array}\right\} = (b^{**}_{21}, \ldots, b^{**}_{2N}),$$

$$\vdots$$

$$\left.\begin{array}{c} [b^*_{H_0+\cdots+H_{r'-1}+1,1}, \ldots, b^*_{H_0+\cdots+H_{r'-1}+1,N}]_Q \\ \vdots \\ [b^*_{H_0+\cdots+H_{r'-1}+H_{r'},1}, \ldots, b^*_{H_0+\cdots+H_{r'-1}+H_{r'},N}]_Q \end{array}\right\} = (b^{**}_{r'1}, \ldots, b^{**}_{r'N}).$$

and $H_0 + \cdots + H_{r'} = H$.

Theorem 3.1. *We have*

$$\lambda_{w^*}(\|AB\|^2) = \sum_{\alpha=0}^{r'} \lambda_{w^{(\alpha)*}}(\|A^{(\alpha)}B^{(\alpha)}\|^2),$$

where $w^{(\alpha)*} = \left\{a^{(\alpha)^*}_{ki}, b^{(\alpha)^*}_{ij}\right\} = \left\{a^*_{k,H_0+\cdots+H_{\alpha-1}+i}, b^{**}_{\alpha j}\right\}_{1\leq k\leq M, 1\leq i\leq H_\alpha, 1\leq j\leq N}$,

$$I = (\ell_1, \dots, \ell_N) \in {\mathbb{N}_{+0}}^N,$$

$$A^{(\alpha)} = \begin{pmatrix} a_{11}^{(\alpha)} & a_{12}^{(\alpha)} & \cdots & a_{1H_\alpha}^{(\alpha)} \\ a_{21}^{(\alpha)} & a_{22}^{(\alpha)} & \cdots & a_{2H_\alpha}^{(\alpha)} \\ & & \vdots & \\ a_{M1}^{(\alpha)} & a_{M2}^{(\alpha)} & \cdots & a_{MH_\alpha}^{(\alpha)} \end{pmatrix}, \quad B_I^{(\alpha)} = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{(\alpha)\ell_j} \\ \prod_{j=1}^N b_{2j}^{(\alpha)\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H_\alpha j}^{(\alpha)\ \ell_j} \end{pmatrix}, \quad \textit{for } \alpha = 0,\ \ r+1 \le \alpha \le r',$$

$$A^{(\alpha)} = \begin{pmatrix} a_{11}^{(\alpha)} & a_{12}^{(\alpha)} & \cdots & a_{1H_\alpha}^{(\alpha)} & a_{1,H+\alpha}^{*} \\ a_{21}^{(\alpha)} & a_{22}^{(\alpha)} & \cdots & a_{2H_\alpha}^{(\alpha)} & a_{2,H+\alpha}^{*} \\ & & \vdots & & \\ a_{M1}^{(\alpha)} & a_{M2}^{(\alpha)} & \cdots & a_{MH_\alpha}^{(\alpha)} & a_{M,H+\alpha}^{*} \end{pmatrix}, \quad B_I^{(\alpha)} = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{(\alpha)\ell_j} \\ \prod_{j=1}^N b_{2j}^{(\alpha)\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{H_\alpha j}^{(\alpha)\ \ell_j} \\ \prod_{j=1}^N b_{\alpha j}^{**\ell_j} \end{pmatrix}, \quad \textit{for } 1 \le \alpha \le r,$$

$B^{(0)} = (B_I^{(0)})_{\ell_1+\cdots+\ell_N=Qn+m, 0\le n\le H_0-1}$, $B^{(\alpha)} = (B_I^{(\alpha)})_{\ell_1+\cdots+\ell_N=n, 0\le n\le H_\alpha-1}$ *for* $r+1 \le \alpha \le r'$ *and* $B^{(\alpha)} = (B_I^{(\alpha)})_{\ell_1+\cdots+\ell_N=n, 0\le n\le H_\alpha}$ *for* $1 \le \alpha \le r$.

$B^{(0)}$, $B^{(\alpha)}(1 \le \alpha \le r)$ *and* $B^{(\alpha)}(r+1 \le \alpha \le r')$ *are* $H_0 \times \sum_{n=0}^{H_0-1} \frac{(Qn+m+N-1)!(Qn+m)!}{(N-1)!}$, $(H_\alpha + 1) \times \sum_{n=0}^{H_\alpha} \frac{(n+N-1)!n!}{(N-1)!}$ *and* $H_\alpha \times \sum_{n=0}^{H_\alpha-1} \frac{(n+N-1)!n!}{(N-1)!}$ *dimensional matrices, respectively.*

Proof. Set

$$\begin{cases} (a_{i1}^{(0)}, \dots, a_{iH_0}^{(0)}) = (a_{i1}, \dots, a_{iH_0}), \\ (a_{i1}^{(1)}, \dots, a_{iH_1}^{(1)}) = (a_{i,H_0+1}, \dots, a_{i,H_0+H_1}), \\ \quad \vdots \\ (a_{i1}^{(r')}, \dots, a_{iH_{r'}}^{(r')}) = (a_{i,H_0+\cdots+H_{r'-1}+1}, \dots, a_{i,H_0+\cdots+H_{r'}}), \end{cases} \qquad \text{for } 1 \le i \le M, \text{ and}$$

$$\begin{cases} (b_{1j}^{(0)}, \dots, b_{H_0 j}^{(0)}) = (b_{1j}, \dots, b_{H_0 j}), \\ (b_{1j}^{(1)}, \dots, b_{H_1 j}^{(1)}) = (b_{H_0+1,j}, \dots, b_{H_0+H_1,j}), \\ \quad \vdots \\ (b_{1j}^{(r')}, \dots, b_{H_{r'} j}^{(r')}) = (b_{H_0+\cdots+H_{r'-1}+1,j}, \dots, b_{H_0+\cdots+H_{r'},j}), \end{cases} \qquad \text{for } 1 \le j \le N.$$

For $\gamma_i(b_{i1}^{(\alpha)}, \dots, b_{iN}^{(\alpha)}) = [b_{i1}^{(\alpha)}, \dots, b_{iN}^{(\alpha)}]_Q$, we again set $a_{ki}^{(\alpha)}$ by $a_{ki}^{(\alpha)}/(\gamma_i)^m$ and $b_{ij}^{(\alpha)}$ by $b_{ij}^{(\alpha)}\gamma_i$, $1 \le j \le N$ and $1 \le k \le M$.

Main parts of the proof is appeared in Appendix 1. By applying Lemma A.4 in Appendix 1, we have this theorem.

Theorem 3.1 shows a kind of an orthogonal relation of the log canonical threshold of Vandermonde matrix type singularities. Usually, r corresponds to the number of elements of a true distribution. It means that the Bayesian learning coefficient related with such singularities is the sum of each for the small model with respect to each element of a true distribution (cf. Sec. 3.2).

Theorem 3.2. *We use the same notations as in Theorem* 3.1. *If* $N = 1$, *we have*:

$$\lambda_{w^*}(\|AB\|^2) = \frac{MQk_0(k_0+1)+2H_0}{4(m+k_0Q)} + \frac{Mr'}{2} + \sum_{\alpha=1}^{r} \frac{Mk_\alpha(k_\alpha+1)+2H_\alpha}{4(1+k_\alpha)}$$

$$+ \sum_{\alpha=r+1}^{r'} \frac{Mk_\alpha(k_\alpha+1)+2(H_\alpha-1)}{4(1+k_\alpha)},$$

where

$$\begin{aligned} k_0 &= \max\{i \in \mathbb{Z};\, 2H_0 \geq M(i(i-1)Q + 2mi)\}, \\ k_\alpha &= \max\{i \in \mathbb{Z};\, 2H_\alpha \geq M(i^2 + i)\}, \\ k_\alpha &= \max\{i \in \mathbb{Z};\, 2(H_\alpha - 1) \geq M(i^2 + i)\}. \end{aligned}$$

For the proof of Theorem 3.2, we use a similar method in the articles (Aoyagi, 2006; Aoyagi and Watanabe, 2005a), where we used recursive blowing ups and toric resolution.

The key point is that $\lambda_0(\|AB\|^2) = \lambda_0(\|AB'\|^2)$ for $N = 1$, where

$$B' = \begin{pmatrix} b_1^m & 0 & 0 & \cdots & 0 \\ 0 & b_2^m(b_2^Q - b_1^Q) & 0 & \cdots & 0 \\ 0 & 0 & b_3^m(b_3^Q - b_1^Q)(b_3^Q - b_2^Q) & \cdots & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & \cdots & b_H^m(b_H^Q - b_1^Q)\cdots(b_H^Q - b_{H-1}^Q) \end{pmatrix},$$

and $|b_H| < |b_{H-1}| < \cdots < |b_2| < |b_1|$.

Recently, we have the explicit values $\lambda_{w^*}(\|AB\|^2)$ for general natural numbers N and M but for $H \leq 2$ (Aoyagi and Nagata, 2008).

3.2. *Normal Mixture Model*

We consider a normal mixture model with unit matrix variances

$$p(x \mid w) = \frac{1}{(2\pi)^{N/2}} \sum_{i=1}^{H} a_i \exp\left(-\frac{\sum_{j=1}^{N}(x_j - b_{ij})^2}{2}\right),$$

where $w = \{a_i, b_{ij}; 1 \leq i \leq H\}, 1 \leq j \leq N\}$ and $\sum_{i=1}^{H} a_i = 1,\ a_i \geq 0$.

Set the true distribution by

$$p(x \mid w_t^*) = \frac{-1}{(2\pi)^{N/2}} \sum_{i=H+1}^{H+r} a_i^* \exp\left(-\frac{\sum_{j=1}^{N}(x_j - b_{ij}^*)^2}{2}\right),$$

where $w_t^* = \{a_i^*, b_{ij}^*; H+1 \leq i \leq H+r, 1 \leq j \leq N\}$ and $\sum_{i=H+1}^{H+r} a_i^* = -1,\ a_i^* < 0$. (In order to simplify the followings, we use the values $a_i^* < 0$ not $a_i^* > 0$.) Suppose that an *a priori* probability density function $\psi(w)$ is a C^∞-function with a compact support W where $\psi(w_t^*) > 0$.

Let $A = (a_1, \ldots, a_H, a^*_{H+1}, \ldots, a^*_{H+r})$, $I = (\ell_1, \ldots, \ell_N) \in \mathbb{N}_{+0}{}^N$,

$$B_I = \left(\prod_{j=1}^{N} b_{1j}^{\ell_j}, \prod_{j=1}^{N} b_{2j}^{\ell_j}, \ldots, \prod_{j=1}^{N} b_{Hj}^{\ell_j}, \prod_{j=1}^{N} b^*_{H+1,j}{}^{\ell_j}, \ldots, \prod_{j=1}^{N} b^*_{H+r,j}{}^{\ell_j} \right)^t$$

and $B = (B_I)_{\ell_1+\cdots+\ell_N=n, 1\le n\le H+r}$ (t denotes the transpose).

Then the Bayesian learning coefficient of the normal mixture model is the largest pole of

$$\int \Psi = \int_{\|AB\|^2<1} \|AB\|^{2z} \prod_{i=1}^{H} \mathrm{d}a_i \prod_{i=1}^{H}\prod_{j=1}^{N} \mathrm{d}b_{ij}, \tag{3}$$

with $\sum_{j=1}^H a_j = 1$, $a_i \ge 0$ and $\sum_{j=H+1}^{H+r} a_j^* = -1$, $a_j^* < 0$ (Watanabe et al., 2004).

Note that we have the relations $\sum_{j=1}^H a_j = 1$, $a_i \ge 0$ and $\sum_{j=H+1}^{H+r} a_j^* = -1$, $a_j^* < 0$. We need to modify the function $\|AB\|^2$ for obtaining the largest pole of $\int \Psi$ by using Vandermonde matrix type singularities. The following theorem is available for such purpose in dimension 1.

Theorem 3.3. *Let w be in a sufficiently small neighborhood of $w^* = \{a_i^*, b_i^*\}_{1\le i\le H}$.*

Let each $b_1^{**}, \ldots, b_{r'}^{**}$ be a different real number in $\{b_i^* : i = 1, \ldots, H + r\}$:

$$\{b_1^{**}, \ldots, b_{r'}^{**}\} = \{b_i^* : i = 1, \ldots, H + r\}.$$

Then $r' \ge r$ and set $b_i^{**} = b^*_{H+i,1}$, for $1 \le i \le r$.

Assume that

$$\begin{aligned} b_1^* &= \cdots = b^*_{H_1} = b_1^{**}, b^*_{H_1+1} = \cdots = b^*_{H_1+H_2} = b_2^{**}, \ldots, b^*_{H_1+\cdots+H_{r'-1}+1} \\ &= \cdots = b^*_{H_1+\cdots+H_{r'-1}+H_{r'}} = b_{r'}^{**} \quad \text{and} \quad H_1 + \cdots + H_{r'} = H. \end{aligned}$$

Then we have

$$\lambda_{w^*}(\|AB\|^2) = \sum_{\alpha=1}^{r'-1} \lambda_{w_1^{(\alpha)*}}\big(a_1^{(\alpha)2}\big) + \sum_{\alpha=1}^{r'} \lambda_{w^{(\alpha)*}}(\|A^{(\alpha)}B^{(\alpha)}\|^2),$$

where $w_1^{(\alpha)*} = \begin{cases} a^*_{H_1+\cdots+H_{\alpha-1}+1} + \cdots + a^*_{H_1+\cdots+H_\alpha} + a^*_{H+\alpha}, & 1 \le \alpha \le r, \\ a^*_{H_1+\cdots+H_{\alpha-1}+1} + \cdots + a^*_{H_1+\cdots+H_\alpha}, & r+1 \le \alpha \le r'-1, \end{cases}$

$$w^{(\alpha)*} = \{a_i^{(\alpha)*}, b_i^{(\alpha)*}\}_{2\le i\le H_\alpha} = \{a^*{}_{H_1+\cdots+H_{\alpha-1}+i}, 0\}_{2\le i\le H_\alpha},$$

$$A^{(\alpha)} = \big(a_2^{(\alpha)}, a_3^{(\alpha)}, \ldots, a_{H_\alpha}^{(\alpha)}, a^*_{H+\alpha}\big), \quad B^{(\alpha)} = \begin{pmatrix} b_1^{(\alpha)} & \cdots & b_1^{(\alpha)H_\alpha} \\ b_2^{(\alpha)} & \cdots & b_2^{(\alpha)H_\alpha} \\ & \vdots & \\ b_{H_\alpha}^{(\alpha)} & \cdots & b_{H_\alpha}^{(\alpha)H_\alpha} \end{pmatrix} \quad \textit{for } 1 \le \alpha \le r,$$

$$A^{(\alpha)} = \left(a_2^{(\alpha)}, a_3^{(\alpha)}, \ldots, a_{H_\alpha}^{(\alpha)}\right), \quad B^{(\alpha)} = \begin{pmatrix} b_1^{(\alpha)} & \cdots & b_1^{(\alpha)H_\alpha-1} \\ b_2^{(\alpha)} & \cdots & b_2^{(\alpha)H_\alpha-1} \\ & \vdots & \\ b_{H_\alpha-1}^{(\alpha)} & \cdots & b_{H_\alpha-1}^{(\alpha)\quad H_\alpha-1} \end{pmatrix} \quad \textit{for } r+1 \le \alpha \le r'.$$

The proof is shown in Appendix 2.

Theorem 3.4. *The average stochastic complexity $F(n)$ in Eq.* (1) *and the generalization error $G(n)$ in Eq.* (2) *are given by using the following largest pole $-\lambda$ of $J(z)$ and its order θ.*

If the true distribution has r peaks,

$$\lambda = r - 1 + \frac{n + n^2 + 2(H-(r-1))}{4(n+1)}, \quad \theta = \begin{cases} 1, & \textit{if } n^2 + n < 2(H-(r-1)), \\ 2, & \textit{if } n^2 + n = 2(H-(r-1)), \end{cases}$$

where $n = \max\{i \in \mathbb{Z}\ ;\ i^2 + i \le 2(H-(r-1))\}$.

Proof. By Theorems 3.2, 3.3 and the result in the article (Aoyagi, 2006; Aoyagi and Watanabe, 2005a), we have, for w^* with $\|AB\|_{w^*} = 0$ as in Theorem 3.3,

$$\lambda_{w^*} = \frac{r'-1}{2} + \sum_{\alpha=1}^{r} \frac{n_\alpha + n_\alpha^2 + 2H_\alpha}{4(n_\alpha+1)} + \sum_{\alpha=r+1}^{r'} \frac{n_\alpha + n_\alpha^2 + 2(H_\alpha - 1)}{4(n_\alpha+1)}, \quad \text{its order } \theta_{w^*} = \#\Theta + 1,$$

where $H_1 + \cdots + H_{r'} = H$, $n_\alpha = \max\{i \in \mathbb{Z}; i^2 + i \le 2H_\alpha\}$ for $1 \le \alpha \le r$, $n_\alpha = \max\{i \in \mathbb{Z}; i^2 + i \le 2(H_\alpha - 1)\}$ for $r+1 \le \alpha \le r'$, and $\Theta = \{n_\alpha; n_\alpha^2 + n_\alpha = 2H_\alpha\}$.

Some computations show that

$$\min\{\lambda_{w^*} : \|AB\|_{w^*} = 0\} = r - 1 + \frac{n + n^2 + 2(H-(r-1))}{4(n+1)},$$

where $n = \max\{i \in \mathbb{Z}; i^2 + i \le 2(H-(r-1))\}$, and so we have the theorem.

Example 3.1. Let $N = 1$ and $H = 5$, that is,

$$p(x \mid w) = \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{5} a_i \exp\left(-\frac{(x-b_i)^2}{2}\right),$$

where $\sum_{i=1}^{5} a_i = 1$, $a_i \ge 0$. Also, let the true distribution has two peaks ($r = 2$):

$$p(x \mid w_t^*) = -\frac{1}{\sqrt{2\pi}} a^* \exp\left(-\frac{(x-b_6^*)^2}{2}\right) - \frac{1}{\sqrt{2\pi}}(1-a^*)\exp\left(-\frac{(x-b_7^*)^2}{2}\right),$$

where $b_6^* \neq b_7^*$. Then, we have $n = \max\{i \in \mathbb{Z};\, i^2 + i \leq 2(5 - (2-1))\} = 2$, and

$$\lambda = 2 - 1 + \frac{2 + 2^2 + 2(5 - (2-1))}{4(2+1)} = \frac{13}{6}, \quad \theta = 1.$$

Similarly, the concrete values of λ and θ for $H, r \leq 5$ are given in the followings.

	r						r				
λ	1	2	3	4	5	θ	1	2	3	4	5
$H = 1$	1/2					$H = 1$	2				
$H = 2$	3/4	3/2				$H = 2$	1	2			
$H = 3$	1	7/4	5/2			$H = 3$	2	1	2		
$H = 4$	7/6	2	11/4	7/2		$H = 4$	1	2	1	2	
$H = 5$	4/3	13/6	3	15/4	9/2	$H = 5$	1	1	2	1	2

4. Conclusion

In this article, we consider the log canonical threshold of Vandermonde matrix-type singularities. Such singularities have recently been recognized that they are connected with Bayesian learning coefficients for several hierarchical learning models. For example, we showed that Bayesian learning coefficients of the three layered neural network (Aoyagi, 2006; Aoyagi and Watanabe, 2005a) and the mixtures of binomial distribution (Yamazaki et al., 2008) are obtained by the singularities. These facts seem to imply that the singularities are essential for learning theory.

By applying the orthogonal relation of them, we show that a Bayesian learning coefficient for normal mixture models is related to the singularities in Theorem 3.3, and then, by using the theorem and by applying techniques of algebraic geometry to learning theory, we obtained Bayesian learning coefficients for normal mixture models with unit matrix variances in dimension one (Theorem 3.4). Moreover, in the recent article (Aoyagi and Nagata, 2008) if the difference between the number of peaks of learning models and that of true distributions are less than one in general dimension, we are obtaining Bayesian learning coefficients. Our future research aims to improve our methods, and to apply them to general cases.

It is well known that the classic model selection methods of regular statistical models such as AIC (Akaike, 1974), TIC (Takeuchi, 1976), HQ (Hannan and Quinn, 1979), NIC (Murata et al., 1994), BIC (Schwarz, 1978), and MDL (Rissanen, 1984), cannot apply to the generalization error for non regular models, since the true parameter set of regular models should be one point and its Fisher matrix function is positive definite. Our theoretical values will be available for constructing a mathematical foundation for model selection methods of non regular models.

Several Bayesian learning coefficients in the article (Aoyagi, 2006; Aoyagi and Watanabe, 2005a) were used by analyzing and developing the precision of the Markov Chain Monte Carlo (Nagata and Watanabe, 2008a). Moreover, (Nagata and Watanabe, 2008b) studied the setting of temperatures for the exchange MCMC method by using such Bayesian learning coefficients. Our theoretical results in this article will also be helpful in these numerical experiments.

Appendix 1

Lemma A.1. *Let U be a neighborhood of $w^* \in \mathbb{R}^d$. Let $\mathcal{I}$ be the ideal generated by $f_1, \ldots, f_n$ which are analytic functions defined on U. If $g_1, \ldots, g_m \in I$, then $\lambda_{w^*}(f_1^2 + \cdots + f_n^2)$ is greater than $\lambda_{w^*}(g_1^2 + \cdots + g_m^2)$. In particular, if $g_1, \ldots, g_m$ generate the ideal $\mathcal{I}$ then*

$$\lambda_{w^*}(f_1^2 + \cdots + f_n^2) = \lambda_{w^*}(g_1^2 + \cdots + g_m^2).$$

Proof. The fact $g_1^2 + \cdots + g_m^2 \le P(f_1^2 + \cdots + f_n^2)$ for $P >> 1$ yields this lemma.

Lemma A.2. *Let* $B' = \begin{pmatrix} b_1^m & b_1^{Q+m} & \cdots & b_1^{Q(H-1)+m} \\ \vdots & & & \vdots \\ b_H^m & b_H^{Q+m} & \cdots & b_H^{Q(H-1)+m} \end{pmatrix}$ *and* $\mathbf{b}'_j = \begin{pmatrix} b_1^{Q(j-1)+m} \\ \vdots \\ b_H^{Q(j-1)+m} \end{pmatrix}$.

Consider a sufficiently small neighborhood of $\{b_i^*\}_{1 \le i \le H}$.

Let $b_i^* = \gamma_i |b_i^*|$.

Set

$$\mathbf{b}''_{ij} = \begin{cases} \gamma_i^m \displaystyle\prod_{|b_k^*|=|b_i^*|, 1 \le k \le j-1} (b_k/\gamma_k - b_i/\gamma_i), & \text{if } b_i^* \ne 0, \\ b_i^m \displaystyle\prod_{b_k^*=0, 1 \le k \le j-1} (b_k^Q - b_i^Q), & \text{if } b_i^* = 0, \end{cases} \quad \text{for } 1 \le j \le i \quad \text{and}$$

$$\mathbf{b}''_j = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \mathbf{b}''_{jj} \\ \vdots \\ \mathbf{b}''_{Hj} \end{pmatrix}, \quad \text{for } 1 \le j \le H.$$

Then there exists a regular matrix R such that $B'R = (\mathbf{b}''_1, \mathbf{b}''_2, \ldots, \mathbf{b}''_H)$.

Proof. We only need to prove that the vector space generated by $\mathbf{b}''_1, \mathbf{b}''_2, \ldots, \mathbf{b}''_H$ is equal to that generated by $\mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_H$.

Some computation shows that the vector space generated by

$$\begin{pmatrix} b_1^m \\ \vdots \\ b_H^m \end{pmatrix}, \begin{pmatrix} 0 \\ b_2^m(b_1^Q - b_2^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q) \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ b_3^m(b_1^Q - b_3^Q)(b_2^Q - b_3^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q)(b_2^Q - b_H^Q) \end{pmatrix}, \ldots,$$

$$\times \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_1^m(b_1^Q - b_H^Q) \cdots (b_{H-1}^Q - b_H^Q) \end{pmatrix}$$

is equal to that generated by $\mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_H$.

Therefore, we may set

$$\mathbf{b}'_1 = \begin{pmatrix} b_1^m \\ \vdots \\ b_H^m \end{pmatrix}, \mathbf{b}'_2 = \begin{pmatrix} 0 \\ b_2^m(b_1^Q - b_2^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q) \end{pmatrix}, \ldots, \mathbf{b}'_H = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_H^m(b_1^Q - b_H^Q)\cdots(b_{H-1}^Q - b_H^Q) \end{pmatrix}.$$

We use an induction.

From now on, denote by $\langle \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_H \rangle$ the vector space generated by vectors $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_H$.

It is easy to check that $\langle \mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_H \rangle = \langle \mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_{H-1}, \mathbf{b}''_H \rangle$.

Let $g_{j,j}(x), g_{j+1,j}(x), \ldots, g_{H,j}(x)$ be polynomials of x, $b_{j-1}, \ldots, b_1$ such that $g_{j',j}(x\gamma_{j'}) = g_{j'',j}(x\gamma_{j''})$ if $|b^*_{j'}| = |b^*_{j''}| \neq 0$ and $g_{j',j}(x) - g_{j'',j}(x')$ can be devided by $x^Q - x'^Q$ if $b^*_{j'} = b^*_{j''} = 0$.

Assume that $\begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j,j}(b_j)\mathbf{b}''_{jj} \\ \vdots \\ g_{H,j}(b_H)\mathbf{b}''_{Hj} \end{pmatrix}$ is an element of $\langle \mathbf{b}''_j, \ldots, \mathbf{b}''_H \rangle$ and that

$$\langle \mathbf{b}'_1, \ldots, \mathbf{b}'_H \rangle = \langle \mathbf{b}'_1, \ldots, \mathbf{b}'_{j-1}, \mathbf{b}''_j, \ldots, \mathbf{b}''_H \rangle.$$

Since

$$\mathbf{b}'_{j-1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_{j-1}^m(b_1^Q - b_{j-1}^Q)\cdots(b_{j-2}^Q - b_{j-1}^Q) \\ \vdots \\ b_H^m(b_1^Q - b_H^Q)\cdots(b_{j-2}^Q - b_H^Q) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j-1,j-1}(b_{j-1})\mathbf{b}''_{j-1,j-1} \\ \vdots \\ g_{H,j-1}(b_H)\mathbf{b}''_{H,j-1} \end{pmatrix},$$

where

$$g_{j-1,j-1}(b_{j-1}) \neq 0, \ldots, g_{H,j-1}(b_H) \neq 0,$$

$g_{j',j-1}(\gamma_{j'}x) = g_{j'',j-1}(\gamma_{j''}x)$ if $|b^*_{j'}| = |b^*_{j''}| \neq 0$ and $g_{j',j-1}(x) - g_{j'',j-1}(x')$ can be divided by $x'^Q - x^Q$ if $b^*_{j'} = b^*_{j''} = 0$, we have:

$$\mathbf{b}'_{j-1} = \mathbf{b}''_{j-1}g_{j-1,j-1}(b_{j-1}) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ (g_{j,j-1}(b_j) - g_{j-1,j-1}(b_{j-1}))\mathbf{b}''_{j,j-1} \\ \vdots \\ (g_{H,j-1}(b_H) - g_{j-1,j-1}(b_{j-1}))\mathbf{b}''_{H,j-1} \end{pmatrix}$$

$$= \mathbf{b}''_{j-1} g_{j-1,j-1}(b_{j-1}) + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j,j}(b_j)\mathbf{b}''_{j,j} \\ \vdots \\ g_{H,j}(b_H)\mathbf{b}''_{H,j} \end{pmatrix},$$

where $\begin{cases} g_{k,j}(b_k) = g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}), & \text{if } |b^*_k| \neq |b^*_{j-1}|, \\ g_{k,j}(b_k) = (g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}))/(b_{j-1}/\gamma_{j-1} - b_k/\gamma_k), & \text{if } |b^*_k| = |b^*_{j-1}| \neq 0, \\ g_{k,j}(b_k) = (g_{k,j-1}(b_k) - g_{j-1,j-1}(b_{j-1}))/(b^Q_{j-1} - b^Q_k) & \text{if } b^*_k = b^*_{j-1} = 0. \end{cases}$

By the inductive assumption, $\begin{pmatrix} 0 \\ \vdots \\ 0 \\ g_{j,j}(b_j)\mathbf{b}''_{j,j} \\ \vdots \\ g_{H,j}(b_H)\mathbf{b}''_{H,j} \end{pmatrix}$ is an element of the vector space generated by $\mathbf{b}''_j, \ldots, \mathbf{b}''_H$.

Therefore, $\langle \mathbf{b}'_1, \ldots, \mathbf{b}'_H \rangle = \langle \mathbf{b}'_1, \ldots, \mathbf{b}'_{j-1}, \mathbf{b}''_j, \ldots, \mathbf{b}''_H \rangle = \langle \mathbf{b}'_1, \ldots, \mathbf{b}'_{j-2}, \mathbf{b}''_{j-1}, \mathbf{b}''_j, \ldots, \mathbf{b}''_H \rangle$.

Lemma A.3. *Let* $B' = \begin{pmatrix} b_1^m & b_1^{Q+m} & \cdots & b_1^{Q(H-1)+m} \\ \vdots & & & \vdots \\ b_H^m & b_H^{Q+m} & \cdots & b_H^{Q(H-1)+m} \end{pmatrix}$ *and* $\mathbf{b}'_j = \begin{pmatrix} b_1^{Q(j-1)+m} \\ \vdots \\ b_H^{Q(j-1)+m} \end{pmatrix}$.

Consider a sufficiently small neighborhood of $\{b^*_i\}_{1 \leq i \leq H}$.

Let $b^*_i = \gamma_i |b^*_i|$.

Let each $|b^{**}_1|, \ldots, |b^{**}_r|$ *be a different real number in* $\{|b^*_i|; |b^*_i| \neq 0\}$:

$$\{|b^{**}_1|, \ldots, |b^{**}_r|; |b^{**}_i| \neq |b^{**}_j|, i \neq j\} = \{|b^*_i|; |b^*_i| \neq 0\}.$$

Also, set $b^{**}_0 = 0$.

Assume that $b^*_1 = \cdots = b^*_{H_0} = b^{**}_0$, $|b^*_{H_0+1}| = \cdots = |b^*_{H_0+H_1}| = |b^{**}_1|, \ldots, |b^*_{H_0+\cdots+H_{r-1}+1}| = \cdots = |b^*_{H_0+\cdots+H_r}| = |b^{**}_r|$.

Set

$$\begin{aligned} (b^{(0)}_1, \ldots, b^{(0)}_{H_0}) &= (b_1, \ldots, b_{H_0}), \\ (b^{(1)}_1, \ldots, b^{(1)}_{H_1}) &= (b_{H_0+1}, \ldots, b_{H_0+H_1}), \\ &\vdots \\ (b^{(r)}_1, \ldots, b^{(r)}_{H_r}) &= (b_{H_0+\cdots+H_{r-1}+1}, \ldots, b_{H_0+\cdots+H_r}). \end{aligned}$$

Let $b^{(\alpha)^*}_i = \gamma^{(\alpha)}_i |b^{(\alpha)^*}_i|$.

Then there exists a regular matrix R such that $B'R = \begin{pmatrix} B^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & B^{(1)} & 0 & \cdots & 0 \\ & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & B^{(r)} \end{pmatrix}$, *where*

$$B^{(0)} = \begin{pmatrix} {b_1^{(0)}}^m & {b_1^{(0)}}^{Q+m} & \cdots & {b_1^{(0)}}^{Q(H_0-1)+m} \\ \vdots & \vdots & & \vdots \\ {b_{H_0}^{(0)}}^m & {b_{H_0}^{(0)}}^{Q+m} & \cdots & {b_{H_0}^{(0)}}^{Q(H_0-1)+m} \end{pmatrix} \quad \textit{and}$$

$$B^{(\alpha)} = \begin{pmatrix} {\gamma_1^{(\alpha)}}^m & {\gamma_1^{(\alpha)}}^m b_1^{(\alpha)}/\gamma_1^{(\alpha)} & {\gamma_1^{(\alpha)}}^m (b_1^{(\alpha)}/\gamma_1^{(\alpha)})^2 & \cdots & {\gamma_1^{(\alpha)}}^m (b_1^{(\alpha)}/\gamma_1^{(\alpha)})^{H_\alpha-1} \\ \vdots & \vdots & & & \vdots \\ {\gamma_{H_\alpha}^{(\alpha)}}^m & {\gamma_{H_\alpha}^{(\alpha)}}^m b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)} & {\gamma_{H_\alpha}^{(\alpha)}}^m (b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)})^2 & \cdots & {\gamma_{H_\alpha}^{(\alpha)}}^m (b_{H_\alpha}^{(\alpha)}/\gamma_{H_\alpha}^{(\alpha)})^{H_\alpha-1} \end{pmatrix}$$

for $1 \le \alpha \le r$.

Proof. Set ${\mathbf{b}''}_1^{(0)} = \begin{pmatrix} {b_1^{(0)}}^m \\ {b_2^{(0)}}^m \\ \vdots \\ {b_{H_0}^{(0)}}^m \end{pmatrix}$ and ${\mathbf{b}''}_j^{(0)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ {b_j^{(0)}}^m \prod_{1\le k\le j-1} \left({b_k^{(0)}}^Q - {b_j^{(0)}}^Q\right) \\ \vdots \\ {b_{H_0}^{(0)}}^m \prod_{1\le k\le j-1} \left({b_k^{(0)}}^Q - {b_{H_0}^{(0)}}^Q\right) \end{pmatrix}$ for $j \ge 2$.

Also, set ${\mathbf{b}''}_j^{(\alpha)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ {\gamma_j^{(\alpha)}}^m \prod_{1\le k\le j-1} \left(b_k^{(\alpha)}/\gamma_k^{(\alpha)} - b_j^{(\alpha)}/\gamma_j^{(\alpha)}\right) \\ \vdots \\ {\gamma_{H_\alpha}^{(\alpha)}}^m \prod_{1\le k\le j-1} \left(b_k^{(\alpha)}/\gamma_k^{(\alpha)} - b_H^{(\alpha)}/\gamma_H^{(\alpha)}\right) \end{pmatrix}$ for $1 \le \alpha \le r$, $2 \le j \le i$.

Then, by Lemma A.2, there exists a regular matrix R such that

$$B'R = \begin{pmatrix} {\mathbf{b}''}_1^{(0)} & {\mathbf{b}''}_2^{(0)} & \cdots & {\mathbf{b}''}_{H_0}^{(0)} & 0 & & & \cdots & & & & & 0 \\ {\mathbf{b}''}_1^{(1)} & {\mathbf{b}''}_1^{(1)} & \cdots & {\mathbf{b}''}_1^{(1)} & {\mathbf{b}''}_1^{(1)} & {\mathbf{b}''}_2^{(1)} & \cdots & {\mathbf{b}''}_{H_1}^{(1)} & 0 & \cdots & & & 0 \\ & & \vdots & & & & \vdots & & & & & & \\ {\mathbf{b}''}_1^{(r)} & {\mathbf{b}''}_1^{(r)} & \cdots & {\mathbf{b}''}_1^{(r)} & {\mathbf{b}''}_1^{(r)} & {\mathbf{b}''}_1^{(r)} & \cdots & {\mathbf{b}''}_1^{(r)} & & \cdots & {\mathbf{b}''}_1^{(r)} & \cdots & {\mathbf{b}''}_{H_r}^{(r)} \end{pmatrix}.$$

Therefore, we have

$$B'RR' = \begin{pmatrix} {\mathbf{b}''}_1^{(0)} & {\mathbf{b}''}_2^{(0)} & \cdots & {\mathbf{b}''}_{H_0}^{(0)} & 0 & & & \cdots & & & & & 0 \\ 0 & 0 & \cdots & 0 & {\mathbf{b}''}_1^{(1)} & {\mathbf{b}''}_2^{(1)} & \cdots & {\mathbf{b}''}_{H_1}^{(1)} & 0 & \cdots & & & 0 \\ & & \vdots & & & & \vdots & & & & & & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & & \cdots & {\mathbf{b}''}_1^{(r)} & \cdots & {\mathbf{b}''}_{H_r}^{(r)} \end{pmatrix},$$

for some regular matrix R'.

By applying Lemma A.2 to $B^{(\alpha)}$, we have the proof.

Lemma A.4. *Let* $B_I = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix}$ *and* $B = (B_I)_{\ell_1+\cdots+\ell_N=Q(n-1)+m, n\in\mathbb{N}}$.

Consider a sufficiently small neighborhood of $\{b_{ij}^*\}_{1\le i\le H, 1\le j\le N}$.

Let each $(b_{11}^{**}, b_{12}^{**}, \ldots, b_{1N}^{**})$, $\ldots$, $(b_{r1}^{**}, b_{r2}^{**}, \ldots, b_{rN}^{**})$ *be a different real vector in*

$$[b_{i1}^*, b_{i2}^*, \ldots, b_{iN}^*]_Q \neq 0, i = 1, \ldots, H+r:$$
$$\{(b_{11}^{**}, \ldots, b_{1N}^{**}), \ldots, (b_{r1}^{**}, \ldots, b_{rN}^{**})\} = \{[b_{i1}^*, \ldots, b_{iN}^*]_Q \neq 0; i = 1, \ldots, H\}.$$

Set $(b_{01}^{**}, b_{02}^{**}, \ldots, b_{0N}^{**}) = (0, \ldots, 0)$.

Assume that

$$\begin{aligned}
&[b_{11}^*, \ldots, b_{1N}^*]_Q = \cdots = [b_{H_0 1}^*, \ldots, b_{H_0 N}^*]_Q = (b_{01}^{**}, \ldots, b_{0N}^{**}),\\
&[b_{H_0+1,1}^*, \ldots, b_{H_0+1,N}^*]_Q = \cdots = [b_{H_0+H_1,1}^*, \ldots, b_{H_0+H_1,N}^*]_Q = (b_{11}^{**}, \ldots, b_{1N}^{**}),\\
&\ldots,\\
&[b_{H_0+\cdots+H_{r-1}+1,1}^*, \ldots, b_{H_0+\cdots+H_{r-1}+1,N}^*]_Q = \cdots = [b_{H_0+\cdots+H_r,1}^*, \ldots, b_{H_0+\cdots+H_r,N}^*]_Q\\
&\quad = (b_{r1}^{**}, \ldots, b_{rN}^{**}).
\end{aligned}$$

Set

$$\begin{aligned}
(b_{1j}^{(0)}, \ldots, b_{H_0 j}^{(0)}) &= (b_{1j}, \ldots, b_{H_0 j}),\\
(b_{1j}^{(1)}, \ldots, b_{H_1 j}^{(1)}) &= (b_{H_0+1,j}, \ldots, b_{H_0+H_1,j}),\\
&\vdots\\
(b_{1j}^{(r)}, \ldots, b_{H_r j}^{(r)}) &= (b_{H_0+\cdots+H_{r-1}+1,j}, \ldots, b_{H_0+\cdots+H_r,j}),
\end{aligned}$$

for $1 \le j \le N$.

Let $I = (\ell_1, \ldots, \ell_N) \in {\mathbb{N}_{+0}}^N$, $B_I^{(\alpha)} = \begin{pmatrix} {\gamma_1^{(\alpha)}}^{m-|I|} \prod_{j=1}^N {b_{1j}^{(\alpha)}}^{\ell_j} \\ {\gamma_2^{(\alpha)}}^{m-|I|} \prod_{j=1}^N {b_{2j}^{(\alpha)}}^{\ell_j} \\ \vdots \\ {\gamma_{H_\alpha}^{(\alpha)}}^{m-|I|} \prod_{j=1}^N {b_{H_\alpha j}^{(\alpha)}}^{\ell_j} \end{pmatrix}$ *and* $B^{(0)} = (B_I^{(0)})_{\ell_1+\cdots+\ell_N=m+Q(n-1), n\in\mathbb{N}}$, $B^{(\alpha)} = (B_I^{(\alpha)})_{\ell_1+\cdots+\ell_N=n, n\in\mathbb{N}_{+0}}$ *for* $1 \le \alpha \le r$, *where*

$$\gamma_i^{(\alpha)}\big({b_{i1}^{(\alpha)}}^*, \ldots, {b_{iN}^{(\alpha)}}^*\big) = \big[{b_{i1}^{(\alpha)}}^*, \ldots, {b_{iN}^{(\alpha)}}^*\big]_Q.$$

Then there exists a regular matrix R such that

$$BR = \begin{pmatrix} B^{(0)} & 0 & 0 & \cdots & 0 \\ 0 & B^{(1)} & 0 & \cdots & 0 \\ & \vdots & & \ddots & \\ 0 & 0 & 0 & \cdots & B^{(r)} \end{pmatrix}.$$

Proof. The key point of the proof is to use

$$\begin{pmatrix}\prod_{j=1}^N {b_{1j}}^{\ell_j}\\ \prod_{j=1}^N {b_{2j}}^{\ell_j}\\ \vdots \\ \prod_{j=1}^N {b_{Hj}}^{\ell_j}\end{pmatrix} = \begin{pmatrix} {b_{11}}^{\ell'_1}\prod_{j=2}^N {b_{1j}}^{\ell_j} & 0 & \cdots & 0\\ 0 & {b_{21}}^{\ell'_1}\prod_{j=2}^N {b_{2j}}^{\ell_j} & \cdots & 0 \\ \vdots & \ddots & & 0\\ 0 & 0 & \cdots & {b_{H1}}^{\ell'_1}\prod_{j=2}^N {b_{Hj}}^{\ell_j}\end{pmatrix}\begin{pmatrix}{b_{11}}^{\ell_1-\ell'_1}\\ {b_{21}}^{\ell_1-\ell'_1}\\ \vdots\\ {b_{H1}}^{\ell_1-\ell'_1}\end{pmatrix},$$

and Lemma A.3.

Appendix 2

Let $A = (a_1, \ldots, a_H, a^*_{H+1}, \ldots, a^*_{H+r})$, $\sum_{i=1}^H a_i = 1$, $\sum_{i=H+1}^r a^*_i = -1$, and

$$B = \begin{pmatrix} b_1 & \cdots & b_1^{H+r}\\ b_2 & \cdots & b_2^{H+r}\\ & \vdots & \\ b_H & \cdots & b_H^{H+r}\\ b^*_{H+1} & \cdots & {b^*_{H+1}}^{H+r}\\ & \vdots & \\ b^*_{H+r} & \cdots & {b^*_{H+r}}^{H+r}\end{pmatrix}.$$

Then we have:

$$(a_1, \ldots, a_H, a^*_{H+1}, \ldots, a^*_{H+r})\begin{pmatrix} b_1^n\\ b_2^n\\ \vdots\\ b_H^n\\ {b^*_{H+1}}^n\\ \vdots\\ {b^*_{H+1}}^n\end{pmatrix} = (a_1, \ldots, a_{H-1}, a^*_{H+1}, \ldots, a^*_{H+r})\begin{pmatrix} b_1^n - b_H^n\\ b_2^n - b_H^n\\ \vdots\\ b_{H-1}^n - b_H^n\\ {b^*_{H+1}}^n - b_H^n\\ \vdots\\ {b^*_{H+r}}^n - b_H^n\end{pmatrix}$$

by using $\sum_{i=1}^H a_i = 1$, $\sum_{i=H+1}^r a^*_i = -1$.

Let

$$B' = \begin{pmatrix} b_1 - b_H & \cdots & b_1^{H+r} - b_H^{H+r}\\ b_2 - b_H & \cdots & b_2^{H+r} - b_H^{H+r}\\ & \vdots & \\ b_{H-1} - b_H & \cdots & b_{H-1}^{H+r} - b_H^{H+r}\\ b^*_{H+1} - b_H & \cdots & {b^*_{H+1}}^{H+r} - b_H^{H+r}\\ & \vdots & \\ b^*_{H+r} - b_H & \cdots & {b^*_{H+r}}^{H+r} - b_H^{H+r}\end{pmatrix}.$$

Since $(b_m^n - b_H^n) - (b_m^{n-1} - b_H^{n-1})b_1 - (b_m - b_H)b_H^{n-1} = (b_m^{n-1} - b_H^{n-1})(b_m - b_1)$, we have a regular matirx R such that

$$B'R = \begin{pmatrix} b_1 - b_H & 0 & \cdots & & 0 \\ b_2 - b_H & (b_2 - b_H)(b_2 - b_1) & 0\cdots & & 0 \\ b_3 - b_H & (b_3 - b_H)(b_3 - b_1) & (b_3 - b_H)(b_3 - b_1)(b_3 - b_2) & \cdots & 0 \\ & & \vdots & & \\ b_{H-1} - b_H & (b_{H-1} - b_H)(b_{H-1} - b_1) & (b_{H-1} - b_H)(b_{H-1} - b_1)(b_{H-1} - b_2) & \cdots & 0 \\ b^*_{H+1} - b_H & (b^*_{H+1} - b_H)(b^*_{H+1} - b_1) & (b^*_{H+1} - b_H)(b^*_{H+1} - b_1)(b^*_{H+1} - b_2) & \cdots & 0 \\ & \vdots & & & \\ b^*_{H+r} - b_H & (b^*_{H+r} - b_H)(b^*_{H+r} - b_1) & (b^*_{H+r} - b_H)(b^*_{H+r} - b_1)(b^*_{H+r} - b_2) & \cdots & (b^*_{H+r} - b_H)\cdots \\ & & & & (b^*_{H+r} - b^*_{H+r-1}) \end{pmatrix},$$

and we have for some regular matirx R',

$$B'R' = \begin{pmatrix} b_1 - b_H & (b_1 - b_H)^2 & \cdots & (b_1 - b_H)^{H+r-1} \\ b_2 - b_H & (b_2 - b_H)^2 & \cdots & (b_2 - b_H)^{H+r-1} \\ & & \vdots & \\ b^*_{H+r} - b_H & (b^*_{H+r} - b_H)^2 & \cdots & (b^*_{H+r} - b_H)^{H+r-1} \end{pmatrix}.$$

Set

$$\begin{aligned} (a_1'^{(1)}, \ldots, a_{H_1}'^{(1)}) &= (a_1, \ldots, a_{H_1}), \\ (a_1'^{(2)}, \ldots, a_{H_2}'^{(2)}) &= (a_{H_1+1}, \ldots, a_{H_1+H_2}), \\ &\vdots \\ (a_1'^{(r')}, \ldots, a_{H_{r'}}'^{(r')}) &= (a_{H_1+\cdots+H_{r'-1}+1}, \ldots, a_{H_1+\cdots+H_{r'}}). \end{aligned}$$

and

$$\begin{aligned} (b_1'^{(1)}, \ldots, b_{H_1}'^{(1)}) &= (b_1 - b_H, \ldots, b_{H_1} - b_H), \\ (b_1'^{(2)}, \ldots, b_{H_2}'^{(2)}) &= (b_{H_1+1} - b_H, \ldots, b_{H_1+H_2} - b_H), \\ &\vdots \\ (b_1'^{(r')}, \ldots, b_{H_{r'}-1}'^{(r')}) &= (b_{H_1+\cdots+H_{r'-1}+1} - b_H, \ldots, b_{H_1+\cdots+H_{r'}-1} - b_H). \end{aligned}$$

Let

$$A'^{(\alpha)} = \begin{cases} (a_1'^{(\alpha)}, a_2'^{(\alpha)}, \ldots, a_{H_\alpha}'^{(\alpha)}, a^*_{H+\alpha}), & \text{for } 1 \leq \alpha \leq r, \alpha \leq r' - 1 \\ (a_1'^{(\alpha)}, a_2'^{(\alpha)}, \ldots, a_{H_\alpha}'^{(\alpha)}, 0), & \text{for } H + 1 \leq \alpha \leq r' - 1, \end{cases}$$

$$A'^{(r')} = \begin{cases} (a_1'^{(r')}, a_2'^{(r')}, \ldots, a_{H_{r'}-1}'^{(r')}, a^*_{H+r'}), & \text{if } r' = r, \\ (a_1'^{(r')}, a_2'^{(r')}, \ldots, a_{H_{r'}-1}'^{(r')}, 0), & \text{if } r' > r, \end{cases}$$

$$B'^{(\alpha)} = \begin{pmatrix} 1 & b_1'^{(\alpha)} & \cdots & b_1'^{(\alpha)^{H_\alpha}} \\ 1 & b_2'^{(\alpha)} & \cdots & b_2'^{(\alpha)^{H_\alpha}} \\ & & \vdots & \\ 1 & b_{H_\alpha}'^{(\alpha)} & \cdots & b_{H_\alpha}'^{(\alpha)^{H_\alpha}} \\ 1 & b_\alpha^{**} - b_H & \cdots & (b_\alpha^{**} - b_H)^{H_\alpha} \end{pmatrix} \quad \text{for } 1 \leq \alpha \leq r' - 1,$$

and

$$B'^{(r')} = \begin{pmatrix} b_1'^{(r')} & b_1'^{(r')^2} & \cdots & b_1'^{(r')^{H_{r'}}} \\ b_2'^{(r')} & b_2'^{(r')^2} & \cdots & b_2'^{(r')^{H_{r'}}} \\ & & \vdots & \\ b_{H_{r'}-1}'^{(r')} & {b_{H_{r'}-1}'^{(r')}}^{2} & \cdots & {b_{H_{r'}-1}'^{(r')}}^{H_{r'}} \\ b_{r'}^{**} - b_H & (b_{r'}^{**} - b_H)^2 & \cdots & (b_{r'}^{**} - b_H)^{H_{r'}} \end{pmatrix}.$$

By Theorem 3.1, we only need to consider the case $\sum_{\alpha=0}^{r'} \|A^{(\alpha)}B^{(\alpha)}\|^2$ instead of $\|AB\|^2$.

Since $b_\alpha^{**} \neq b_H^* = b_{r'}^{**}$ for $\alpha = 1, \ldots, r-1$, we have:

$$B'^{(\alpha)} = \begin{pmatrix} 1 & \frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H} & \cdots & \left(\frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H}\right)^{H_\alpha} \\ 1 & \frac{b_2'^{(\alpha)}}{b_\alpha^{**}-b_H} & \cdots & \left(\frac{b_2'^{(\alpha)}}{b_\alpha^{**}-b_H}\right)^{H_\alpha} \\ & & \vdots & \\ 1 & \frac{b_{H_\alpha}'^{(\alpha)}}{b_\alpha^{**}-b_H} & \cdots & \left(\frac{b_{H_\alpha}'^{(\alpha)}}{b_\alpha^{**}-b_H}\right)^{H_\alpha} \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & b_\alpha^{**} - b_H & \cdots & 0 \\ & \vdots & & \\ 0 & 0 & \cdots & (b_\alpha^{**} - b_H)^{H_\alpha} \end{pmatrix}.$$

We have, therefore, a regular matrix R'' such that

$$B'^{(\alpha)}R'' = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & \frac{b_2'^{(\alpha)}}{b_\alpha^{**}-b_H} - \frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H} & \cdots & \left(\frac{b_2'^{(\alpha)}}{b_\alpha^{**}-b_H} - \frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H}\right)^{H_\alpha} \\ & & \vdots & \\ 1 & \frac{b_{H_\alpha}'^{(\alpha)}}{b_\alpha^{**}-b_H} - \frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H} & \cdots & \left(\frac{b_{H_\alpha}'^{(\alpha)}}{b_\alpha^{**}-b_H} - \frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H}\right)^{H_\alpha} \\ 1 & 1 - \frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H} & \cdots & \left(1 - \frac{b_1'^{(\alpha)}}{b_\alpha^{**}-b_H}\right)^{H_\alpha} \end{pmatrix}.$$

Set

$$a_1^{(\alpha)} = \begin{cases} a_1'^{(\alpha)} + a_2'^{(\alpha)} + \cdots + a_{H_\alpha}'^{(\alpha)} + a_i^*, & \text{for } 1 \leq \alpha \leq r, \alpha \leq r' - 1, \\ a_1'^{(\alpha)} + a_2'^{(\alpha)} + \cdots + a_{H_\alpha}'^{(\alpha)}, & \text{for } H + 1 \leq \alpha \leq r' - 1, \end{cases}.$$

$$A^{(\alpha)} = \begin{cases} \big(a_2^{(\alpha)}, a_3^{(\alpha)}, \ldots, a_{H_\alpha}^{(\alpha)}, a_{H+\alpha}^{*}\big) & \text{for } 1 \le \alpha \le r, \alpha \le r' - 1 \\ \quad = \big(a_2^{\prime(\alpha)}, a_3^{\prime(\alpha)}, \ldots, a_{H_\alpha}^{\prime(\alpha)}, a_{H+\alpha}^{*}\big), & \\ \big(a_2^{(\alpha)}, a_3^{(\alpha)}, \ldots, a_{H_\alpha}^{(\alpha)}\big) = \big(a_2^{\prime(\alpha)}, a_3^{\prime(\alpha)}, \ldots, a_{H_\alpha}^{\prime(\alpha)}\big), & \text{for } H+1 \le \alpha \le r' - 1, \end{cases}.$$

$$A^{(r')} = \begin{cases} (a_2^{(r')}, a_3^{(r')}, \ldots, a_{H_{r'}}^{(r')}, a_{H+\alpha}^{*}) = (a_1^{\prime(r')}, a_2^{\prime(r')}, \ldots, a_{H_\alpha-1}^{\prime(r')}, a_{H+\alpha}^{*}), & \text{if } r = r', \\ (a_2^{(r')}, a_3^{(r')}, \ldots, a_{H_{r'}}^{(r')}) = (a_1^{\prime(r')}, a_2^{\prime(r')}, \ldots, a_{H_\alpha-1}^{\prime(r')}), & \text{if } r < r', \end{cases}.$$

and

$$\begin{pmatrix} b_1^{(\alpha)} \\ \vdots \\ b_{H_\alpha}^{(\alpha)} \end{pmatrix} = \begin{pmatrix} \frac{b_2^{\prime(\alpha)}}{b_\alpha^{**}-b_H} - \frac{b_1^{\prime(\alpha)}}{b_\alpha^{**}-b_H} \\ \vdots \\ \frac{b_{H_\alpha}^{\prime(\alpha)}}{b_\alpha^{**}-b_H} - \frac{b_1^{\prime(\alpha)}}{b_\alpha^{**}-b_H} \\ 1 - \frac{b_1^{\prime(\alpha)}}{b_\alpha^{**}-b_H} \end{pmatrix}, \quad \begin{pmatrix} b_1^{(r')} \\ \vdots \\ b_{H_{r'}}^{(r')} \end{pmatrix} = \begin{pmatrix} b_1^{\prime(r')} \\ \vdots \\ b_{H_{r'}-1}^{\prime(r')} \\ b_{r'}^{**} - b_H \end{pmatrix}.$$

Then we have Theorem 3.3.

Acknowledgment

The author would like to thank the Editor and the reviewer for valuable comments improving the article.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Contr.* 19:716–723.

Akaike, H. (1980). Likelihood and Bayes procedure. In: Bernald J.M., ed. *Bayesian Statistics*. Valencia, Spain: University Press, pp. 143–166.

Amari, S., Fujita, N., Shinomoto, S. (1992). Four types of learning curves. *Neur. Computat.* 4:608–618.

Amari, S., Murata, N. (1993). Statistical theory of learning curves under entropic loss. *Neur. Computat.* 5:140–153.

Aoyagi, M. (2006). The zeta function of learning theory and generalization error of three layered neural perceptron. *RIMS Kokyuroku, Recent Topics on Real and Complex Singularities* 1501:153–167.

Aoyagi, M., Nagata, K. (2008). Learning coefficient of generalization error of three layered neural networks and normal mixture models in Bayesian estimation. Preprint.

Aoyagi, M., Watanabe, S. (2005a). Resolution of singularities and the generalization error with Bayesian estimation for layered neural network. *IEICE Trans. J88-D-II*, 10:2112–2124 (English version: *Systems and Computers in Japan*, New York: John Wiley & Sons Inc., in press).

Aoyagi, M., Watanabe, S. (2005b). Stochastic complexities of reduced rank regression in Bayesian estimation. *Neur. Netw.* 18:924–933.

Fukumizu, K. (1996). A regularity condition of the information matrix of a multilayer perceptron network. *Neur. Netw.* 9(5):871–879.

Fulton, W. (1993). Introduction to toric varieties. *Ann. Math. Stu.*, Princeton University Press, p. 131.

Hagiwara, K., Toda, N., Usui, S. (1993). On the problem of applying AIC to determine the structure of a layered feed-forward neural network. *Proc. IJCNN Nagoya Japan* 3:2263–2266.

Hannan, E. J., Quinn, B. G. (1979). The determination of the order of an autoregression. *J. Roy. Statis. Soc. Ser. B* 41:190–195.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. *Proc. Berkeley Conf. Honor of J. Neyman and J. Kiefer* 2:807–810.

Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. *Ann. Math.* 79:109–326.

Kollár, J. (1997). Singularities of pairs, Algebraic geometry-Santa Cruz 1995. *Proc. Symp. Pure Math.*, Vol 62. Providence, RI: American Mathematics Society, pp. 221–287.

Levin, E., Tishby, N., Solla, S. A. (1990). A statistical approaches to learning and generalization in layered neural networks. *Proc. of IEEE* 78(10):1568–1674.

Mackay, D. J. (1992). Bayesian interpolation. *Neur. Computat.* 4(2):415–447.

Murata, N. J., Yoshizawa, S. G., Amari, S. (1994). Network information criterion – determining the number of hidden units for an artificial neural network model. *IEEE Trans. Neur. Netw.* 5(6):865–872.

Mustata, M. (2002). Singularities of pairs via jet schemes. *J. Amer. Math. Soc.* 15:599–615.

Nagata, K., Watanabe, S. (2008a). Exchange Monte Carlo sampling from Bayesian posterior for singular learning machines. *IEEE Trans. Neur. Netw.* 19(7):1253–1266.

Nagata, K., Watanabe, S. (2008b). Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *Int. J. Neur. Netw.* 21(7):980–988.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. IEEE Trans. Inform. Theor. 30(4):629–636.

Rissanen, J. (1986). Stochastic complexity and modeling. *Ann. Statist.* 14:1080–1100.

Rusakov, D., Geiger, D. (2005). Asymptotic model selection for naive Bayesian networks. *J. Machine Learning Res.* 6:1–35.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statis.* 6(2):461–464.

Sussmann, H. J. (1992). Uniqueness of the weights for minimal feed-forward nets with a given input-output map. *Neur. Netw.* 5:589–593.

Takeuchi, K. (1976). Distribution of an information statistic and the criterion for the optimal model. *Mathemat. Sci.* 153:12–18.

Watanabe, S. (2001a). Algebraic analysis for nonidentifiable learning machines. *Neur. Computa.* 13(4):899–933.

Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neur. Netw.* 14(8):1049–1060.

Watanabe, S., Yamazaki, K., Aoyagi, M. (2004). Kullback information of normal mixture is not an analytic function. *Technical Report of IEICE*, NC2004, pp. 41–46.

Yamanishi, K. (1998). A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. Inform. Theor.* 44(4):1424–1439.

Yamazaki, K., Aoyagi, M., Watanabe, S. (2008). Asymptotic analysis of Bayesian generalization error with Newton diagram. *Neural Networks* 23:35–43.

Yamazaki, Y., Watanabe, S. (2003). Singularities in mixture models and upper bounds of stochastic complexity. *Int. J. Neur. Netw.* 16:1029–1038.