Learning coefficients and information criteria

Miki AOYAGI [a,1],

[a] *Department of Mathematics, College of Science & Technology, Nihon University, 1-8-14, Surugadai, Kanda, Chiyoda-ku, Tokyo 101-8308, Japan; E-mail: aoyagi.miki@math.cst.nihon-u.ac.jp*

Abstract. In recent studies, image or speech recognition, psychology, and economics, etc. in real big data have been analyzed by learning systems. It is one of important problems to approximate an unknown true density function from training data selected from the true density function independently and identically using learning models. In a stochastic model, many hierarchical learning models for analyzing real data have been proposed, and proved to be effective. They are, however, singular, which classic theories for regular models cannot apply to. Therefore, the need for appropriate model selection methods for singular models has increased and several information criteria for singular models have been developed. For example, singular Bayesian information criterion, widely applicable information criterion, widely applicable Bayesian information criterion, and cross-validation have been considered based on mathematical theorems in algebraic analysis and geometry . In this paper, we consider learning coefficients in learning theory, which serve to measure the main term of learning efficiency in singular learning models. These coefficients have an important role in information criteria and are mathematically equal to the log canonical thresholds of Kullback functions. We show several mathematical theorems for obtaining these coefficients, and apply these theorems to Poisson distribution mixture models.

Keywords. learning coefficient, Kullback information, singular learning models, construction of blow ups

Introduction

Recently, many hierarchical learning models, for example, layered neural network, reduced rank regression, the Boltzmann machine, and the normal mixture model, have been used to analyze real data. These models are singular, which are not classical regular ones, and thus the need to analyze singular ones has increased.

In this section, we introduce several information criteria for singular models and their results.

We denote by $q(x)$ a true probability density function of variables $x \in \mathbf{R}^N$ and set $x^n = \{x_i\}_{1 \le i \le n}$ as n training samples distributed from $q(x)$ independently and identically.

We first introduce Kullback information $K(q||p)$ for density functions $p(x), q(x)$:

[1] Corresponding Author: Miki Aoyagi, Department of Mathematics, College of Science & Technology, Nihon University, 1-8-14, Surugadai, Kanda, Chiyoda-ku, Tokyo 101-8308, Japan; E-mail: aoyagi.miki@math.cst.nihon-u.ac.jp.

$$K(q||p) = \int \log \frac{q(x)}{p(x)} q(x)dx.$$

This function is a pseudo-distance, because $K(p||q) \geq 0$ and satisfies $q(x) = p(x)$, if and only if, $K(q||p) = 0$. Also we define empirical Kullback information $K_n(q||p)$ as

$$K_n(q||p) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(x_i)}{p(x_i)},$$

which satisfies $E[K_n(q||p)] = K(q||p)$.

Consider a learning model $p(x|w)$, which is a probability density function of variables $x \in \mathbf{R}^N$ with parameter $w \in W \subset \mathbf{R}^d$.

In Bayesian estimation, one of our goals of the learning system is to approximate unknown true density function $q(x)$ from data x^n using $p(x|w)$.

Let us consider an *a priori* probability density function $\psi(w)$ on parameter set W and the *a posteriori* probability density function $p(w|x^n)$:

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

where

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

for inverse temperature β. We typically set $\beta = 1$.

Let

$$F_n(\beta) = -\log Z_n(\beta).$$

$F_n(1)$ is called free energy. Additionally, $F_n(1)$ is known as the Bayesian criterion[1], stochastic complexity in universal coding[2,3], Akaike's Bayesian criterion[4], and evidence in neural network learning[5].

Watanabe [6,7,8,9,10,11,12] proved that

$$E[F_n(1)] = L(w_0) + \lambda \log n - (\theta - 1) \log \log n + O(1)$$

for learning coefficient $\lambda \in \mathbf{Q}$ and its order θ, which provide the learning efficiencies, where $L(w) = -E_x[\log p(x|w)]$ and $w_0 \in W_0 = \{w_0 \in W \mid L(w_0) = \min_{w \in W} L(w)\}$.

By analyzing this relation and using the unique solution to certain equation system, the "singular Bayesian information criterion" (sBIC) [13] is obtained.

"Widely applicable Bayesian information criterion" (WBIC) [14] is also obtained, and denoted by

$$WBIC = -E_w^{\beta}[\sum_{i=1}^{n} p(X_i|w)]$$

for $\beta = 1/\log n$, where

$$E_w^\beta[f(w)] = \frac{\int dw f(w) \prod_{i=1}^n p(x_i|w)^\beta \psi(w)}{\int dw \prod_{i=1}^n p(x_i|w)^\beta \psi(w)}.$$

These information criteria are extensions of BIC for singular models.

Next we introduce the widely applicable information criterion [6,7,8,9,10,11,12] and the cross-validation loss.

We have the predictive density function, i.e., the average inference in Bayes estimation, $p(x|x^n) = E_w^\beta[p(x|w)]$.

We define Bayes training loss T_n and Bayes generalization loss G_n as follows:

$$T_n = -\frac{1}{n}\sum_{i=1}^n \log p(x_i|x^n)$$

and

$$G_n = -\int q(x) \log p(x|x^n) dx.$$

Then we have $E[T_n] = G_n$ and $E[G_n] = E[F_{n+1}] - E[F_n] (\beta = 1)$ for $n \in \mathbf{N}$ [15,16,17].

Let

$$V_w^\beta[f(w)] = E_w^\beta[f(w)^2] - E_w^\beta[f(w)]^2.$$

Additionally, we define Bayesian generalization error B_g and Bayesian training error B_t as follows:

$$B_g = K(q(x)||p(x|x^n))$$

and

$$B_t = K_n(q(x)||p(x|x^n)).$$

Then we have

$$\begin{aligned} B_g &= G_n - S, \\ B_t &= T_n - S_n \end{aligned}$$

for average entropy $S = -\int q(x) \log q(x) dx$ and empirical entropy $S_n = -\frac{1}{n}\sum_{i=1}^n \log q(x_i)$ of the true density function. Value B_g describes how precisely the predictive density function $p(x|x^n)$ approximates the true density function $q(x)$.

We define $x^n \backslash x_i = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$. The widely applicable information criterion [6,7,8,9,10,11,12] is denoted by

$$W_n = T_n + \frac{\beta}{n}\sum_{i=1}^n V_w^\beta[\log p(x_i|w)]$$

and cross-validation loss is denoted by

$$C_n = -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|x^n \backslash x_i)$$

for $n \geq 2$.

Watanabe proved the following relations:

$$\begin{aligned}
E[G_n] &= L(w_0)+\frac{1}{n\beta}(\lambda+\frac{\beta-1}{\beta}\nu)+o(\frac{1}{n\beta}),\\
E[T_n] &= L(w_0)+\frac{1}{n\beta}(\lambda-\frac{\beta+1}{\beta}\nu)+o(\frac{1}{n\beta}),\\
E[W_n] &= L(w_0)+\frac{1}{n\beta}(\lambda+\frac{\beta-1}{\beta}\nu)+o(\frac{1}{n\beta}),\\
E[C_n] &= L(w_0)+\frac{1}{n\beta}(\lambda+\frac{\beta-1}{\beta}\nu)+o(\frac{1}{n\beta})
\end{aligned}$$

for singular fluctuation $\nu \in \mathbf{R}$.

Value ν is obtained by theoretically using learning coefficient λ and its order θ as follows:

$$\nu = \frac{1}{2}E_\xi \frac{\int_0^\infty dt \sum_{u^*} \int t^{\lambda-1/2} e^{\beta\sqrt{t}\xi(u)-\beta t}\xi(u)du}{\int_0^\infty dt \sum_{u^*} \int t^{\lambda-1/2} e^{\beta\sqrt{t}\xi(u)-\beta t}du},$$

where $\xi(u)$ is obtained by an empirical process $K_n(q||p)$ defined on the smooth manifold using a resolution of singularities, and u^* denotes a local coordinate that attain the learning coefficient λ and its order θ.

In this paper, we consider value λ and determine the exact values of Poisson distribution mixture models [18], which are very useful models with discrete input values in learning theory. These coefficients are equal to the log canonical thresholds of the Kullback function mathematically introduced in Definition 1.

In recent studies, we determined the exact values or bounds of several learning coefficients [19,20,21,22,23,24,25,26]. Additionally, in [27,28] and [29], the learning coefficients for naive Bayesian networks and directed tree models with hidden variables were obtained, respectively. Drton et al.[30] considered these coefficients of a Gaussian latent tree and forest models.

1. Main results

Denote constants, such as w^*, a^* and b^* using suffix $*$ and define the norm of matrix $A = (a_{ij})$ as $||A|| = \sqrt{\sum_{i,j}|a_{ij}|^2}$. Set $\mathbf{N}_{+0} = \{0, 1, 2, \cdots\}$.

Definition 1 *Assume that $f(w)$ is an analytic function in a small neighborhood U of w^* and $\psi(w)$ is a C^∞ function having a compact support.*

We define log canonical threshold $-\lambda_{w^*}(f, \psi)$ as the largest pole of $\int_U |f(w)|^z\psi(w)dw$, and also define $\theta_{w^*}(f, \psi)$ by its order.

In this paper, we use the notations $\lambda_{w^*}(f)$ and $\theta_{w^*}(f)$ instead of $\lambda_{w^*}(f,\psi)$ and $\theta_{w^*}(f,\psi)$ respectively, because if $\psi(w^*)\neq 0$, then these values are independent of ψ.

Example 2 *(1) If $f(w)=f(w_1,w_2)=w_1^4w_2^6$ and $U=(-1,1)\times(-1,1)$, then $\lambda_{(0,0)}(f)=\frac{1}{6}$ and $\theta_{(0,0)}(f)=1$ since $\int_U |w_1^4w_2^6|^z dw=\frac{2}{(6z+1)(4z+1)}$.*

(2) If $f(w)=f(w_1,w_2)=w_1^6+w_1^2w_2^4$ and $U=(-1,1)\times(-1,1)$, then $\lambda_{(0,0)}(f)=\frac{2}{6}=\frac{1}{3}$ and $\theta_{(0,0)}(f)=1$ since

$$\int_U |w_1^6+w_1^2w_2^4|^z dw=\int_{U'} |{w'_1}^6(1+{w'_2}^4)|^z w'_1 dw' + \int_{U''} |{w''_2}^6{w''_1}^2({w''_1}^4+1)|^z w''_2 dw''$$

$$=\frac{g(z)}{(6z+2)}+\frac{h(z)}{(6z+2)(2z+1)},$$

where $w_1=w'_1, w_2=w'_1w'_2$ on $U'=(-1,1)\times(-1,1)$, $w_1=w''_1w''_2, w_2=w''_2$ on $U''=(-1,1)\times(-1,1)$ and $g(z),h(z)$ are holomorphic functions.

Hironaka's Theorem [31] establishes the existence of maps from a smooth manifold to obtain these log canonical thresholds by resolution of singularities. The map $w_1=w'_1, w_2=w'_1w'_2$ on $U'=(-1,1)\times(-1,1)$, $w_1=w''_1w''_2, w_2=w''_2$ on $U''=(-1,1)\times(-1,1)$ in Example 2 (2) is one of such desingularization maps. However, generally, because of complicated singularities of Kullback functions, to obtain such maps is very difficult in learning theory. Therefore, we need to construct several mathematical theories for the purpose.

Lemma 3 ([22,23,32]) *Assume that $\mathcal{J}$ is the ideal generated by $f_1(w), f_2(w),\cdots,f_n(w)$, which are analytic functions defined on a neighborhood U of $w^*\in\mathbf{R}^d$.*
(1) If $\sum_{i=1}^m g_i^2\leq\sum_{i=1}^n f_i^2$, then $\lambda_{w^*}(\sum_{i=1}^m g_i^2)\leq\lambda_{w^*}(\sum_{i=1}^n f_i^2)$.
(2) If $g_1,g_2,\cdots,g_m\in\mathcal{J}$, then $\lambda_{w^*}(\sum_{i=1}^m g_i^2)\leq\lambda_{w^*}(\sum_{i=1}^n g_i^2)$. In particular, if $g_1,g_2,\cdots,g_m$ generate $\mathcal{J}$, then $\lambda_{w^*}(\sum_{i=1}^n f_i^2)=\lambda_{w^*}(\sum_{i=1}^m g_i^2)$.

Consider the mixture of N dimensional Poisson distributions with H components and assume the true distribution with r components. An input value of N dimensional Poisson distributions is $x=(x_j)\in\mathbf{Z}_{\geq 0}^N$ and we have

$$p(x|w)=\sum_{i=1}^{H} a_{1i}\prod_{j=1}^{N}\exp(-b_{ij})\frac{b_{ij}^{x_j}}{x_j!},$$

where $w=\{a_{1i},b_{ij}|1\leq i\leq H, 1\leq j\leq H\}$, $b_{ij}>0$ and $\sum_{i=1}^H a_{1i}=1, a_{1i}\geq 0$.

Also we have the true distribution:

$$p(x|w_t^*)=-\sum_{i=H+1}^{H+r} a_{1i}^*\prod_{j=1}^{N}\exp(-b_{ij}^*)\frac{{b_{ij}^*}^{x_j}}{x_j!},$$

where $w_t^*=\{a_{1i}^*,b_{ij}^*|H+1\leq i\leq H+r, 1\leq j\leq H\}$, $b_{ij}^*>0$ and $\sum_{i=H+1}^{H+r}a_{1i}^*=-1$, $a_{1i}^*<0$. (We use the values $a_{1i}^*<0$, not $a_{1i}^*>0$, in order to simplify the following.)

Theorem 4 *Consider the ideal $\mathcal{J}$ generated by $p(x|w) - p(x|w_t^*)$ for $x \in \mathbf{Z}^N_{\geq 0}$. Then the generators of $\mathcal{J}$ are*

$$\sum_{i=1}^{H} a_{1i} \prod_{j=1}^{N} {b_{ij}}^{x_j} + \sum_{i=H+1}^{H+r} a^*_{1i} \prod_{j=1}^{N} {b^*_{ij}}^{x_j} \ (x \in \mathbf{Z}^N_{\geq 0}).$$

We have its proof since $\exp x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.

The generators of the mixture of N dimensional Poisson distributions with H components and the true distribution with r components are obtained by Vandermonde matrix type singularity in Definition 6.

Definition 5 *Let: $[b_1^*, \cdots, b_N^*]_Q = \xi_i(0, \cdots, 0, b_i^*, \cdots, b_N^*)$ for $b_s^* = 0$ $s = 1, \cdots, i-1$, $b_i^* \neq 0$, and $\xi_i = \begin{cases} 1 & \text{if } Q \text{ is odd,} \\ \mathrm{sign}(b_i^*) & \text{if } Q \text{ is even.} \end{cases}$*

Definition 6 (Vandermonde matrix type singularity) *Set $Q \in \mathbf{N}$ and fix it.*

$$\text{Let } A_{M,H,r} = \begin{pmatrix} a_{11} & \cdots & a_{1H} & a^*_{1,H+1} & \cdots & a^*_{1,H+r} \\ a_{21} & \cdots & a_{2H} & a^*_{2,H+1} & \cdots & a^*_{2,H+r} \\ \vdots & & & \vdots & & \\ a_{M1} & \cdots & a_{MH} & a^*_{M,H+1} & \cdots & a^*_{M,H+r} \end{pmatrix},$$

$$B_{H,N,r,I} = \Big(\prod_{j=1}^{N} b_{1j}^{\ell_j}, \prod_{j=1}^{N} b_{2j}^{\ell_j}, \cdots, \prod_{j=1}^{N} b_{Hj}^{\ell_j}, \prod_{j=1}^{N} {b^*_{H+1,j}}^{\ell_j}, \cdots, \prod_{j=1}^{N} {b^*_{H+r,j}}^{\ell_j}\Big)^t$$

for $I = (\ell_1, \ldots, \ell_N) \in {\mathbf{N}_{+0}}^N$ and

$$B^{(Q)}_{H,N,r} = (B_{H,N,r,I})_{\sum_{j=1}^{N} \ell_j = 1+Qn, n \geq 0},$$

where t denotes the transpose.

Variables a_{ki} and b_{ij} $(k = 1, \ldots, M, i = 1, \ldots, H, j = 1, \ldots, N)$ are defined in a neighborhood of constants a^*_{ki} and b^*_{ij}.

Set $\mathcal{J}$ be the ideal generated by all elements of $A_{M,H,r} B^{(Q)}_{H,N,r}$ and then its singularities are called Vandermonde matrix-type singularities.

For simplicity, we assume that

$$\prod_{k}^{M} a^*_{k,H+j} \neq 0, \prod_{\ell=1}^{N} b^*_{H+j,\ell} \neq 0$$

for $1 \leq j \leq r$ and $[b^*_{H+j,1}, \ldots, b^*_{H+j,N}]_Q \neq [b^*_{H+j',1}, \ldots, b^*_{H+j',N}]_Q$ for $j \neq j'$.

We use $w = \{a_{ki}, b_{ij}\}_{i=1,\cdots,H}$ instead of $w = \{a_{ki}, b_{ij}\}_{k=1,\cdots,M, i=1,\cdots,H, j=1,\cdots,N}$, since in this section we always have $k = 1, \cdots, M$, $j = 1, \cdots, N$.

Theorem 7 *The singularity of the mixture of N dimensional Poisson distributions with H components and the true distribution with r components, corresponds to the Vandermonde matrix type singularity with* $M = 1, Q = 1$ *and* $\sum_{i=1}^{H} a_{1i} = 1, a_{1i} > 0$.

These log canonical thresholds of Vandermonde matrix-type singularities provide the learning coefficients of three-layered neural networks, normal mixture models, and mixtures of binomial distributions[33], which are known as effective learning models and widely used. This fact shows that these singularities are essential and generic in learning theory.

Example 8 *If* $Q = N = M = r = 1$, *then we have*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & -1 \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{11}^2 & \cdots & b_{11}^{H+1} \\ b_{21} & b_{21}^2 & \cdots & b_{21}^{H+1} \\ \vdots & & & \vdots \\ b_{H+1,1}^{*} & b_{H+1,1}^{*2} & \cdots & b_{H+1,1}^{*H+1} \end{pmatrix}.$$

These matrices A, B *correspond to the mixture of* 1 *dimensional Poisson distributions with H components:*

$$p(x|w) = \sum_{i=1}^{H} a_{1i} \exp(-b_{i1}) \frac{b_{i1}^{x_1}}{x_1!},$$

and the true distribution:

$$p(x|w) = \exp(-b_{H+1,1}^{*}) \frac{b_{H+1,1}^{*x_1}}{x_1!}.$$

Example 9 *If* $H = 2, N = 2, Q = r = M = 1$, *then we have*

$$A = \begin{pmatrix} a_{11} & a_{12} & -1 \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} & b_{11}^2 & b_{11}b_{12} & b_{12}^2 & b_{11}^3 & b_{11}b_{12}^2 & b_{11}^2 b_{12} & b_{12}^3 \\ b_{21} & b_{22} & b_{21}^2 & b_{21}b_{22} & b_{22}^2 & b_{21}^3 & b_{21}b_{22}^2 & b_{21}^2 b_{22} & b_{22}^3 \\ b_{31}^{*} & b_{32}^{*} & b_{31}^{*2} & b_{31}^{*}b_{32}^{*} & b_{32}^{*2} & b_{31}^{*3} & b_{31}^{*}b_{32}^{*2} & b_{31}^{*2}b_{32}^{*} & b_{32}^{*3} \end{pmatrix}.$$

These matrices A, B *correspond to the model*

$$p(x|w) = a_{11} \exp(-b_{11} - b_{12}) \frac{b_{11}^{x_1} b_{12}^{x_2}}{x_1! x_2!} + a_{12} \exp(-b_{21} - b_{22}) \frac{b_{21}^{x_1} b_{22}^{x_2}}{x_1! x_2!}.$$

and the true distribution:

$$p(x|w) = \exp(-b_{31}^{*} - b_{32}^{*}) \frac{b_{31}^{*x_1} b_{32}^{*x_2}}{x_1! x_2!}.$$

Theorem 10 ([23]) *Consider variables* $w = \{a_{ki}, b_{ij}\}_{1 \le i \le H}$ *in a sufficiently small neighborhood U of*

$$w^{*} = \{a_{ki}^{*}, b_{ij}^{*}\}_{1 \le i \le H}.$$

Set $(b^{**}_{01}, b^{**}_{02}, \cdots, b^{**}_{0N}) = (0, \ldots, 0)$. Assume each $(b^{**}_{11}, b^{**}_{12}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, b^{**}_{r'2}, \cdots, b^{**}_{r'N})$ be a different real vector in

$$[b^*_{i1}, b^*_{i2}, \cdots, b^*_{iN}]_Q \neq 0, \text{ for } i = 1, \ldots, H{+}r;$$

that is,

$$\{(b^{**}_{11}, \cdots, b^{**}_{1N}), \ldots, (b^{**}_{r'1}, \cdots, b^{**}_{r'N}) \mid [b^*_{i1}, \cdots, b^*_{iN}]_Q \neq 0, i = 1, \ldots, H{+}r\}.$$

The value $r' \geq r$ is determined uniquely by Definition 6. Let $(b^{**}_{i1}, \cdots, b^{**}_{iN}) = [b^*_{H+i,1}, \cdots, b^*_{H+i,N}]_Q$ for $i = 1, \cdots, r$.

$$\text{Set } [b^*_{i1}, \cdots, b^*_{iN}]_Q = \begin{cases} \quad 0, & i = 1, \cdots, H_0 \\ (b^{**}_{11}, \cdots, b^{**}_{1N}), & i = H_0{+}1, \cdots, H_0{+}H_1, \\ (b^{**}_{21}, \cdots, b^{**}_{2N}), & i = H_0{+}H_1{+}1, \cdots, H_0{+}H_1{+}H_2, \\ \quad \vdots & \\ (b^{**}_{r'1}, \cdots, b^{**}_{r'N}), & i = H_0{+}\cdots{+}H_{r'-1}{+}1, \cdots, H_0{+}\cdots{+}H_{r'}, \end{cases}$$

and $H_0{+}\cdots{+}H_{r'} = H$. Then, we have

$$\lambda_{w^*}(\|A_{M,H,r}B^{(Q)}_{H,N,r}\|^2) = \frac{Mr'}{2} + \lambda_{w_1^{(0)*}}(\|A_{M,H_0,0}B^{(Q)}_{H_0,N,0}\|^2)$$

$$+ \sum_{\alpha = 1}^{r} \lambda_{w_1^{(\alpha)*}}(\|A_{M,H_\alpha-1,1}B^{(1)}_{H_\alpha,N,0}\|^2) + \sum_{\alpha = r+1}^{r'} \lambda_{w_1^{(\alpha)*}}(\|A_{M,H_\alpha-1}B^{(1)}_{H_\alpha-1,N,0}\|^2),$$

where $w_1^{(0)*} = \{a^*_{k,i}, 0\}_{1 \leq i \leq H_\alpha}$, $w_1^{(\alpha)*} = \{a^*_{k,H_0+\cdots+H_{\alpha-1}+i}, 0\}_{2 \leq i \leq H_\alpha}$, $\mathbf{a}^{(\alpha)^*} = \begin{pmatrix} a^*_{1,H+\alpha} \\ \vdots \\ a^*_{M,H+\alpha} \end{pmatrix}$ and

$A_{M,H_\alpha-1,1} = (A_{M,H_\alpha-1,0}, \mathbf{a}^{(\alpha)^*})$ for $\alpha \geq 1$.

Theorem 11 *[18] We use the same notation as in Theorem 10. Assume that $Q = 1$, $r = 1$, $\sum_{i=1}^{H} a_{ki} = 1$, $a^*_{iH+1} = -1$ and $a_{ki} \geq 0$. Then, the ideal is generated by*

$$\sum_{i=1}^{H} a_{ki} \prod_{j=1}^{N} (b_{ij} - b^*_{i,H+1})^{\ell_j} \; (I = (\ell_1, \cdots, \ell_N) \in \{0, 1\}^N, |I| \neq 0),$$

$$a_{ki}(b_{ij} - b^*_{i,H+1})^2 \; (1 \leq k \leq M, 1 \leq i \leq H, 1 \leq j \leq N).$$

By Theorem 10, we can see that the case $r = 1$ in Theorem 11 is an essential part.

Theorem 12 *We use the same notation as in Theorem 10.*

Assume that $Q = 1$, $r = 1$, $\sum_{i=1}^{H} a_i = 1$, $a^*_{H+1} = -1$ and $a_i \geq 0$.

We have the following:

$$\begin{aligned}&\lambda_{w^*}(\|A_{1,H,1}B^{(1)}_{H,N,1}\|^2)\\&=\min\{\frac{N(H-\beta)+\beta}{2}(0\le\beta\le H-1),\frac{N(H-\beta')+2\beta'+N}{4}(0\le\beta'\le H-1)\}\\&=\begin{cases}(H+1)/4 & \text{if } N=1\\(H-1+N)/2 & \text{if } N\ge 2.\end{cases}\end{aligned}$$

2. Conclusion

In this paper, we considered the case when $Q=1$ and the elements of matrix A are non-negative in Vandermonde matrix-type singularities (Definition 6). Theorem 12 determines the explicit values of the log canonical thresholds. These results are related to Poisson distribution mixture models and also a normal mixture model with identity matrix variances [23]. Since the log canonical thresholds of Vandermonde matrix-type singularities have been still obtained partially and many of learning coefficients seem to be related to such singularities, these results in this paper are useful for obtaining log canonical thresholds for other cases.

Drton and Plummer [13] have used the learning coefficients from our recent results very effectively in sBIC. Furthermore, our theoretical and mathematical values will be helpful in numerical experiments such as the Markov chain Monte Carlo. In the papers [34,35], Nagata have constructed the mathematical foundation for developing and analyzing the precision of the Markov chain Monte Carlo method by using our theoretical values of marginal likelihoods.

Acknowledgements

This research was funded by the Ministry of Education, Culture, Sports, Science and Technology in Japan, Grant-in-Aid for Scientific Research of the ideal: 18K11479.

A. Proof of Theorem 12

We use the following theorem in the proof.

Theorem 13 (Method to add variables [26]) *Let* $f_1(w_1,\ldots,w_d)$, $\ldots$, $f_m(w_1,\ldots,w_d)$ *be homogeneous functions of* $w_1,\cdots,w_d$. *Set* $f_1'(w_2,\ldots,w_d)=f_1(1,w_2,\ldots,w_d)$, $\ldots$, $f_m'(w_2,\ldots,w_d)=f_m(1,w_2,\ldots,w_d)$. *If* $w_1^*\neq 0$, *then we have:*

$$\lambda_{(w_1^*,\cdots,w_d^*)}(f_1^2+\cdots+f_m^2)=\lambda_{(w_2^*/w_1^*,\cdots,w_d^*/w_1^*)}(f_1'^2+\cdots+f_m'^2).$$

Let us consider the generators of the ideals

$$\mathcal{J}=\left\langle\begin{pmatrix}a_1\cdots a_H\end{pmatrix}\begin{pmatrix}b_{11}^{\ell_1}\cdots b_{1N}^{\ell_N}\\b_{21}^{\ell_1}\cdots b_{2N}^{\ell_N}\\\vdots\\b_{H1}^{\ell_1}\cdots b_{HN}^{\ell_N}\end{pmatrix}\mid\sum_{i=1}^{N}\ell_i=nQ+1,n\ge 0\right\rangle.$$

Because we assume that $Q = 1$ and $a_i \geq 0$, we can set

$$\mathcal{J} = \left\langle \begin{pmatrix} a_1 \cdots a_H \end{pmatrix} \begin{pmatrix} b_{11}^{\ell_1} \cdots b_{1N}^{\ell_N} \\ b_{21}^{\ell_1} \cdots b_{2N}^{\ell_N} \\ \vdots \\ b_{H1}^{\ell_1} \cdots b_{HN}^{\ell_N} \end{pmatrix} \mid \sum_{i=1}^{N} \ell_i = n{+}1, n \geq 0 \right\rangle$$

$$+ \left\langle a_i b_{ij}^2 \mid 1 \leq i \leq H, 1 \leq j \leq N \right\rangle .$$

By constructing the blowup repeatedly, we have the following:

Set $\alpha[i] \in \{1, \cdots, N\}$ and set $b_{ij} = v_1 \cdots v_i b'_{ij}$ for $1 \leq i \leq H, 1 \leq j \leq N$, and set $b'_{i\alpha[i]} = 1$. Then we have

$$\mathcal{J} = \left\langle v_1^2 a_1, v_1^2 v_2^2 a_2, \cdots, v_1^2 v_2^2 \cdots v_H^2 a_H, \right\rangle$$

$$+ \left\langle \begin{pmatrix} a_1 v_1 & a_2 v_1 v_2 & \cdots & a_H v_1 v_2 \cdots v_H \end{pmatrix} \begin{pmatrix} b'_{1,j} \\ b'_{2,j} \\ \vdots \\ b'_{Hj} \end{pmatrix} \mid 1 \leq j \leq N \right\rangle .$$

Set $a_H = 1 - a_1 - \cdots - a_{H-1}$. Then we have

$$\mathcal{J} = \left\langle v_1^2 a_1, v_1^2 v_2^2 a_2, \cdots, v_1^2 v_2^2 \cdots v_H^2 \right\rangle$$

$$+ \left\langle \begin{pmatrix} a_1 v_1 & a_2 v_1 v_2 & \cdots & v_1 v_2 \cdots v_H \end{pmatrix} \begin{pmatrix} b'_{1,j} - v_2 \cdots v_H b'_{Hj} \\ b'_{2,j} - v_3 \cdots v_H b'_{Hj} \\ \vdots \\ b'_{H-1,j} - v_H b'_{Hj} \\ b'_{Hj} \end{pmatrix} \mid 1 \leq j \leq N \right\rangle .$$

Set $\begin{pmatrix} b''_{1,j} \\ b''_{2,j} \\ \vdots \\ b''_{H-1,j} \\ b''_{Hj} \end{pmatrix} = \begin{pmatrix} b'_{1,j} - v_2 \cdots v_H b'_{Hj} \\ b'_{2,j} - v_3 \cdots v_H b'_{Hj} \\ \vdots \\ b'_{H-1,j} - v_H b'_{Hj} \\ b'_{Hj} \end{pmatrix}$, $1 \leq \beta_1 \leq H, a_1 = v_2 \cdots v_{\beta_1} a'_1 u_1, a_2 = v_3 \cdots v_{\beta_1} a'_2 u_1,$ $\cdots, a_{\beta_1 - 1} = v_{\beta_1} a'_{\beta_1 - 1} u_1$, $a_{\beta_1} = a'_{\beta_1} u_1$, and $v_{\beta_1 + 1} = v'_{\beta_1 + 1} u_1$. Also set $a'_{i_1} = 1 (i_1 \leq \beta_1)$. By Theorem 13, we can assume $b''_{i_1 j}$ is a variable.

Then by setting $b'''_{i_1 j} = \begin{pmatrix} a'_1 & a'_2 & \cdots & a'_{\beta_1} & a'_{\beta_1+1} v'_{\beta_1+1} & \cdots & v'_{\beta_1+1} \cdots v_H \end{pmatrix} \begin{pmatrix} b''_{1,j} \\ b''_{2,j} \\ \vdots \\ b''_{Hj} \end{pmatrix}$ we have

$$\begin{aligned}\mathcal{J} &= \left\langle v_1^2 v_2 \cdots v_{\beta_1} a_1'' u_1, v_1^2 v_2^2 v_3 \cdots v_{\beta_1} a_2'' u_1, \cdots, v_1^2 v_2^2 \cdots v_{i_1}^2 v_{i_1+1} \cdots v_{\beta_1} a_{i_1}'' u_1 \right\rangle \\ &\quad + \left\langle u_1 v_1 v_2 \cdots v_{\beta_1} b_{i_1 j}''', 1 \leq j \leq N \right\rangle .\end{aligned}$$

The Jacobian is

$$v_1^{NH} v_2^{N(H-1)+1} \cdots v_{\beta_1}^{N(H-\beta_1+1)+\beta_1-1} u_1^{N(H-\beta_1)+\beta_1} \times v_{\beta_1+1}^{N(H-\beta_1)} \cdots v_H^N \times 1/(v_1 \cdots v_H u_1)$$

References

[1] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* **1978**, *6(2)*, 461–464.

[2] Rissanen, J. Stochastic complexity and modeling. *Annals of Statistics* **1986**, *14*, 1080–1100.

[3] Yamanishi, K. A decision-theoretic extension of stochastic complexity and its applications to learning. *IEEE Trans. on Information Theory* **1998**, *44(4)*, 1424–1439.

[4] Akaike, H. Likelihood and the Bayes procedure. *In J. M. Bernald, editor, Bayesian Statistics. University Press, Valencia, Spain* **1980**, 143–166.

[5] Mackay, D. J. Bayesian interpolation. *Neural Computation* **1992**, 415–447.

[6] Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723.

[7] Watanabe, S. Algebraic analysis for nonidentifiable learning machines. *Neural Comput.* **2001**, *13*, 899–933.

[8] Watanabe, S. Algebraic geometrical methods for hierarchical learning machines. *Neural Netw.* **2001**, *14*, 1049–1060.

[9] Watanabe, S. Algebraic geometry of learning machines with singularities and their prior distributions. *J. Jpn. Soc. Artif. Intell.* **2001**, *16*, 308–315.

[10] Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*; Cambridge University Press: New York, NY, USA, 2009; Volume 25.

[11] Watanabe, S. Equations of states in singular statistical estimation. *Neural Netw.* **2010**, *23*, 20–34.

[12] Watanabe, S. *Mathematical Theory of Bayesian Statistics*; CRC Press: New York, NY, USA, 2018.

[13] Drton, M.; Plummer, M. A Bayesian information criterion for singular models. *J. R. Statist. Soc. B* **2017**, *79*, 1–38.

[14] Watanabe, S. *A widely applicable bayesian information criterion*; *Journal of Machine Learning Research* **2013**, *14*, 867–897.

[15] Levin, E., Tishby, N. and Solla, S. A. *A statistical approaches to learning and generalization in layered neural networks*; *In Proceedings of IEEE* **1990**, *78*, 1568–1674.

[16] Amari, S., Fujita, N. and Shinomoto, S. *Four types of learning curves*. *Neural Computation* **1992**, *4(4)*, 608–618.

[17] Amari, S. and Murata, N. *Statistical theory of learning curves under entropic loss criterion*. *Neural Computation* **1993**, *5*, 140–153.

[18] Sato, K. and Watanabe, S. *Real log canonical threshold and bayesian generalization error of mixture of poisson distributions*. *IBISML* **2018**, *14(3)*, 1–6.

[19] Aoyagi, M.; Watanabe, S. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Netw.* **2005**, *18*, 924–933.

[20] Aoyagi, M.; Watanabe, S. Resolution of singularities and the generalization error with Bayesian estimation for layered neural network. *IEICE Trans. J88-D-II* **2005**, *10*, 2112–2124.

[21] Aoyagi, M. The zeta function of learning theory and generalization error of three layered neural perceptron. *RIMS Kokyuroku Recent Top. Real Complex Singul.* **2006**, *1501*, 153–167.

[22] Aoyagi, M. Log canonical threshold of Vandermonde matrix type singularities and generalization error of a three layered neural network. *Int. J. Pure Appl. Math.* **2009**, *52*, 177–204.

[23] Aoyagi, M. A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities. *Commun. Stat. Theory Methods* **2010**, *39*, 2667–2687.

[24] Aoyagi, M.; Nagata, K. Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix type singularity. *Neural Comput.* **2012**, *24*, 1569–1610.

[25] Aoyagi, M. Learning coefficient in Bayesian estimation of restricted Boltzmann machine. *J. Algebr. Stat.* **2013**, *4*, 30–57.

[26] Aoyagi, M. Consideration on singularities in learning theory and the learning coefficient. *Entropy* **2013**, *15*, 3714–3733.

[27] Rusakov, D.; Geiger, D. Asymptotic Model Selection for Naive Bayesian Networks. In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, Alberta, AB, Canada, 1–4 August 2002; pp. 438–445.

[28] Rusakov, D.; Geiger, D. Asymptotic model selection for naive Bayesian networks. *J. Mach. Learn. Res.* **2005**, *6*, 1–35.

[29] Zwiernik, P. An asymptotic behavior of the marginal likelihood for general Markov models. *J. Mach. Learn. Res.* **2011**, *12*, 3283–3310.

[30] Drton, M.; Lin, S.; Weihs, L.; Zwiernik, P. Marginal likelihood and model selection for Gaussian latent tree and forest models. *Bernoulli* **2017**, *23*, 1202–1232.

[31] Hironaka, H. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Ann. Math.* **1964**, *79*, 109–326.

[32] Lin, S. Asymptotic approximation of marginal likelihood integrals. *arXiv* **2010**, arXiv:1003.5338v2.

[33] Yamazaki, K.; Aoyagi, M.; Watanabe, S. Asymptotic analysis of Bayesian generalization error with Newton diagram. *Neural Network.* **2010**, *23*, 35–43.

[34] Nagata, K.; Watanabe, S. Exchange Monte Carlo Sampling from Bayesian posterior for singular learning machines. *IEEE Trans. Neural Netw.* **2008**, *19*, 1253–1266.

[35] Nagata, K.; Watanabe, S. Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *Int. J. Neural Netw.* **2008**, *21*, 980–988.