

Article

Learning Coefficient of Vandermonde Matrix-Type Singularities in Model Selection

Miki Aoyagi 

Department of Mathematics, College of Science & Technology, Nihon University, 1-8-14, Surugadai, Kanda, Chiyoda-ku, Tokyo 101-8308, Japan; aoyagi.miki@math.cst.nihon-u.ac.jp; Tel.: +81-3-3259-0883

Received: 11 March 2019; Accepted: 29 May 2019; Published: 4 June 2019



Abstract: In recent years, selecting appropriate learning models has become more important with the increased need to analyze learning systems, and many model selection methods have been developed. The learning coefficient in Bayesian estimation, which serves to measure the learning efficiency in singular learning models, has an important role in several information criteria. The learning coefficient in regular models is known as the dimension of the parameter space over two, while that in singular models is smaller and varies in learning models. The learning coefficient is known mathematically as the log canonical threshold. In this paper, we provide a new rational blowing-up method for obtaining these coefficients. In the application to Vandermonde matrix-type singularities, we show the efficiency of such methods.

Keywords: learning coefficient; Kullback function; singular learning machine; resolution of singularities

1. Introduction

In recent studies, real data associated with, for example, image or speech recognition, psychology, and economics, have been analyzed by learning systems. Hence, many learning models have been proposed, and thus, the need for appropriate model selection methods has increased.

In this section, we first introduce the widely-applicable information criterion (WAIC) [1–7] and cross-validation in Bayesian model selection.

Let $q(x)$ be a true probability density function of variables, $x \in \mathbb{R}^N$, and let $x^n := \{x_i\}_{i=1}^n$ be n training samples selected from $q(x)$ independently and identically. Consider a learning model that is written in probabilistic form as $p(x|w)$, where $w \in W \subset \mathbb{R}^d$ is a parameter.

Suppose that the purpose of the learning system is to estimate the unknown true density function $q(x)$ from x^n using $p(x|w)$ in Bayesian estimation. Let $\psi(w)$ be an a priori probability density function on parameter set W and $p(w|x^n)$ be the a posteriori probability density function:

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

where:

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

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

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

and:

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

We then have predictive density function $p(x|x^n) = E_w^\beta[p(x|w)]$, which is the average inference of the Bayesian density function.

We next introduce Kullback function $K(q||p)$ and empirical Kullback function $K_n(q||p)$ for density functions $p(x), q(x)$:

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

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

Function $K(p||q)$, which always has a non-negative value and satisfies $K(q||p) = 0$, if and only if $q(x) = p(x)$, is a pseudo-distance between density functions $p(x), q(x)$. 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:

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

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:

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

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 function approximates the true density function.

We define $x^n \setminus x_i = \{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n\}$. The WAIC is denoted by:

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

and the cross-validation loss is denoted by:

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

for $n \geq 2$.

Watanabe [2,3,6,7] proved the following four relations:

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

for learning coefficient $\lambda \in \mathbb{Q}$ and singular fluctuation $\nu \in \mathbb{R}$, where $L(w) = -E_x[\log p(x|w)]$ and $w_0 \in W_0 = \{w_0 \in W \mid L(w_0) = \min_{w \in W} L(w)\}$. For singular models, the set W_0 is typically not a singleton set. Nevertheless, the WAIC and the cross-validation can estimate the Bayesian generalization error without any knowledge of the true probability density function.

These values are calculated from training samples x_i using learning model p . In real applications or experiments, we typically do not know the true distribution, but only the values of the training errors. Our purpose is to show that these methods are effective. We can select a suitable model from several statistical models by observing these values.

In this paper, we consider the value λ , which is equal to the log canonical threshold introduced in Definition 1. This coefficient is not needed to evaluate the WAIC and the cross-validation in practice, while the learning coefficients from our recent results have been used very effectively by Drton and Plummer [8] for model selection using a method called the “singular Bayesian information criterion (sBIC)”. The method sBIC even works using the bounds of learning coefficients.

It is known that $\lambda = \nu = d/2$ holds, where d is the dimension of the parameter space for regular models. Value λ is obtained by a blowing-up process, which is the main tool in the desingularization of an algebraic variety. The following theorem is an analytic version of Hironaka’s theorem [9] used by Atiyah [10].

Theorem 1 (Desingularization [9]). *Let f be an analytic function in a neighborhood of $w^* = (w_1^*, \dots, w_d^*) \in \mathbf{R}^d$ with $f(w^*) = 0$. There exists an open set $U \ni w^*$, an analytic manifold M , and a proper analytic map μ from M to U such that: (1) $\mu : M - \mathcal{E} \rightarrow U - f^{-1}(0)$ is an isomorphism, where $\mathcal{E} = \mu^{-1}(f^{-1}(0))$, and: (2) for each $u \in M$, there is a local analytic coordinate system (u_1, \dots, u_n) such that $f(\mu(u)) = \pm u_1^{s_1} u_2^{s_2} \dots u_n^{s_n}$, where s_1, \dots, s_n are non-negative integers.*

The theorem establishes the existence of the desingularization map; however, generally, it is still difficult to obtain such maps for Kullback functions because the singularities of these functions are very complicated. From learning coefficient λ and its order θ , value ν is obtained theoretically as follows: Let $\zeta(u)$ be an empirical process defined on the manifold obtained by a resolution of singularities and \sum_{u^*} denote the sum of local coordinates that attain the minimum λ and maximum θ . We then have:

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

where $\zeta(u)$ is a random variable of a Gaussian process with mean zero, and covariance $E_{\zeta}[\zeta(w)\zeta(u)] = E_X[a(x,w)a(x,u)]$ for the analytic function $a(x,w)$ obtained by the resolution of singularities using $\log(q(x)/p(x|w))$.

Our purpose in this paper is to obtain λ . In recent studies, we determined the learning coefficients for reduced rank regression [11], the three-layered neural network with one input unit and one output unit [12,13], normal mixture models with a dimension of one [14], and the restricted Boltzmann

machine [15]. Additionally, Rusakov and Geiger [16,17] and Zwiernik [18], respectively, obtained the learning coefficients for naive Bayesian networks and directed tree models with hidden variables. Drton et al. [19] considered these coefficients of the Gaussian latent tree and forest models.

The papers [20,21] derived bounds on the learning coefficients for Vandermonde matrix-type singularities and explicit values under some conditions.

The remainder of this paper is structured as follows: In Section 2, we introduce log canonical thresholds in algebraic geometry. In Section 3, we summarize key theorems for obtaining learning coefficients for learning theory. In Section 4, we present our main results. We consider the log canonical thresholds of Vandermonde matrix-type singularities (Definition 3). We present our conclusions in Section 5.

2. Log Canonical Threshold

Definition 1. Let f be an analytic function in neighborhood U of w^* . Let ψ be a C^∞ function with a compact support. Define log canonical threshold $\lambda_{w^*}(f, \psi)$ as the largest pole of $\int_U |f|^{2z} \psi dw$ over \mathbb{C} or $\int_U |f|^z \psi dw$ over \mathbb{R} . Additionally, define $\theta_{w^*}(f, \psi)$ by its order. If $\psi(w^*) \neq 0$, then we define $\lambda_{w^*}(f) = \lambda_{w^*}(f, \psi)$ and $\theta_{w^*}(f) = \theta_{w^*}(f, \psi)$ because the log canonical threshold and its order are independent of ψ .

Applying Hironaka’s Theorem 1 to function $f(w)$, we have the proper analytic map μ from manifold M to neighborhood U of w^* that satisfies Hironaka’s Theorems (1) and (2). Then, integration $\int_U |f|^z \psi(w) dw$ is equal to $\int_M |f \circ \mu|^z \psi(\mu(u)) |\mu'(u)| du$, which is the sum of $\int_{U_M} |u_1^{s_1} u_2^{s_2} \cdots u_d^{s_d}|^z \psi(\mu(u)) |\mu'(u)| du$, where (u_1, \dots, u_d) is a local analytic coordinate system on $U_M \subset M$. Therefore, the poles can be obtained. Note that for each w^* with $f(w^*) \neq 0$, there exists neighborhood U such that $f(w) \neq 0$ for all $w \in U$. Thus, $\int_U |f|^z \psi(w) dw$ has no poles. The learning coefficient is the log canonical threshold of the Kullback function (relative entropy) over the real field.

3. Main Theorems

In this section, several theorems are introduced for obtaining real log canonical thresholds over the real field. Theorem 2 (the method for determining the deepest singular point), Theorem 3 (the method to add variables), and Theorem 4 (the rational blowing-up method) are very helpful for obtaining the log canonical threshold. These theorems over the real field are useful for reducing the number of changes of coordinates via blow-ups.

We denote constants, such as a^*, b^* , and w^* , by suffix $*$. Define the norm of a matrix $C = (c_{ij})$ as $\|C\| = \sqrt{\sum_{i,j} |c_{ij}|^2}$. Set $\mathbb{N}_{+0} = \mathbb{N} \cup \{0\}$.

Lemma 1 ([14,22,23]). Let U be a neighborhood of $w^* \in \mathbb{R}^d$. Consider the ring of analytic functions on U . Let \mathcal{J} be the ideal generated by f_1, \dots, f_n , which are analytic functions defined on U . (1) If $g_1^2 + \dots + g_m^2 \leq f_1^2 + \dots + f_n^2$, then $\lambda_{w^*}(g_1^2 + \dots + g_m^2) \leq \lambda_{w^*}(f_1^2 + \dots + f_n^2)$. (2) If $g_1, \dots, g_m \in \mathcal{J}$, then $\lambda_{w^*}(g_1^2 + \dots + g_m^2) \leq \lambda_{w^*}(f_1^2 + \dots + f_n^2)$. In particular, if g_1, \dots, g_m generate ideal \mathcal{J} , then $\lambda_{w^*}(f_1^2 + \dots + f_n^2) = \lambda_{w^*}(g_1^2 + \dots + g_m^2)$.

The following lemma is also used in the proofs.

Lemma 2 ([15]). Let $\mathcal{J}, \mathcal{J}'$ be the ideals generated by $f_1(w), \dots, f_n(w)$ and $g_1(w'), \dots, g_m(w')$, respectively. If w and w' are different variables, then:

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

Theorem 2 (Method for determining the deepest singular point [21]). Let $f_1(w_1, \dots, w_d), \dots, f_m(w_1, \dots, w_d)$ be homogeneous functions of w_1, \dots, w_j ($j \leq d$). Furthermore, let ψ be a C^∞ function

such that $\psi(0, \dots, 0, w_{j+1}^*, \dots, w_d^*) \geq \psi(w_1^*, \dots, w_d^*)$ and ψ_w is a homogeneous function of w_1, \dots, w_j in a small neighborhood of $(0, \dots, 0, w_{j+1}^*, \dots, w_d^*)$. Then, we have:

$$\lambda_{(0, \dots, 0, w_{j+1}^*, \dots, w_d^*)}(f_1^2 + \dots + f_m^2, \psi) \leq \lambda_{(w_1^*, \dots, w_j^*, w_{j+1}^*, \dots, w_d^*)}(f_1^2 + \dots + f_m^2, \psi)$$

Theorem 3 (Method to add variables [21]). Let $f_1(w_1, \dots, w_d), \dots, f_m(w_1, \dots, w_d)$ be homogeneous functions of w_1, \dots, w_d . Set $f_1'(w_2, \dots, w_d) = f_1(1, w_2, \dots, w_d), \dots, f_m'(w_2, \dots, w_d) = f_m(1, w_2, \dots, w_d)$. If $w_1^* \neq 0$, then we have:

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

Resolutions of singularities are obtained by constructing the blow-up along the smooth submanifold. In this paper, we use the blow-up method along some singular varieties as explained below for obtaining log canonical thresholds.

Theorem 4 (Rational blow-up process). Let $m \leq d$ and $(\alpha_1, \dots, \alpha_m) \in \mathbf{R}^m, \alpha_1, \dots, \alpha_m > 0$. Consider the set $U = \{w = (w_1, \dots, w_d) \in \mathbf{R}^d \mid 0 \leq w_1, \dots, w_d \leq 1\}$.

We set $g_i(u_i) = (g_{1i}(u_i), \dots, g_{di}(u_i))$ for $g_{ii}(u_i) = u_{ii}, g_{ji}(u_i) = u_{ji}u_{ii}^{\alpha_j/\alpha_i} (1 \leq j \leq m, j \neq i), g_{ji}(u_i) = u_{ji} (1 \leq j \leq d)$.

Let f, ψ be analytic functions defined on U and $f_i(u_{1i}, \dots, u_{di}) = f(g_i(u_i))$. Then, we have:

$$\min_{1 \leq i \leq m} \{ \lambda_0(f_i(u_i), u_{ii}^{\sum_{j=1}^m \alpha_j/\alpha_i - 1} \psi(g_i(u_i))) \} = \lambda_0(f, \psi).$$

Proof. The proof of this theorem uses a resolution of singularities along a smooth submanifold.

Set $w'_i = w_i^{1/\alpha_i}$, and construct the blow-up along the submanifold $\{w'_1 = \dots = w'_m = 0\}$. Then, we have for $1 \leq i \leq m, w'_i = v, w'_j = vw''_j (1 \leq j \leq m, j \neq i)$, and $w'_j = w''_j (m+1 \leq j \leq d)$.

Consider the set $U_i = \{u_i = (u_{1i}, \dots, u_{di}) \in \mathbf{R}^d \mid 0 \leq u_{1i}, \dots, u_{di} \leq 1\}$ for $1 \leq i \leq m$.

Set $u_{ii} = v^{\alpha_i}, u_{ji} = w''_j \alpha_j (j \neq i)$, then we have $w_i = u_{ii}, w_j = u_{ii}^{\alpha_j/\alpha_i} u_{ji} (1 \leq j \leq m, j \neq i)$, and $w_j = u_{ji} (m+1 \leq j \leq d)$.

The Jacobian $\partial w / \partial u_i$ is $\prod_{1 \leq j \leq m, j \neq i} u_{ii}^{\alpha_j/\alpha_i}$. \square

The theorem is simple; however, it is a useful tool for obtaining log canonical thresholds.

4. Main Results

In this section, we apply the theorems in Section 3 to Vandermonde matrix-type singularities, which are generic and essential in learning theory. Their associated log canonical thresholds provide the learning coefficients of, for example, three-layered neural networks in Section 4.1, normal mixture models in Section 4.2, and mixtures of binomial distributions [24].

4.1. Three-Layered Neural Network

Consider the three-layered neural network with N input units, H hidden units, and M output units, which is trained for estimating the true distribution with r hidden units. Denote an input value by $z = (z_j) \in \mathbf{R}^N$ with a probability density function $q(z)$. Then, an output value $y = (y_k) \in \mathbf{R}^M$ of the three layered neural network is given by $y_k = f_k(z, w) + (\text{noise})$, where $w = \{a_{ki}, b_{ij}\}_{1 \leq i \leq H}$ and:

$$f_k(z, w) = \sum_{i=1}^H a_{ki} \tanh\left(\sum_{j=1}^N b_{ij} z_j\right).$$

Consider a statistical model:

$$p(y|z, w) = \frac{1}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2} \|y - f(z, w)\|^2\right),$$

and $p(z, y|w) = p(y|z, w)q(z)$. Assume that the true distribution:

$$p(y|z, w_t^*) = \frac{1}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2} \|y - f(z, w_t^*)\|^2\right),$$

is included in the learning model, where $w_t^* = \{a_{k,H+i}^*, b_{H+i,j}^*\}_{1 \leq i \leq r}$ and $f_k(z, w_t^*) = \sum_{i=1}^r (-a_{k,H+i}^*) \tanh(\sum_{j=1}^N b_{H+i,j}^* x_j)$.

We have:

$$p(z, y|w) = \frac{1}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2} \|y - f(z, w)\|^2\right)q(z),$$

and the notation (z, y) for the three-layered neural network corresponds to x in Sections 1 and 2.

4.2. Normal Mixture Models

We consider a normal mixture model [14] with identity matrix variances:

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

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

Set the true distribution by:

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

where $w_t^* = \{a_i^*, b_{ij}^*\}_{H+1 \leq i \leq H+r}$ and $\sum_{i=H+1}^{H+r} a_i^* = -1, a_i^* < 0$. (In order to simplify the following, we use the values $a_i^* < 0$, not $a_i^* > 0$.)

4.3. Vandermonde Matrix-Type Singularities

Definition 2. Fix $Q \in \mathbb{N}$. Define: $[b_1^*, b_2^*, \dots, b_N^*]_Q = \gamma_i(0, \dots, 0, b_i^*, \dots, b_N^*)$ if $b_1^* = \dots = b_{i-1}^* = 0, b_i^* \neq 0$, and $\gamma_i = \begin{cases} 1 & \text{if } Q \text{ is odd,} \\ \text{sign}(b_i^*) & \text{if } Q \text{ is even.} \end{cases}$

For simplicity, we use the notation $w = \{a_{ki}, b_{ij}\}_{1 \leq i \leq H}$ instead of $w = \{a_{ki}, b_{ij}\}_{1 \leq k \leq M, 1 \leq i \leq H, 1 \leq j \leq N}$ because we always have $1 \leq k \leq M$ and $1 \leq j \leq N$ in this section.

Definition 3. Fix $Q \in \mathbb{N}$ and $m \in \mathbb{N}_{+0}$.

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

$$B_I = \left(\prod_{j=1}^N b_{1j'}^{\ell_j}, \prod_{j=1}^N b_{2j'}^{\ell_j}, \dots, \prod_{j=1}^N b_{Hj'}^{\ell_j}, \prod_{j=1}^N b_{H+1,j}^{\ell_j}, \dots, \prod_{j=1}^N b_{H+r,j}^{\ell_j} \right)^t,$$

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

$$B = (B_I)_{\ell_1 + \dots + \ell_N = Qn+m, n \geq 0} = (B_{(m,0,\dots,0)}, B_{(m-1,1,\dots,0)}, \dots, B_{(0,0,\dots,m)}, B_{(m+Q,0,\dots,0)}, \dots)$$

(t denotes the transpose).

a_{ki} and b_{ij} ($1 \leq k \leq M, 1 \leq i \leq H, 1 \leq j \leq N$) are the variables in a neighborhood of a_{ki}^* and b_{ij}^* , where a_{ki}^* and b_{ij}^* are fixed constants.

Let \mathcal{J} be the ideal generated by the elements of AB .

We call singularities of \mathcal{J} Vandermonde matrix-type singularities.

To simplify, we usually assume that:

$$(a_{1,H+j}^*, a_{2,H+j}^*, \dots, a_{M,H+j}^*)^t \neq 0, (b_{H+j,1}^*, b_{H+j,2}^*, \dots, b_{H+j,N}^*) \neq 0,$$

for $1 \leq j \leq r$ and:

$$[b_{H+j,1}^*, b_{H+j,2}^*, \dots, b_{H+j,N}^*]_Q \neq [b_{H+j',1}^*, b_{H+j',2}^*, \dots, b_{H+j',N}^*]_Q,$$

for $j \neq j'$.

Example 1. If $m = N = M = r = 1, Q = 2, H = 3$, then we have $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14}^* \end{pmatrix}$,

$$B = \begin{pmatrix} b_{11} & b_{11}^3 & b_{11}^5 & b_{11}^7 \\ b_{21} & b_{21}^3 & b_{21}^5 & b_{21}^7 \\ b_{31} & b_{31}^3 & b_{31}^5 & b_{31}^7 \\ b_{41}^* & b_{41}^{*3} & b_{41}^{*5} & b_{41}^{*7} \end{pmatrix}. \text{ These matrices } A, B \text{ correspond to the three-layered neural network:}$$

$$p(y|z, w) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}(y - a_{11} \tanh(b_{11}z) - a_{12} \tanh(b_{21}z) - a_{13} \tanh(b_{31}z))^2\right),$$

and the true distribution:

$$p(y|x, w_i^*) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2}(y + a_{14}^* \tanh(b_{41}^*z))^2\right).$$

Example 2. If $Q = r = M = 1, H = 2, N = 2$, then we have $A = \begin{pmatrix} a_{11} & a_{12} & a_{13}^* \end{pmatrix}$,

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

If $a_{13}^* = -1$, these matrices A, B correspond to a normal mixture model with identity matrix variances:

$$p(x|w) = \frac{a_{11}}{2\pi} \exp\left(-\frac{(z_1 - b_{11})^2 + (z_2 - b_{12})^2}{2}\right) + \frac{a_{12}}{2\pi} \exp\left(-\frac{(z_1 - b_{21})^2 + (z_2 - b_{22})^2}{2}\right),$$

$\sum_{i=1}^2 a_{1i} = 1, a_{1i} \geq 0$, and the true distribution is:

$$p(x|w_i^*) = \frac{1}{2\pi} (-a_{13}^*) \exp\left(-\frac{(z_1 - b_{31}^*)^2 + (z_2 - b_{32}^*)^2}{2}\right), -a_{13}^* = 1.$$

In this paper, we denote:

$$A_{M,H} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1H} \\ a_{21} & a_{22} & \dots & a_{2H} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{MH} \end{pmatrix}, B_{H,N,I} = \begin{pmatrix} \prod_{j=1}^N b_{1j}^{\ell_j} \\ \prod_{j=1}^N b_{2j}^{\ell_j} \\ \vdots \\ \prod_{j=1}^N b_{Hj}^{\ell_j} \end{pmatrix} \text{ and :}$$

$$B_{H,N}^{(Q)} = (B_{H,N,I})_{\ell_1+\dots+\ell_N=Qn+1, 0 \leq n \leq H-1}.$$

Furthermore, we denote: $\mathbf{a}^* = \begin{pmatrix} a_{1,H+1}^* \\ \vdots \\ a_{M,H+1}^* \end{pmatrix}$ and:

$$(A_{M,H}, \mathbf{a}^*) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1H} & a_{1,H+1}^* \\ a_{21} & a_{22} & \cdots & a_{2H} & a_{2,H+1}^* \\ & & \vdots & & \\ a_{M1} & a_{M2} & \cdots & a_{MH} & a_{M,H+1}^* \end{pmatrix}$$

Theorem 5 ([14]). Consider sufficiently small neighborhood U of:

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

and variables $w = \{a_{ki}, b_{ij}\}_{1 \leq i \leq H}$ in set U . Set $(b_{01}^{**}, b_{02}^{**}, \dots, b_{0N}^{**}) = (0, \dots, 0)$. Let each $(b_{11}^{**}, b_{12}^{**}, \dots, b_{1N}^{**}), \dots, (b_{r'1}^{**}, b_{r'2}^{**}, \dots, b_{r'N}^{**})$ be a different real vector in:

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

that is,

$$\{(b_{11}^{**}, \dots, b_{1N}^{**}), \dots, (b_{r'1}^{**}, \dots, b_{r'N}^{**}); [b_{i1}^*, \dots, b_{iN}^*]_Q \neq 0, i = 1, \dots, H+r\}.$$

Then, r' is uniquely determined, and $r' \geq r$ by the assumption in Definition 3. Set $(b_{i1}^{**}, \dots, b_{iN}^{**}) = [b_{H+i,1}^*, \dots, b_{H+i,N}^*]_Q$, for $1 \leq i \leq r$. Assume that $[b_{i1}^*, \dots, b_{iN}^*]_Q =$

$$\begin{cases} 0, & 1 \leq i \leq H_0 \\ (b_{11}^{**}, \dots, b_{1N}^{**}), & H_0 + 1 \leq i \leq H_0 + H_1, \\ (b_{21}^{**}, \dots, b_{2N}^{**}), & H_0 + H_1 + 1 \leq i \leq H_0 + H_1 + H_2, \\ \vdots \\ (b_{r'1}^{**}, \dots, b_{r'N}^{**}), & H_0 + \dots + H_{r'-1} + 1 \leq i \leq H_0 + \dots + H_{r'}, \end{cases}$$

and $H_0 + \dots + H_{r'} = H$. We then have:

$$\begin{aligned} \lambda_{w^*}(\|AB\|^2) &= \frac{Mr'}{2} + \lambda_{w_1^{(0)*}}(\|A_{M,H_0} B_{H_0,N}^{(Q)}\|^2) \\ &+ \sum_{\alpha=1}^{r'} \lambda_{w_1^{(\alpha)*}}(\|(A_{M,H_\alpha-1}, \mathbf{a}^{(\alpha)*}) B_{H_\alpha,N}^{(1)}\|^2) + \sum_{\alpha=r'+1}^{r'} \lambda_{w_1^{(\alpha)*}}(\|A_{M,H_\alpha-1} B_{H_\alpha-1,N}^{(1)}\|^2), \end{aligned}$$

where $w_1^{(0)*} = \{a_{ki}^*, 0\}_{1 \leq i \leq H_\alpha}$,

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

Theorem 6. We use the same notation as in Theorem 5. Set $\lambda = \lambda_0(\|A_{M,H} B_{H,N}^{(Q)}\|^2)$.

We have the following:

1. $H = 1$. $\lambda = \min\{\frac{M}{2}, \frac{N}{2}\}$.
2. $H = 2$. $\lambda = \min\{\frac{\beta N + (2 - \beta)M}{2}, \beta = 0, 1, 2, \frac{2N + Q(N - 1 + M)}{2Q + 2}\}$.
3. $H = 3$.
 $\lambda = \min\{\frac{\beta N + (3 - \beta)M}{2}, \beta = 0, 1, 2, 3,$

$$\frac{\beta N + (3 - \beta)M + Q(\alpha(N + \alpha - \beta) + (3 - \alpha)M)}{2(Q + 1)}, \alpha = 1, \dots, \beta - 1, \beta = 2, 3,$$

$$\frac{3N + Q(3N - 3 + 3M)}{2(2Q + 1)}\}$$

4. $H = 4$.

$$\lambda = \min\left\{\frac{\beta N + (4 - \beta)M}{2}, \beta = 0, 1, 2, 3, 4,\right.$$

$$\frac{\beta N + (4 - \beta)M + Q(\alpha(N + \alpha - \beta) + (4 - \alpha)M)}{2(Q + 1)}, \alpha = 1, \dots, \beta - 1, \beta = 2, 3, 4,$$

$$\frac{4N + Q(\alpha N - \alpha - 1 + (8 - \alpha)M)}{2(2Q + 1)}, \alpha = 2, 3, 4, \frac{4N + Q(5N - 5 + 3M)}{2(2Q + 1)},$$

$$\frac{3N + M + Q(\alpha N - \alpha + (8 - \alpha)M)}{2(2Q + 1)}, \alpha = 2, 3,$$

$$\left.\frac{4N + Q(6N - 6 + 6M)}{2(3Q + 1)}\right\}.$$

Its proof appears in Appendix A.

In paper [22], we had exact values for $N = 1$:

$$\lambda_0(\|A_{M,H}B_{H,1}^{(Q)}\|^2) = \frac{MQk(k + 1) + 2H}{4(1 + kQ)}$$

where: $k = \max\{i \in \mathbb{Z} : 2H \geq M(i(i - 1)Q + 2i)\}$, and we had:

$$\theta = \begin{cases} 1, & \text{if } 2H > M(k(k - 1)Q + 2k) \\ 2, & \text{if } 2H = M(k(k - 1)Q + 2k) \end{cases}$$

5. Conclusions

In this paper, we proposed a new method of “rational blowing-up” (Theorem 4), and we applied the method to Vandermonde matrix-type singularities and demonstrated its effectiveness. Theorem 6 determines the explicit values of log canonical thresholds for $H = 1, 2, 3, 4$. Our future research aim is to improve our methods and obtain explicit values for the general model.

These theoretical values introduce a mathematical measure of preciseness for numerical calculations in information criteria in Section 1. Furthermore, our theoretical results will be helpful in numerical experiments such as the Markov chain Monte Carlo (MCMC). In the papers [25,26], the mathematical foundation for analyzing and developing the precision of the MCMC method was constructed by using the theoretical values of marginal likelihoods.

We will also consider these applications in the future.

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

Acknowledgments: We thank Maxine Garcia, PhD, from Edanz Group for editing a draft of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

We demonstrate Theorem 6 by blowing-up processes.

Let $j[\alpha], j[\alpha, \beta] \in \{1, \dots, N\}$ and $i[\alpha], i[\alpha, \beta] \in \{1, \dots, H\}$.

By constructing the blow-up along $\{b_{ij} = 0, 1 \leq i \leq H, 1 \leq j \leq N\}$ and by choosing one branch of the blow-up process, we assume that $b_{ij} = v_1 b'_{ij}$ for $i, j \geq 1, j[1] = 1$ and $b'_{1j[1]} \neq 0$. We also set

$$b'_{ij} = b''_{ij} + \frac{b'_{ij[1]}}{b'_{1j[1]}} b'_{1j} \quad (i \geq 2, j \neq j[1]), \quad b''_{1j} = b'_{1j}, \quad b''_{ij[1]} = b'_{ij[1]} \quad (i, j \geq 1) \quad \text{and} \quad a'_{k1} = \sum_{i=1}^H a_{ki} b_{ij[1]}.$$

Then, we have:

$$\begin{aligned} f_{kI} &= \sum_{i=1}^H a_{ki} \prod_{j=1}^N b_{ij}^{\ell_j}, \\ &= v_1^{|I|} (a_{k1} \prod_{j=1}^N b'_{1j}{}^{\ell_j} + \sum_{i=2}^H a_{ki} b'_{ij[1]}{}^{\ell_{j[1]}} \prod_{j \neq j[1]} (b''_{ij} + \frac{b'_{ij[1]} b'_{1j}}{b'_{1j[1]}}{}^{\ell_j}), \end{aligned}$$

for $I = (\ell_1, \dots, \ell_N)$.

$$\begin{aligned} \mathcal{J} &= \left\langle f_{kI} \mid I = (\ell_1, \dots, \ell_N) \in \mathbb{N}_{+0}^N, |I| = nQ + 1, n \in \mathbb{N}_{+0}, 1 \leq k \leq M \right\rangle \\ &= \left\langle v_1^{nQ+1} \sum_{i=1}^H a_{ki} b''_{ij[1]}{}^{nQ+1}, v_1^{|I|} \sum_{i=2}^H a_{ki} \prod_{j=1}^N b''_{ij}{}^{\ell_j} \right\rangle. \\ &= \left\langle v_1 a'_{k1}, v_1^{(n+1)Q+1} \sum_{i=2}^H a_{ki} b''_{ij[1]}{}^{nQ+1} (b''_{ij[1]}{}^Q - b''_{1j[1]}{}^Q), v_1^{|I|} \sum_{i=2}^H a_{ki} \prod_{j=1}^N b''_{ij}{}^{\ell_j} \right\rangle. \end{aligned}$$

For simplicity, we set $b''_{ij} = b_{ij}$ and $a'_{k1} = a_{k1}$ again.

By constructing the blow-up along $\{b_{ij} = 0, 2 \leq j \leq N, 2 \leq i \leq H\}$ and by choosing one branch of the blow-up process, set $b_{ij} = v_{21} b'_{ij}$ for $i, j \geq 2$ and $b'_{2j[2]} \neq 0$.

1. If $j[1] \neq j[2]$, then set $b'_{ij} = b''_{ij} + \frac{b'_{ij[2]} b'_{2j}}{b'_{2j[2]}} (i \geq 3, j \neq j[2])$, $b''_{2j} = b'_{2j}$, $b''_{ij[2]} = b'_{ij[2]}$ ($i, j \geq 2$), and $a'_{k2} = \sum_{i=2}^H a_{ki} b_{ij[2]}$.

$$\begin{aligned} \mathcal{J} &= \left\langle f_{kI} \mid I = (\ell_1, \dots, \ell_N) \in \mathbb{N}_{+0}^N, |I| = nQ + 1, n \in \mathbb{N}_{+0}, 1 \leq k \leq M \right\rangle \\ &= \left\langle v_1 a_{k1}, v_1^{(n+1)Q+1} \sum_{i=2}^H a_{ki} b_{ij[1]}{}^{nQ+1} (b_{ij[1]}{}^Q - b_{1j[1]}{}^Q), v_1^{|I|} \sum_{i=2}^H a_{ki} \prod_{j=1}^N b_{ij}{}^{\ell_j} \right\rangle \\ &= \left\langle v_1 a_{k1}, v_1 v_{21} a'_{k2}, v_1^{(n+1)Q+1} v_{21}^{nQ+1} \sum_{i=3}^H a_{ki} b''_{ij[1]}{}^{nQ+1} (b''_{ij[1]}{}^Q v_{21}^Q - b''_{1j[1]}{}^Q), \right. \\ &\quad \left. v_1^{|I|} v_{21}^{|I|} \sum_{i=3}^H a_{ki} \prod_{j=1}^N b''_{ij}{}^{\ell_j}, v_1^{(n+1)Q+1} v_{21}^{(n+1)Q+1} \sum_{i=3}^H a_{ki} b''_{ij[2]}{}^{nQ+1} (b''_{ij[2]}{}^Q - b''_{2j[2]}{}^Q), \right\rangle. \end{aligned}$$

For simplicity, we set $b''_{ij} = b_{ij}$ and $a'_{k1} = a_{k1}$ again.

2. If $j[1] = j[2]$, then, by constructing the blow-up along $\{b_{ij} = 0, 2 \leq i \leq H, 2 \leq j \leq N, j \neq j[2]\}$ and by choosing one branch of the blow-up process, we set $b'_{ij} = v_{22} b''_{ij}$ for $i, j \geq 2, j \neq j[2]$. Assume that $b''_{i[2,2],j[2,2]} \neq 0$.

- (a) By constructing the rational blow-up along $\{v_1^Q = 0, v_{22} = 0\}$, we have the following (i) and (ii).

- i. Set $v_{22} = v_1^Q v'_{22}$.

$$b'_{ij} = b''_{ij} + \frac{b'_{ij[2]} (b'_{ij[2]}{}^Q - b'_{1j[2]}{}^Q)}{b'_{2j[2]} (b'_{2j[2]}{}^Q - b'_{1j[2]}{}^Q)} b'_{2j} \text{ for } i \geq 3, j \neq j[2], b''_{2j} = b'_{2j}, b''_{ij[2]} = b'_{ij[2]} \text{ for } i, j \geq 2,$$

$$\text{and } a'_{k2} = \sum_{i=2}^H a_{ki} b_{ij[2]} (b_{ij[2]}{}^Q v_2^Q - b_{1j[2]}{}^Q).$$

$$\begin{aligned}
 \mathcal{J} &= \left\langle f_{kI} \mid I = (\ell_1, \dots, \ell_N) \in \mathbb{N}_{+0}^N, |I| = nQ + 1, n \in \mathbb{N}_{+0}, 1 \leq k \leq M \right\rangle \\
 &= \left\langle v_1 a_{k1}, v_1^{(n+1)Q+1} \sum_{i=2}^H a_{ki} b_{ij[1]}^{nQ+1} (b_{ij[1]}^Q - b_{1j[1]}^Q), v_1^{|I|} \sum_{i=2}^H a_{ki} \prod_{j=1}^N b_{ij}^{\ell_j} \right\rangle \\
 &= \left\langle v_1 a_{k1}, v_1^{(n+1)Q+1} v_{21}^{nQ+1} \sum_{i=2}^H a_{ki} b_{ij[1]}^{nQ+1} (b_{ij[1]}^{Q} v_2^Q - b_{1j[1]}^{Q}), \right. \\
 &\quad \left. v_1^{|I|} v_{21}^{|I|} v_{22}^{|I|-\ell_{j[2]}} \sum_{i \geq 2}^H a_{ki} \prod_{j=1}^N b_{ij}^{\ell_j}, \text{ for } |I| - \ell_{j[2]} > 0 \right\rangle \\
 &= \left\langle v_1 a_{k1}, v_1^{(n+2)Q+1} v_{21}^{(n+1)Q+1} \sum_{i=3}^H a_{ki} b_{ij[1]}^{nQ+1} (b_{ij[1]}^{Q} v_2^Q - b_{1j[1]}^{Q}) (b_{ij[1]}^{Q} - b_{2j[2]}^{Q}), \right. \\
 &\quad \left. v_1^{Q+1} v_{21} a'_{k2}, v_1^{|I|(Q+1)-Q\ell_{j[2]}} v_{21}^{|I|} v_{22}^{|I|-\ell_{j[2]}} \sum_{i \geq 3}^H a_{ki} \prod_{j=1}^N b_{ij}^{\ell_j}, \text{ for } |I| - \ell_{j[2]} > 0 \right\rangle.
 \end{aligned}$$

ii. Set $v_1 = v_{22}^{1/Q} v'_1$.

Set $b'_{ij} = b''_{ij} + \frac{b'_{ij[2,2]}}{b'_{i[2,2]j[2,2]}} b'_{i[2,2]j}$ for $i \geq 2, i \neq [2,2], j \neq j[2,2], j[2]$. and $a'_{ki[2,2]} = \sum_{i=2}^H a_{ki} b_{ij[2,2]}$.

$$\begin{aligned}
 \mathcal{J} &= \left\langle f_{kI} \mid I = (\ell_1, \dots, \ell_N) \in \mathbb{N}_{+0}^N, |I| = nQ + 1, n \in \mathbb{N}_{+0}, 1 \leq k \leq M \right\rangle \\
 &\quad \left. v_{22}^{|I|/Q} v_1^{|I|} v_{21}^{|I|} \sum_{i=2}^H a_{ki} \prod_{j=1}^N b_{ij}^{\ell_j} \right\rangle \\
 &= \left\langle v_1' v_{22}^{1/Q} a_{k1}, v_{22}^{(n+1)+1/Q} v_1'^{(n+1)Q+1} \sum_{i=2}^H a_{ki} b_{ij[1]}^{nQ+1} (b_{ij[1]}^Q - b_{1j[1]}^Q), \right. \\
 &\quad \left. v_{22}^{|I|/Q} v_1^{|I|} v_{21}^{|I|} v_{22}^{|I|-\ell_{j[2]}} \sum_{i=2}^H a_{ki} b'_{ij[2]}^{\ell_{j[2]}} b'_{ij[2,2]}^{\ell_{j[2,2]}}, \text{ for } |I| = \ell_{j[2]} + \ell_{j[2,2]}, \ell_{j[2,2]} > 0, \right. \\
 &\quad \left. v_{22}^{|I|/Q} v_1^{|I|} v_{22}^{|I|-\ell_{j[2]}} \sum_{i \geq 2, i \neq [2,2]}^H a_{ki} \prod_{j=1}^N b_{ij}^{\ell_j}, \text{ for } |I| - \ell_{j[2]} - \ell_{j[2,2]} > 0 \right\rangle \\
 &= \left\langle v_1' v_{22}^{1/Q} a_{k1}, v_1' v_{22}^{1/Q} v_{22} v_{21} a_{ki[2,2]}, v_{22}^{(n+1)+1/Q} v_1'^{(n+1)Q+1} \sum_{i=2}^H a_{ki} b_{ij[1]}^{nQ+1} (b_{ij[1]}^Q - b_{1j[1]}^Q), \right. \\
 &\quad \left. v_{22}^{|I|/Q} v_1^{|I|} v_{21}^{|I|} v_{22}^{|I|-\ell_{j[2]}} \sum_{i=2}^H a_{ki} b'_{ij[2]}^{\ell_{j[2]}} b'_{ij[2,2]}^{\ell_{j[2,2]}}, \text{ for } |I| = \ell_{j[2]} + \ell_{j[2,2]}, \ell_{j[2,2]} > 0, \right. \\
 &\quad \left. v_{22}^{|I|/Q} v_1^{|I|} v_{21}^{|I|} v_{22}^{|I|-\ell_{j[2]}} \sum_{i \geq 2, i \neq [2,2]}^H a_{ki} \prod_{j=1}^N b_{ij}^{\ell_j}, \text{ for } |I| - \ell_{j[2]} - \ell_{j[2,2]} > 0 \right\rangle.
 \end{aligned}$$

By constructing blow-ups, we have the theorem.

References

1. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [[CrossRef](#)]
2. Watanabe, S. Algebraic analysis for nonidentifiable learning machines. *Neural Comput.* **2001**, *13*, 899–933. [[CrossRef](#)] [[PubMed](#)]

3. Watanabe, S. Algebraic geometrical methods for hierarchical learning machines. *Neural Netw.* **2001**, *14*, 1049–1060. [[CrossRef](#)]
4. Watanabe, S. Algebraic geometry of learning machines with singularities and their prior distributions. *J. Jpn. Soc. Artif. Intell.* **2001**, *16*, 308–315.
5. Watanabe, S. *Algebraic Geometry and Statistical Learning Theory*; Cambridge University Press: New York, NY, USA, 2009; Volume 25.
6. Watanabe, S. Equations of states in singular statistical estimation. *Neural Netw.* **2010**, *23*, 20–34. [[CrossRef](#)]
7. Watanabe, S. *Mathematical Theory of Bayesian Statistics*; CRC Press: New York, NY, USA, 2018.
8. Drton, M.; Plummer, M. A Bayesian information criterion for singular models. *J. R. Statist. Soc. B* **2017**, *79*, 1–38. [[CrossRef](#)]
9. Hironaka, H. Resolution of singularities of an algebraic variety over a field of characteristic zero. *Ann. Math.* **1964**, *79*, 109–326. [[CrossRef](#)]
10. Atiyah, M.F. Resolution of singularities and division of distributions. *Commun. Pure Appl. Math.* **1970**, *13*, 145–150. [[CrossRef](#)]
11. Aoyagi, M.; Watanabe, S. Stochastic complexities of reduced rank regression in Bayesian estimation. *Neural Netw.* **2005**, *18*, 924–933. [[CrossRef](#)]
12. 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.
13. 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.
14. Aoyagi, M. A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities. *Commun. Stat. Theory Methods* **2010**, *39*, 2667–2687. [[CrossRef](#)]
15. Aoyagi, M. Learning coefficient in Bayesian estimation of restricted Boltzmann machine. *J. Algebr. Stat.* **2013**, *4*, 30–57. [[CrossRef](#)]
16. 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.
17. Rusakov, D.; Geiger, D. Asymptotic model selection for naive Bayesian networks. *J. Mach. Learn. Res.* **2005**, *6*, 1–35.
18. Zwiernik, P. An asymptotic behavior of the marginal likelihood for general Markov models. *J. Mach. Learn. Res.* **2011**, *12*, 3283–3310.
19. 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. [[CrossRef](#)]
20. Aoyagi, M.; Nagata, K. Learning coefficient of generalization error in Bayesian estimation and Vandermonde matrix type singularity. *Neural Comput.* **2012**, *24*, 1569–1610. [[CrossRef](#)] [[PubMed](#)]
21. Aoyagi, M. Consideration on singularities in learning theory and the learning coefficient. *Entropy* **2013**, *15*, 3714–3733. [[CrossRef](#)]
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. Lin, S. Asymptotic approximation of marginal likelihood integrals. *arXiv* **2010**, arXiv:1003.5338v2.
24. Yamazaki, K.; Aoyagi, M.; Watanabe, S. Asymptotic analysis of Bayesian generalization error with Newton diagram. *Neural Netw.* **2010**, *23*, 35–43. [[CrossRef](#)] [[PubMed](#)]
25. Nagata, K.; Watanabe, S. Exchange Monte Carlo Sampling from Bayesian posterior for singular learning machines. *IEEE Trans. Neural Netw.* **2008**, *19*, 1253–1266. [[CrossRef](#)]
26. Nagata, K.; Watanabe, S. Asymptotic behavior of exchange ratio in exchange Monte Carlo method. *Int. J. Neural Netw.* **2008**, *21*, 980–988. [[CrossRef](#)] [[PubMed](#)]

