



Title	Learning coefficients in semiregular models I: properties
Author(s)	Kurumadani, Yuki
Citation	Japanese Journal of Statistics and Data Science. 2025, 8, p. 1051-1079
Version Type	AM
URL	https://hdl.handle.net/11094/104333
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

<https://ir.library.osaka-u.ac.jp/>

The University of Osaka

Learning Coefficients in Semiregular Models

I : properties

Yuki Kurumadani

*Graduate School of Engineering Science, The University of Osaka, 1
Chome-3 Machikaneyamacho, Toyonaka, 560-0043, Osaka, Japan.

Corresponding author(s). E-mail(s):
kurumadani@sigmath.es.osaka-u.ac.jp;

Abstract

In Bayesian inference, the asymptotic behavior of evaluation metrics for predictive accuracy, such as generalization loss and free energy, is determined by a model-specific rational number known as the learning coefficient (real log canonical threshold) (Watanabe, 2009).

Learning coefficients are already known for models that satisfy regularity conditions (Watanabe, 2009). However, for singular models that do not satisfy these conditions, specific values of learning coefficients have been provided for certain models such as reduced-rank regression (Aoyagi and Watanabe, 2005; Aoyagi, 2024). Nevertheless, a general formula for learning coefficients that broadly applies to singular models has not yet been established.

In this paper, we derive a formula for the learning coefficients of statistical models that satisfy a condition slightly less stringent than the regularity condition, which we refer to as *semiregular models*.

This formula reveals that a natural number $\mathbf{m}$, determined by the partial derivatives of the log-likelihood ratio function, significantly affects the learning coefficient.

Our main results are applicable to relatively simple singular models. As a specific example, we provide the learning coefficients for mixture distribution models with a constant mixing proportion, which had not been elucidated before.

Keywords: resolution map, singular learning theory, real log canonical threshold, algebraic geometry

1 Introduction

Hierarchical models such as neural networks, reduced-rank regression, normal mixture models, and Boltzmann machines are referred to as singular learning models. These models are widely used in the analysis of real data, but their use in theoretical analysis has not been sufficiently developed. These models possess nonpositive definite Fisher information matrices and therefore cannot be analyzed by classic theory (Watanabe, 2009). For example, model selection methods such as AIC (Akaike, 1974), TIC (Takeuchi, 1976), and BIC (Schwarz, 1978), which require the Fisher information matrix to be positive definite and a unique point that minimizes a log loss function, cannot be applied to singular models.

Watanabe (2009) established a theory of learning coefficients using algebraic geometry for singular models. However, only universal laws were presented, and further mathematical considerations are necessary to obtain specific learning coefficients.

Based on the free energy demonstrated by Watanabe (2009), widely applicable Bayesian information criteria (WBIC) (Watanabe, 2013) and the singular Bayesian information criterion (sBIC) (Drton and Plummer, 2017) have been developed. The sBIC effectively uses the learning coefficient λ and its multiplicity $\mathbf{m}$ with a fixed point equation system of marginal likelihood. Moreover, the learning coefficients λ and their multiplicities $\mathbf{m}$ provide mathematical indicators for analyzing the accuracy of model selection methods such as WAIC (Akaike, 1974; Watanabe, 2001a, 2001b, 2001c, 2009, 2010, 2018) and cross-validation in Bayesian estimation approaches.

The learning coefficient λ is equal to the real log canonical threshold of the Kullback-Leibler function in algebraic geometry, and its specific value can be obtained using the recursive blow-up process (Hironaka, 1964). However, this result is an abstract existence theorem, and it is not clear which specific blow-ups should be performed. Furthermore, the theory of real log canonical thresholds has fewer achievements in the real number field compared to the complex number field or algebraically

closed fields (Kashiwara, 1976; Mustata, 2002). Therefore, obtaining real log canonical thresholds is a matter of interest in various fields, including mathematics.

Aoyagi and Watanabe (2005); Aoyagi (2024) obtained the specific values of the learning coefficients for reduced-rank regression models. Furthermore, Aoyagi (2006, 2013a, 2019a, 2019b) obtained specific values or upper bounds of learning coefficients for models such as three-layer neural networks and normal mixture models related to Vandermonde matrix-type singularities. Additionally, Aoyagi (2013b) obtained specific values for restricted Boltzmann machines. Moreover, Rusakov and Geiger (2002, 2005) , and Drton et al (2017) obtained learning coefficients for naive Bayesian networks, directed tree models with hidden variables, and Gaussian latent tree and forest models, respectively. While there is previous research on learning coefficients λ for each statistical model, a formula for learning coefficients λ that can be widely applied to singular models has not been identified.

The aim of this paper is to generalize the formula for the learning coefficient λ for regular models and to obtain a formula for λ that can be applied to relatively simple singular models. Regular models refer to cases where the Fisher information matrix is positive definite and there is a formula $\lambda = d/2$ using the number of parameters d . We consider cases where the Fisher information matrix is not positive definite and its rank r satisfies $0 < r < d$. Assuming that a certain type of linear independence holds, we derive the formula

$$\lambda = \frac{d - r + rm}{2m},$$

where m is a natural number determined by the order of differentiability of the log-likelihood ratio function.

This generalizes the regular case. In particular, we clarify that this natural number m , along with the number of parameters d and the rank r of the Fisher information matrix, is a factor that determines the learning coefficient λ . We utilize this formula to

obtain the learning coefficients for a specific mixture distribution model with a fixed mixing ratio, which had previously remained unresolved.

In this paper, we outline the concepts of singularity resolution and real log canonical thresholds in algebraic geometry in Section 2, and we describe the main results and their applications in Section 3. In Section 4, we demonstrate application examples of the main results and their consistency with numerical experiments. We provide the proofs of the main results in Section 5, and we discuss the conclusions in Section 6.

Definitions and Assumptions

Throughout this paper, we consider statistical models $p(x|\theta)$ with continuous parameters $\theta = (\theta_1, \dots, \theta_d) \in \Theta \subset \mathbb{R}^d$ where $d \geq 1$, and we denote the true distribution by $q(x)$. We denote the set of all possible data x by χ . We assume that the statistical model is *realizable*, meaning that there exists a parameter θ_* such that $q(x) = p(x|\theta_*)$ almost surely. We refer to such parameters θ_* as realization parameters and denote the entire set of realization parameters by Θ_* . We assume that the prior distribution of the parameters $\varphi(\theta)$ satisfies $\varphi(\theta_*) > 0$ for any realization parameter θ_* . We denote a random variable following the true distribution q by X , and we use $\mathbb{E}_X[\cdot]$ to represent the expectation over the random variable X . In this paper, we assume that the operations of taking expectations and taking partial derivatives with respect to θ are interchangeable.

We define the log-likelihood ratio function f and the Kullback-Leibler divergence $K(\theta)$ respectively as

$$f(x|\theta) := \log \frac{q(x)}{p(x|\theta)}, \quad K(\theta) := \mathbb{E}_X \left[\log \frac{q(X)}{p(X|\theta)} \right],$$

and we assume they are analytic around $\theta = \theta_*$.

For fixed data x , we define the m -th order terms in the Taylor expansion of the log-likelihood ratio function $f(x|\theta)$ around $\theta = \theta_*$, involving $\theta_1, \dots, \theta_s$ ($s \leq d$). These

terms are given by

$$F_m(x|\theta_1, \dots, \theta_s) := \sum_{\substack{i_1 + \dots + i_s = m \\ i_1, \dots, i_s \in \mathbb{Z}_{\geq 0}}} \frac{1}{i_1! \dots i_s!} \frac{\partial^m f(x|\theta)}{\partial \theta_1^{i_1} \dots \partial \theta_s^{i_s}} \Big|_{\theta=\theta_*} (\theta_1 - \theta_{1*})^{i_1} \dots (\theta_s - \theta_{s*})^{i_s}.$$

Furthermore, we define

$$I = \text{Cov} \left(\frac{\partial f(X|\theta)}{\partial \theta_i} \Big|_{\theta=\theta_*}, \frac{\partial f(X|\theta)}{\partial \theta_j} \Big|_{\theta=\theta_*} \right)_{i,j=1,\dots,d}, \quad J = \left(\frac{\partial^2 K(\theta)}{\partial \theta_i \partial \theta_j} \Big|_{\theta=\theta_*} \right)_{i,j=1,\dots,d}.$$

The matrix I is the Fisher information matrix at $\theta = \theta_*$. Generally, I and J do not coincide, but under the conditions of this paper, they do coincide (see Remark A.1).

We denote the rank of this matrix by r .

2 Real log canonical thresholds

It is known that learning coefficients correspond to the concept of *real log canonical thresholds* in algebraic geometry. Real log canonical thresholds are defined using a resolution of singularities. Here, the resolution of singularities refers to transforming the analytic function F into a normal crossing, as specified by Theorem 2.1.

Theorem 2.1 (Resolution of Singularities). (*Hironaka, 1964*)(*Watanabe, 2009, Theorem 2.3*) *Let $F(x)$ be a real analytic function defined near the origin in $\mathbb{R}^d$ that is not identically zero, with $F(0) = 0$. Then, there exists an open set $W \subset \mathbb{R}^d$ containing the origin, a real analytic manifold U , and a proper analytic map $g : U \rightarrow W$ satisfying:*

- (1) *Let $W_0 := F^{-1}(0)$ and $U_0 := g^{-1}(W_0)$; then, $g : U \setminus U_0 \rightarrow W \setminus W_0$ is an analytic isomorphism.*

(2) At any point Q in U_0 , taking local coordinates $u = (u_1, \dots, u_d)$ of U at Q as the origin, it can be expressed as:

$$\begin{aligned} F(g(u)) &= a(u)u_1^{k_1}u_2^{k_2}\cdots u_d^{k_d}, \\ |g'(u)| &= \left| b(u)u_1^{h_1}u_2^{h_2}\cdots u_d^{h_d} \right|, \end{aligned} \tag{2.1}$$

where k_i, h_i ($i = 1, \dots, d$) are nonnegative integers and $a(u), b(u)$ are real analytic functions with $a(u) \neq 0, b(u) \neq 0$.

Expressions such as (2.1) are referred to as *normal crossings*.

Definition 2.1 (Real Log Canonical Threshold). Let F be a real analytic function defined on an open set O in $\mathbb{R}^d$, and let C be a compact subset containing O . For each point P in C where $F(P) = 0$, we perform a translation so that point P coincides with the origin in $\mathbb{R}^d$, and Theorem 2.1 applies, fixing (W, U, g) as guaranteed by Theorem 2.1(2). We denote the nonnegative integers h_i, k_i given by Theorem 2.1(2) near any point $Q \in U_0$ as $h_i^{(Q)}, k_i^{(Q)}$.

(1) The *real log canonical threshold* λ_P at point P , and the *real log canonical threshold* λ for the compact set C , are defined respectively as:

$$\lambda_P := \inf_{Q \in U_0} \left\{ \min_{i=1, \dots, d} \frac{h_i^{(Q)} + 1}{k_i^{(Q)}} \right\}, \quad \lambda := \inf_{P \in C} \lambda_P,$$

with $(h_i^{(Q)} + 1)/k_i^{(Q)} = \infty$ defined for cases where $k_i^{(Q)} = 0$ (Watanabe, 2009, Definition 2.7).

(2) For the point $P \in C$ giving the minimum in (1), the maximum number of i satisfying $\lambda_P = (h_i^{(Q)} + 1)/k_i^{(Q)}$ is called the *multiplicity* $\mathbf{m}$. If there are multiple such points P , the multiplicity is defined as the maximum number among them.

We consider a case where the analytic function F corresponds to the Kullback-Leibler divergence $K(\theta)$. Under the condition that the prior distribution $\varphi(\theta_*) > 0$, it is known that *the learning coefficient λ is equal to the real log canonical threshold for the compact set $\Theta_* = \{\theta \in \Theta \mid K(\theta) = 0\}$* (Watanabe, 2009, Theorem 6.6, Definition 6.4).

Theorem 2.1 ensures the existence of the function g , which can be constructed by a finite number of repetitions of an operation known as blow-up (Watanabe, 2009, Definition 3.12). This result is demonstrated in Hironaka (1964) and Watanabe (2009), Theorem 3.6. However, the existence of g is only proven abstractly, and it is usually difficult to find a specific construction method for g . In this paper, we construct the coordinate transformation g through a concrete sequence of blow-ups to calculate the learning coefficient λ .

3 Main Theorem

Without loss of generality, we assume $\theta_* = 0$ throughout the following discussion.

Definition 3.1 (Semiregularity). A statistical model $p(x|\theta)$ is said to be *semiregular*¹ at a parameter θ_* (which realizes the true model $q(x)$) if the following conditions are satisfied:

- The set of realizing parameters Θ_* contains only the single element θ_* and is contained within the interior of Θ .
- The rank r of the Fisher information matrix $I \in \mathbb{R}^{d \times d}$ is greater than zero.
- The model satisfies Assumption 1.

Assumption 1. Divide the d parameters $\theta = (\theta_1, \dots, \theta_d)$ into $\theta_1, \dots, \theta_r$ and $\theta_{r+1}, \dots, \theta_d$, and assume:

¹This term is specific to this paper and is not generally established.

- (1) There exists a maximal integer m (with $2 \leq m < \infty$) such that for all multiple indices $(i_{r+1}, \dots, i_d)$ satisfying $1 \leq i_{r+1} + \dots + i_d \leq m - 1$, the following condition holds:

$$\left. \frac{\partial^{i_{r+1} + \dots + i_d} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} = 0 \quad \text{a.s.}$$

- (2) For all $(\theta_{r+1}, \dots, \theta_d) \neq 0$, the function $F_m(X|\theta_{r+1}, \dots, \theta_d)$ and the r random variables

$$\left. \frac{\partial f(X|\theta)}{\partial \theta_1} \right|_{\theta=0}, \dots, \left. \frac{\partial f(X|\theta)}{\partial \theta_r} \right|_{\theta=0} \quad (3.1)$$

are linearly independent².

This definition generalizes the concept of regularity, where Θ_* contains only a single element within the interior of Θ , and I is positive definite. Since it is known from regular models that $\lambda = d/2$ (Watanabe, 2009), we exclude such cases from our discussion. Therefore, we assume $1 \leq r \leq d - 1$ throughout this paper.

Remark 3.1. We now explain the meaning of the assumption of semiregularity. The assumptions other than Assumption 1 are the same as those in the regularity condition, and Assumption 1 generalizes the regularity condition that the Fisher information matrix is positive definite.

As for Assumption 1(1), since $1 \leq r \leq d - 1$, the first derivatives of f with respect to $\theta_{r+1}, \dots, \theta_d$ are linearly dependent on the r random variables in (3.1). As explained in the subsection “Derivation Method for m ”, by applying an appropriate coordinate transformation, all such linearly dependent derivatives can be made zero. Thus, Assumption 1(1) can always be satisfied.

As for Assumption 1(2), it is a condition necessary to guarantee normal crossings (see the proof of the Main Theorem). Unlike (1), there exist statistical models that do not satisfy this condition. Some examples are provided in the Supplementary Information.

Example 3.1. We present specific statistical models that satisfy Assumption 1.

²The random variables being linearly independent means that they are so with probability one.

- (1) Let $\tilde{p}(\theta)$ represent the binomial distribution with $M(\geq 2)$ trials and success probability θ , the Poisson distribution with mean θ , or the normal distribution with mean θ and variance 1. Consider a mixture model with constant mixing proportions $T_i(> 0)$:

$$\begin{cases} p(x|\theta) := \sum_{i=1}^H T_i \tilde{p}(x|\theta_i) \\ q(x) := \tilde{p}(x|\theta_*). \end{cases}$$

This model satisfies semiregularity for the realized parameters $(\theta_1, \dots, \theta_H) = (\theta_*, \dots, \theta_*)$ when $(d, r, m) = (H, 1, 2)$. Details are provided in Examples 4.1 and Appendix C.

- (2) Let $\mathcal{N}$ be a random variable that is subject to the standard normal distribution. The probability distribution $q(x)$ of X is assumed to have compact support. Consider the following statistical model in which a random variable Y for a given X is defined by

$$Y = A_1 \sigma(\theta_1 X) + A_2 \sigma(\theta_2 X) + \mathcal{N},$$

where A_1, A_2 are nonzero constants. This is a layered neural network with one input unit, two hidden units, and one output unit. The true distribution of Y is assumed to be

$$Y = 0 + \mathcal{N}.$$

The activation function $\sigma(x)$ is an analytic function satisfying $\sigma(0) = 0, \sigma'(0) \neq 0$, and the smallest natural number $m(\geq 2)$ such that $\sigma^{(m)}(0) \neq 0$ is defined as m . For instance, $m = 2$ for $\sigma(x) = e^x - 1$, $m = 3$ for $\sigma(x) = \sin(x), \arctan(x), \tanh(x)$, and m for $\sigma(x) = x^m + x$.

We assume that the constants A_1, A_2 satisfy $A_1^{m-1} + (-1)^m A_2^{m-1} \neq 0$ and that the random variables X and X^m are linearly independent. Under these conditions, the statistical model satisfies Assumption 1 for the realized parameters $(\theta_1, \theta_2) = (0, 0)$ when $(d, r, m) = (2, 1, m)$.

Specifically, the log-likelihood ratio function f is given by

$$f(x, y|\theta) = -\{A_1\sigma(\theta_1 x) + A_2\sigma(\theta_2 x)\}y + \frac{1}{2}\{A_1\sigma(\theta_1 x) + A_2\sigma(\theta_2 x)\}^2.$$

By performing the coordinate transformation $\theta'_1 = \theta_1 + (A_2/A_1)\theta_2$, we obtain

$$\begin{aligned}\left.\frac{\partial f}{\partial \theta'_1}\right|_{(\theta'_1, \theta_2)=0} &= -A_1\sigma'(0)xy, \\ \left.\frac{\partial f}{\partial \theta_2^n}\right|_{(\theta'_1, \theta_2)=0} &= 0 \quad (1 \leq n < m), \\ \left.\frac{\partial f}{\partial \theta_2^m}\right|_{(\theta'_1, \theta_2)=0} &= -\frac{A_2}{A_1^{m-1}}\{A_1^{m-1} + (-1)^m A_2^{m-1}\}\sigma^{(m)}(0)x^m y.\end{aligned}$$

Since $x^m y$ and xy are linearly independent by assumption, the model satisfies Assumption 1 for $(d, r, m) = (2, 1, m)$.

Remark 3.2. We extend Example 3.1(2) to the case where the hidden layer has $H \geq 3$ units. The statistical model is expressed as $Y = \sum_{i=1}^H A_i \sigma(\theta_i X) + \mathcal{N}$. The conditions on the activation function σ and the random variable X are the same as those in Example 3.1(2). The details are provided in the Supplementary Information.

- (1) First, we assume that the true distribution is $Y = 0 + \mathcal{N}$. When $m = 2$, if all A_i are positive constants, the statistical model satisfies Assumption 1 with $(d, r, m) = (H, 1, 2)$. On the other hand, if there are constants A_i that are nonpositive, then there are cases in which Assumption 1(2) is not satisfied. When $m = 3$, Assumption 1(2) is not satisfied for general constants A_i .
- (2) Next, we assume that the true distribution is given by $Y = \sum_{i=1}^{H-2} A_i \sigma(\theta_{i*} X) + \mathcal{N}$, where $\theta_{1*}, \dots, \theta_{H-2*}$ are distinct nonzero constants. We assume that all constants $A_1, \dots, A_H$ are nonzero and satisfy $A_{H-1}^{m-1} + (-1)^m (A_H)^{m-1} \neq 0$. Furthermore, we assume that for the random variable X , the H random variables $\sigma'(\theta_{1*} X)X, \dots, \sigma'(\theta_{H-2*} X)X, X, X^m$ are linearly independent. Under these

conditions, we verify that this statistical model satisfies Assumption 1 at the realizable parameters $(\theta_1, \dots, \theta_{H-2}, \theta_{H-1}, \theta_H) = (\theta_{1*}, \dots, \theta_{H-2*}, 0, 0)$ when $(d, r, m) = (H, H-1, m)$. Note that this assumption of linear independence does not hold when the activation function σ is $\sigma(x) = x^m + x$.

Main Theorem. *The learning coefficient for a semiregular model is given by the following formula (with multiplicity $\mathbf{m} = 1$):*

$$\lambda = \frac{d - r + rm}{2m} \quad (3.2)$$

Remark 3.3. In the case of regular models, substituting $r = d$ into (3.2) results in

$$\lambda = \frac{d - d + d \times m}{2 \times m} = \frac{d}{2},$$

which matches the known result for regular models. Thus, this formula generalizes the regular case. Previous research has generally shown that $\lambda \leq d/2$ (Watanabe, 2009, Theorem 7.2), and

$$\frac{d - r + rm}{2m} \leq \frac{d}{2},$$

so the Main Theorem does not contradict this finding.

Derivation Method for m

To satisfy Assumption 1(1), we can explicitly construct the variable transformation as follows. Given the semiregularity, we have r parameters $\theta_1, \dots, \theta_r$ for which the derivatives

$$\left. \frac{\partial f}{\partial \theta_1} \right|_{\theta=0}, \dots, \left. \frac{\partial f}{\partial \theta_r} \right|_{\theta=0} \quad (3.3)$$

are linearly independent. For the remaining parameters $\theta_{r+1}, \dots, \theta_d$, starting from $l = 1$, if the derivative can be expressed as

$$\left. \frac{\partial^l f}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} = \sum_{k=1}^r a_k \left. \frac{\partial f}{\partial \theta_k} \right|_{\theta=0} \quad \text{a.s.,} \quad a_k \in \mathbb{R},$$

then we perform the coordinate transformation

$$\begin{cases} \theta'_1 := \theta_1 + \frac{a_1}{i_{r+1}! \dots i_d!} \theta_{r+1}^{i_{r+1}} \dots \theta_d^{i_d} \\ \vdots \\ \theta'_r := \theta_r + \frac{a_r}{i_{r+1}! \dots i_d!} \theta_{r+1}^{i_{r+1}} \dots \theta_d^{i_d} \end{cases} \quad \begin{cases} \theta'_{r+1} := \theta_{r+1} \\ \vdots \\ \theta'_d := \theta_d \end{cases}$$

which makes

$$\left. \frac{\partial^l f}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta'=0} = 0 \quad \text{a.s.} \quad (3.4)$$

without changing other derivatives of order l or lower (this fact is proven in Proposition B.1). We repeat this process until we first encounter a derivative that is linearly independent from (3.3); the order of this derivative is m .

Remark 3.4. We must have $m < \infty$. If $m = \infty$, that is, if for any $i_{r+1}, \dots, i_d$, condition (3.4) is satisfied; then, in the Taylor expansion of f around $\theta = 0$, there would be no terms consisting solely of $\theta_{r+1}, \dots, \theta_d$. This means that for sufficiently small arbitrary values of $(\theta_{r+1}, \dots, \theta_d)$,

$$f(X|\theta_1 = 0, \dots, \theta_r = 0, \forall \theta_{r+1}, \dots, \forall \theta_d) = 0,$$

resulting in the statistical model $p(x|\theta)$ realizing the true distribution $q(x)$ at multiple points near the origin. This contradicts the assumption of semiregularity, which presumes realization at only one point.

4 Experiments

Example 4.1. Let $\tilde{p}(x|\theta)$ be one of the following probability distributions:

- The binomial distribution $\text{Bin}(M, \theta)$ with number of trials $M \geq 2$ and success probability θ .
- The Poisson distribution $\text{Po}(\theta)$ with mean θ .
- The normal distribution $N(\theta, 1)$ with mean θ and variance 1.

Consider a mixture distribution model with $H \geq 2$ mixture components, where H is a constant. Let $(T_1, \dots, T_{H-1})$ be constants satisfying $0 < T_i < 1$ and $\sum_{i=1}^{H-1} T_i < 1$, representing mixture proportions. The model has H parameters $(\theta_1, \dots, \theta_H)$ and is defined as:

$$p(x|\theta) := T_1 \tilde{p}(x|\theta_1) + \dots + T_{H-1} \tilde{p}(x|\theta_{H-1}) + \left(1 - \sum_{i=1}^{H-1} T_i\right) \tilde{p}(x|\theta_H)$$

Assuming the true distribution is $q(x) = \tilde{p}(x|\theta_*)$, where θ_* is a constant within the interior of the parameter space Θ , this statistical model constitutes a semiregular model satisfying Assumption 1 with $(d, r, m) = (H, 1, 2)$. According to the Main Theorem, the learning coefficient is

$$\lambda = \frac{d - r + rm}{2m} = \frac{H + 1}{4}$$

with multiplicity $\mathbf{m} = 1$ (details are provided in Appendix C).

We verify the above results through numerical experiments. Here, we use the estimator $\widehat{\lambda}^{100}$ proposed by Imai (2019):

$$\widehat{\lambda}^{100} := \frac{1}{100} \sum_{k=1}^{100} \widehat{\lambda}, \quad \widehat{\lambda} := t^2 \mathbb{V}_{\theta}^t \{\log p(X^n|\theta)\}, \quad t := \frac{1}{\log(n)}$$

$$\mathbb{E}_\theta^t\{h(\theta)\} := \frac{\int h(\theta)p(X^n|\theta)^t\varphi(\theta)d\theta}{\int p(X^n|\theta)^t\varphi(\theta)d\theta}, \quad \mathbb{V}_\theta^t\{h(\theta)\} := \mathbb{E}_\theta^t\{h(\theta)^2\} - (\mathbb{E}_\theta^t\{h(\theta)\})^2$$

where $p(X^n|\theta)$ represents the likelihood function for n independent and identically distributed samples following the true distribution q .

We conducted the numerical experiments with $n = 10000$, using the Hamiltonian Monte Carlo method implemented in the R package *RStan* to obtain the posterior distribution.

The assumptions for the prior distribution $\varphi(\theta)$ are as follows: for the parameter of the binomial distribution, $\theta \sim \text{Unif}(0, 1)$; for the parameter of the Poisson distribution, $\log(\theta) \sim \text{N}(0, 0.8)$; and for the parameter of the normal distribution, $\theta \sim \text{N}(0, 1)$.

The results of the simulations are shown in Table 1 and Figure 1. In each case, the results closely match the theoretical values.

Table 1 Estimated learning coefficients for different distributions

	1	2	3	H 4	5	6	7	8
Theoretical value	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
$\text{Bin}(2, \theta)$	0.500	0.751	1.003	1.251	1.488	1.738	1.991	2.240
$\text{Po}(\theta)$	0.499	0.765	1.025	1.274	1.518	1.766	1.973	2.184
$N(\theta, 1)$	0.499	0.754	1.003	1.245	1.483	1.722	1.961	2.199

5 Proof of the Main Theorem

In the proof of the Main Theorem, we use the following theorem (for the proof, see Appendix A):

Theorem 5.1. *For a semiregular model, the Taylor expansion of $K(\theta)$ at $\theta = 0$ can be expressed as:*

$$K(\theta) = \frac{1}{2}\mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 \right] + (\text{higher order terms})$$

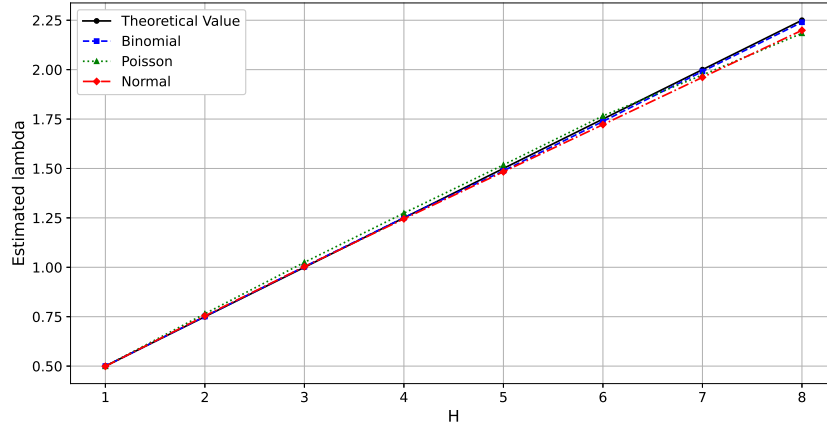


Fig. 1 Graph of estimated learning coefficients for each distribution

where the higher order terms do not include the following:

- Terms of degree $2m$ or less consisting only of $\theta_{r+1}, \dots, \theta_d$.
- Terms of first degree in $\theta_1, \dots, \theta_r$ and up to m -th degree in $\theta_{r+1}, \dots, \theta_d$.
- Terms of second degree consisting only of $\theta_1, \dots, \theta_r$.

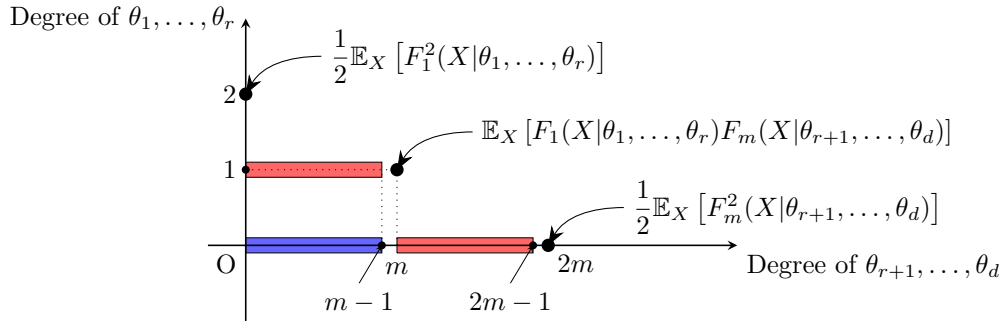


Fig. 2 According to Theorem 5.1, when the derivatives of the log-likelihood ratio function are assumed to be zero (almost surely) in the blue area (including endpoints), the derivatives in the red area (including endpoints) become random variables with zero expected value. That is, the coefficients in the Taylor expansion of $K(\theta)$ in these regions are all zero

Example 5.1. Consider a random variable X following a binomial distribution $\text{Bin}(2, \theta)$ with parameter θ , where $0 < \theta < 1$:

$$\tilde{p}(X = x|\theta) = \binom{2}{x} \theta^x (1 - \theta)^{2-x} = \begin{cases} (1 - \theta)^2, & (x = 0) \\ 2\theta(1 - \theta), & (x = 1) \\ \theta^2, & (x = 2) \end{cases}$$

Consider the following mixture distribution model with parameters (θ_1, θ_2) :

$$p(X = x|\theta_1, \theta_2) := \frac{1}{2} \tilde{p}\left(X = x \left| \theta_1 - \theta_2 + \frac{1}{2} \right.\right) + \frac{1}{2} \tilde{p}\left(X = x \left| \theta_2 + \frac{1}{2} \right.\right) \quad (x = 0, 1, 2) \quad (5.1)$$

Assuming the true distribution is $\tilde{p}(X|1/2)$, this model realizes the true distribution at parameters $(\theta_1, \theta_2) = 0$. The following derivatives exist at $(\theta_1, \theta_2) = 0$:

$$\left. \frac{\partial f(X|\theta)}{\partial \theta_1} \right|_{(\theta_1, \theta_2)=0}, \quad \left. \frac{\partial f(X|\theta)}{\partial \theta_2} \right|_{(\theta_1, \theta_2)=0}, \quad \left. \frac{\partial^2 f(X|\theta)}{\partial \theta_2^2} \right|_{(\theta_1, \theta_2)=0}$$

resulting, respectively, in:

$$\begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} -8 \\ 8 \\ -8 \end{bmatrix}.$$

Each column vector corresponds to $x = 0$, $x = 1$, and $x = 2$. The first and third vectors are linearly independent. This verifies that Assumption 1 is satisfied with $(d, r, m) =$

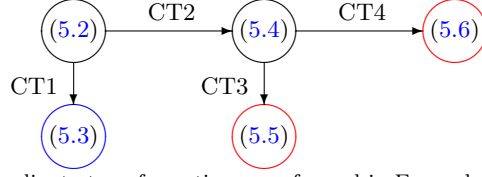


Fig. 3 The procedure of coordinate transformation g performed in Example 5.1. Local coordinates displayed in red represent the minimal real log canonical thresholds

$(2, 1, 2)$. Thus,

$$F_1(X|\theta_1) + F_2(X|\theta_2) = \begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix} \theta_1 + \frac{1}{2} \begin{bmatrix} -8 \\ 8 \\ -8 \end{bmatrix} \theta_2^2 = \begin{bmatrix} 2\theta_1 - 4\theta_2^2 \\ 0 + 4\theta_2^2 \\ -2\theta_1 - 4\theta_2^2 \end{bmatrix}$$

Then,

$$\begin{aligned} & \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1) + F_2(X|\theta_2)\}^2 \right] \\ &= \frac{1}{2} \left\{ (2\theta_1 - 4\theta_2^2)^2 \cdot \frac{1}{4} + (0 + 4\theta_2^2)^2 \cdot \frac{1}{2} + (-2\theta_1 - 4\theta_2^2)^2 \cdot \frac{1}{4} \right\} = \theta_1^2 + 8\theta_2^4 \end{aligned}$$

On the other hand, we obtain:

$$K(\theta) = \theta_1^2 + 8\theta_2^4 + 8\theta_1^2\theta_2^2 - 16\theta_1\theta_2^3 + \dots \quad (5.2)$$

where terms of order five and higher are omitted. Therefore, as stated in Theorem 5.1, $K(\theta)$ can be expressed as

$$K(\theta) = \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1) + F_2(X|\theta_2)\}^2 \right] + (\text{higher order terms}).$$

Next, we verify that the Main Theorem holds by performing the coordinate transformation g , known as a "blow-up" (CT1)-(CT4) as illustrated in Figure 3. This transformation ensures that a normal crossing is obtained for each local coordinate.

(CT1) Transforming with $\theta_2 = \theta_1 \theta'_2$ in (5.2) gives:

$$K(\theta) = \theta_1^2 \{1 + 8\theta_1^2 \theta_2'^4 + \theta_1 \times \dots\} = \theta_1^2 a_1(\theta_1, \theta'_2), \quad a_1(0, \theta'_2) = 1 \neq 0 \quad (5.3)$$

Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_1 = 0$ is satisfied. Therefore, a normal crossing is achieved in this local coordinate system (θ_1, θ'_2) .

(CT2) Transforming with $\theta_1 = \theta_2 \theta'_1$ in (5.2) results in:

$$K(\theta) = \theta_2^2 \{\theta_1'^2 + 8\theta_2^2 + \theta_2 \times \dots\} = \theta_2^2 a_2(\theta'_1, \theta_2), \quad a_2(\theta'_1, 0) = \theta_1'^2 \quad (5.4)$$

Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_2 = 0$ is satisfied. Therefore, a normal crossing is achieved in this local coordinate system (θ'_1, θ_2) except at the point $(\theta'_1, \theta_2) = 0$.

To determine the normal crossing at this point $(\theta'_1, \theta_2) = 0$, we perform another blow-up centered at this point. Under the transformations (CT3): $\theta_2 = \theta'_1 \theta_2''$ and (CT4): $\theta'_1 = \theta_2 \theta_1''$, we obtain:

$$K(\theta) = \theta_1'^4 \theta_2''^2 \{1 + 8\theta_2''^2 + \theta'_1 \times \dots\} = \theta_1'^4 \theta_2''^2 a_3(\theta'_1, \theta_2''), \quad a_3(0, \theta_2'') = 1 + 8\theta_2''^2 \neq 0 \quad (5.5)$$

$$K(\theta) = \theta_2^4 \{\theta_1''^2 + 8 + \theta_2 \times \dots\} = \theta_2^4 a_4(\theta_1'', \theta_2), \quad a_4(\theta_1'', 0) = \theta_1''^2 + 8 \neq 0 \quad (5.6)$$

Therefore, normal crossings are achieved in these local coordinates.

We compile the normal crossings for each local coordinate into Table 2, where the symbols $g, k_i^{(Q)}, h_i^{(Q)}$ refer to Definition 2.1. Thus, we determine that the real log canonical threshold is $3/4$ (with multiplicity $\mathfrak{m} = 1$).

Table 2 Normal crossings for each local coordinate

	u	$g(u)$	$K(g(u))$	$ g'(u) $	$k_i^{(Q)}$	$h_i^{(Q)}$	λ
(5.3)	(θ_1, θ'_2)	$(\theta_1, \theta_1 \theta'_2)$	$\theta_1^2 a_1(\theta_1, \theta'_2)$	$ \theta_1 $	$(2, 0)$	$(1, 0)$	1
(5.5)	(θ'_1, θ''_2)	$(\theta_1'^2 \theta''_2, \theta'_1 \theta''_2)$	$\theta_1'^4 \theta_2''^2 a_3(\theta'_1, \theta''_2)$	$ \theta_1'^2 \theta_2'' $	$(4, 2)$	$(2, 1)$	3/4
(5.6)	(θ''_1, θ_2)	$(\theta_1'' \theta_2^2, \theta_2)$	$\theta_2^4 a_4(\theta''_1, \theta_2)$	$ \theta_2^2 $	$(0, 4)$	$(0, 2)$	3/4

Proof of Main Theorem. Using Theorem 5.1, the Taylor expansion of $K(\theta)$ at $(\theta_1, \dots, \theta_d) = 0$ can be expressed as:

$$K(\theta) = \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 \right] + (\text{higher order terms})$$

where the higher order terms are specifically expressed as the sum of the following four terms:

- $f_0(\theta_{r+1}, \dots, \theta_d)$: terms of degree $2m + 1$ and higher
- $f_1(\theta_1, \dots, \theta_d)$: terms of first degree homogeneous in $\theta_1, \dots, \theta_r$ and of degree $m + 1$ and higher in $\theta_{r+1}, \dots, \theta_d$
- $f_2(\theta_1, \dots, \theta_d)$: terms of second degree and higher in $\theta_1, \dots, \theta_r$ and of first degree and higher in $\theta_{r+1}, \dots, \theta_d$
- $f_3(\theta_1, \dots, \theta_r)$: terms of third degree and higher.

We now consider these terms at their lowest degrees, namely:

- $f_0(\theta_{r+1}, \dots, \theta_d)$: terms of degree $2m + 1$ homogeneous
- $f_1(\theta_1, \dots, \theta_d)$: terms of first degree homogeneous in $\theta_1, \dots, \theta_r$ and $m + 1$ degree homogeneous in $\theta_{r+1}, \dots, \theta_d$
- $f_2(\theta_1, \dots, \theta_d)$: terms of second degree homogeneous in $\theta_1, \dots, \theta_r$ and first degree homogeneous in $\theta_{r+1}, \dots, \theta_d$
- $f_3(\theta_1, \dots, \theta_r)$: terms of third degree homogeneous.

The proof for the general case proceeds similarly.

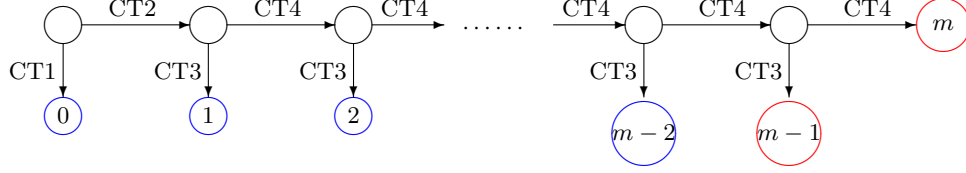


Fig. 4 The procedure of coordinate transformations g performed in the proof of the Main Theorem. Local coordinates displayed in red represent the minimal real log canonical thresholds

We perform the following four types of coordinate transformations, CT1–CT4 (see Figure 4).³ We demonstrate that normal crossings are achieved in all the resulting local coordinates.

CT1 For $1 \leq s \leq r$, $\{\theta_j \rightarrow \theta_s \theta'_j \mid 1 \leq j \leq d, j \neq s\}$

CT2 For $r+1 \leq t \leq d$, $\{\theta_j \rightarrow \theta_t \theta'_j \mid 1 \leq j \leq d, j \neq t\}$

CT3 For the t selected in CT2 and for $1 \leq s \leq r$, $\{\theta'_j \rightarrow \theta_s \theta''_j, \theta_t \rightarrow \theta_s \theta'_t \mid 1 \leq j \leq r, j \neq s\}$

CT4 For the t selected in CT2, $\{\theta'_j \rightarrow \theta_t \theta''_j \mid 1 \leq j \leq r\}$

Initially, we perform CT1 and CT2. The first transformation achieves normal crossings, but the second does not. Therefore, we proceed to perform further iterations involving CT3 and CT4. In these iterations, CT3 achieves normal crossings, but CT4 does not. By repeating this process up to $m-1$ times (resulting in a total of m iterations), we eventually achieve normal crossings. The details of this procedure are provided below.

Performing CT1 Once

We can assume $s = 1$ without loss of generality. In CT1, we perform $g = \{\theta_j = \theta_1 \theta'_j \mid 2 \leq j \leq d\}$.

³In terms of algebraic geometry, CT1 and CT2 correspond to “blow-ups centered at the origin of $\mathbb{R}^d$ ”. CT3 and CT4 correspond to “blow-ups centered at the subvariety $\{(\theta_1, \dots, \theta_d) \mid \theta_1 = \dots = \theta_r = \theta_t = 0\}$ when the exceptional surface from CT1 and CT2 is $\{\theta_t = 0\}$ ”.

$$\begin{aligned}
K(\theta) &= \frac{1}{2} \mathbb{E}_X \left[\{F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d)\}^2 \right] \\
&\quad + f_0(\theta_{r+1}, \dots, \theta_d) + f_1(\theta_1, \dots, \theta_d) + f_2(\theta_1, \dots, \theta_d) + f_3(\theta_1, \dots, \theta_r) \\
&= \frac{1}{2} \theta_1^2 \left\{ \mathbb{E}_X \left[\{F_1(X|1, \theta'_2, \dots, \theta'_r) + \theta_1^{m-1} F_m(X|\theta'_{r+1}, \dots, \theta'_d)\}^2 \right] \right. \\
&\quad + \theta_1^{2m-1} f_0(\theta'_{r+1}, \dots, \theta'_d) + \theta_1^m f_1(1, \theta'_2, \dots, \theta'_d) \\
&\quad \left. + \theta_1 f_2(1, \theta'_2, \dots, \theta'_d) + \theta_1 f_3(1, \theta'_2, \dots, \theta'_r) \right\} = \frac{1}{2} \theta_1^2 a(\theta_1, \theta'_2, \dots, \theta'_d)
\end{aligned}$$

where a is an analytic function. Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_1 = 0$ is satisfied. From Assumption 1(2), we have

$$a(0, \theta'_2, \dots, \theta'_d) = \mathbb{E}_X [F_1^2(X|1, \theta'_2, \dots, \theta'_r)] > 0,$$

and thus, a normal crossing is achieved.

The Jacobian of this coordinate transformation is θ_1^{d-1} , so

$$\inf_{Q \in U_0} \min_{j=1, \dots, d} \left\{ \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} \right\} = \frac{(d-1) + 1}{2} = \frac{d}{2}$$

(with multiplicity $\mathfrak{m} = 1$). The symbols $k_j^{(Q)}$ and $h_j^{(Q)}$ refer to Definition 2.1.

Sequence of Transformations: CT2 Once, CT4 $i (\leq m - 2)$ Times, and CT3

Assume without loss of generality that $t = d$. First, perform CT2: $g = \{\theta_j \rightarrow \theta_d \theta'_j \mid 1 \leq j \leq d - 1\}$, which transforms $K(\theta)$ as follows:

$$\begin{aligned} K(\theta) = \frac{1}{2} \theta_d^2 & \left\{ \mathbb{E}_X \left[\left\{ F_1(X|\theta'_1, \dots, \theta'_r) + \theta_d^{m-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) \right\}^2 \right] \right. \\ & + \theta_d^{2m-1} f_0(\theta'_{r+1}, \dots, \theta'_{d-1}, 1) + \theta_d^m f_1(\theta'_1, \dots, \theta'_{d-1}, 1) \\ & \left. + \theta_d f_2(\theta'_1, \dots, \theta'_{d-1}, 1) + \theta_d f_3(\theta'_1, \dots, \theta'_r) \right\} = \frac{1}{2} \theta_d^2 a(\theta'_1, \dots, \theta'_{d-1}, \theta_d), \end{aligned}$$

where a is an analytic function. Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_d = 0$ is satisfied. We have

$$a(\theta'_1, \dots, \theta'_{d-1}, 0) = \mathbb{E}_X [F_1^2(X|\theta'_1, \dots, \theta'_r)].$$

For any point Q on U_0 with $(\theta'_1, \dots, \theta'_r) \neq 0$, the normal crossing of $K(\theta)$ is achieved, and

$$\inf_Q \min_{j=1, \dots, d} \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} = \frac{d}{2}$$

(with multiplicity $\mathbf{m} = 1$). Therefore, we only need to consider points on U_0 where $(\theta'_1, \dots, \theta'_r) = 0$.

After applying CT4: $\{\theta'_j \rightarrow \theta_d \theta''_j \mid 1 \leq j \leq r\}$ i times, $K(\theta)$ can be expressed as:

$$\begin{aligned} K(\theta) = \frac{1}{2} \theta_d^{2(i+1)} & \left\{ \mathbb{E}_X \left[\left\{ F_1(X|\theta''_1, \dots, \theta''_r) + \theta_d^{m-i-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) \right\}^2 \right] \right. \\ & + \theta_d^{2(m-i)-1} f_0(\theta'_{r+1}, \dots, \theta'_{d-1}, 1) + \theta_d^{m-i} f_1(\theta''_1, \dots, \theta'_{d-1}, 1) \\ & \left. + \theta_d f_2(\theta''_1, \dots, \theta'_{d-1}, 1) + \theta_d^{i+1} f_3(\theta''_1, \dots, \theta''_r) \right\}. \end{aligned}$$

Additionally, we only need to consider the point where $(\theta''_1, \dots, \theta''_r, \theta_d) = (0, \dots, 0, 0)$.

We denote $\theta''_1, \dots, \theta''_r$ as $\theta'_1, \dots, \theta'_r$.

Next, perform CT3: $\{\theta'_j \rightarrow \theta'_1 \theta''_j, \theta_d \rightarrow \theta'_1 \theta'_d \mid 2 \leq j \leq r\}$, which transforms $K(\theta)$ into:

$$\begin{aligned} K(\theta) &= \frac{1}{2} \theta_1'^{2(i+2)} \theta_d'^{2(i+1)} \times \\ &\quad \left\{ \mathbb{E}_X \left[\{F_1(X|1, \theta''_2, \dots, \theta''_r) + \theta_1'^{m-i-2} \theta_d'^{m-i-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1)\}^2 \right] \right. \\ &\quad + \theta_1'^{2(m-i)-3} \theta_d'^{2(m-i)-1} f_0(\theta'_{r+1}, \dots, \theta'_{d-1}, 1) + \theta_1'^{m-i-1} \theta_d'^{m-i} f_1(1, \theta''_2, \dots, \theta'_{d-1}, 1) \\ &\quad \left. + \theta'_1 \theta'_d f_2(1, \theta''_2, \dots, \theta'_{d-1}, 1) + \theta_1'^{i+2} \theta_d'^{i+1} f_3(1, \theta''_2, \dots, \theta''_r) \right\} \\ &= \frac{1}{2} \theta_1'^{2(i+2)} \theta_d'^{2(i+1)} a(\theta'_1, \theta''_2, \dots, \theta'_d), \end{aligned}$$

where a is an analytic function. Considering points on $U_0 = g^{-1}(O)$, the condition $\theta'_1 = 0$ is satisfied. We have

$$\begin{aligned} a(0, \theta''_2, \dots, \theta'_d) &= \\ &\quad \begin{cases} \mathbb{E}_X \left[\{F_1(X|1, \theta''_2, \dots, \theta''_r) + \theta_d'^{m-i-1} F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1)\}^2 \right] & \text{if } i = m-2, \\ \mathbb{E}_X \left[F_1(X|1, \theta''_2, \dots, \theta''_r)^2 \right] & \text{if } i < m-2. \end{cases} \end{aligned}$$

According to Assumption 1(2), this expression is positive. Therefore, in this local coordinate, the normal crossing of $K(\theta)$ is achieved at any point Q on U_0 . The Jacobian of this coordinate transformation is $\theta_1'^{d+r(i+1)-1} \theta_d'^{d+ri-1}$, leading to:

$$\inf_{Q \in U_0} \min_{j=1, \dots, d} \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} = \min \left\{ \frac{d+r(i+1)-1+1}{2(i+2)}, \frac{d+ri-1+1}{2(i+1)} \right\} = \frac{d-r+r(i+2)}{2(i+2)}$$

(with multiplicity $\mathbf{m} = 1$).

After Performing CT2 Once, Then Applying CT4 $m - 1$ Times

Following the above method, after performing the coordinate transformation g and using the analytic function a , $K(\theta)$ can be expressed as:

$$\begin{aligned} K(\theta) = \frac{1}{2}\theta_d^{2m} & \left\{ \mathbb{E}_X \left[\{F_1(X|\theta''_1, \dots, \theta''_r) + F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1)\}^2 \right] \right. \\ & + \theta_d f_0(\theta'_{r+1}, \dots, \theta'_{d-1}, 1) + \theta_d f_1(\theta''_1, \dots, \theta'_{d-1}, 1) \\ & \left. + \theta_d f_2(\theta''_1, \dots, \theta'_{d-1}, 1) + \theta_d^m f_3(\theta''_1, \dots, \theta''_r) \right\} = \frac{1}{2}\theta_d^{2m} a(\theta''_1, \dots, \theta'_{d-1}, \theta_d). \end{aligned}$$

Considering points on $U_0 = g^{-1}(O)$, the condition $\theta_d = 0$ is satisfied. In accordance with Assumption 1(2), we have

$$a(\theta''_1, \dots, \theta'_{d-1}, 0) = \mathbb{E}_X \left[\{F_1(X|\theta''_1, \dots, \theta''_r) + F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1)\}^2 \right] > 0.$$

This conclusion can be understood as follows. The condition $a(\theta''_1, \dots, \theta'_{d-1}, 0) = 0$ holds if and only if

$$\begin{aligned} & F_1(X|\theta''_1, \dots, \theta''_r) + F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) = 0 \quad \text{a.s.} \\ \Leftrightarrow & \sum_{k=1}^r \theta''_k \cdot \frac{\partial f(X|\theta)}{\partial \theta_k} \Big|_{(\theta_1, \dots, \theta_d)=0} + F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) = 0 \quad \text{a.s..} \end{aligned}$$

However, this equivalence does not hold because, by Assumption 1(2), the random variable $F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1)$ is linearly independent of the r random variables

$$\frac{\partial f(X|\theta)}{\partial \theta_1} \Big|_{(\theta_1, \dots, \theta_d)=0}, \dots, \frac{\partial f(X|\theta)}{\partial \theta_r} \Big|_{(\theta_1, \dots, \theta_d)=0}.$$

Thus, the normal crossing of $K(\theta)$ is achieved. The Jacobian of this coordinate transformation is $\theta_d^{d+r(m-1)-1}$, so

$$\inf_{Q \in U_0} \min_{j=1, \dots, d} \frac{h_j^{(Q)} + 1}{k_j^{(Q)}} = \frac{d + r(m-1) - 1 + 1}{2m} = \frac{d - r + rm}{2m}$$

(with multiplicity $\mathbf{m} = 1$). Therefore, we achieve normal crossings at all local coordinates, and we determine the real log canonical threshold as:

$$\min_{1 \leq m' \leq m} \left\{ \frac{d - r + rm'}{2m'} \right\} = \frac{d - r + rm}{2m}.$$

□

Remark 5.1. In the proof of the Main Theorem, after performing CT2 once and CT4 $m - 1$ times, we consider the change in the log-likelihood ratio function f before and after the coordinate transformation g . It can be expressed as

$$\begin{aligned} f(X|\theta) &= F_1(X|\theta_1, \dots, \theta_r) + F_m(X|\theta_{r+1}, \dots, \theta_d) + (\text{higher order terms}) \\ &= \theta_d^m \{ F_1(X|\theta_1'', \dots, \theta_r'') + F_m(X|\theta_{r+1}', \dots, \theta_{d-1}', 1) + b_1(\theta_d) \}, \end{aligned}$$

where $b_1(0) = 0$. Meanwhile, $K(\theta)$ can be expressed as:

$$K(\theta) = \frac{1}{2} \theta_d^{2m} \left\{ \mathbb{E}_X \left[\{ F_1(X|\theta_1'', \dots, \theta_r'') + F_m(X|\theta_{r+1}', \dots, \theta_{d-1}', 1) \}^2 \right] + b_2(\theta_d) \right\},$$

where $b_2(0) = 0$. From this, it follows that:

$$\lim_{\theta \rightarrow 0} \frac{\mathbb{E}_X[f(X|\theta)^2]}{K(\theta)}$$

$$= 2 \lim_{\theta_d \rightarrow 0} \frac{\mathbb{E}_X \left[\left\{ F_1(X|\theta''_1, \dots, \theta''_r) + F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) + b_1(\theta_d) \right\}^2 \right]}{\mathbb{E}_X \left[\left\{ F_1(X|\theta''_1, \dots, \theta''_r) + F_m(X|\theta'_{r+1}, \dots, \theta'_{d-1}, 1) \right\}^2 \right] + b_2(\theta_d)} = 2$$

This is consistent with Watanabe (2009), Theorem 6.3.

Remark 5.2. When considering the coordinate transformation g performed in the Main Theorem, we can deduce that the only point in the neighborhood of the origin in $\mathbb{R}^d$ where $K(\theta) = 0$ is the origin itself. For example, after performing CT2 once and CT4 $m-1$ times, in the local coordinates $u = (\theta''_1, \dots, \theta'_{d-1}, \theta_d)$, we have $K(\theta) = K(g(u)) = 0 \Rightarrow \theta_d = 0$ for all $\theta''_1, \dots, \theta'_{d-1}$, and it follows that $g(\theta''_1, \dots, \theta'_{d-1}, \theta_d = 0) = (0, \dots, 0)$.⁴ Therefore, even without assuming the condition that “ Θ_* contains only the single element” in the definition of semiregularity, it follows that Θ_* consists of only one point in a neighborhood of $\theta = 0$, provided that Assumption 1 is satisfied.

6 Conclusion

In this paper, we defined semiregular models as a generalization of regular models and provided a formula for the learning coefficient λ for semiregular models. As an example, we presented a formula for the learning coefficient of a specific mixture distribution with constant mixture proportions.

In deriving the main theorem, we used the lower-order terms in the Taylor expansion of the Kullback-Leibler divergence (Theorem 5.1). We found that the natural number m in Assumption 1(1) significantly influences the value of the learning coefficient λ .

For future research, we intend to further generalize the methods used in this study and derive formulas for the learning coefficient λ for simple singular models where Θ_* is not a single point.

⁴In the terminology of algebraic geometry, this follows because the centers of the blow-ups conducted are all contained in the exceptional surfaces of the previous blow-ups.

Acknowledgements. I am grateful to Professor Joe Suzuki of Osaka University for teaching me the basics of Bayesian theory and for providing a research topic that bridges algebraic geometry and statistics. I also thank him for his valuable comments, advice, and for checking this manuscript. I express my gratitude to Professor Sumio Watanabe, who proposed the concept of the learning coefficient.

Declarations

The author has no relevant financial or non-financial interests to disclose.

Appendix A Proof of Theorem 5.1

Let $i_1, \dots, i_k \in \{1, \dots, d\}$. We define

$$Y_{i_1, \dots, i_k} := -\frac{\partial^k f(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_k}},$$

where we aim to identify patterns in the partial derivatives of the log-likelihood ratio function f . We refer to $Y_{i_1, \dots, i_k}$ as indeterminate symbols of weight k . Setting $\theta = (\theta_1, \dots, \theta_d)$, we obtain the following:

$$\begin{aligned} \frac{\partial f(x|\theta)}{\partial \theta_{i_1}} &= -\frac{1}{p(x|\theta)} \frac{\partial p(x|\theta)}{\partial \theta_{i_1}}, \\ \frac{\partial^2 f(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2}} &= -\frac{1}{p(x|\theta)} \frac{\partial^2 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2}} + Y_{i_1} Y_{i_2}, \\ \frac{\partial^3 f(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} &= -\frac{1}{p(x|\theta)} \frac{\partial^3 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3}} + Y_{i_1} Y_{i_2} Y_{i_3} + Y_{i_1, i_2} Y_{i_3} + Y_{i_2, i_3} Y_{i_1} + Y_{i_3, i_1} Y_{i_2}, \\ \frac{\partial^4 f(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3} \partial \theta_{i_4}} &= -\frac{1}{p(x|\theta)} \frac{\partial^4 p(x|\theta)}{\partial \theta_{i_1} \partial \theta_{i_2} \partial \theta_{i_3} \partial \theta_{i_4}} + Y_{i_1} Y_{i_2} Y_{i_3} Y_{i_4} \\ &\quad + Y_{i_1, i_2} Y_{i_3} Y_{i_4} + Y_{i_1, i_3} Y_{i_2} Y_{i_4} + Y_{i_1, i_4} Y_{i_2} Y_{i_3} \\ &\quad + Y_{i_2, i_3} Y_{i_1} Y_{i_4} + Y_{i_2, i_4} Y_{i_1} Y_{i_3} + Y_{i_3, i_4} Y_{i_1} Y_{i_2} \\ &\quad + Y_{i_1, i_2} Y_{i_3, i_4} + Y_{i_1, i_3} Y_{i_2, i_4} + Y_{i_1, i_4} Y_{i_2, i_3} \end{aligned} \tag{A1}$$

$$+Y_{i_1,i_2,i_3}Y_{i_4} + Y_{i_1,i_2,i_4}Y_{i_3} + Y_{i_1,i_3,i_4}Y_{i_2} + Y_{i_2,i_3,i_4}Y_{i_1}.$$

Assuming we can interchange the order of partial differentiation and integration, the first term on the right-hand side becomes zero at $\theta = 0$ when taking the expectation with respect to X . Therefore, the partial derivatives of $K(\theta)$ at $\theta = 0$ depend primarily on the terms from the second term onward.

Remark A.1. By taking the expectation with respect to X in (A1) and substituting $\theta = 0$, we find that the Fisher information matrices I and J coincide.

The subsequent terms on the right-hand side are sums of products of two or more indeterminate symbols Y . Each term has a coefficient of 1, and the total weight of each product equals the order of differentiation. This pattern holds in general.

Proposition A.1.

$$G_n(i_1, \dots, i_n) := \frac{\partial^n f(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_n}} + \frac{1}{p(x|\theta)} \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_n}}$$

is a polynomial consisting of indeterminate symbols $Y_{i_1, \dots, i_k}$ with weights less than n . Each term in the polynomial has a coefficient of 1, and each term is a product of two or more indeterminates, with a total weight of n . The polynomial is the sum of all such terms with these weights.

Proof. We prove this by induction on n . For $n = 1$, $G_1(i_1) = 0$ satisfies the condition since there are no terms. Assuming $G_n(i_1, \dots, i_n)$ is such a polynomial for some $n \geq 1$, we obtain

$$\begin{aligned} & \frac{\partial^{n+1} f(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_{n+1}}} \\ &= \frac{\partial}{\partial \theta_{i_{n+1}}} \left\{ -\frac{1}{p(x|\theta)} \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_n}} + G_n(i_1, \dots, i_n) \right\} \\ &= -\frac{1}{p(x|\theta)} \frac{\partial^{n+1} p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_{n+1}}} + \frac{1}{p(x|\theta)} \frac{\partial^n p(x|\theta)}{\partial \theta_{i_1} \dots \partial \theta_{i_n}} Y_{i_{n+1}} + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}} \end{aligned}$$

$$= -\frac{1}{p(x|\theta)} \frac{\partial^{n+1} p(x|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_{n+1}}} + G_n(i_1, \dots, i_n) Y_{i_{n+1}} + Y_{i_1, \dots, i_n} Y_{i_{n+1}} + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}}.$$

Thus,

$$G_{n+1}(i_1, \dots, i_{n+1}) = G_n(i_1, \dots, i_n) Y_{i_{n+1}} + Y_{i_1, \dots, i_n} Y_{i_{n+1}} + \frac{\partial G_n(i_1, \dots, i_n)}{\partial \theta_{i_{n+1}}}.$$

By the induction hypothesis, this is a polynomial consisting of indeterminate symbols Y with weights less than or equal to n , where each term has a coefficient of 1 and the sum of weights in each term equals $n + 1$. Moreover, the terms containing the indeterminate $Y_{i_{n+1}}$ are derived from the first and second terms on the right-hand side, while the terms that do not contain $Y_{i_{n+1}}$ are derived from the third term on the right-hand side. Thus, we have shown that each term is a product of two or more indeterminate symbols, summing to a total weight of $n + 1$. $\square$

Remark A.2. The statement of Proposition A.1 can be written explicitly as follows:

$$G_n(i_1, \dots, i_n) = \sum_{T \in S_{i_1, \dots, i_n}} \prod_{U \in T} Y_U, \quad (\text{A2})$$

where $S_{i_1, \dots, i_n}$ is the set of all disjoint partitions of the set $\{i_1, \dots, i_n\}$. A disjoint partition satisfies

$$\bigcup_i U_i = \{i_1, \dots, i_n\}, \quad U_i \cap U_j = \{\} \text{ for } i \neq j, \quad U_i \neq \{\}, \{i_1, \dots, i_n\},$$

and each partition is denoted as $T = \{U_1, U_2, \dots\}$.

For example, when $n = 2$, the disjoint partition of $\{i_1, i_2\}$ consists of a single element, $T_1 = \{\{i_1\}, \{i_2\}\}$, so $S_{i_1, i_2} = \{T_1\}$. Substituting this into equation (A2), we

have

$$G_2(i_1, i_2) = \prod_{U \in T_1} Y_U = Y_{i_1} Y_{i_2}.$$

Similarly, when $n = 3$, the disjoint partitions of $\{i_1, i_2, i_3\}$ are given by $T_1 = \{\{i_1\}, \{i_2\}, \{i_3\}\}$, $T_2 = \{\{i_1, i_2\}, \{i_3\}\}$, $T_3 = \{\{i_2, i_3\}, \{i_1\}\}$, $T_4 = \{\{i_3, i_1\}, \{i_2\}\}$, so $S_{i_1, i_2, i_3} = \{T_1, T_2, T_3, T_4\}$. Substituting into equation (A2), we have

$$\begin{aligned} G_3(i_1, i_2, i_3) &= \prod_{U \in T_1} Y_U + \prod_{U \in T_2} Y_U + \prod_{U \in T_3} Y_U + \prod_{U \in T_4} Y_U \\ &= Y_{i_1} Y_{i_2} Y_{i_3} + Y_{i_1, i_2} Y_{i_3} + Y_{i_2, i_3} Y_{i_1} + Y_{i_3, i_1} Y_{i_2}. \end{aligned}$$

Corollary A.1.

- (1) *If all indeterminate symbols of weight $m - 1$ or less are zero almost surely at $\theta = 0$, then G_k (for $k \leq 2m - 1$) will also be zero almost surely at $\theta = 0$.*
- (2) *When the interchange of differentiation and integration is permissible,*

$$\left. \frac{\partial^n K(\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} \right|_{\theta=0} = \mathbb{E}_X \left[\left. \frac{\partial^n f(X|\theta)}{\partial \theta_{i_1} \cdots \partial \theta_{i_n}} \right|_{\theta=0} \right]$$

equals the expectation over X of a polynomial. This polynomial is obtained by summing all terms where at least two indeterminates $Y_{i_1, \dots, i_k}$ are multiplied together, the total weight is n , and the coefficients of all terms are 1.

Proof. (1) follows because terms with a sum of weights up to $2m - 1$ are products of two or more terms, so at least one term will have a weight of $m - 1$ or less and thus will be zero almost surely at $\theta = 0$. (2) follows directly from the previous discussions. $\square$

Proof of Theorem 5.1. We need to prove the following five points about the Taylor expansion of $K(\theta)$ at $\theta = 0$:

- (1) Coefficients of terms of order zero in $(\theta_1, \dots, \theta_r)$ and of order $2m - 1$ or less in $(\theta_{r+1}, \dots, \theta_d)$ are zero.
- (2) Coefficients of terms of first order in $(\theta_1, \dots, \theta_r)$ and of order $m - 1$ or less in $(\theta_{r+1}, \dots, \theta_d)$ are zero.
- (3) Terms of first order in $(\theta_1, \dots, \theta_r)$ and of m -th order in $(\theta_{r+1}, \dots, \theta_d)$ are represented by:

$$\mathbb{E}_X [F_1(X|\theta_1, \dots, \theta_r) F_m(X|\theta_{r+1}, \dots, \theta_d)].$$

- (4) Terms of second order in $(\theta_1, \dots, \theta_r)$ and of zero order in $(\theta_{r+1}, \dots, \theta_d)$ are represented by:

$$\frac{1}{2} \mathbb{E}_X [F_1^2(X|\theta_1, \dots, \theta_r)].$$

- (5) Terms of zero order in $(\theta_1, \dots, \theta_r)$ and of $2m$ -th order in $(\theta_{r+1}, \dots, \theta_d)$ are represented by:

$$\frac{1}{2} \mathbb{E}_X [F_m^2(X|\theta_{r+1}, \dots, \theta_d)].$$

For (1), it suffices to show that for $i_{r+1} + \dots + i_d \leq 2m - 1$,

$$\left. \frac{\partial^{i_{r+1} + \dots + i_d} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0}$$

is a random variable with an expected value of zero. This has already been demonstrated in Corollary A.1(1). For (2), for $1 \leq j \leq r$ and $i_{r+1} + \dots + i_d \leq m - 1$,

$$\left. \frac{\partial^{1+i_{r+1} + \dots + i_d} f(X|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0}$$

is shown to have an expected value of zero. Among the products of two or more indeterminates claimed in Corollary A.1(2), those that do not include θ_j are zero with probability 1 by assumption. Therefore, their expected value is zero.

For (3), using a similar argument as in (2), consider a case where $1 \leq j \leq r$ and $i_{r+1} + \dots + i_d = m$. For

$$\left. \frac{\partial^{m+1} f(X|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0},$$

the nonzero products of two or more indeterminates claimed in Corollary A.1(2) are limited to

$$\left. \frac{\partial f(X|\theta)}{\partial \theta_j} \right|_{\theta=0} \left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0}.$$

Therefore,

$$\mathbb{E}_X \left[\left. \frac{\partial^{m+1} f(X|\theta)}{\partial \theta_j \partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} \right] = \mathbb{E}_X \left[\left. \frac{\partial f(X|\theta)}{\partial \theta_j} \right|_{\theta=0} \left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} \right],$$

validating the claim in (3).

Since (4) is a special case of (5), it suffices to demonstrate (5) alone. For $i_{r+1} + \dots + i_d = 2m$,

$$\left. \frac{\partial^{2m} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0}$$

is restricted to the sum over the set of nonzero products of two or more indeterminates claimed in Corollary A.1(2), given by

$$\left\{ \left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{j_{r+1}} \dots \partial \theta_d^{j_d}} \right|_{\theta=0} \cdot \left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}-j_{r+1}} \dots \partial \theta_d^{i_d-j_d}} \right|_{\theta=0} \mid j_{r+1} + \dots + j_d = m, i_h \geq j_h \geq 0 \right\}.$$

Note that each element in this set appears as

$$\binom{i_{r+1}}{j_{r+1}} \dots \binom{i_d}{j_d} \times \begin{cases} \frac{1}{2} & \text{if } (2j_{r+1}, \dots, 2j_d) = (i_{r+1}, \dots, i_d), \\ 1 & \text{otherwise.} \end{cases}$$

Thus, we can express

$$\begin{aligned} \left. \frac{\partial^{2m} K(\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} &= \mathbb{E}_X \left[\left. \frac{\partial^{2m} f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} \right] \\ &= \frac{1}{2} \sum_{\substack{j_{r+1} + \dots + j_d = m \\ i_h \geq j_h \geq 0}} \binom{i_{r+1}}{j_{r+1}} \dots \binom{i_d}{j_d} \mathbb{E}_X \left[\left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{j_{r+1}} \dots \partial \theta_d^{j_d}} \right|_{\theta=0} \cdot \left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{i_{r+1}-j_{r+1}} \dots \partial \theta_d^{i_d-j_d}} \right|_{\theta=0} \right]. \end{aligned}$$

Transforming to the target expression, we find:

$$\begin{aligned} \frac{1}{2} \mathbb{E}_X [F_m^2(X|\theta_{r+1}, \dots, \theta_d)] &= \sum_{\substack{k_{r+1} + \dots + k_d = m \\ k_h \geq 0}} \sum_{\substack{l_{r+1} + \dots + l_d = m \\ l_h \geq 0}} \frac{\theta_{r+1}^{k_{r+1}+l_{r+1}} \dots \theta_d^{k_d+l_d}}{2k_{r+1}! \dots k_d! l_{r+1}! \dots l_d!} \\ &\quad \times \mathbb{E}_X \left[\left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{k_{r+1}} \dots \partial \theta_d^{k_d}} \right|_{\theta=0} \left. \frac{\partial^m f(X|\theta)}{\partial \theta_{r+1}^{l_{r+1}} \dots \partial \theta_d^{l_d}} \right|_{\theta=0} \right] \\ &= \sum_{\substack{i_{r+1} + \dots + i_d = 2m \\ i_h \geq 0}} \frac{1}{i_{r+1}! \dots i_d!} \left. \frac{\partial^{2m} K(\theta)}{\partial \theta_{r+1}^{i_{r+1}} \dots \partial \theta_d^{i_d}} \right|_{\theta=0} \theta_{r+1}^{i_{r+1}} \dots \theta_d^{i_d}. \end{aligned}$$

The proof is completed using the variable substitutions $k_h + l_h = i_h$, $k_h = j_h$ for $h = r+1, \dots, d$. Thus, the claim is demonstrated. $\square$

Remark A.3. In Theorem 5.1, semiregularity is not essential; for example, it does not necessitate Assumption 1(2) or that the realization parameters consist solely of a single point. Although it is beyond the scope of this paper, it is possible to generalize from the degree $(1, m)$ to a general degree (n, m) . Furthermore, it is also possible to generalize to sets containing three or more natural numbers.

Appendix B Derivation Method for m

Proposition B.1 (Properties of the Coordinate Transformation). *Fix nonnegative integers $i_{r+1}, \dots, i_d$ with $i_{r+1} + \dots + i_d \geq 1$. For $k = 1, \dots, r$, consider the coordinate*

transformation $\varphi : (\theta_1, \dots, \theta_d) \mapsto (\theta'_1, \dots, \theta'_d)$ defined by

$$a_k(\theta_{r+1}, \dots, \theta_d) := \frac{c_k}{i_{r+1}! \dots i_d!} \cdot \theta_{r+1}^{i_{r+1}} \dots \theta_d^{i_d}, \quad c_k \in \mathbb{R},$$

where

$$\begin{aligned} \theta'_1 &= \theta_1 + a_1(\theta_{r+1}, \dots, \theta_d), \quad \dots, \quad \theta'_r = \theta_r + a_r(\theta_{r+1}, \dots, \theta_d), \\ \theta'_{r+1} &= \theta_{r+1}, \quad \dots, \quad \theta'_d = \theta_d. \end{aligned}$$

This transformation satisfies the following conditions:

- (1) $\varphi(0) = 0$, $|\det \varphi'(0)| = 1$, and φ is bijective and analytic in a neighborhood of 0.
- (2) For $j = 1, \dots, r$, we have $\frac{\partial}{\partial \theta'_j} = \frac{\partial}{\partial \theta_j}$.
- (3) For any $(h_{r+1}, \dots, h_d) \in \mathbb{Z}_{\geq 0}^{d-r}$ with $h_{r+1} + \dots + h_d \leq i_{r+1} + \dots + i_d$, the following holds:

$$\begin{aligned} & \left. \frac{\partial^{h_{r+1} + \dots + h_d}}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{(\theta'_1, \dots, \theta'_d)=0} \\ &= \left. \frac{\partial^{h_{r+1} + \dots + h_d}}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{\theta=0} - \begin{cases} 0 & \text{if } (h_{r+1}, \dots, h_d) \neq (i_{r+1}, \dots, i_d), \\ \sum_{k=1}^r c_k \frac{\partial}{\partial \theta_k} \Big|_{\theta=0} & \text{if } (h_{r+1}, \dots, h_d) = (i_{r+1}, \dots, i_d). \end{cases} \end{aligned}$$

Proof. (1) is straightforward. (2) follows from the relationship of partial derivatives before and after the transformation:

$$\begin{aligned} \frac{\partial}{\partial \theta'_j} &= \sum_{k=1}^d \frac{\partial}{\partial \theta_k} \cdot \frac{\partial \theta_k}{\partial \theta'_j} = \sum_{k=1}^r \frac{\partial}{\partial \theta_k} \cdot \frac{\partial \theta_k}{\partial \theta'_j} + \sum_{k=r+1}^d \frac{\partial}{\partial \theta_k} \cdot \delta_{k,j} \\ &= \frac{\partial}{\partial \theta_j} - \sum_{k=1}^r \frac{\partial a_k}{\partial \theta_j} \cdot \frac{\partial}{\partial \theta_k} \quad (j = r+1, \dots, d), \end{aligned}$$

$$\frac{\partial}{\partial \theta'_j} = \frac{\partial}{\partial \theta_j} \quad (j = 1, \dots, r).$$

For (3), let $m := i_{r+1} + \dots + i_d \geq 1$ and $l := h_{r+1} + \dots + h_d \leq m$. We have:

$$\begin{aligned} \frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} &= \prod_{j=r+1}^d \left(\frac{\partial}{\partial \theta_j} - \sum_{k=1}^r \frac{\partial a_k}{\partial \theta_j} \cdot \frac{\partial}{\partial \theta_k} \right)^{h_j} \\ &= \prod_{j=r+1}^d \left(\frac{\partial^{h_j}}{\partial \theta_j^{h_j}} - \sum_{k=1}^r \frac{\partial^{h_j} a_k}{\partial \theta_j^{h_j}} \cdot \frac{\partial}{\partial \theta_k} + \dots \right) \end{aligned} \quad (\text{B1})$$

$$= \frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} - \sum_{k=1}^r \frac{\partial^l a_k}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \cdot \frac{\partial}{\partial \theta_k} + \dots \quad (\text{B2})$$

We focus on the terms indicated by “...” in equations (B1) and (B2), which involve derivatives of $a_k(\theta_{r+1}, \dots, \theta_d)$ of order less than or equal to $l - 1$ with respect to $\theta_{r+1}, \dots, \theta_d$. Since $a_k(\theta_{r+1}, \dots, \theta_d)$ is a homogeneous polynomial of degree m in $\theta_{r+1}, \dots, \theta_d$, these terms evaluate to zero at $(\theta_1, \dots, \theta_d) = 0$. Therefore, we have:

$$\left. \frac{\partial^l a_k}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{\theta=0} = \begin{cases} 0 & \text{if } (h_{r+1}, \dots, h_d) \neq (i_{r+1}, \dots, i_d), \\ c_k & \text{if } (h_{r+1}, \dots, h_d) = (i_{r+1}, \dots, i_d). \end{cases}$$

Consequently,

$$\begin{aligned} &\left. \frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{(\theta'_1, \dots, \theta'_d)=0} \\ &= \left. \frac{\partial^l}{\partial \theta_{r+1}^{h_{r+1}} \dots \partial \theta_d^{h_d}} \right|_{\theta=0} - \begin{cases} 0 & \text{if } (h_{r+1}, \dots, h_d) \neq (i_{r+1}, \dots, i_d), \\ \sum_{k=1}^r c_k \frac{\partial}{\partial \theta_k} \Big|_{\theta=0} & \text{if } (h_{r+1}, \dots, h_d) = (i_{r+1}, \dots, i_d), \end{cases} \end{aligned}$$

as desired. $\square$

Appendix C Detailed Analysis of Example 4.1

We verify that the statistical model in Example 4.1 is a semiregular model satisfying Assumption 1 when $(d, r, m) = (H, 1, 2)$.

First, from Lemma C.1, we note that the set of realization parameters Θ_* consists only of the point $(\theta_1, \dots, \theta_H) = (\theta_*, \dots, \theta_*)$. We redefine the statistical model by applying a translation that moves the realization parameter to the origin:

$$p(x|\theta) := T_1 \tilde{p}(x|\theta_1 + \theta_*) + \dots + T_{H-1} \tilde{p}(x|\theta_{H-1} + \theta_*) + \left(1 - \sum_{i=1}^{H-1} T_i\right) \tilde{p}(x|\theta_H + \theta_*),$$

and obtain

$$\begin{aligned} \left. \frac{\partial f}{\partial \theta_i} \right|_{(\theta_1, \dots, \theta_H)=0} &= -\frac{T_i}{q(x)} \frac{\partial \tilde{p}}{\partial \theta}(x|\theta_*) \quad (i = 1, \dots, H-1), \\ \left. \frac{\partial f}{\partial \theta_H} \right|_{(\theta_1, \dots, \theta_H)=0} &= -\frac{1 - \sum_{i=1}^{H-1} T_i}{q(x)} \frac{\partial \tilde{p}}{\partial \theta}(x|\theta_*). \end{aligned}$$

These derivatives satisfy the linear dependence relation:

$$\left. \frac{\partial f}{\partial \theta_i} \right|_{(\theta_1, \dots, \theta_H)=0} = \frac{T_i}{1 - \sum_{i=1}^{H-1} T_i} \left. \frac{\partial f}{\partial \theta_H} \right|_{(\theta_1, \dots, \theta_H)=0} \quad (i = 1, \dots, H-1).$$

By applying the corresponding coordinate transformation

$$\theta'_H := \theta_H + \frac{\sum_{i=1}^{H-1} T_i \theta_i}{1 - \sum_{i=1}^{H-1} T_i},$$

we have

$$\left. \frac{\partial f}{\partial \theta_i} \right|_{(\theta_1, \dots, \theta_{H-1}, \theta'_H)=0} = 0 \quad (i = 1, \dots, H-1),$$

$$\left. \frac{\partial f}{\partial \theta'_H} \right|_{(\theta_1, \dots, \theta_{H-1}, \theta'_H)=0} = -\frac{1 - \sum_{i=1}^{H-1} T_i}{q(x)} \frac{\partial \tilde{p}}{\partial \theta}(x|\theta_*). \quad (\text{C1})$$

The second-order differential $F_2(x|\theta_1, \dots, \theta_{H-1})$ can be expressed as the following quadratic form:

$$F_2(x|\theta_1, \dots, \theta_{H-1}) = \frac{-1}{1 - \sum_{i=1}^{H-1} T_i} \cdot \frac{1}{2q(x)} \cdot \frac{\partial^2 \tilde{p}}{\partial \theta^2}(x|\theta_*) \cdot (\theta_1 \dots \theta_{H-1}) \Sigma (\theta_1 \dots \theta_{H-1})^\top,$$

where the symmetric matrix $\Sigma = (\sigma_{i,j})_{1 \leq i,j \leq H-1} \in \mathbb{R}^{(H-1) \times (H-1)}$ is defined as

$$\sigma_{i,j} := \begin{cases} T_i \left(1 - \sum_{k=1}^{H-1} T_k + T_i \right) & \text{if } i = j, \\ T_i T_j & \text{if } i \neq j. \end{cases} \quad (\text{C2})$$

It is clear that $\frac{\partial \tilde{p}}{\partial \theta}(x|\theta_*)$ and $\frac{\partial^2 \tilde{p}}{\partial \theta^2}(x|\theta_*)$ are linearly independent. Furthermore, as Lemma C.2 shows, Σ is a positive definite matrix. Therefore,

$$\forall (\theta_1, \dots, \theta_{H-1}) \neq 0, \quad (\theta_1 \dots \theta_{H-1}) \Sigma (\theta_1 \dots \theta_{H-1})^\top \neq 0,$$

and $F_2(x|\theta_1, \dots, \theta_{H-1})$ and (C1) are linearly independent. Thus, we confirm that this statistical model satisfies Assumption 1 when $(d, r, m) = (H, 1, 2)$.

Remark C.1. The Taylor expansion of

$$p(x|\theta) := \sum_{i=1}^{H-1} T_i \tilde{p}(x|\theta_i + \theta_*) + \left(1 - \sum_{i=1}^{H-1} T_i \right) \tilde{p} \left(x \left| \theta_H - \frac{\sum_{i=1}^{H-1} T_i \theta_i}{1 - \sum_{i=1}^{H-1} T_i} + \theta_* \right. \right)$$

at $\theta = 0$ is given by

$$p(x|\theta) = q(x) + \left(1 - \sum_{i=1}^{H-1} T_i \right) \frac{\partial \tilde{p}}{\partial \theta}(x|\theta_*) \cdot \theta_H$$

$$\begin{aligned}
& + \frac{1}{2(1 - \sum_i T_i)} \frac{\partial^2 \tilde{p}}{\partial \theta^2}(x|\theta_*) \cdot (\theta_1 \dots \theta_{H-1}) \Sigma(\theta_1 \dots \theta_{H-1})^\top \\
& + (\text{higher order terms}).
\end{aligned}$$

Here, the higher order terms do not include the following:

- Terms of degree 2 or less involving only $\theta_1, \dots, \theta_{H-1}$,
- Terms of degree 1 involving only θ_H .

Lemma C.1. *Under the assumptions of Example 4.1, the set of realization parameters Θ_* consists only of the point $(\theta_1, \dots, \theta_H) = (\theta_*, \dots, \theta_*)$.*

Proof. Consider the equation:

$$T_1 \tilde{p}(x|\theta_1) + \dots + T_{H-1} \tilde{p}(x|\theta_{H-1}) + \left(1 - \sum_{i=1}^{H-1} T_i\right) \tilde{p}(x|\theta_H) = \tilde{p}(x|\theta_*) \quad \text{a.s.} \quad (\text{C3})$$

Multiplying both sides of (C3) by x and summing over $x \in \chi$ yields:

$$T_1 \theta_1 + \dots + T_{H-1} \theta_{H-1} + \left(1 - \sum_{i=1}^{H-1} T_i\right) \theta_H = \theta_*.$$

Similarly, multiplying both sides of (C3) by x^2 and summing gives:

$$T_1 \theta_1^2 + \dots + T_{H-1} \theta_{H-1}^2 + \left(1 - \sum_{i=1}^{H-1} T_i\right) \theta_H^2 = \theta_*^2.$$

This implies that:

$$\sum_{i=1}^{H-1} T_i \theta_i^2 + \left(1 - \sum_{i=1}^{H-1} T_i\right) \theta_H^2 = \theta_*^2 = \left\{ \sum_{i=1}^{H-1} T_i \theta_i + \left(1 - \sum_{i=1}^{H-1} T_i\right) \theta_H \right\}^2.$$

Since equality holds in Jensen's inequality, it follows that $\theta_1 = \dots = \theta_H = \theta_*$. $\square$

Lemma C.2. Let N be an integer greater than or equal to 2, and consider variables $T_1, \dots, T_N$ with $0 < T_i < 1$ and $\sum_{i=1}^N T_i < 1$. We define the symmetric matrix $\Sigma_N = (\sigma_{i,j})_{1 \leq i,j \leq N} \in \mathbb{R}^{N \times N}$ as specified in (C2), substituting N for $H - 1$. Then, Σ_N is positive definite.

Proof. We prove this by induction on N . For $N = 2$, the assertion clearly holds. We assume that the statement holds up to $N - 1$. Let $\tilde{\Sigma}_N \in \mathbb{R}^{(N-1) \times (N-1)}$ be the matrix obtained from Σ_N by removing the N -th row and the N -th column. We consider $\mathbf{S}_N := (s_1, \dots, s_N)^\top \in \mathbb{R}^N$. By completing the square with respect to s_N in the following quadratic form, we obtain

$$\begin{aligned} & \mathbf{S}_N^\top \Sigma_N \mathbf{S}_N \\ &= T_N \left(1 - \sum_{i=1}^{N-1} T_i \right) s_N^2 + 2 \sum_{i=1}^{N-1} T_i T_N s_i s_N + \mathbf{S}_{N-1}^\top \tilde{\Sigma}_N \mathbf{S}_{N-1} \\ &= T_N \left(1 - \sum_{i=1}^{N-1} T_i \right) \left(s_N + \frac{\sum_{i=1}^{N-1} T_i s_i}{1 - \sum_{i=1}^{N-1} T_i} \right)^2 + \frac{1 - \sum_{i=1}^{N-1} T_i}{1 - \sum_{i=1}^{N-1} T_i} \mathbf{S}_{N-1}^\top \tilde{\Sigma}_N \mathbf{S}_{N-1} > 0. \end{aligned}$$

Therefore, Σ_N is positive definite. $\square$

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- 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. (2013a). Consideration on singularities in learning theory and the learning coefficient. *Entropy*, 15(9), 3714–3733.

- Aoyagi, M. (2013b). Learning coefficient in bayesian estimation of restricted boltzmann machine. *Journal of Algebraic Statistics*, 4(1), 30–57.
- Aoyagi, M. (2019a). Learning coefficient of vandermonde matrix-type singularities in model selection. *Entropy (Information Theory, Probability and Statistics)*, 21(6-561), 1–12.
- Aoyagi, M. (2019b). Learning coefficients and information criteria. *Frontiers in Artificial Intelligence and Applications*, 320, 351–362.
- Aoyagi, M. (2024). Consideration on the learning efficiency of multiple-layered neural networks with linear units. *Neural Networks*, 172, 106132.
- Aoyagi, M., & Watanabe, S. (2005). Stochastic complexities of reduced rank regression in bayesian estimation. *Neural Networks*, 18, 924–933.
- Drton, M., & Plummer, M. (2017). A bayesian information criterion for singular models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(2), 323–380.
- Drton, M., Lin, S., Weihs, L., & Zwiernik, P. (2017). Marginal likelihood and model selection for gaussian latent tree and forest models. *Bernoulli*, 23(2), 1202–1232.
- Hironaka, H. (1964). Resolution of singularities of an algebraic variety over a field of characteristic zero. *Annals of Mathematics*, 79, 109–326.
- Imai, T. (2019). Estimating real log canonical thresholds. ArXiv preprint [arXiv:1906.01341](https://arxiv.org/abs/1906.01341).
- Kashiwara, M. (1976). B-functions and holonomic systems. *Inventions Mathematics*, 38, 33–53.

- Mustata, M. (2002). Singularities of pairs via jet schemes. *Journal of the American Mathematical Society*, 15, 599–615.
- Rusakov, D., & Geiger, D. (2002). Asymptotic model selection for naive bayesian networks. Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (pp. 438-445).
- Rusakov, D., & Geiger, D. (2005). Asymptotic model selection for naive bayesian networks. *Journal of Machine Learning Research*, 6, 1–35.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.
- Takeuchi, K. (1976). Distribution of an information statistic and the criterion for the optimal model. *Mathematical Science*, 153, 12–18.
- Watanabe, S. (2001a). Algebraic analysis for nonidentifiable learning machines. *Neural Computation*, 13(4), 899–933.
- Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, 14(8), 1049–1060.
- Watanabe, S. (2001c). Algebraic geometry of learning machines with singularities and their prior distributions. *Journal of Japanese Society of Artificial Intelligence*, 16(2), 308–315.
- Watanabe, S. (2009). *Algebraic geometry and statistical learning theory* (vol. 25). Cambridge University Press, New York, USA.
- Watanabe, S. (2010). Equations of states in singular statistical estimation. *Neural Networks*, 23(1), 20–34.

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

Watanabe, S. (2018). *Mathematical theory of Bayesian statistics*. CRC Press.