# On the Error Probability of Model Selection for Classification

Joe Suzuki*

## Abstract

We consider model selection based on information criteria for classification. The information criterion is expressed in the form of the empirical entropy plus a compensation term $(k(g)/2)d(n)$, where $k(g)$ is the number of independent parameters in a model $g$, $d(n)$ is a function of $n$, and $n$ is the number of examples. First of all, we derive for arbitrary $d(\cdot)$ the asymptotically exact error probabilities in model selection. Although it was known for linear/autoregression processes that $d(n) = \log\log n$ is the minimum function of $n$ such that the model selection satisfies strong consistency, the problem whether the same thing holds for classification has been open. We solve this problem affirmatively. Additionally, we derive for arbitrary $d(\cdot)$ the expected Kullback-leibler divergence between a true conditional probability and the conditional probability estimated by the model selection and the Laplace estimators. The derived value is $k(g^*)/(2n)$, where $g^*$ is a true model, and the accumulated value over $n$ time instances is $(k(g^*)/2)\log n + O(1)$, which implies the optimality of a predictive coding based on the model selection.
Keywords: model selection, error probability, strong consistency, Kullback-Leibler divergence, minimum description length principle, Hannan/Quinn's procedure, unseparated/separated models, Kolmogorov's law of the iterated logarithm..

## 1 Introduction

We estimate a conditional probability $P(y|x)$ of each class $y \in Y$ given each attribute $x \in X$ from training examples, where $X$ and $Y$ are respectively infinite and finite sets. The estimated conditional probability is used for classification in which a class $y$ is supposed to be guessed from a given attribute $x$. We first assume that

**Assumption 1** *Training examples $z^n = z_1 z_2 \cdots z_n$ of pairs $z_i = (x_i, y_i)$, $x_i \in X$, $y_i \in Y$, $i = 1, 2, \cdots, n$, are emitted independently according to a conditional probability $\prod_{i=1}^n P(y_i|x_i)$, where $P(y|x)$, $x \in X$, $y \in Y$, is expressed by a model and stochastic parameters.*

The models that we deal with are defined as functions $g$ that map each attribute $x \in X$ into an element (state) $s = g(x)$ in a finite set (state set) $S(g)$. We secondly assume that

*Information Systems Laboratory, Stanford University, 135 Durand, Stanford, CA 94305-4055 (E-mail: jsuzuki@isl.stanford.edu)

**Assumption 2** *The model that expresses a true conditional probability is unknown but known to be included in a finite set $G$.*

By the parameters, we mean probabilities $p[y, s, g]$ of $y \in Y$ given $s \in S(g)$ when a model $g$ is assumed. Once a model $g$ is selected by a suitable procedure, all we have to do is to estimate those parameters. We thirdly assume

**Assumption 3** *We select the model that minimizes an information criterion.*

The information criterion is defined as the quantity

$$L(g, z^n) = H(g, z^n) + \frac{k(g)}{2} d(n) \ , \tag{1}$$

where

$$H(g, z^n) = \sum_{s \in S(g)} \sum_{y \in Y} c_n[y, s, g] \log \frac{c_n[s, g]}{c_n[y, s, g]} \ , \tag{2}$$

$$k(g) = (|Y| - 1)|S(g)| \ , \tag{3}$$

$$c_n[y, s, g] = \sum_{t=1}^{n} I[g(x_t) = s, y_t = y] \ , \tag{4}$$

$$c_n[s, g] = \sum_{t=1}^{n} I[g(x_t) = s] \ , \tag{5}$$

$I$ is an indicator. i.e., $I[E] = 1$ if an event $E$ is true and $I[E] = 0$ otherwise, and $d(n)$ is a function of $n$. For the cases of $d(n) = 2$ and $d(n) = \log n$, the information criteria are respectively said to be Akaike's information criterion (AIC) [1] and the minimum description length (MDL) principle (Rissanen [10]). We finally assume

**Assumption 4** *For any $s \in S(g)$ and $g \in G$, $c_n[s, g] \to \infty$ as $n \to \infty$,*

which implies from the strong law of large numbers

1. for any $s \in S(g)$ and $g \in G$

$$c_n[s, g]/n \to \pi[s, g] \ \ a.s., \tag{6}$$

   where $\pi[s, g]$ is the stationary probability of a state $s$ when a model $g$ is assumed; and

2. for any $y \in Y$, $s \in S(g)$, and $g \in G$

$$c_n[y, s, g]/c_n[s, g] \to p[y, s, g] \ \ a.s. \tag{7}$$

In this paper, we derive (in Section 2) the asymptotic exact error probability in model selection for arbitrary function $d(\cdot)$ which determines the procedure as well as the information criterion in Eq. (1).

The order identification based on information criteria (Finesso [4]) for ergodic Markov models satisfies the above conditions, which is shown as follows. A sequence $y_1 y_2 \cdots y_n$ with $y_t \in Y$, $t = 1, 2, \cdots, n$, is said to be emitted according to a Markov model of order $T$ if the conditional probability $P(y_{t+1}|y_1^t)$ of $y_{t+1} \in Y$ given $y_1^t \in Y^t$ reduces to $P(y_{t+1}|y_{t-T+1} \cdots y_t)$ when $t \geq T$. Then, we can define $|Y|^T$ states each of which is determined by the previous sequence $x_{t+1} = y_{t-T+1} \cdots y_t \in X$. ($X = Y^d$ is rather a finite set in this case.) A Markov model is said to be ergodic if the transition matrix probability $A = (a_{s,s'})$ among the states satisfies $A^k > 0$ for an integer $k$, where $a_{s,s'}$ is a transition probability from one state $s$ to another $s'$, and $B > 0$ denotes that all the elements of a matrix $B$ are positive. If a Markov model is ergodic, Assumption 4 is satisfied because each state is realized infinitely many times. In this case, the identification of an order $T$ corresponds to a model selection. When the true order is $T$, although the states for[1] $x_1 = \lambda, x_2 = y_1, \cdots, x_d = y_1 \cdots y_{d-1}$ is not generally decided, we assume knowledge of the sequence $y_{-\infty}^0 \in Y^\infty$. Since we mainly focus on the cases with $n$ large, the effect of $y_{-\infty}^0$ is neglectable.

Several results are known on model selection, in particular for linear/autoregressive processes in which the definitions of $H(g, z^n)$ and $k(g)$ are different whereas the information criteria are given in the form of Eq. (1). (Shibata 76 [12]) pointed out that, for linear/autoregressive processes, AIC asymptotically provides an efficient estimator albeit it does not satisfy even weak consistency of model selection (the property that a selected model asymptotically coincides with the true model in probability). For linear/autoregression processes, the conditions for weak consistency are $\liminf_n d(n) = \infty$ and $\liminf_n d(n)/n = 0$ [12]. On the other hand, (Hannan and Quinn [5]) derived for linear/autoregressive processes the slowest function $d(\cdot)$ that satisfies strong consistency (the property that a selected model almost surely coincides with the true model) from the law of the iterated logarithm (LIL), which suggests that for linear/autoregression processes the conditions for strong consistency are $\liminf_n d(n)/(2 \log\log n) > 1$ and $\liminf_n d(n)/n = 0$. We focus on classification rather than linear/autoregressive processes.

Also, some properties of model selection which is not based on Assumption 3 have been reported for the order identification of Markov models. Let $T^*$ be the true order of a Markov model. Merhav, Gutman, and Ziv [8]) attempted to minimize the error probability for models of order $T < T^*$ while keeping the error probability exponent for models of order $T > T^*$ at a given prescribed level $\alpha > 0$. However, if $\alpha$ is large, even weak consistency is not satisfied. If Eq. (1) is applied, for the first error probability to be $\Omega(e^{-\alpha n})$, in our case, $d(n)$ should be $\Omega(n)$, which, as seen in Section 2, leads to inconsistency of model selection. In this sense, the first error probability being $\Omega(e^{-\alpha n})$ seems to be too demanding to make the second error probability desirable. Merhav [9] extended the result into the case of independent identically distributed (IID) exponential family of distributions which includes linear/autoregression processes. Furthermore, Ziv and Merhav [15] extended the result into the case of finite state models and also dealt with the identification of the number of states for hidden Markov models.

On the other hand, Liu and Narayan [7] proposed an improvement of Merhav, Gutman, and Ziv's scheme [8] in the sense that strong consistency is satisfied while the first error probability is $O(n^{-3})$. Also, Finesso [4] and Liu and Narayan [7] dealt with the identification of the number of states for hidden Markov models based on the maximum redundancy (Csiszar [3]) for hidden Markov models.

Although the extension to hidden Markov models is not considered, we find that, if $\liminf_n d(n) = \infty$ and $\liminf_n d(n)/n = 0$, the first error probability could be upperbounded by an arbitrary function of $n$ which does not diminish exponentially in $n$, and that the second error probability diminishes exponentially in $n$.

The climax of this paper (in Section 3) is in the derivation of the slowest function $d(\cdot)$ that satisfies strong consistency for classification, i.e., the classification counterpart of Hannan and Quinn's information criterion [5] which proved for linear/autoregression processes. The problem is whether $d(n) = 2 \log\log n$ makes model selection strongly consistent for classification as well as for linear/autoregression processes. We solve this problem affirmatively.

For the same problem, Finesso [4] derived a similar result: the model $g$ that minimizes

$$L(g, z^n) = H(g, z^n) + h(g) \log\log n \ ,$$

for each $n$ almost surely converges to a true model $g^*$, where the function $h(g)$, $g \in G$, satisfies

$$h(g') - h(g) \geq D k(g')$$

for all $g, g' \in G$ such that $h(g') > h(g) \geq 0$, and $D$ is a constant. However, the compensation term $h(g) \log\log n$ is larger than $k(g) \log\log n$ whereas the orders of both compensation terms are $O(\log\log n)$. Worse, the constant $D$ depends on $k(g^*)$, i.e., the number of independent parameters for the true model $g^*$. Instead, Finesso [4] also showed the model selection that minimizes

$$\bar{L}(g, z^n) = H(g, z^n) + \bar{h}(g) \log n \ ,$$

where $\bar{h}(g)$ is any strictly increasing function of $k(g)$, satisfies strong consistency of model selection. In this sense, the property has been proved for the MDL principle $d(n) = \log n$.

Similar result has been done by Kieffer [6] in which the minimization of

$$\tilde{L}(g, z^n) + \tilde{h}(g) \log n$$

[1] $\lambda$ denotes an empty sequence.

is considered, where $\tilde{L}(z^n, g)$ is the length of the maximum likelihood code (Shtarkov [13])

$$\tilde{L}(g, z^n) = H(g, z^n) + \frac{k(g^*)}{2} \log n + O(1) ,$$

and $\tilde{h}(g)$ is a strictly increasing function of $k(g)$. Kieffer [6] proved strong consistency of the model selection for Markov and hidden Markov models. The model selection does not provide us any suggestion to the problem since the compensation terms are $O(\log n)$.

Finally, we derive (in Section 4) an upperbound of the expected Kullback-Leibler divergence (KLD) between a true conditional probability and the conditional probabilities obtained by the model selection based on Eq. (1) and the Laplace estimators which are defined in Section 4. The analysis takes into account model selection as well as parameter estimation, which is much harder to prove but more useful than just considering parameter estimation. Assuming $\liminf_n d(n) = \infty$ and $\liminf_n d(n)/n = 0$, the expected KLD is $k(g^*)/(2n) + o(1/n)$. Because the expected KLD is $k(g^*)/(2n) + o(1/n)$ even if the model $g^*$ is known, it can be said that the amplification of estimation loss by considering model selection as well as parameter estimation can be neglectable.

On the other hand, the sum of $k(g^*)/(2t) + o(1/t)$ over $t = 1, 2, \cdots, n$ is $(k(g^*)/2) \log n + O(1)$. Since the expected redundancy is asymptotically at least $(k(g^*)/2) \log n$ (Rissanen 86), the length of the following coding procedure with estimation (if $d(t) = \log t$, the length reduces to Rissanen's predictive MDL [11].) is optimal up to $O(1)$: for each $t = 1, 2, \cdots, n$

1. a model $g$ is selected such that $L(g, z^t)$ is minimized;
2. parameters are estimated by a Laplace estimator; and
3. a suitable coding procedure such as arithmetic coding is applied to the estimated model and parameters

## 2 The Error Probabilities

The selected model does not always coincide with a true model. Two kinds of errors in model selection should be considered: errors of selecting models unseparated from a true model; and errors of selecting models separated from a true model, where (Atkinson [2]) for models $g_1, g_2$

1. $g_1$ is unseparated from $g_2$ if any conditional probability based on $g_2$ is expressed by a conditional probability based on $g_1$ by setting parameters of $g_1$ to some values; and
2. otherwise, $g_1$ is separated from $g_2$.

**Example 1** *Suppose* $Y = \{0,1\}$, $X = [0,1]$, *and* $G = \{g_1, g_2\}$, *where*

1. $g_1$: $y_t \in Y$ *is independent from* $x_t \in X$.
2. $g_2$: $y_t \in Y$ *depends on* $s = g_2(x_t)$, $x_t \in X$, *where* $g_2(x) = 0$ *if* $0 \le x < 1/2$ *and* $= 1$ *otherwise, and* $S(g_2) = \{0,1\}$.

If $g_1$ expresses a true model, $g_2$ is unseparated from $g_1$. Similarly, if $g_2$ expresses a true model, $g_1$ is separated from $g_2$.

**Example 2** *Suppose* $Y = \{0,1\}$, $X = \{0,1\}$, *and* $G = \{g_1, g_2\}$, *where*

1. $g_1$: $y_t \in Y$ *is independent from* $x_t = y_{t-1} \in X$.
2. $g_2$: $y_t \in Y$ *depends on* $s = g_2(x_t)$, $x_t = y_{t-1} \in X$, $y_0 = 0$, $g_2(x) = x$, $x \in X$, *and* $S(g_2) = \{0,1\}$.

If $g_1$ expresses a true model, $g_2$ is unseparated from $g_1$. Similarly, if $g_2$ expresses a true model, $g_1$ is separated from $g_2$.

In the following, the sets of unseparated and separated models from a true model $g^*$ are denoted as $UNS(g^*)$ and $SEP(g^*)$, respectively. If $g \in UNS(g^*)$, for each $x_t \in X$, $t = 1, 2, \cdots, n$, $p[y, g(x_t), g] = p[y, g^*(x_t), g^*]$ is required. Let $S(s^*, g)$, $s^* \in S(g^*)$, $g \in UNS(g^*)$, be the set of $s \in S(g)$ that satisfies

$$p[y, s, g] = p[y, s^*, g^*] . \tag{8}$$

Then, the occurrences $c_n[y, s, g]$, $y \in Y$, $s \in S(g)$, should meet the following constraint $C$:

$$\sum_{s \in S(s^*, g)} c_n[y, s, g] = c_n[y, s^*, g^*] \tag{9}$$

for each $y \in Y$, which also implies

$$\sum_{s \in S(s^*, g)} c_n[s, g] = c_n[s^*, g^*] . \tag{10}$$

In the following, we derive $P\{z^n : L(g, z^n) < L(g^*, z^n)\}$ both for $g \in UNS(g^*)$ and for $g \in SEP(g^*)$.

**Theorem 1** *For*[2] $g \in UNS(g^*)$

$$\begin{aligned} & P\{z^n : L(g, z^n) < L(g^*, z^n)\} \\ \simeq\ & 1 - \frac{\Gamma_{\frac{k(g)-k(g^*)}{2} d(n)}(\frac{k(g)-k(g^*)}{2})}{\Gamma(\frac{k(g)-k(g^*)}{2})} \end{aligned} \tag{11}$$

*where* $\Gamma_\cdot(\cdot)$ *is the incomplete Gamma function*

$$\Gamma_x(\alpha) = \int_0^x t^{\alpha-1} e^{-t} dt$$

*and* $\Gamma(\cdot) = \Gamma_\infty(\cdot)$ *is the Gamma function.*

Proof of Theorem 1: It suffices to derive[3] for large $n$

$$2[H(g^*, z^n) - H(g, z^n)] \sim \chi^2_{k(g)-k(g^*)} ,$$

where $\chi^2_l$ denotes the $\chi^2$ distribution with $l$ degrees of freedom, because

[2] $a_n \simeq b_n$ denotes $\lim_{n\to\infty} a_n/b_n = 1$.
[3] $X \sim d$ denotes that r. v. $X$ is according to distribution $d$.

1. the LHS of Eq. (11) can be written as

$$Pr\{z^n : H(g^*, z^n) - H(g, z^n) > \frac{k(g) - k(g^*)}{2} d(n)\} \ ;$$

and

2. the RHS of Eq. (11) is obtained substituting $l = k(g) - k(g^*)$ and $x = [k(g) - k(g^*)]d(n)$ into

$$\int_x^\infty f_l(t)dt = 1 - \frac{\Gamma_{x/2}(l/2)}{\Gamma(l/2)}$$
$$= \frac{(x/2)^{l/2-1}}{e^{x/2}\Gamma(l/2)} + \frac{(x/2)^{l/2-2}}{e^{x/2}\Gamma(l/2-1)} + \cdots , \quad (12)$$

where

$$f_l(x) = \begin{cases} \frac{1}{2^{l/2}\Gamma(l/2)} x^{(l/2)-1} e^{-x/2} & (x > 0) \\ 0 & (x \le 0) \end{cases}$$

is the probability density function of $\chi_l^2$ distribution.

Note

$$\begin{aligned} & H(g^*, z^n) - H(g, z^n) \\ = & \sum_{s \in S(g^*)} \sum_{y \in Y} c_n[y, s, g^*] \log \frac{c_n[s, g^*]}{c_n[y, s, g^*]} \\ & - \sum_{s \in S(g)} \sum_{y \in Y} c_n[y, s, g] \log \frac{c_n[s, g]}{c_n[y, s, g]} \\ = & - \sum_{s \in S(g^*)} \sum_{y \in Y} c_n[y, s, g^*] \log p[y, s, g^*] \\ & + \sum_{s \in S(g)} \sum_{y \in Y} c_n[y, s, g] \log p[y, s, g] \\ & + \frac{1}{2} \sum_{s \in S(g^*)} \chi^2(s, g, z^n) - \sum_{s \in S(g^*)} h(s, g^*, z^n) \\ & + \sum_{s \in S(g)} h(s, g, z^n) \end{aligned} \quad (13)$$

where the first and second terms in the RHS of Eq. (13) are cancelled out from Eqs. (8) and (9), $\chi^2(s^*, g, z^n)$ and $h(s, g, z^n)$ are respectively

$$\begin{aligned} \chi^2(s^*, g, z^n) = & \sum_{s \in S(s^*, g)} \sum_{y \in Y} \frac{(c_n[y, s, g] - c_n[s, g]p[y, s, g])^2}{c_n[s, y, g]p[y, s, g]} \\ & - \sum_{y \in Y} \frac{(c_n[y, s^*, g^*] - c_n[s^*, g^*]p[y, s^*, g^*])^2}{c_n[s^*, g^*]p[y, s^*, g]} \end{aligned}$$

and

$$\begin{aligned} & h(s, g, z^n) \\ = & \sum_{y \in Y} \frac{c_n[s, g]p[y, s, g](\frac{c_n[y, s, g]}{c_n[s, g]p[y, s, g]} - 1)^3}{6[1 + \delta(\frac{c_n[y, s, g]}{c_n[s, g]p[y, s, g]} - 1) \cdot (\frac{c_n[y, s, g]}{c_n[s, g]p[y, s, g]} - 1)]^2} , \end{aligned} \quad (14)$$

and $0 < \delta(x) < 1$ is the function of $x$ $(x > -1)$ that satisfies

$$(1 + x)\log(1 + x) = x + \frac{x^2}{2} - \frac{x^3}{6[1 + \delta(x)x]} .$$

Since for the last two terms

**Lemma 1** *$h(s, g, z^n) \to 0$ a.s. for any $s \in S(g)$ and $g \in G$*

(See Appendix A for the proof.), we obtain

$$H(g^*, z^n) - H(g, z^n) - \frac{1}{2} \sum_{s \in S(g^*)} \chi^2(s, g, z^n) \to 0 \ \ a.s. \quad (15)$$

Therefore, it suffices to show

$$\chi^2(s^*, g, z^n) \sim \chi^2_{(|S(s^*, g)|-1)(|Y|-1)} .$$

Then, since $\chi^2_{(|S(s^*, g)|-1)(|Y|-1)}$, $s^* \in S(g^*)$, are independent each other, $2[H(g, z^n) - H(g^*, z^n)]$ has the $\chi^2$ distribution with degree

$$\begin{aligned} & \sum_{s^* \in S(g^*)} (|S(s^*, g)| - 1)(|Y| - 1) \\ = & \sum_{s^* \in S(g^*)} |S(s^*, g)|(|Y| - 1) \\ & - \sum_{s^* \in S(g^*)} (|Y| - 1) = k(g) - k(g^*) . \end{aligned}$$

In the following, denoting $\alpha = |S(s^*, g)|$ and $\beta = |Y|$, we derive for arbitrary positive constants $a, b > 0$ and under constraint $C$

$$\lim_{n \to \infty} P\{z^n : a < \chi^2(s^*, g, z^n) < b \mid C\} = \int_a^b f_{(\alpha-1)(\beta-1)}(y)dy . \quad (16)$$

The proof consists of four steps. In the proof, since under $g \in UNS(g^*)$ Eq. (8) holds, we use $p[y, s, g]$, $s \in S(s^*, g)$, and $p[y, s^*, g^*]$ interchangeably.

Step 1. Let

$$v[y, s, g] = (c_n[y, s, g] - c_n[s, g]p[y, s, g])/\sqrt{c_n[s, g]p[y, s, g]} \quad (17)$$

for $y \in Y$ and $s \in S(s^*, g)$, thus

$$\chi^2(s^*, g, z^n) = \sum_{s \in S(s^*, g)} \sum_{y \in Y} v[y, s, g]^2 - \sum_{y \in Y} v[y, s^*, g^*]^2. \quad (18)$$

We show

$$\chi^2(s^*, g, z^n) = trace(V^T V - V^T R V) \quad (19)$$

where $V$ is a point in the coordinates of dimension $\alpha\beta$ whose elements are expressed by

$$V = \begin{bmatrix} v[y^{(1)}, s^{(1)}, g] & v[y^{(2)}, s^{(1)}, g] & \cdots & v[y^{(\beta)}, s^{(1)}, g] \\ v[y^{(1)}, s^{(2)}, g] & v[y^{(2)}, s^{(2)}, g] & \cdots & v[y^{(\beta)}, s^{(2)}, g] \\ \cdots & \cdots & \cdots & \cdots \\ v[y^{(1)}, s^{(\alpha)}, g] & v[y^{(2)}, s^{(\alpha)}, g] & \cdots & v[y^{(\beta)}, s^{(\alpha)}, g] \end{bmatrix} ,$$

$S(s^*, g) = \{s^{(1)}, s^{(2)}, \cdots, s^{(\alpha)}\}$, $Y = \{y^{(1)}, y^{(2)}, \cdots, y^{(\beta)}\}$, and $R$ is defined as

$$R = \left(\frac{\sqrt{c_n[s^{(i)}, g] c_n[s^{(j)}, g]}}{c_n[s^*, g^*]}\right)$$

$i, j = 1, 2, \cdots, \alpha$. From constraint $C$, we obtain

$$v[y, s^*, g^*] = \sum_{s \in S(s^*, g)} \sqrt{\frac{c_n[s, g]}{c_n[s^*, g^*]}} v[y, s, g] \ . \tag{20}$$

Combining Eqs. (17) and (20) with Eq. (18), we obtain the claim.

Step 2. We show

$$\begin{aligned} &\lim_{n \to \infty} P\{z^n : a < \chi^2(s^*, g, z^n) < b \mid C\} \\ &= A_1 \int_{D_1} e^{-\frac{trace(V^T V - V^T R V)}{2}} dV \ , \end{aligned} \tag{21}$$

where $A_1$ is a constant, and $D_1$ is the range of $V$ that satisfies $a < trace(V^T V - V^T R V) < b$ and constraint $C$.

To this end, we apply the following lemma

**Lemma 2** *The distribution of $\chi^2(s^*, g, z^n)$ under constraint $C$ is*

$$\begin{aligned} &P\{z^n : a < \chi^2(s^*, g, z^n) < b \mid C\} \\ &\simeq A_2 \sum_{a < \chi^2(s^*, g, z^n) < b} e^{-\frac{\chi^2(s^*, g, z^n)}{2}} c_n[s^*, g^*]^{-\frac{(\alpha-1)(\beta-1)}{2}} \end{aligned} \tag{22}$$

*where $A_2$ is a constant.*

(See Appendix B for the proof.) We consider a cubic for $V$ whose vertexes are

$$\begin{bmatrix} v[y^{(1)}, s^{(1)}, g] & & v[y^{(2)}, s^{(1)}, g] \\ v[y^{(1)}, s^{(2)}, g] & v[y^{(2)}, s^{(2)}, g] \pm \frac{1}{2\sqrt{c_n[s^{(2)}, g] p[y^{(2)}, s^{(2)}, g]}} & \\ \cdots & & \cdots \\ v[y^{(1)}, s^{(\alpha)}, g] & v[y^{(2)}, s^{(\alpha)}, g] \pm \frac{1}{2\sqrt{c_n[s^{(\alpha)}, g] p[y^{(2)}, s^{(\alpha)}, g]}} & \\ \cdots & & v[y^{(\beta)}, s^{(1)}, g] \\ \cdots & v[y^{(\beta)}, s^{(2)}, g] \pm \frac{1}{2\sqrt{c_n[s^{(2)}, g] p[y^{(\beta)}, s^{(2)}, g]}} & \\ \cdots & & \cdots \\ \cdots & v[y^{(\beta)} s^{(\alpha)}, g] \pm \frac{1}{2\sqrt{c_n[s^{(\alpha)}, g] p[y^{(\beta)}, s^{(\alpha)}, g]}} & \end{bmatrix} .$$

Note the cubics for each $V$ are exclusive each other and cover all the region $[0, 1]^{\alpha\beta}$. Then, the volume of each cubic $\sigma(V)$ is

$$\prod_{s \in S(g)} \{c_n[s, g] p[y, s, g]\}^{-\frac{\beta-1}{2}} \simeq A_3 c_n[s^*, g^*]^{-\frac{(\alpha-1)(\beta-1)}{2}}$$

because, from Eq. (6), $c_n[s, g] / c_n[s^*, g^*]$ converges to a constant $(\pi[s, g] / \pi[s^*, g^*])$ a.s., where $A_3$ is a constant. The volume cannot be zero for finite $n$ because $V$ ranges over an $(\alpha-1)(\beta-1)$ dimensional space. Since the symbol $\lim_{n \to \infty} \sum$ can be replaced by $\int dV$, and from Eq. (19), we obtain Eq. (21).

Step 3. We show that Eq. (21) also can be written as

$$\begin{aligned} &\lim_{n \to \infty} P\{z^n : a < \chi^2(s^*, g, z^n) < b \mid C\} \\ &\simeq A_4 \int_{D_2} e^{-\frac{trace(Q^T Q)}{2}} dQ \ , \end{aligned} \tag{23}$$

where $Q$ is an $(\alpha - 1) \times \beta$ matrix, $A_4$ is a constant, and $D_2$ is the range of $Q$ that satisfies $a < trace(Q^T Q) < b$ and constraint $C$. Let $U$ is an $\alpha \times \alpha$ matrix $(u[0], u[1], \cdots, u[\alpha - 1])$, $u[i] = (u[s^{(1)}, i], u[s^{(2)}, i], \cdots, u[s^{(\alpha)}, i])^T$, $i = 0, 1, \cdots, \alpha - 1$, such that[4]

$$U^T U = U U^T = E \tag{24}$$

and $U^T (E - R) U$ is diagonalized where $E$ is an $\alpha \times \alpha$ identity matrix. (Such a matrix $U$ always exists since $E - R$ is symmetric and so positive definite.) Indeed, the matrix $E - R$ has two kinds of eigenvalues:

1. zero with multiplicity 1 (the eigenvector is $u[0]$); and
2. one with multiplicity $\alpha - 1$ (the eigenvectors are $u[1]$ through $u[\alpha - 1]$),

where

$$u[0] = \left(\sqrt{\frac{c_n[s^{(1)}, g]}{c_n[s^*, g^*]}}, \sqrt{\frac{c_n[s^{(2)}, g]}{c_n[s^*, g^*]}}, \cdots, \sqrt{\frac{c_n[s^{(\alpha)}, g]}{c_n[s^*, g^*]}}\right) . \tag{25}$$

Thus,

$$U^T (E - R) U = \tilde{E} \ , \tag{26}$$

where $\tilde{E} = (\tilde{\delta}_{kl})$,

$$\tilde{\delta}_{kl} = \begin{cases} 1 & (\text{if } 1 \le k = l \le \alpha - 1) \\ 0 & (\text{otherwise}) \ . \end{cases} \tag{27}$$

Let

$$\bar{Q} = U^T V \ . \tag{28}$$

From Eqs. (20) and (25), the first row of $\bar{Q}$ is

$$q[0, y] = v[y, s^*, g^*] \ , \ y \in Y \ . \tag{29}$$

We put the remaining rows as $Q = (q[i, y])$, $i = 1, 2, \cdots, \alpha - 1$, $y \in Y$. Then,

$$\bar{Q}^T \tilde{E} \bar{Q} = Q^T Q \tag{30}$$

since from Eqs. (27) and (29) for any $y, y' \in Y$

$$\begin{aligned} &\sum_{k=0}^{\alpha-1} \sum_{l=0}^{\alpha-1} q[k, y] \tilde{\delta}_{kl} q[l, y'] = \sum_{k=1}^{\alpha-1} \sum_{l=1}^{\alpha-1} q[k, y] \tilde{\delta}_{k,l} q[l, y'] \\ &= \sum_{k=1}^{\alpha-1} q[k, y] q[k, y'] \ . \end{aligned}$$

From Eqs. (24), (26), (28), and (30), we obtain

$$V^T V - V^T R V = Q^T Q \ . \tag{31}$$

[4] $T$ denotes a transpose.

Additionally, the Jacobian used in the transformation from $V$ to $Q$ is a constant because any element in $U$ is a constant. Therefore, $dV/dQ$ is a constant.

Step 4. We show Eq. (16). If we put $u = trace(Q^T Q)$, the volume element $dQ$ is expressed as $A_5 u^{\frac{(\alpha-1)(\beta-1)}{2}-1} du$ since any volume can be expressed in the form of $A_6 u^{\frac{(\alpha-1)(\beta-1)}{2}}$, where $A_5$ and $A_6$ are constants. Then,

$$\lim_{n\to\infty} P\{z^n : a < \chi^2(s^*, g, z^n) < b \mid C\} = A_7 \int_a^b e^{-u/2} u^{\frac{(\alpha-1)(\beta-1)}{2}-1} du, \quad (32)$$

where $A_7$ is a constant. $A_7$ is computed as $A_7 = [2^{\frac{(\alpha-1)(\beta-1)}{2}} \Gamma(\frac{(\alpha-1)(\beta-1)}{2})]^{-1}$ since

$$1 = A_7 \int_0^\infty e^{-\frac{u}{2}} u^{\frac{(\alpha-1)(\beta-1)}{2}} du = A_7 2^{\frac{(\alpha-1)(\beta-1)}{2}} \Gamma(\frac{(\alpha-1)(\beta-1)}{2})$$

by $a \to 0$ and $b \to \infty$ in Eq. (32).

Q.E.D

**Theorem 2** *For $g \in SEP(g^*)$,*

$$P\{z^n : L(g, z^n) < L(g^*, z^n)\} \simeq \Phi(\frac{n\mu + \frac{k(g)-k(g^*)}{2} d(n)}{\sqrt{n}\sigma}) \quad (33)$$

*where $\mu$ is defined by*

$$\mu = \sum_{s\in S(g)} \sum_{y\in Y} -\pi[s,g] p[y,s,g] \log p[y,s,g] - \sum_{s\in S(g^*)} \sum_{y\in Y} -\pi[s,g^*] p[y,s,g^*] \log p[y,s,g^*] \quad (34)$$

*$\sigma^2$ is a constant irrespective of $n$, and*

$$\Phi(u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \simeq \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \{\frac{1}{u} - \frac{1}{u^3} + \frac{1\cdot 3}{u^5} - \frac{1\cdot 3\cdot 5}{u^7} - \cdots (-1)^k \frac{1\cdot 3\cdots(2k-1)}{u^{2k+1}}\} . \quad (35)$$

Proof of Theorem 2: Abbreviated.

In Theorem 1, we usually consider the case of $d(n) \to \infty$ as $n \to 0$. Otherwise, like AIC $d(n) = 2$, the information criterion would make the error probability a positive constant value. On the other hand, in Theorem 2, we usually consider the case of $d(n) = o(n)$. Otherwise, the error probability might approach to one as $n \to \infty$. In the remainder of this paper, we assume

**Assumption 5**

$$\liminf_n d(n) = \infty \quad (36)$$

*and*

$$\liminf_n d(n)/n = 0 , \quad (37)$$

Note Assumption 5 assures weak consistency of model selection for classification as well as for linear/autoregression processes.

**Corollary 1** *For $g \in UNS(g^*)$,*

$$P\{z^n : L(g, z^n) < L(g^*, z^n)\} \simeq \frac{\{\frac{k(g)-k(g^*)}{2} d(n)\}^{\frac{k(g)-k(g^*)}{2}-1}}{\exp\{\frac{k(g)-k(g^*)}{2} d(n)\} \Gamma(\frac{k(g)-k(g^*)}{2})} . \quad (38)$$

Proof of Corollary 1: Abbreviated.

**Corollary 2** *For $g \in SEP(g^*)$,*

$$P\{z^n : L(g, z^n) < L(g^*, z^n)\} \simeq \frac{1}{\sqrt{2\pi n}\mu/\sigma} \exp[-\frac{\mu^2}{2\sigma^2} n] . \quad (39)$$

Proof of Corollary 2: Abbreviated.

# 3 Strong Consistency of Model Selection

Theorem 3 suggests that $d(n) = 2\xi \log\log n$ ($\xi > 1$) is the slowest function of $n$ for classification as well as for linear/autoregressive processes.

**Theorem 3** *For $d(n) = 2\xi \log\log n$ and $g \neq g^*$, where $\xi > 1$, $L(g^*, z^n) < L(g, z^n)$ a.s.*

Proof of Theorem 3: We show for $g \in UNS(g^*)$

$$H(g^*, z^n) - H(g, z^n) < (k(g) - k(g^*)) \log\log n \ \ a.s., \quad (40)$$

which implies

$$\begin{aligned} & L(g, z^n) - L(g^*, z^n) \\ = \ & H(g, z^n) - H(g^*, z^n) + \xi(k(g) - k(g^*)) \log\log n \\ > \ & (\xi - 1)(k(g) - k(g^*)) \log\log n \ \ a.s. \end{aligned}$$

for $g \in UNS(g^*)$. For $g \in SEP(g^*)$, $L(g, z^n) > L(g^*, z^n)$ a.s. Indeed, as seen in Theorem 2, since the error probabilities for each $n$ are $o(1/n)$ and summable, from Borel-Cantelli's lemma, strong consistency holds for $g \in SEP(g^*)$.

Note that it suffices to show that, for each element of matrix $Q$ which was defined in the proof of Theorem 1,

$$\limsup_{n\to\infty} \frac{q[i,y]^2}{2\log\log n} = 1 - p[y, s, g^*] \ \ a.s., \quad (41)$$

where $s^* \in S(g^*)$. Indeed, then, from

$$\chi^2(s^*, g, z^n) = trace(Q^T Q) = \sum_{i=1}^{\alpha-1} \sum_{y\in Y} q[i,y]^2 ,$$

we have

$$\limsup_{n\to\infty}\frac{\chi^2(s^*,g,z^n)}{2\log\log n}=\limsup_{n\to\infty}\frac{\sum_{i=1}^{\alpha-1}\sum_{y\in Y}q[i,y]^2}{2\log\log n}$$
$$=\sum_{i=1}^{\alpha-1}(1-p[y,s,g])=(|S(s^*,g)|-1)(|Y|-1)\ .$$

Also since

$$\sum_{s^*\in S(g^*)}(|S(s^*,g)|-1)(|Y|-1)$$
$$=(|S(g)|-|S(g^*)|)(|Y|-1)=k(g)-k(g^*)\ ,$$

and from Eq. (15), the equation

$$\limsup_{n\to\infty}\frac{H(g^*,z^n)-H(g,z^n)}{2\log\log n}=k(g)-k(g^*)\ \ a.s. \tag{42}$$

follows.

The proof is based on Kolmogorov's law of the iterated logarithm (See Stout [14]):

**Lemma 3** *Let $\{K_n, n\geq 1\}$ be positive constants such that $K_n\to 0$ as $n\to\infty$. Suppose $s_n^2=\sum_{t=1}^n E[Z_t^2]\to\infty$ as $n\to\infty$ and*

$$|Z_n|\leq\frac{K_n s_n}{\sqrt{\log\log s_n^2}}\ \ a.s. \tag{43}$$

*for each $n$. Then,*

$$\limsup_{n\to\infty}\frac{|S_n|}{\sqrt{2s_n^2\log\log s_n^2}}=1\ \ a.s., \tag{44}$$

*where $S_n=\sum_{t=1}^n Z_t$.*

We put $Z_n$ as

$$Z_n=\sum_{s\in S(g^*)}u[s,i]\frac{I[g(x_n)=s](I[y_n=y]-p[y,s^*,g^*])}{r[s,g]p[y,s^*,g^*](1-p[y,s^*,g^*])}\ , \tag{45}$$

where $u[s,i]$ is the $s$-th element of vector $u[i]$ which was defined in the proof of Theorem 1. Then we have

$$\begin{aligned}S_n&=\sum_{t=1}^n Z_t\\&=\sum_{s\in S(s^*,g)}u[s,i]\{\sum_{t=1}^n I[g(x_t)=s,y_t=y]\\&\quad-\sum_{t=1}^n I[g(x_t)=s]p[y,s^*,g^*]\}\\&\quad\cdot\{r[s,g]p[y,s^*,g^*](1-p[y,s^*,g^*])\}^{-1/2}\\&=\sum_{s\in S(s^*,g)}u[s,i]\frac{c_n[y,s,g]-c_n[s,g]p[y,s^*,g^*]}{\sqrt{r[s,g]p[y,s^*,g^*](1-p[y,s^*,g^*])}}\ ,\end{aligned}$$

where Eqs. (4) and (5) have been applied. From the definition of matrix $V$,

$$S_n=\sum_{s\in S(s^*,g)}u[s,i]\frac{v[y,s,g]\sqrt{c_n[s,g]p[y,s^*,g^*]}}{\sqrt{r[s,g]p[y,s^*,g^*](1-p[y,s^*,g^*])}}\ .$$

From Eq. (7) and the definition of matrix $Q$,

$$\begin{aligned}S_n&\simeq\sqrt{\frac{n}{1-p[y,s^*,g^*]}}\sum_{s\in S(s^*,g)}u[s,i]v[y,s,g]\ \ a.s.\\&=\sqrt{\frac{n}{1-p[y,s^*,g^*]}}q[i,y]\ .\end{aligned} \tag{46}$$

On the other hand, we can prove $s_n^2=\sum_{t=1}^n EZ_t^2\simeq n$ a.s. In fact,

$$\begin{aligned}&s_n^2\\&=\sum_{t=1}^n EZ_t^2\\&=\sum_{t=1}^n E\{\sum_{s\in S(s^*,g)}u[s,i]\frac{I[g(x_t)=s](I[y_t=y]-p[y,s^*,g^*])}{\sqrt{r[s,g]p[y,s,g](1-p[y,s^*,g^*])}}\}^2\end{aligned}$$

Since $\{Z_n, n\geq 1\}$ is independent,

$$\begin{aligned}&s_n^2\\&=n\sum_{s\in S(s^*,g)}\frac{u[s,i]^2E\{I[g(x_t)=s](I[y_t=y]-p[y,s,g])\}^2}{r[s,g]p[y,s^*,g^*](1-p[y,s^*,g^*])}\ .\end{aligned}$$

From the facts

$$EI[g(x_t)=s]^2=EI[g(x_t)=s]=r[s,g]$$

and

$$E(I[y_t=y]-p[y,s,g])^2=p[y,s,g](1-p[y,s,g])\ ,$$

we obtain

$$\begin{aligned}&s_n^2\\&=n\sum_{s\in S(s^*,g)}\frac{u[s,i]^2EI[g(x_t)=s]E(I[y_t=y]-p[y,s,g])^2}{r[s,g]p[y,s^*,g^*](1-p[y,s^*,g^*])}\\&=n\sum_{s\in S(s^*,g)}u[s,i]^2\end{aligned}$$

Also from Eq. (24),

$$s_n^2=n \tag{47}$$

holds.

Now we apply Eqs. (45), (46), and (47) to Lemma 2. If we put $K_n=n^{-1/3}$, the conditions for Lemma 2 are satisfied:

1. $K_n\to 0$ as $n\to\infty$;
2. $s_n^2=n\to\infty$ as $n\to\infty$; and
3. $|Z_n|<\infty$ and
$$K_n s_n/\sqrt{2\log\log s_n^2}=n^{1/6}/\sqrt{2\log\log n}\to\infty$$
as $n\to\infty$.

hold. Therefore, from Eqs. (45), (46), and (47), we obtain

$$\limsup_{n\to\infty}\frac{\sqrt{n/(1-p[y,s^*,g^*])}q[i,y]}{\sqrt{2n\log\log n}}=1\ \ a.s.$$

Q.E.D

## 4 The Kullback-Leibler Divergence

In this section, we derive the expected Kullback-Leibler divergence (KLD) when parameters are obtained by the Laplace estimators with respect to a selected model.

Let $\theta[p, g^*]$ be the true conditional probability whose model and parameters are respectively $g^*$ and $p[y, s, g^*]$, $y \in Y$, $s \in S(g^*)$. Also let $\theta[\hat{p}_n, g]$ be a conditional probability whose model is $g$ and whose parameters are given by a Laplace estimator

$$\hat{p}_n[y, s, g] = \frac{c_n[y, s, g] + a}{c_n[s, g] + \beta a} , \tag{48}$$

where $a > 0$ is a constant, and $\beta = |Y|$. The KLD between $\theta[p, g^*]$ and $\theta[\hat{p}_n, g]$ is defined by

$$\begin{aligned} & D(\theta[p, g^*] || \theta[\hat{p}_n, g]) \\ = & \sum_{s \in S(g^*)} \sum_{y \in Y} \pi[s, g^*] p[y, s, g^*] \log p[y, s, g^*] \\ & - \sum_{s \in S(g)} \sum_{y \in Y} \pi[s, g] p[y, s, g] \log \frac{c_n[s, g] + \beta a}{c_n[y, s, g] + a} . \end{aligned} \tag{49}$$

Then, the expected KLD between $\theta[p, g^*]$ and $\theta[\hat{p}_n, \hat{g}_n]$, where $\hat{g}_n$ is the model that minimizes $L[\cdot, z^n]$ among $G$, is upper-bounded by the following:

**Theorem 4**

$$E[D(\theta[p, g^*] || \theta[\hat{p}_n, \hat{g}_n])] \leq \frac{k(g^*)}{2n} + o(1/n) . \tag{50}$$

Proof of Theorem 4 (Sketch): Note $E[D(\theta[p, g^*] || \theta[\hat{p}_n, \hat{g}_n])]$ is upperbounded by

$$\begin{aligned} & E[D(\theta[p, g^*] || \theta[\hat{p}_n, \hat{g}_n])] \\ \leq & \sum_{g \in G} \int_0^\infty x P\{z^n : D(\theta[p, g^*] || \theta[\hat{p}_n, g]) = x, \\ & L(g) < L(g^*)\} dx . \end{aligned} \tag{51}$$

Theorem 4 follows from the equation

$$\begin{aligned} & \int_0^\infty x P\{z^n : D(\theta[p, g^*] || \theta[\hat{p}_n, g]) = x, L(g) \leq L(g^*)\} dx \\ \leq & \begin{cases} \dfrac{\{\frac{k(g) - k(g^*)}{2} d(n)\}^{k(g)/2}}{n \exp\{\frac{k(g) - k(g^*)}{2} d(n)\} \Gamma(\frac{k(g)}{2}\}} & (g \in UNS(g^*))] \\ \eta \exp[-\dfrac{\mu^2}{2\sigma^2} n] & (g \in SEP(g^*)) \\ \frac{k(g^*)}{2n} & (g = g^*) \end{cases} \end{aligned} \tag{52}$$

for large $n$. The proof of Eq. (52) is done for each cases of $g \in UNS(g^*)$, $g \in SEP(g^*)$, and $g = g^*$. (The detail is abbreviated.)

## References


[1] H. Akaike, "A new look at the statistical model identification", *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 716-723. 1974

[2] A. C. Atkinson, "A Method for discriminating between Models", *J. Roy. Statist., Soc. Ser.*, Vol. B32, pp. 323-353, 1970.

[3] I. Csiszar, "Information theoretical methods in statistics", class notes, Univ. Maryland, College Park, Spring 1990.

[4] L. Finesso, "Consistent estimation of the order for Markov and hidden Markov chains", *Ph. D thesis, University of Maryland*, Maryland, 1990.

[5] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression", *J. Roy. Statist. Soc.*, Ser. B. 41, pp. 190-195. 1979.

[6] J. C. Kieffer, "Strongly Consistent Code-based Identification and Order Estimation for Constrained Finite-State Model Classes", *IEEE Trans. on Information Theory*, Vol. 39, pp. 893-902, 1993.

[7] "Order estimation and sequential universal data compression of a hidden Markov source by the method of mixtures", *IEEE Trans. on Information Theory*, Vol. IT-40, pp. 1167-1180, 1994.

[8] "On the estimation of the order of a Markov chain and universal data compression", *IEEE Trans. on Information Theory*, Vol. IT-35, pp. 1014-1019, 1989.

[9] N. Merhav, "The estimation of model order in exponential families", *IEEE Trans. on Information Theory*, Vol. IT-35, pp. 1109-1113, 1989.

[10] J. Rissanen, "Modeling by shortest data description", *Automatica*, vol. 14, pp. 465-471. 1978;

[11] J. Rissanen, "Stochastic complexity and modeling", *Ann. Statist*, vol. 14, pp. 1080-1100. 1986;

[12] R. Shibata, "Selection of the order of autoregressive model by Akaike's information criterion", *Biometrika*, vol. 63, pp. 117-126. 1976.

[13] Y. M. Shtarkov, "Universal sequential coding of single messages", *Probl. Inform. Transm.*, Vol. 16, pp. 175-186, 1987.

[14] W. F. Stout, *Almost Sure Convergence*, Academic Press, 1974.

[15] J. Ziv and N. Merhav, "Estimating the number of states of a finite state source", *IEEE Trans. on Information Theory*, Vol. IT-38, pp. 61-65, 1992.