

The Hannan-Quinn Proposition for Linear Regression

Joe Suzuki¹

¹ Department of Mathematics, Osaka University, Osaka, Japan

Correspondence: Joe Suzuki, Department of Mathematics, Osaka University, 1-1 Machikaneyama-cho, Toyonaka, Osaka 560-0043, Japan. Tel: 81-6-6850-5315. E-mail: suzuki@math.sci.osaka-u.ac.jp

Received: September 5, 2012 Accepted: September 26, 2012 Online Published: October 17, 2012

doi:10.5539/ijsp.v1n2p179 URL: <http://dx.doi.org/10.5539/ijsp.v1n2p179>

Abstract

We consider the variable selection problem in linear regression. Suppose that we have a set of random variables $X_1, \dots, X_m, Y, \epsilon$ ($m \geq 1$) such that $Y = \sum_{k \in \pi} \alpha_k X_k + \epsilon$ with $\pi \subseteq \{1, \dots, m\}$ and reals $\{\alpha_k\}_{k=1}^m$, assuming that ϵ is independent of any linear combination of $X_1, \dots, X_m$. Given n examples $\{(x_{i,1}, \dots, x_{i,m}, y_i)\}_{i=1}^n$ actually independently emitted from $(X_1, \dots, X_m, Y)$, we wish to estimate the true π based on information criteria in the form of $H + (k/2)d_n$, where H is the likelihood with respect to π multiplied by -1 , and $\{d_n\}$ is a positive real sequence. If d_n is too small, we cannot obtain consistency because of overestimation. For autoregression, Hannan-Quinn proved that the rate $d_n = 2 \log \log n$ is the minimum satisfying strong consistency. This paper solves the statement affirmative for linear regression. Thus far, there was no proof for the proposition while $d_n = c \log \log n$ for some $c > 0$ was shown to be sufficient.

Keywords: Hannan-Quinn, linear regression, the law of iterated logarithms, strong consistency, information criteria, model selection

1. Introduction

We consider model selection based on information criteria such as AIC (Akaike's information criterion) and BIC (Bayesian information criterion).

For example, given independently and identically distributed (i.i.d.) random variables $\{\epsilon_i\}_{i=-\infty}^{\infty}$ and nonnegative reals $\{\alpha_i\}_{i=1}^k$ ($k \geq 0$), we can define random variables $\{X_i\}_{i=-\infty}^{\infty}$ such that

$$X_i = \sum_{j=1}^k \alpha_j X_{i,j} + \epsilon_i$$

(autoregression). Suppose that we wish to know the minimum true k as well as the values of $\{\alpha_i\}_{i=1}^k$ from a number of examples $\{x_i\}_{i=1}^n$ ($n \geq 1$) emitted from $\{X_i\}_{i=1}^n$. Then, one way to estimate the order k is to prepare a positive real sequence $\{d_n\}_{n=1}^{\infty}$ and to choose k minimizing the information criterion

$$n \log S_k + \frac{k}{2} d_n$$

with respect to d_n , where S_k is the estimated variance based on the Yule-Walker algorithm.

The sequence $\{d_n\}_{n=1}^{\infty}$ balances fitness of the examples to the model and simplicity of the model. If d_n is too small and too large, the estimated model will be overestimated and underestimated, respectively. The information criteria are said AIC and BIC if $d_n = 2$ and $d_n = \log n$, respectively.

In this paper, we consider consistency of model selection: the estimation is weakly and strongly consistent if the true model is obtained as $n \rightarrow \infty$ in probability and almost surely, respectively. For autoregression, Hannan and Quinn (1979) proved strong consistency for $d_n = (2 + \epsilon) \log \log n$ with arbitrary $\epsilon > 0$ based on the law of iterated logarithms. They also showed the converse: $d_n = (2 - \epsilon) \log \log n$ does not satisfy the property.

For linear regression, we can draw a similar scenario: given random variables $\{X_i\}_{i=1}^m$ and ϵ that is independent of any linear combination of $\{X_i\}_{i=1}^m$, we can define

$$Y = \sum_{j=1}^k \alpha_j X_j + \epsilon,$$

where $0 \leq k \leq m$ and $\{\alpha_j\}_{j=1}^k$ are reals. We wish to know the minimum true k as well as the values of $\{\alpha_j\}_{j=1}^k$ from n examples $\{[y_i, x_{i1}, \dots, x_{im}]\}_{i=1}^n$ independently emitted from $(Y, X_1, \dots, X_m)$. Similarly, we can define information criteria

$$n \log S_k + \frac{k}{2} d_n$$

such as $d_n = 2$ (AIC) and $d_n = \log n$ (BIC), where S_k is the empirical square error of the n examples.

However, currently, we do not know whether $d_n = (2 + \epsilon) \log \log n$ with arbitrary $\epsilon > 0$ achieves strong consistency for linear regression. In fact, no proof was given for the proposition. Wu and Zen (1999) suggested that $d_n = c \log \log n$ with some $c > 0$ realizes strong consistency. However, they did not obtain either the exact value of c or any converse result.

On the other hand, for the problem of classification rules which has many applications such as Markov order estimation, data mining, and pattern recognition, Suzuki (2006) proved the Hannan-Quinn proposition.

The main purpose of this paper is to prove the Hannan-Quinn proposition for linear regression. We do not assume that the noise ϵ to be normal in the final result.

Section 2 gives preliminary for linear regression such as idempotent matrices and eigenspaces. In Section 3, we derive the asymptotic error probability of model selection in linear regression when information criteria are applied, which will be an important step to prove the main result. In Section 4, we give a proof of the Hannan-Quinn proposition for linear regression. Section 5 summarizes the results in this paper and gives a future problem.

Throughout the paper, we denote by $X(\Omega)$ the image $\{X(\omega) | \omega \in \Omega\}$ of a random variable $X : \Omega \rightarrow \mathbb{R}$, where Ω is the underlying sample space.

2. Linear Regression

Let $X_1, \dots, X_m$ be random variables, $\epsilon \sim \mathcal{N}(0, \sigma^2)$ a normal random variable with expectation zero and variance $\sigma^2 > 0$, and

$$Y := \sum_{j=1}^p \alpha_j X_j + \epsilon,$$

where $\alpha := [\alpha_1, \dots, \alpha_p]^T \in \mathbb{R}^p$ ($0 \leq p \leq m$). We assume that ϵ is independent of any linear combination of $X_1, \dots, X_m$.

Suppose we do not know the values of order p and coefficients α , and that we are given independently emitted n examples

$$z^n := \{[y_i, x_{i1}, \dots, x_{im}]\}_{i=1}^n$$

with

$$y_i \in Y(\Omega), [x_{i1}, \dots, x_{im}] \in X_1(\Omega) \times \dots \times X_m(\Omega),$$

where $\{[x_{1j}, \dots, x_{nj}]\}_{j=1}^m$ are to be linearly independent. If we define

$$X_p := \begin{bmatrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{bmatrix}, \mathbf{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\epsilon} := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

we can write $\mathbf{y} = X_p \alpha + \boldsymbol{\epsilon}$. Suppose that we estimate p by q ($0 \leq q \leq m$). If we wish to minimize the quantity $\sum_{i=1}^n (y_i - \sum_{j=1}^q \hat{\alpha}_{jq} x_{ij})^2$ given the n examples, then $\hat{\alpha}_q = [\hat{\alpha}_{1,q}, \dots, \hat{\alpha}_{q,q}]^T := (X_q^T X_q)^{-1} X_q^T \mathbf{y}$ is the exact solution (minimum square error estimation), where

$$X_q := \begin{bmatrix} x_{1,1} & \dots & x_{1,q} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,q} \end{bmatrix}$$

Suppose $p \leq q$. If we define $P_q := X_q (X_q^T X_q)^{-1} X_q^T$, we have

$$P_q^2 = P_q$$

and

$$(I - P_q)^2 = I - P_q,$$

so that the square error is expressed by

$$S_q := \sum_{i=1}^n (y_i - \sum_{j=1}^q \hat{\alpha}_{j,q} x_{i,j})^2 = \|\mathbf{y} - \mathbf{X}_q \hat{\boldsymbol{\alpha}}_q\|^2 = \|(I - P_q)\mathbf{y}\|^2 = \mathbf{y}^T (I - P_q) \mathbf{y}.$$

Similarly, if $q = p$, for $P_p := \mathbf{X}_p (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T$ and $\hat{\boldsymbol{\alpha}}_p = [\hat{\alpha}_{1,p}, \dots, \hat{\alpha}_{p,p}]^T := (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T \mathbf{y}$, the square error is expressed by

$$S_p = \mathbf{y}^T (I - P_p) \mathbf{y}.$$

Thus, the difference between the square errors is

$$S_p - S_q = \mathbf{y}^T (I - P_q) \mathbf{y} - \mathbf{y}^T (I - P_p) \mathbf{y} = \mathbf{y}^T (P_q - P_p) \mathbf{y}.$$

On the other hand, we have

$$P_q^T = (\mathbf{X}_q^T)^T \{(\mathbf{X}_q^T \mathbf{X}_q)^{-1}\}^T \mathbf{X}_q^T = \mathbf{X}_q \{(\mathbf{X}_q^T \mathbf{X}_q)^{-1}\}^T \mathbf{X}_q^T = P_q$$

and $P_p^T = P_p$. From $P_q \mathbf{X}_p = \mathbf{X}_p$ and $P_p \mathbf{X}_p = \mathbf{X}_p$, we obtain

$$P_q P_p = P_q \mathbf{X}_p (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T = \mathbf{X}_p (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T = P_p$$

and

$$P_p P_q = P_p^T P_q^T = (P_q P_p)^T = P_p^T = P_p.$$

Thus, not just for $P_p, I - P_p$ but also for $P_q - P_p$, the property

$$(P_q - P_p)^2 = P_q^2 - P_q P_p - P_p P_q + P_p^2 = P_q - P_p$$

holds. Such square matrices satisfying the property are called idempotent matrices (Chatterjee-Hadi, 1987).

In general, for idempotent matrix $P \in \mathbb{R}^{n \times n}$, the inner product $(Px, (I - P)x) = 0$ for any $x = Px + (I - P)x \in \mathbb{R}^n$, so that the eigenspaces are

- 1) $V_1 := \{Px | x \in \mathbb{R}^n\}$ with $\dim(V_1) = \text{rank}(P)$, and
- 2) $V_0 := \{(I - P)x | x \in \mathbb{R}^n\}$ with $\dim(V_0) = n - \text{rank}(P)$.

Since the eigenvalues are one and zero, the multiplicity of eigenvalue one is the same as the trace. Notice that for $(\mathbf{X}_q^T \mathbf{X}_q) = [y_{jk}]$ and $(\mathbf{X}_q^T \mathbf{X}_q)^{-1} = [z_{jk}]$,

$$\text{trace}(P_q) = \text{trace}(\mathbf{X}_q (\mathbf{X}_q^T \mathbf{X}_q)^{-1} \mathbf{X}_q^T) = \sum_{i=1}^n \sum_{j=1}^q \sum_{k=1}^q x_{ij} z_{jk} x_{ki} = \sum_{j=1}^q \sum_{k=1}^q y_{kj} z_{jk} = \sum_{k=1}^q 1 = q,$$

and $\text{trace}(P_p) = p$, so that we have the following table.

P	$\text{trace}(P)$	$\dim(V_1)$	$\dim(V_0)$	$\text{rank}(P)$
P_p	p	p	$n - p$	p
$I - P_p$	$n - p$	$n - p$	p	$n - p$
$P_q - P_p$	$q - p$	$q - p$	$n - q + p$	$q - p$

3. Error Probability in Model Selection

3.1 Overestimation

Proposition 1 If $p < q$, $\frac{S_p - S_q}{S_p/n}$ asymptotically obeys the χ^2 distribution with freedom $q - p$.

Proof. Given $\mathbf{X}_p$, we choose an orthogonal matrix $U = [\mathbf{u}_1, \dots, \mathbf{u}_n]$ of $I - P_p$ so that $U_1 = \langle \mathbf{u}_1, \dots, \mathbf{u}_{n-p} \rangle$ and $U_0 = \langle \mathbf{u}_{n-p+1}, \dots, \mathbf{u}_n \rangle$ are the eigenspaces of eigenvalues one and zero, respectively. Notice that

$$(I - P_p)\mathbf{y} = \mathbf{y} - (\mathbf{X}_p \boldsymbol{\alpha} + P_p \boldsymbol{\epsilon}) = \boldsymbol{\epsilon} - P_p \boldsymbol{\epsilon} = (I - P_p)\boldsymbol{\epsilon}. \quad (1)$$

For $j = 1, \dots, n - p$, multiplying $\mathbf{u}_j^T$ in both hands from left, we get a normal random variable

$$z_j := \mathbf{u}_j^T \mathbf{y} = \mathbf{u}_j^T \boldsymbol{\epsilon}.$$

Since the expectation and variance of ϵ_i are zero and σ^2 (independent), and

$$\mathbf{u}_j^T \mathbf{u}_k = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases}$$

we have $E[z_j] = 0$ and

$$E[z_j z_k] = E[\mathbf{u}_j^T \boldsymbol{\epsilon} \cdot \mathbf{u}_k^T \boldsymbol{\epsilon}] = \sigma^2 \mathbf{u}_j^T \mathbf{u}_k = \begin{cases} \sigma^2, & j = k, \\ 0, & j \neq k. \end{cases}$$

Thus, from the strong law of large numbers, with probability one as $n \rightarrow \infty$,

$$\frac{1}{n} S_p = \frac{1}{n} \sum_{j=1}^{n-p} z_j^2 \rightarrow \sigma^2. \quad (2)$$

On the other hand, given X_q , we choose an orthogonal matrix $V = [\mathbf{v}_1, \dots, \mathbf{v}_n]$ of $P_q - P_p$ so that $V_1 = \langle \mathbf{v}_1, \dots, \mathbf{v}_{q-p} \rangle$ and $V_0 = \langle \mathbf{v}_{q-p+1}, \dots, \mathbf{v}_n \rangle$ are the eigenspaces of eigenvalues one and zero, respectively. Notice that from (1), we have

$$(P_q - P_p)\mathbf{y} = P_q(I - P_p)\mathbf{y} = P_q(I - P_p)\boldsymbol{\epsilon} = (P_q - P_p)\boldsymbol{\epsilon}.$$

For $j = 1, \dots, q - p$, multiplying $\mathbf{v}_j$ in both hands from left, we get a normal random variable

$$r_j := \mathbf{v}_j^T \mathbf{y} = \mathbf{v}_j^T \boldsymbol{\epsilon}.$$

Since the expectation and variance of ϵ_i are zero and σ^2 (independent), and

$$\mathbf{v}_j^T \mathbf{v}_k = \begin{cases} 1, & j = k, \\ 0, & j \neq k, \end{cases}$$

we have $E[r_j] = 0$ and

$$E[r_j r_k] = E[\mathbf{v}_j^T \boldsymbol{\epsilon} \cdot \mathbf{v}_k^T \boldsymbol{\epsilon}] = \sigma^2 \mathbf{v}_j^T \mathbf{v}_k = \begin{cases} \sigma^2, & j = k, \\ 0, & j \neq k. \end{cases}$$

Hence, as $n \rightarrow \infty$,

$$\frac{S_p - S_q}{\sigma^2} = \sum_{j=1}^{q-p} \frac{r_j^2}{\sigma^2} \sim \chi_q^2 \quad (3)$$

where the fact that the square sum of $q - p$ independent random variables with the standard normal distribution obeys the χ^2 distribution of freedom $q - p$ has been applied. Equations (2)(3) imply Proposition 1.

In the sequel, for $\pi \subseteq \{1, \dots, m\}$, we write the square error of $\{X_j\}_{j \in \pi}$ and Y by $S(\pi)$, and put

$$L(z^n, \pi) := n \log S(\pi) + \frac{k(\pi)}{2} d_n$$

and $k(\pi) = |\pi|$, given $z^n = \{[y_i, x_{i,1}, \dots, x_{i,m}]\}_{i=1}^n$. Let $\pi_* \subseteq \{1, \dots, m\}$ be the true π .

Theorem 1 For $\pi \supset \pi_*$, the probability of $L(z^n, \pi) < L(z^n, \pi_*)$ is

$$\int_{n\{1 - \exp[-\frac{k(\pi) - k(\pi_*)}{2n} d_n]\}}^{\infty} f_{k(\pi) - k(\pi_*)}(x) dx,$$

where f_l is the probability density function of the χ^2 distribution of freedom l .

Proof. Notice that

$$2\{L(z^n, \pi) - L(z^n, \pi_*)\} = 2n \log \frac{S(\pi)}{S(\pi_*)} + \{k(\pi) - k(\pi_*)\} d_n = 2n \log(1 - \frac{S(\pi_*) - S(\pi)}{S(\pi_*)}) + \{k(\pi) - k(\pi_*)\} d_n,$$

so that

$$L(z^n, \pi) < L(z^n, \pi_*) \iff \frac{S(\pi_*) - S(\pi)}{S(\pi_*)/n} > n\{1 - \exp[-\frac{k(\pi) - k(\pi_*)}{2n}d_n]\}. \quad (4)$$

From Proposition 2, we obtain Theorem 1.

3.2 Underestimation

Hereafter, we do not assume that ϵ to be normally distributed.

Theorem 2 For $\pi \not\subseteq \pi_*$, $L(z^n, \pi) > L(z^n, \pi_*)$ with probability one as $n \rightarrow \infty$.

Proof. Suppose $q < p$. Given X_p , we choose an orthogonal matrix $W := [w_1, \dots, w_n]$ of $P_p - P_q$ so that $W_1 = \langle w_1, \dots, w_{p-q} \rangle$ and $W_0 = \langle w_{p-q+1}, \dots, w_n \rangle$ are the eigenspaces of eigenvalue one and zero, respectively.

If we define $t_i := \sum_{k=q+1}^p x_{ik}\alpha_k + \epsilon_i$ and

$$s_j := \sum_{i=1}^n w_{ij}y_i = \sum_{i=1}^n w_{ij}(\sum_{k=1}^p x_{ik}\alpha_k + \epsilon_i) = \sum_{i=1}^n w_{ij}(\sum_{k=q+1}^p x_{ik}\alpha_k + \epsilon_i) = (\mathbf{w}_j, \mathbf{t})$$

for $\mathbf{w}_j = [w_{1j}, \dots, w_{nj}]^T$ and $\mathbf{t} = [t_1, \dots, t_n]^T$, then $\|\mathbf{w}_j\|^2 = \sum_{i=1}^n w_{ij}^2 = 1$, and $\|\mathbf{t}\|^2/n = \sum_{i=1}^n t_i^2/n$ converges to a positive constant with probability one. Otherwise, $\epsilon = -\sum_{k=q+1}^p X_k\alpha_k$ with probability one (contradiction). If $\mathbf{w}_j$ and $\mathbf{t}$ are orthogonal, the $\mathbf{t}$ should be in the form

$$(\sum_{k=1}^q x_{1k}\beta_k, \dots, \sum_{k=1}^q x_{nk}\beta_k)$$

for some $[\beta_1, \dots, \beta_q] \in \mathbf{R}^q$, which means that $\epsilon = \sum_{k=1}^q X_k\beta_k - \sum_{k=q+1}^p X_k\alpha_k$ with probability one (contradiction). Hence,

$$\frac{1}{n}(S_q - S_p) = \frac{1}{n} \sum_{j=q+1}^p s_j^2 = \sum_{j=q+1}^p \|\mathbf{w}_j\|^2 \|\mathbf{t}\|^2/n \cdot \cos(\mathbf{w}_j, \mathbf{t})^2 \quad (5)$$

converges to a positive value, which implies the theorem when $\pi \subset \pi_*$. Suppose $\pi \not\subset \pi_*$. In the same way, since (5) converges to a positive value even for $q = |\pi \cap \pi_*|$, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \{S(\pi \cap \pi_*) - S(\pi_*)\} > 0. \quad (6)$$

Furthermore, if we replace π_* by $\pi \cap \pi_*$, from a similar discussion as in Theorem 1, we have

$$\lim_{n \rightarrow \infty} \frac{1}{n} \{S(\pi) - S(\pi \cap \pi_*)\} = 0. \quad (7)$$

The statements (6)(7) imply the theorem.

4. Proof of the Hannan-Quinn Proposition

In this section, we do not assume that $\epsilon \sim \mathcal{N}(0, \sigma^2)$ but that ϵ is an independently identically distributed random variable with expectation zero and variance σ^2 .

Proposition 2 If $q > p$, with probability one,

$$1 \leq \limsup_{n \rightarrow \infty} \left\{ \frac{S_p - S_q}{S_p/n} / \log \log n \right\} \leq q - p \quad (8)$$

Proof. The notation is similar to Proposition 2, and let $p+1 \leq j \leq q$.

Let $\lambda_1, \dots, \lambda_{q-p}$ and $[\beta_{1,1}, \dots, \beta_{1,q}]^T, \dots, [\beta_{q-p,1}, \dots, \beta_{q-p,q}]^T$ be the nonzero eigenvalues and corresponding unit eigenvectors of

$$X_{p,q} := \left(\frac{1}{n} X_q^T X_q \right)^{-1} - \begin{pmatrix} (\frac{1}{n} X_p^T X_p)^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Then, from

$$n(S_q - S_p) = n\epsilon^T(P_q - P_p)\epsilon = n(\epsilon^T X_q)X_{p,q}(\epsilon^T X_q)^T = \sum_{j=1}^{q-p} \lambda_j \left(\sum_{k=1}^q \beta_{jk} \sum_{i=1}^n x_{i,k} \epsilon_i \right)^2$$

and $E[r_j^2] = E[\epsilon_i^2] = \sigma^2$, we require $\lambda_j = [\sum_{i=1}^n (\sum_{k=1}^q \beta_{jk} x_{ik})^2]^{-1}$, thus,

$$v_{ij} = \frac{\sum_k \beta_{jk} x_{ik}}{\sqrt{\sum_{i=1}^n (\sum_l \beta_{jl} x_{il})^2}}.$$

Since $X_{p,q}$ converges to a constant matrix as $n \rightarrow \infty$ and $\lim_{n \rightarrow \infty} \beta_{jk}$ exists with probability one, so does

$$\gamma_{jk} := \lim_{n \rightarrow \infty} \frac{\beta_{jk}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (\sum_l \beta_{jl} x_{il})^2}}.$$

Let $Z_i = \sum_k \gamma_{jk} x_{ik} \epsilon_i / \sigma$. Then, $E[\sum_{i=1}^n Z_i] = 0$, $E[\sum_{i=1}^n Z_i^2] = n$, and $\{Z_i\}_{i=1}^n$ are independent. From the law of iterated logarithms (Stout 1974), we have

$$\limsup_{n \rightarrow \infty} \frac{\sum_i \sqrt{n} v_{ij} \epsilon_i / \sigma}{\sqrt{n \log \log n}} = \limsup_{n \rightarrow \infty} \frac{\sum_i (\sum_k \gamma_{jk} x_{ik}) \epsilon_i / \sigma}{\sqrt{n \log \log n}} = \limsup_{n \rightarrow \infty} \frac{\sum_{i=1}^n Z_i}{\sqrt{n \log \log n}} = 1,$$

namely,

$$\limsup_{n \rightarrow \infty} \frac{r_j}{\sigma \sqrt{\log \log n}} = 1$$

with probability one. Since

$$\limsup_{n \rightarrow \infty} \frac{r_j}{\sigma \sqrt{\log \log n}} \leq \limsup_{n \rightarrow \infty} \sum_{j=p+1}^q \frac{r_j}{\sigma \sqrt{\log \log n}} \leq \sum_{j=p+1}^q \limsup_{n \rightarrow \infty} \frac{r_j}{\sigma \sqrt{\log \log n}},$$

and from (2) and (3), we have (8) with probability one.

The following equation is useful in derivation of the final result.

$$\frac{1}{2} \{k(\pi) - k(\pi_*)\} d_n - \frac{1}{4n} [\{k(\pi) - k(\pi_*)\} d_n]^2 \leq n [1 - \exp\{-\frac{k(\pi) - k(\pi_*)}{2n} d_n\}] \leq \frac{1}{2} \{k(\pi) - k(\pi_*)\} d_n \quad (9)$$

Theorem 3 For $d_n := (2 + \epsilon) \log \log n$ ($\epsilon > 0$), $L(z^n, \pi) > L(z^n, \pi_*)$ with probability one.

Proof. From Theorem 2, the error for $\pi_* \not\subset \pi$ is almost surely zero as long as $\frac{d_n}{n} \rightarrow 0$ ($n \rightarrow \infty$), so that we only need to consider the case $\pi_* \subset \pi$. However, $d_n = (2 + \epsilon) \log \log n$ with $\epsilon > 0$ implies the left hand side of (9) is strictly larger than $(q - p) \log \log n$ with $p = k(\pi_*)$ and $q = k(\pi)$ for large n , which from Proposition 2 and (4) implies Theorem 3.

Theorem 4 For $d_n := (2 - \epsilon) \log \log n$ ($\epsilon > 0$), $L(z^n, \pi) \leq L(z^n, \pi_*)$ with nonzero probability as $n \rightarrow \infty$ for π such that $k(\pi) = k(\pi_*) + 1$.

Proof. $d_n = (2 - \epsilon) \log \log n$ with $\epsilon > 0$ implies the right hand side of (9) is strictly smaller than $(q - p) \log \log n$ with $p = k(\pi_*)$ and $q = k(\pi) = p + 1$ for large n , which from Proposition 2 and (4) implies Theorem 4.

For example, suppose $p = 0$ and $q = 1$. Then, $X_{0,1} = \frac{1}{n} \sum_{i=1}^n x_{i1}^2$, $v_{i1} = \frac{x_{i1}}{\sqrt{\frac{1}{n} \sum_{h=1}^n x_{h1}^2}}$, and $\gamma_{11} = E[X_1^2]^{-1/2}$. In this

case, $S_0 - S_1 = \frac{\sum_{i=1}^n x_{i1}^2 \epsilon_i^2}{\sum_{h=1}^n x_{h1}^2}$ and $\frac{S_0}{n} = \frac{1}{n} \sum_{i=1}^n \epsilon_i^2$. Thus, with probability one,

$$\frac{S_0 - S_1}{S_0/n} = \frac{n \sum_{i=1}^n x_{i1}^2 \epsilon_i^2}{\sum_{h=1}^n x_{h1}^2 \sum_{l=1}^n \epsilon_l^2}$$

exceeds $(1 + \epsilon) \log \log n$ finitely many times and $(1 - \epsilon) \log \log n$ with nonzero probability, so that the model selection procedure makes wrong results at most finitely many times.

5. Conclusion

We proved that the Hannan-Quinn proposition is true for linear regression as well as for auto regression (Hannan-Quinn, 1979) and for classification (Suzuki, 2006): the minimum rate of d_n satisfying strong consistency is $(2 + \epsilon) \log \log n$ for arbitrary $\epsilon > 0$.

The future problems contain finding strong consistency conditions that are good for all the cases including linear regression, auto regression, and classification. Making clear why the same $d_n = 2 \log \log n$ is the crucial rate for those problems would be the first step to solve the problem.

References

- Chatterjee, S., & Hadi, A. S. (1988). *Sensitivity Analysis In Linear Regression*. New York: John Wiley & Sons. <http://dx.doi.org/10.1002/jae.3950050108>
- Hannan, E. J., & Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, B*, 41, 190-195. <http://dx.doi.org/10.1086/260800>
- Stout, W. (1974). *Almost Sure Convergence*. New York: Academic Press.
- Suzuki, J. (2006). On Strong Consistency of Model Selection in Classification. *IEEE Transactions on Information Theory*, 52(11), 4767-4774. <http://dx.doi.org/10.1109/TIT.2006.883611>
- Wu, Y., & Zen, M. (1999). A strongly consistent information criterion for linear model selection based on M-estimation. *Probab. Theory Relat. Fields*, 113, 599-625. <http://dx.doi.org/10.1007/s004400050219>