情報論的学習理論テクニカルレポート 2009
Technical Report on Information-Based Induction Sciences 2009 (IBIS2009)

Hannan-Quinnの命題は、線形回帰でも、ガウス型Baysianネットワークの構造推定でも正しい

The Hannan-Quinn proposition is true for learning linear regression and Bayesian network structures

鈴木譲*
Joe Suzuki

Abstract: This paper proves $d_n = 2\log\log n$ is the smallest $\{d_n\}_{n=1}^{\infty}$ such that the information criterion $H + (k/2)d_n$, where H is the empirical entropy of n examples and k is the number of parameters that express the probability distribution, satisfies consistency for the problem of learning linear regression. Thus far, the problem was solved only for the problems of learning ARMA (autoregressive moving average, Hanann-Quinn, 1979) and conditional probabilities (Suzuki, 2006). The new result is good for learning the structure of Gaussian Bayesian networks as well.

Keywords:

1 まえがき

AIC や MDL/BIC といった情報量基準は、統計科学の諸問題でよく適用されている。ここで、例の個数を n とすれば、尤度の最大値にマイナスをつけたもの (経験的エントロピー) にパラメータ数もしくはパラメータ数を加えた値を AIC、それに $\times(1/2)\log n$ を加えた値を MDL/BIC といっている。一般的に、$d_n/n \to 0$ となる正実数列を $\{d_n\}_{n=1}^{\infty}$、モデル g の経験的エントリピーおよびパラメータ数をそれぞれ $H(g), k(g)$ とおいて

$$H(g) + \frac{k(g)}{2} d_n$$

(情報量基準) を最小にするモデルを選択する問題として定式化できる。AIC なら $d_n = 2$、MDL/BIC なら $d_n = \log n$ というふうにである。つまり、情報量基準の $\{d_n\}$ の選び方は無数にある。

モデル選択で、$\{d_n\}$ の選び方によって、$n \to \infty$ で正しいモデルを見出す性質 (一致性) がどのようにかわるかが、よく議論される。

1. 各 n で正しいモデルを選択する確率が 1 に収束する (弱一致性)
2. 正しくないモデルが有限回しか選択されない例の無限列の集合に対して確率測度 1 が割当てられる (強一致性)

MDL/BIC($d_n = \log n$) では両方の性質が満足されるが、AIC($d_n = 2$) では両方とも満足されない。一般に、d_n が小さいと過学習が原因で、強一致性が満足されない。

無論、モデル選択において一致性だけが要求されるすべての性能ではないが、本論文では、強一致性を満足する最小の $\{d_n\}$ とな何かについて議論する。

情報量基準は、適用する問題によって、経験的エントロピーとパラメータ数の定義が異なる。1979 年に Hannan-Quinn は、ARMA(Autoregression Moving Average, 自己回帰移動平均) の場合、「$d_n = 2\log\log n$ が強一致性を満足する最小の $\{d_n\}$ である」(本論文では、Hannan-Quinn の命題とよぶ) ことを証明した。そして、$d_n = 2\log\log n$ とする情報量基準 HQ が、ARMA 以外の問題にも適用されるようになった。しかしながら、Hannan-Quinn のオリジナル論文からわかることであるが、Hannan-Quinn の命題の証明が、ARMA という問題の性質に強く依存しており、線形回帰をはじめとする他の問題で同じ性質が真であるか否かは、証明がされていない。ただ、そういうことに頓着せずに、ARMA 以外の問題に HQ を無造作に使う人は多い。

ただ、ごく最近になって、データマニングやパター

*大阪大学大学院理学研究科, 560-0043 豊中市待兼山, tel. 06-6850-5315, e-mail suzuki@math.sci.osaka-u.ac.jp,
Osaka University, Yoyonaka, Osaka, 560-0043, Japan.

ン認識でよく用いられる条件付確率の学習について、Hannan-Quinn の命題が正しいことが証明された (Suzuki, 2006)。このことは、有限型 Bayesian ネットワークの構造推定においても、Hannan-Quinn の命題が正しいことを意味する。

本論文では、線形回帰においても Hannan-Quinn の命題が正しいことを証明する。この問題の解決は、非常に大きな意義があると思われる。そうでないと、情報量基準を適用する際に、Hannan-Quinn の命題が証明されないと、HQ を適用する意味が見出されないからである。

強一致性は例では検証できないため、線形回帰で Hannan-Quinn の命題が真であるか否か、予想することは容易ではなかった。線形回帰で、HQ 以外では、

$$\frac{d_n}{\log\log n} \to \infty$$

が過去に知られた強一致性を満足する最小の $\{d_n\}$ の十分条件である (Rao-Wu, 1989)。

2 節で、ARMA と条件付確率で Hannan-Quinn の命題がどのように証明されたかの概略を簡潔に述べる。3 節で、まず、線形回帰で情報量基準を適用した場合のモデル選択の漸近的な誤り率の公式を導く (3.1 節)。これは副産物ではあるが、主結果の証明への重要なステップとなる。3.2 節で Hannan-Quinn の命題の証明をおこなう。さらに 4 節で、線形回帰とガウス型 Bayesian ネットワークの構造学習の問題との関係をのべ、後者においても Hannan-Quinn の命題が正しいことを証明する。5 節で、本論文のまとめをおこなう。

2 Hannan-Quinnの命題が証明された例

2.1 ARMA

以下では、$\{W_i\}_{i=-\infty}^{\infty}$ を平均 0 分散 1 の独立同一分布にしたがう確率変数の列、$\{X_i\}_{i=-\infty}^{\infty}$ を平均 0、

$$X_i = \sum_{j=1}^{k} \lambda_j X_{i-j} + W_i$$

および実数 $\{\lambda_i\}_{i=1}^{k}$ で定義される確率変数の列とする (自己回帰移動平均, autoregressive moving average)。ここで、$E[X_i] = 0$ であり、また $\{X_i\}$ は定常であるので、$m \geq 0$ として、$\gamma_m := EX_iX_{i+m}$ は i によらず、また以下の方程式 (Yule-Waker 方程式) が得られる。

$$\gamma_m = \sum_{j=1}^{k} \lambda_j \gamma_{m-j} + \delta_{0m}\sigma^2$$

Cramer の公式を用いて、$\{\gamma_m\}_{m=0}^{k}$ の値から、$(k+1)\times(k+1)$ の連立一次方程式の解として、σ^2 および $\{\lambda_m\}_{m=1}^{k}$ の値が得られる。

一般に、$\{\gamma_m\}_{m=0}^{k}$ の値は、未知であるので、その実現値

$$x = (x_1, \cdots, x_n) \in X_1(\Omega) \times \cdots \times X_n(\Omega)$$

から、

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

および

$$\hat{\gamma}_m := \hat{\gamma}_{-m} := \frac{1}{n}\sum_{i=1}^{n-m}(x_i - \bar{x})(x_{i+m} - \bar{x})$$

を計算して推定する。したがって、Yule-Walker 方程式は、以下のようにかける。

$$\begin{bmatrix} -1 & \hat{\gamma}_1 & \hat{\gamma}_2 & \cdots & \hat{\gamma}_k \\ 0 & \hat{\gamma}_0 & \hat{\gamma}_1 & \cdots & \hat{\gamma}_{k-1} \\ 0 & \hat{\gamma}_1 & \hat{\gamma}_0 & \cdots & \hat{\gamma}_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \hat{\gamma}_{k-1} & \hat{\gamma}_{k-2} & \cdots & \hat{\gamma}_0 \end{bmatrix} \begin{bmatrix} \hat{\sigma}_k^2 \\ \hat{\lambda}_{1,k} \\ \hat{\lambda}_{2,k} \\ \vdots \\ \lambda_{k,k} \end{bmatrix} = \begin{bmatrix} -\hat{\gamma}_0 \\ -\hat{\gamma}_1 \\ -\hat{\gamma}_2 \\ \vdots \\ -\hat{\gamma}_k \end{bmatrix},$$

特に、次数 k が未知であれば、各 k について上記の連立方程式を解き、

$$L(x^n, k) = \frac{n}{2}\log\hat{\sigma}_k^2 + \frac{k}{2}d_n$$

の値を計算して、この値を最小にする k を推定値とする。この k を求める操作を ARMA の学習とよぶ。

一般に、

$$\hat{\sigma}_k^2 = \{1 - \hat{\lambda}_{k,k}^2\}\hat{\sigma}_{k-1}^2$$

が成立する。したがって、$k = 1, 2, \cdots$ に対して、

$$\begin{aligned} & 2\{L(x^n, k) - L(x^n, k-1)\} \\ = \;& n\log\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2} + d_n \\ \leq \;& -n(1 - \frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2}) + d_n \qquad (1) \\ = \;& -n\hat{\lambda}_{k,k}^2 + d_n \end{aligned}$$

なる変形ができる。また、$n \to \infty$ で、$k \leq k^*$ に対しては、$\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2}$ は 1 より小さな値に概収束する。したがって、(1) より、

$$L(x^n, 0) > L(x^n, 1) > \cdots > L(x^n, k^* - 1) > L(x^n, k^*)$$

が確率 1 で成立する。他方、$k \geq k^* + 1$ に対しては、$\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2}$ は 1 に概収束する。Hannan-Quinn(1979) は、重複対数の法則から、

$$\frac{\hat{\lambda}_{k,k}^2}{2n^{-1}\log\log n} \leq 1$$

が確率 1 で成立することを証明し、$d_n = 2c\log\log n$ $(c > 1)$ について

$$L(x^n, k^*) < L(x^n, k^* + 1) < \cdots$$

が確率 1 で成立することを証明した。

2.2 条件付確率

$X, Y : \Omega \to \mathbb{R}$ を確率空間 $(\Omega, \mathcal{F}, \mu)$ における確率変数とし、Y が有限 $(|Y(\Omega)| < \infty)$ であることを仮定する。以下では、σ-集合体 $\mathcal{G}_1, \mathcal{G}_2$ が $A \in \mathcal{G}_1 \Longrightarrow A \in \mathcal{G}_2$ を満足するとき、$\mathcal{G}_1 \subseteq \mathcal{G}_2$ とかくものとする。また、μ は確率測度であり、$\mu(\Omega) = 1$ を満足する測度である。

以下では、事象 $(X = x)$ のもとでの、事象 $(Y = y)$ の条件付確率を $\mu_{Y|X}(y|x)$, $x \in X(\Omega)$, $y \in Y(\Omega)$ で表記する。$X(\Omega)$ における同値関係 $\sim$ を、$x, x' \in X(\Omega)$ について

$$\mu_{Y|X}(y|x) = \mu_{Y|X}(y|x'), y \in Y(\Omega) \Longleftrightarrow x \sim x'$$

で定義する。実際、

1. $x \sim x$
2. $x \sim x' \Longleftrightarrow x' \sim x$
3. $x \sim x', x' \sim x'' \Longrightarrow x \sim x''$.

が成立する。この同値類で生成される事象の集合を $\mathcal{G} \subseteq \mathcal{F}$ とかく。条件付確率の学習では、n 個の独立に発生した例 $z^n = (z_1, \cdots, z_n)$,

$$z_i := (x_i, y_i) \in X(\Omega) \times Y(\Omega)\ ,\ i = 1, 2, \cdots, n$$

から、モデル $\mathcal{G}$ を見出す問題として定義されるものとする。

本稿では、例 $z^n \in X^n(\Omega) \times Y^n(\Omega)$ から

$$L(z^n, \mathcal{G}) := H(z^n, \mathcal{G}) + \frac{k(\mathcal{G})}{2} d_n\ ,$$

(情報量基準) の値を比較して、それを最小にする $\mathcal{G}$ を、学習結果とよぶ。ただし、$H(z^n, \mathcal{G})$ を経験的エントロピー、$k(\mathcal{G})$ をそのパラメータの数とした。

学習の誤りは、真の $\mathcal{G}$ を $\mathcal{G}^*$、推定された $\mathcal{G}$ を $\hat{\mathcal{G}}_n$ とかくと、

1. $\hat{\mathcal{G}}_n = \mathcal{G}^*$
2. $\hat{\mathcal{G}}_n \supset \mathcal{G}^*$ (過学習)
3. $\hat{\mathcal{G}}_n \not\supseteq \mathcal{G}^*$ (未学習)

に大別できる。f_l で自由度 l の χ^2 分布の確率密度関数をあらわすものとする。

命題 1 (Suzuki, 2006) 1. $\mathcal{G} \supset \mathcal{G}^*$ への誤り確率は $\int_{\{k(\mathcal{G})-k(\mathcal{G}_*)\}d_n}^{\infty} f_{k(\mathcal{G})-k(\mathcal{G}_*)}(x)dx$

2. 確率 1 で、$\mathcal{G}$ $(\not\supseteq \mathcal{G}^*)$ への誤り確率は 0
3. $d_n = 2c\log\log n$ $(c > 1)$ で、確率 1 で誤り確率は 0

$X^{(1)}, \cdots, X^{(N)}$ を有限の値をとる確率変数とし、$\pi^{(i)} \subseteq \{1, \cdots, i-1\} (i = 1, \cdots, N)$ について、

1. $X^{(i)}$ と $\{X^{(j)}\}_{j \in \{1,\cdots,i-1\} \setminus \pi^{(i)}}$ が $\{X^{(j)}\}_{j \in \pi^{(i)}}$ のもとで、条件付独立であり、
2. $\pi^{(i)}$ の真部分集合について、条件付独立でない

この条件を満足する $\pi := (\pi^{(1)}, \cdots, \pi^{(N)})$ は一意にきまる (有限型 Bayesian ネットワークの構造)。$|X^{(i)}(\Omega)| = \alpha^{(i)}$ とおくと、確率パラメータの数は、$k(\pi^{(i)}) := (\alpha^{(i)} - 1)\prod_{j \in \pi^{(i)}} \alpha^{(j)}$ となる。学習の誤りは、真の $\pi^{(i)}$ を $\pi_*^{(i)}$、推定された $\pi^{(i)}$ を $\hat{\pi}_n^{(i)}$ とかくと、

1. $\hat{\pi}_n^{(i)} = \pi_*^{(i)}$
2. $\hat{\pi}_n^{(i)} \supset \pi_*^{(i)}$ (過学習)
3. $\hat{\pi}_n^{(i)} \not\supseteq \pi_*^{(i)}$ (未学習)

に大別できる。

系 1 1. $\pi^{(i)}$ $(\supset \pi_*^{(i)})$ への誤り確率は、

$$\int_{\{k(\pi^{(i)})-k(\pi_*^{(i)})\}d_n}^{\infty} f_{k(\pi^{(i)})-k(\pi_*^{(i)})}(x)dx$$

2. 確率 1 で、$\pi^{(i)}$ $(\not\supseteq \pi_*^{(i)})$ への誤り確率は 0
3. $d_n = 2c\log\log n$ $(c > 1)$ で、確率 1 で誤り確率は 0

3　線形回帰の場合の証明

以下では、標本空間を Ω とし、ϵ を正規分布 $\mathcal{N}(0,\sigma^2)$ にしたがう確率変数、$Y, X_1, \cdots, X_p$ を

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

にしたがう確率変数、$X_{p+1}, \cdots, X_m$ を $Y - \sum_{j=1}^{p} \alpha_j X_j$ とは独立な確率変数とする $(0 \leq p \leq m)$。

独立に生成した n 個の例 $\{(y_i, x_{i,1}, \cdots, x_{i,m})\}_{i=1}^{n}$

$$y_i \in Y(\Omega), (x_{i,1}, \cdots, x_{i,m}) \in X_1(\Omega) \times \cdots \times X_m(\Omega)$$

が得られたものとする。$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}$ を未知として、

$\epsilon_i := y_i - \sum_{j=1}^{m} \alpha_j x_{i,j}$

$$\boldsymbol{X}_m := \begin{bmatrix} x_{1,1} & \cdots & x_{1,m} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,m} \end{bmatrix}, \mathbf{y} := \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\epsilon} := \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{bmatrix}$$

とおくと、$\boldsymbol{y} = \boldsymbol{X}_m \boldsymbol{\alpha} + \boldsymbol{\epsilon}$ とかける。これを解くと、$\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_1, \cdots, \hat{\alpha}_m]^T := (\boldsymbol{X}_m^T \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m^T \boldsymbol{y}$ が得られる。他方、$P_m := \boldsymbol{X}_m (\boldsymbol{X}_m^T \boldsymbol{X}_m)^{-1} \boldsymbol{X}_m^T$ とおくと、

$$P_m^2 = P_m$$

$$(I - P_m)^2 = I - P_m$$

がいえるので、このときの 2 乗誤差は

$$\begin{aligned} S_m &:= \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} \hat{\alpha}_j x_{i,j})^2 \\ &= ||\boldsymbol{y} - \boldsymbol{X}_m \hat{\boldsymbol{\alpha}}||^2 \\ &= ||(I - P_m)\boldsymbol{y}||^2 \\ &= \boldsymbol{y}^T (I - P_m) \boldsymbol{y} \end{aligned}$$

とかける。同様に、n 個の例 $\{(y_i, x_{i,1}, \cdots, x_{i,p})\}_{i=1}^{n}$ から出発する場合、

$$\boldsymbol{X}_p := \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}, P_p := \boldsymbol{X}_p (\boldsymbol{X}_p^T \boldsymbol{X}_p)^{-1} \boldsymbol{X}_p^T$$

とおくと、2 乗誤差は

$$S_p := \boldsymbol{y}^T (I - P_p) \boldsymbol{y}$$

となる。

したがって、2 乗誤差の差は

$$S_p - S_m = \boldsymbol{y}^T (I - P_p)\boldsymbol{y} - \boldsymbol{y}^T (I - P_m)\boldsymbol{y} = \boldsymbol{y}^T (P_m - P_p)\boldsymbol{y}$$

となる。ここで、

$$\begin{aligned} P_m^T &= (X_m^T)^T \{(X_m^T X_m)^{-1}\}^T X_m^T \\ &= X_m \{(X_m^T X_m)^T\}^{-1} X_m^T = P_m \end{aligned}$$

同様に $P_p^T = P_p$ が成立する。また、$P_p X_p = X_p$, $P_p X_p = X_p$ より、

$$P_m P_p = P_m X_p (X_p^T X_p)^{-1} X_p^T = X_p (X_p^T X_p)^{-1} X_p^T = P_p$$

$$P_p P_m = P_p^T P_m^T = (P_m P_p)^T = P_p^T = P_p$$

が成立する。したがって、$P_p, I - P_p$ だけではなく、$P_m - P_p$ についても、

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

が成立する。このような性質を満足する正方行列は冪等 (Idempotent) 行列とよばれ、固有値として 0 および 1 しかもたないことが知られている (Chatterjee-Hadi, 1987)。

3.1　モデル選択の誤り確率

命題 2 $\dfrac{S_p - S_m}{S_p / n}$ は、自由度 $q := m - p$ の χ^2 分布に従う

証明: 簡単な計算から、行列 P_p のトレースが p となるので、$I - P_p$, $P_m - P_p$ のトレースはそれぞれ $n-p, q$ となる。また、固有値 1 の重複度もそれぞれ、$n-p$, q となる (固有値 0 の重複度はそれぞれ、p, $n-q$ となる)。

まず、事象 $\boldsymbol{X}_p$ のもとで、$I - P_p$ の直交行列 $U = (u_{i,j})$ を、$\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{n-p}$ の固有値が 1、$\boldsymbol{u}_{n-p+1}, \cdots, \boldsymbol{u}_n$ の固有値が 0 になるように選ぶと、Y の分散が σ^2、$\sum_{i=1}^{n} u_{j,i}^2 = 1$ より、$z_j := \boldsymbol{u}_j^T \boldsymbol{y} = \sum_{i=1}^{n} u_{j,i} y_i$ は、

1. $1 \leq j \leq n-p$ に関して、$\mathcal{N}(0, \sigma^2)$ にしたがう。
2. $n - p + 1 \leq j \leq n$ に関して、値が 0

したがって、大数の強法則より、確率 1 で

$$\frac{1}{n} S_p = \frac{1}{n} \sum_{j=1}^{n-p} z_j^2 \to \sigma^2 \tag{2}$$

次に、事象 $\boldsymbol{X}_m$ のもとで、$P_m - P_p$ の直交行列 $V = (v_{i,j})$ を、$\boldsymbol{v}_1, \cdots, \boldsymbol{v}_q$ の固有値が 1、$\boldsymbol{v}_{q+1}, \cdots, \boldsymbol{v}_n$ の固有値が 0 になるように選ぶと、Y の分散が σ^2、$\sum_{i=1}^{n} v_{j,i}^2 = 1$、$\boldsymbol{v}_j^T \boldsymbol{v}_k = 0 (j \neq k)$ より、$r_j := \boldsymbol{v}_j^T \boldsymbol{y} = \sum_{i=1}^{n} v_{j,i} y_i$ は、

1. $1 \leq j \leq q$ に関して、$\mathcal{N}(0, \sigma^2)$ にしたがう (相互に独立)。

2. $q+1 \leq j \leq n$ に関して、値が 0

したがって、$n \to \infty$ で、

$$\frac{S_p - S_m}{\sigma^2} = \sum_{j=1}^{q} \frac{z_j^2}{\sigma^2} \sim \chi_q^2 \tag{3}$$

ここで、独立な標準正規分布にしたがう q 個の確率変数の 2 乗和が自由度 q の χ^2 分布にしたがうことを用いた。(2)(3) は、命題 2 を意味する。

(証明終)

以下では、$\pi \subseteq \{1, \cdots, N\}$ について、$\{X_j\}_{j \in \pi}, Y$ の 2 乗誤差を $S(\pi)$ であらわし、

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

および $k(\pi) = |\pi|$ とおく。ただし、$\pi_* \subseteq \{1, \cdots, N\}$ を真の π であるとする。また、

定理 1 $\pi \supset \pi_*$ について、$L(z^n, \pi) < L(z^n, \pi_*)$ となる確率は、

$$\int_{\{k(\pi)-k(\pi_*)\}d_n}^{\infty} f_{k(\pi)-k(\pi_*)}(x) dx$$

証明: まず、

$$\begin{aligned} & 2\{L(z^n, \pi) - L(z^n, \pi_*)\} \\ = \; & n \log \frac{S(\pi)}{S(\pi_*)} + \{k(\pi) - k(\pi_*)\} d_n \qquad (4) \\ = \; & n \log(1 - \frac{S(\pi_*) - S(\pi)}{S(\pi_*)}) + \{k(\pi) - k(\pi_*)\} d_n \\ = \; & -\frac{S(\pi_*) - S(\pi)}{S(\pi_*)/n} - R_n + \{k(\pi) - k(\pi_*)\} d_n \end{aligned} \tag{5}$$

の第 2 項

$$R_n := \frac{1}{2n} \frac{(nh)^2}{(\theta h - 1)^2}, h := \frac{S(\pi_*) - S(\pi)}{S(\pi_*)}, 0 < \theta < 1$$

は、$n \to \infty$ で 0 に概収束する。したがって、

$$\begin{aligned} & L(z^n, \pi) < L(z^n, \pi_*) \\ \iff \; & \frac{S(\pi_*) - S(\pi)}{S(\pi_*)/n} > \{k(\pi) - k(\pi_*)\} d_n \end{aligned} \tag{6}$$

命題 2 より、これは定理 1 を意味する。

(証明終)

定理 2 $\pi \not\supseteq \pi_*$ について、確率 1 で $L(z^n, \pi) > L(z^n, \pi_*)$

証明: 確率 1 で $\frac{S(\pi)}{S(\pi_*)} > 0$ となり、また $d_n/n \to 0$ が成立する。(4) より、これは定理 2 を意味する。

(証明終)

3.2 Hannan-Quinn の命題の証明

命題 3

$$\frac{S_p - S_m}{S_p} \leq q \log \log n \tag{7}$$

証明: 記法は命題 2 と同様とし、$p+1 \leq j \leq m$ とする。大数の強法則より、$\sqrt{n} v_{i,j}$ は確率 1 で n によらない一定値 $\delta_{i,j}$ に収束する。$Z_i := \frac{\delta_{i,j} y_i}{\sigma}$ とおくと、$\sqrt{n} \boldsymbol{v}_j^T \boldsymbol{y}$ と

$$\sigma \sum_{i=1}^{n} \frac{\delta_{i,j} y_i}{\sigma} = \sigma \sum_{i=1}^{n} Z_i$$

の比が確率 1 で 1 になる。$E[Z_i] = 0$, $E[Z_i^2] = 1$ であるので、重複対数の法則を適用すると、

$$\frac{\sqrt{n} \boldsymbol{v}_j^T \boldsymbol{y} / \sigma}{\sqrt{n \log \log n}} \leq 1$$

すなわち

$$\frac{\boldsymbol{v}_j^T \boldsymbol{y}}{\sigma} \leq \sqrt{\log \log n}$$

が確率 1 で成立する。これは、

$$\frac{S_p - S_m}{S_p/n} \leq q \log \log n$$

が確率 1 で成立することを意味する。

(証明終)

定理 3 $d_n := 2c \log \log n$ $(c > 1)$ のとき、確率 1 で $L(z^n, \pi) > L(z^n, \pi_*)$

証明: (6) と命題 3 は、定理 3 の成立を意味する。

(証明終)

4 ガウス型 Bayesian ネットワークの構造学習の場合の証明

条件付確率の学習を、有限型 Bayesian ネットワークの構造学習に適用したのと同様に、線形回帰の学習を、ガウス型 Bayesian ネットワークの構造学習に適用することができる。$X^{(1)}, \cdots, X^{(N)}$ を確率変数、$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma_i^2)$ として、

$$X^{(i)} = \sum_{j \in \pi^{(i)}} \alpha^{(j)} X^{(j)} + \epsilon^{(i)}$$

を満足する $\pi^{(i)}$ を見出す問題となる。

定理 4

1. $\pi^{(i)} (\supset \pi_*^{(i)})$ への誤り率は、

$$\int_{\{k(\pi^{(i)})-k(\pi_*^{(i)})\}d_n}^{\infty} f_{k(\pi^{(i)})-k(\pi_*^{(i)})}(x) dx$$

2. 確率 1 で、$\pi^{(i)} (\not\supseteq \pi_*^{(i)})$ への誤り確率は 0

3. $d_n = 2c \log \log n$ $(c > 1)$ で、確率 1 で誤り確率は 0

5　まとめ

線形回帰の学習、およびその帰結として、ガウス型 Bayesian ネットワークの学習に関して、Hannan-Quinn の命題が真であることを証明した。そもそも、ARMA でしか成立していなかったが、条件付確率 (Suzuki, 2006) の場合に引き続いて、本論文によって、線形回帰の場合でも解決された。長年未解決であり、喜んでいただけるものと思われる。

過学習の誤り確率の計算で、条件付確率および線形回帰のいずれにおいても、経験的エントロピーの差がパラメータ数の差を自由度にもつ χ^2 分布にしたがうことがわかった。そして、ともにその性質を利用して、Hannan-Quinn の命題が真となることが証明された。では、条件付確率、線形回帰を含むどのような一般的な条件で Hannan-Quinn の命題が真となるのであろうか。大変興味深い。

参考文献

[1] Akaike, H. (1974): "A New Look at the Statistical Model Identification," I.E.E.E. Transactions on Automatic Control, AC 19, 716-723

[2] Schwarz, G. (1978): "Estimating the Dimension of a Model," Annals of Statistics, 6, 461-464.

[3] Hannan, E. J., and B. G. Quinn (1979): "The Determination of the Order of an Autoregression," Journal of the Royal Statistical Society, B, 41, 190-195.

[4] J. Suzuki (2006): On Strong Consistency of Model Selection in Classification. IEEE Transactions on Information Theory 52(11): 4767-4774

[5] Rao, C.R., Wu, Y., (1989): A strongly consistent procedure for model selection in a regression problem. Biometrika 76, 369-374

[6] Chatterjee, S. and Hadi, A. S. (1988), Sensitivity Analysis In Linear Regression, New York: John Wiley & Sons.

[7] 鈴木譲 (2009): ベイジアンネットワーク入門 (培風館)