鈴木讓「ベイジアンネットワーク入門」(培風館)
3. 統計的学習
3.6 Gauss 型 Bayesian ネットワークの学習

鈴木讓

大阪大学

2010 年 7 月 7 日

あらまし

ARMA

線形回帰

ガウス型 Bayesian ネットワークの構造学習

自己回帰移動平均, AutoRegressive Moving Average

$\{W_i\}_{i=-\infty}^{\infty}$: 平均 0 分散 1 の独立同一分布にしたがう確率変数の列
$\{\lambda_i\}_{i=1}^{k}$: 実数の列
$\{X_i\}_{i=-\infty}^{\infty}$: 平均 0 の確率変数の列

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

Yule-Waker 方程式

$\{X_i\}$: 定常 より、$\gamma_m := E[X_i X_{i+m}]$, $m \geq 0$ が、i によらない

Yule-Waker 方程式

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

$\{\gamma_m\}_{m=0}^{k}$ の値から、$\lambda_0 := \sigma^2$, $\{\lambda_m\}_{m=1}^{k}$ の値を求める

$$\begin{bmatrix} -1 & \gamma_1 & \gamma_2 & \cdots & \gamma_k \\ 0 & \gamma_0 & \gamma_1 & \cdots & \gamma_{k-1} \\ 0 & \gamma_1 & \gamma_0 & \cdots & \gamma_{k-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \gamma_{k-1} & \gamma_{k-2} & \cdots & \gamma_0 \end{bmatrix} \begin{bmatrix} \sigma_k^2 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_k \end{bmatrix} = \begin{bmatrix} -\gamma_0 \\ -\gamma_1 \\ -\gamma_2 \\ \vdots \\ -\gamma_k \end{bmatrix},$$

k: 既知、$\{\gamma_m\}_{m=0}^{k}$: 未知 のとき

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

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

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

を計算して γ_m の推定値とする

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

k: 未知、$\{\gamma_m\}_{m=0}^{k}$: 未知 のとき

1.

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

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

を計算して、この値を最小にする k ($\hat{k}$ とおく) を推定値

2. そのときの連立方程式の解 $\hat{\lambda}_{0,k} := \hat{\sigma}_{\hat{k}}^2$, $\{\hat{\lambda}_{m,\hat{k}}\}_{m=1}^{\hat{k}}$

Hannan-Quinn, 1979

一般に、

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

$k = 1, 2, \cdots$ が成立し、

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

Hannan-Quinn, 1979 (続)

k^*: 真の k

$n \longrightarrow \infty$

$k \leq k^*$ 確率 1 で $\dfrac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2} \to \exists K < 1$

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

$k \geq k^* + 1$ 確率 1 で $\dfrac{\hat{\sigma}_k^2}{\hat{\sigma}_{k-1}^2} \to 1$

重複対数の法則から、$\dfrac{\hat{\lambda}_{k,k}^2}{2n^{-1} \log \log n} \leq 1$ を導出

$$d_n = 2c \log \log n, c > 1 \Longrightarrow L(x^n, k^*) < L(x^n, k^*+1) < \cdots$$

線形回帰

$\epsilon \sim \mathcal{N}(0, \sigma^2)$
$Y, X_1, \cdots, X_p$: $Y = \sum_{j=1}^{p} \alpha_j X_j + \epsilon$ にしたがう確率変数
$X_{p+1}, \cdots, X_m$: $Y - \sum_{j=1}^{p} \alpha_j X_j$ と独立な確率変数 $(0 \leq p \leq m)$

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

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

仮定

$\begin{bmatrix} x_{1,j} \\ \vdots \\ x_{n,j} \end{bmatrix}, j = 1, \cdots, m$: 確率 1 で一次独立

最小 2 乗法

$$\boldsymbol{\alpha} := \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} \text{(未知)}$$

$\mathbf{y} = \mathbf{X}_m \boldsymbol{\alpha} + \epsilon$

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

$$\Longrightarrow \quad \hat{\boldsymbol{\alpha}} = \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_m \end{bmatrix} := (\mathbf{X}_m^T \mathbf{X}_m)^{-1} \mathbf{X}_m^T \mathbf{y}$$

2 乗誤差の差

$P_m := \mathbf{X}_m(\mathbf{X}_m^T\mathbf{X}_m)^{-1}\mathbf{X}_m^T$

$$P_m^2 = P_m$$

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

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

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

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

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

冪等 (Idempotent) 行列

$$P_m^T = (X_m^T)^T\{(X_m^T X_m)^{-1}\}^T X_m^T = X_m\{(X_m^T X_m)^T\}^{-1} X_m^T = P_m$$

$$P_p^T = P_p$$

$P_p X_p = X_p$, $P_m X_p = X_p$ より、

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

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

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

$$\frac{S_p - S_m}{S_p/n} \sim \chi^2_q,\ q := m - p$$

一般に、$P^2 = P \Longrightarrow x = Px + (I-P)x\ ,\ (Px, (I-P)x) = 0$
$V_1 := \{Px | x \in \mathbb{R}^n\}$: 次元 $\mathrm{rank}\ P$
$V_0 := \{(I-P)x | x \in \mathbb{R}^n\}$: 次元 $n - \mathrm{rank}\ P$

- P_p: 固有値 0,1 のみ、rank p
- P_m: 固有値 0,1 のみ、rank m

	rank	V_1 の次元	V_0 の次元
P_p	p	p	$n-p$
$I - P_p$	$n-p$	$n-p$	p
$P_m - P_p$	q	q	$n-q$

$\frac{S_p - S_m}{S_p/n} \sim \chi_q^2$, $q := m - p$ (続)

$I - P_p$ の直交行列 $U = (u_{i,j})$:

- $\mathbf{u}_1, \cdots, \mathbf{u}_{n-p}$ の固有値が 1
- $\mathbf{u}_{n-p+1}, \cdots, \mathbf{u}_n$ の固有値が 0

$V[Y] = \sigma^2$、$\sum_{i=1}^n u_{i,j}^2 = 1$、$\mathbf{u}_j^T \mathbf{u}_k = 0 (j \neq k)$ より、$1 \leq j \leq n - p$ で、

$$z_j := \mathbf{u}_j^T \mathbf{y} = \sum_{i=1}^n u_{i,j} y_i \sim \mathcal{N}(0, \sigma^2)$$

大数の強法則より、確率 1 で

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

$\dfrac{S_p - S_m}{S_p/n} \sim \chi_q^2$, $q := m - p$ (続)

$P_m - P_p$ の直交行列 $V = (v_{i,j})$:

- $\mathbf{v}_1, \cdots, \mathbf{v}_q$ の固有値が 1
- $\mathbf{v}_{q+1}, \cdots, \mathbf{v}_n$ の固有値が 0

$V[Y] = \sigma^2$、$\sum_{i=1}^n v_{i,j}^2 = 1$、$\mathbf{v}_j^T \mathbf{v}_k = 0 (j \neq k)$ より、$1 \leq j \leq q$ で

$$r_j := \mathbf{v}_j^T \mathbf{y} = \sum_{i=1}^n v_{i,j} y_i \sim \mathcal{N}(0, \sigma^2)$$

大数の強法則より、確率 1 で

$$\frac{S_p - S_m}{S_p/n} \to \frac{S_p - S_m}{\sigma^2} = \sum_{j=1}^q \frac{r_j^2}{\sigma^2} \sim \chi_q^2$$

過学習の誤り率

$\pi \subseteq \{1, \cdots, N\}$
$S(\pi)$: $\{X_j\}_{j \in \pi}, Y$ の 2 乗誤差を

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

$k(\pi) := |\pi|$
$\pi_* \subseteq \{1, \cdots, N\}$: 真の π

$L(z^n, \pi) < L(z^n, \pi_*)$ となる確率

$$\pi \supset \pi_* \Longrightarrow \int_{\frac{k(\pi)-k(\pi_*)}{2} d_n}^{\infty} f_{k(\pi)-k(\pi_*)}(x) dx$$

過学習の誤り率 (続)

証明:

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

確率 1 で、$R_n := \frac{1}{n}\frac{(nh)^2}{(\theta h-1)^2} \to 0$

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

$$L(z^n,\pi)<L(z^n,\pi_*) \Longleftrightarrow \frac{S(\pi_*)-S(\pi)}{S(\pi_*)/n} > \frac{k(\pi)-k(\pi_*)}{2}d_n$$

未学習の誤り率

$L(z^n,\pi) < L(z^n,\pi_*)$ となる確率

$\pi \not\supseteq \pi_* \Longrightarrow$ 確率 1 で $L(z^n,\pi) > L(z^n,\pi_*)$

証明: 確率 1 で $\dfrac{S(\pi)}{S(\pi_*)} > 0$ となり、また $d_n/n \to 0$

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

$p+1 \leq j \leq m$ を固定して、$Z_i := \frac{\sqrt{n} v_{i,j} y_i}{\sigma}$

$\sum_{i=1}^{n} Z_i = \frac{\sqrt{n} r_j}{\sigma} \sim \mathcal{N}(0, n)$

$E[\sum_{i=1}^{n} Z_i] = 1,\ E[(\sum_{i=1}^{n} Z_i)^2] = n$

Z_i は i.i.d. なので、$E[Z_i] = 0,\ E[Z_i^2] = 1$、重複対数の法則を適用

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

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

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

が確率 1 で成立

Hannan-Quinn の命題は線形回帰でも成立

Hannan-Quinn の命題

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

Hannan-Quinn の命題は線形回帰でも成立

$X^{(1)}, \cdots, X^{(N)}$: 確率変数
$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma_i^2)$

ガウス型 Bayesian ネットワークの構造学習

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

なる $\pi^{(i)}$ を見出す

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

$$\int_{\frac{k(\pi^{(i)}) - k(\pi_*^{(i)})}{2} d_n}^{\infty} f_{k(\pi^{(i)}) - k(\pi_*^{(i)})}(x) dx$$

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