Technical Report of the 2nd Workshop on Latent Dynamics (Jun 22, 2011, Tokyo, Japan)

不完全データの一部に興味がある場合の情報量規準

原 照雅* 下平英寿†
Hara Terumasa Shimodaira Hidetoshi

Abstract: ある変数やある部分が直接観測できないデータを不完全データと呼ぶ。不完全データは音声認識における隠れマルコフモデル、時系列解析における状態空間モデルなど、様々な分野で登場する。本研究では、観測データと潜在データからなる完全データの一部に興味があり、その特定の部分を重視したモデル選択を試みた。まず赤池情報量規準 (AIC) の拡張としてデータの特定の部分を重視した AICp を導出した。そして、不完全データにおける情報量規準 PDIO(Shimodaira 1994) にも同様の拡張を行い、完全データの一部に興味がある場合に不完全データから予測誤差を計算する情報量規準、PDIOp を導出した。この量は周辺分布の平均対数尤度とフィッシャー情報行列を用いて計算することができる。

Keywords: 情報量規準　モデル選択

1　はじめに

データの予測のための最良のモデルを選ぶ手段として、情報量規準によるモデル選択がある．一般にモデル選択では観測されたデータ（観測データ）へのモデルの当てはまりを評価するが、直接観測できないデータ（潜在データ）への当てはまりも評価したい場合がある．潜在データが欠けたデータを不完全データと呼ぶ．不完全データが発生する枠組みとして、状態空間モデル、混合正規分布、隠れマルコフモデルなどがある．また、心理統計の欠損データ分析への応用も期待される。

モデル選択における情報量規準として赤池情報量規準（AIC）が幅広い分野で用いられているが、AIC を不完全データに適用すると潜在データを上手く評価できないという問題が指摘されている．この問題を解消し、潜在部分も含めた評価を行う情報量規準として、不完全データにおける情報量規準（Predictive Divergence for Indirect Ovservation、PDIO）が Shimodaira(1994,[1]) によって提案された．

しかし、潜在データはしばしばいくつかの部分から成り立つ．例えば、経済時系列においてはトレンド成分、季節成分、定常 AR 成分などの和で潜在データを表現することがある．このとき、潜在データの特定の部分（例えば、トレンド成分のみ）に興味があり、その部分への当てはまりを重視したモデル選択を行いたい場合を考える．このために、まずは AIC を拡張して一部分を重視した情報量規準 AICp を提案する．次に、これを不完全データに適応させるために PDIO を拡張したのが、不完全データの一部に興味がある場合の情報量規準 PDIOp である．

本研究では、これらの情報量規準 AICp,PDIOp を導出し、PDIOp についてシミュレーションを行った．

2　不完全データと情報量規準

不完全データは、以下のように表される．

- 観測可能なデータを観測データ又は不完全データと呼び、**Y** と表記する
- 観測不可能なデータを潜在データと呼び、**Z** と表記する

*東京工業大学 情報理工学研究科 数理・計算科学専攻
〒152-8552　東京都目黒区大岡山 2-12-1-W8-46
e-mail hara.t.aa@m.titech.ac.jp
Dept.of Mathematical and Computing Sciences, Tokyo Institute of Technology
W8-46 2-12-1 Ookayama Meguro-ku Tokyo 152-8552 Japan
†東京工業大学 情報理工学研究科 数理・計算科学専攻
〒152-8552　東京都目黒区大岡山 2-12-1-W8-46
e-mail shimo@is.titech.ac.jp
Dept.of Mathematical and Computing Sciences, Tokyo Institute of Technology
W8-46 2-12-1 Ookayama Meguro-ku Tokyo 152-8552 Japan

- 観測データと潜在データを合わせたものを完全データと呼び、$\mathbf{X}$ と表記する

なお、完全データは $\mathbf{X} = (\mathbf{Y}, \mathbf{Z})$ と分割できると考える．AIC は $\mathbf{Y}$ への当てはまりしか評価できないが、PDIO は $\mathbf{X}$ 全体への当てはまりを評価できる．

本研究では、潜在データ $\mathbf{Z}$ を興味がある部分 $\mathbf{Z}_1$と興味がない部分 $\mathbf{Z}_2$、観測部分 $\mathbf{Y}$ を興味がある部分 $\mathbf{Y}_1$と興味がない部分 $\mathbf{Y}_2$ に分割できるとする．$\mathbf{Z} = \phi$ の場合、$\mathbf{X}_1$ と $\mathbf{Y}_1$ は一致し、$\mathbf{X}_1$ は直接観測できる．このとき、AICp は $\mathbf{X}_1$ への当てはまりを評価する情報量規準として導かれる．$\mathbf{Z} \neq \phi$ の場合、PDIOp は $\mathbf{X}_1 = (\mathbf{Y}_1, \mathbf{Z}_1)$ への当てはまりを評価する情報量規準として導かれる．

AIC,PDIO,AICp,PDIOp の具体的な形は以下の通り．

$$\begin{aligned}
\mathrm{AIC} &= -2l(\hat{\theta}(\mathbf{y})) + 2m \\
\mathrm{PDIO} &= -2l(\hat{\theta}(\mathbf{y})) + 2\mathbf{tr}(\mathbf{Hx}{\mathbf{H_Y}}^{-1}) \\
\mathrm{AICp} &= -2l(\hat{\theta}(\mathbf{y}_1)) + 2\mathbf{tr}(\mathbf{Hx}_1{\mathbf{Hx}}^{-1}) \\
\mathrm{PDIOp} &= -2l(\hat{\theta}(\mathbf{y}_1)) + 2\mathbf{tr}(\mathbf{Hx}_1{\mathbf{H_Y}}^{-1})
\end{aligned}$$

ただし、$l(\hat{\theta}(\mathbf{y}))$ は観測データ $\mathbf{Y}$ での最大対数尤度、$l(\hat{\theta}(\mathbf{y}_1))$ は $\mathbf{Y}_1$ での周辺分布の対数尤度、m は自由パラメタ数、$\mathbf{Hx}, \mathbf{H_Y}, \mathbf{Hx}_1$ はそれぞれ $\mathbf{X}, \mathbf{Y}, \mathbf{X}_1$ でのフィッシャー情報行列である．

それぞれ第 1 項はモデルへの当てはまりを評価する項で、第 2 項はモデルの複雑さに罰則を与える項と見なせる．これらを最小にするモデルを選択することでそれぞれ目的とする最良のモデルが選択されることが期待される．

3　情報量規準 PDIOp の導出

PDIOp の導出を行う．真の分布 q と候補モデル p 間の隔たりを表す量として次の量を用いる．

$$\mathcal{L}(q(\mathbf{x}), p(\mathbf{x})) \equiv -\int q(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

この値が小さいほど p は q に近く、良い近似だとみなせる．興味ある部分の完全データ $\mathbf{X}_1 = (\mathbf{Y}_1, \mathbf{Z}_1)$ の予測分布の良さ $\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1))$ を評価する情報量規準を求めたい。最適なパラメータとして、次のようなパラメータ $\bar{\theta}_{x_1}$、$\bar{\theta}_y$ を定義する。

$$\begin{aligned}
\bar{\theta}_{x_1} &= \arg\left\{\min_{\theta\in\Theta} \mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\theta))\right\} \\
\bar{\theta}_y &= \arg\left\{\min_{\theta\in\Theta} \mathcal{L}(q(\mathbf{Y}), p(\mathbf{Y}|\theta))\right\}
\end{aligned}$$

以下のような仮定を置く。

$$\begin{aligned}
q(\mathbf{X}_1) &\approx p(\mathbf{X}_1|\bar{\theta}_y) \\
p(\mathbf{X}_1|\bar{\theta}_y) &\approx p(\mathbf{X}_1|\bar{\theta}_{x_1})
\end{aligned} \tag{1}$$

(1) から

$$\mathcal{L}(\hat{q}(\mathbf{Y}_1), p(\mathbf{Y}_1)) = \min \mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1))$$

がいえる．なお、$\hat{q}$ は q の経験分布関数である．

$\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta}))$ の期待値を $\theta = \bar{\theta}_y$ 近傍で Taylor 展開すると

$$\begin{aligned}
&E_q\left[\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta}))\right] \\
&\approx E_q\left[\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\bar{\theta}_y)) + \frac{1}{2}(\hat{\theta}-\bar{\theta}_y)^T\mathbf{H}_{\mathbf{X_1}}(\hat{\theta}-\bar{\theta}_y)\right] \\
&\approx E_q\left[\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\bar{\theta}_y))\right] + \frac{1}{2n}\mathbf{tr}\{\mathbf{H}_{\mathbf{X_1}}{\mathbf{H_Y}}^{-1}\}
\end{aligned}$$

展開後の第 1 項に注目すると

$$\begin{aligned}
&E_q\left[\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\bar{\theta}_y))\right] \\
&= E_q[\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\bar{\theta}_y)) - \mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1|\bar{\theta}_y)) \\
&\quad +\mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1|\bar{\theta}_y)) - \mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta})) \\
&\quad +\mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta}))] \\
&\quad \approx E_q\left[0 + \frac{1}{2}(\hat{\theta}-\bar{\theta}_y)^T\mathbf{H}_{\mathbf{X_1}}(\hat{\theta}-\bar{\theta}_y) + \mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta}))\right] \\
&\quad = \frac{1}{2n}\mathbf{tr}\{\mathbf{H}_{\mathbf{X_1}}{\mathbf{H_Y}}^{-1}\} + E_q\left[\mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta}))\right]
\end{aligned}$$

とさらに展開できる．以上をまとめると

$$\begin{aligned}
&E_q\left[\mathcal{L}(q(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta}))\right] \\
&\approx E_q\left[\mathcal{L}(\hat{q}(\mathbf{X}_1), p(\mathbf{X}_1|\hat{\theta}))\right] + \frac{1}{n}\mathbf{tr}\{\mathbf{H}_{\mathbf{X_1}}{\mathbf{H_Y}}^{-1}\} \\
&\approx E_q\left[\mathcal{L}(\hat{q}(\mathbf{Y}_1), p(\mathbf{Y}_1|\hat{\theta}))\right] + \frac{1}{n}\mathbf{tr}\{\mathbf{H}_{\mathbf{X_1}}{\mathbf{H_Y}}^{-1}\}
\end{aligned}$$

が得られた．上式の第 1 項の推定量として $-\frac{1}{n}l(\hat{\theta}(\mathbf{y}_1))$ を用い、これらを $2n$ 倍したものが PDIOp である．

4　数値実験

PDIOp が機能するか見るために、回帰モデルによるシミュレーションを行った．

- 観測データ $\mathbf{Y}$ は $\mathbf{Z}_1$と $\mathbf{Z}_2$ の和にノイズを加えたものとする
- $\mathbf{Z}_1$ は単調増加する成分にノイズを加えたものとする
- $\mathbf{Z}_2$ は周期をもつ成分で、三角関数の和にノイズを加えたものとする

- 前半のデータでパラメタ推定とモデル選択を行い、後半のデータで真の $\mathbf{Z}_1$、$\mathbf{Z}_2$、$\mathbf{Y}$ との予測二乗誤差を計算する．これを AIC、PDIO、PDIOp それぞれについて行う．
- ここでは $\mathbf{Z}_1$ への当てはまりを重視したいものとする

このシミュレーションを 10000 回行った．結果は以下のとおりである．表 1 はそれぞれの情報量規準による平均予測二乗誤差である．これを見ると、$\mathbf{Z}_1$ 以外への当てはまりは PDIO が良いものの、今回重視している $\mathbf{Z}_1$ への当てはまりは PDIOp が一番良いことがわかる．

表 1　回帰モデルによる平均予測二乗誤差　() 内は標準誤差　「基準」は真のパラメタを使った場合の平均予測二乗誤差

	$\mathbf{Z}_1$	$\mathbf{Z}_2$	$\mathbf{Y}$
AIC	62.8(0.55)	42.7	154.1
PDIO	42.9(0.54)	33.3	112.2
PDIOp	38.5(0.34)	38.7	114.9
基準	29.9	30.0	90.2

1 回のシミュレーションについて、$\mathbf{Z}_1$ のそれぞれの予測を図示したのが図 1 である．これを見ると下から 2 番目の PDIOp による予測が、一番下の真の予測に極めて近いことが分かる．この図からも、PDIOp による予測が優れていることが分かる．

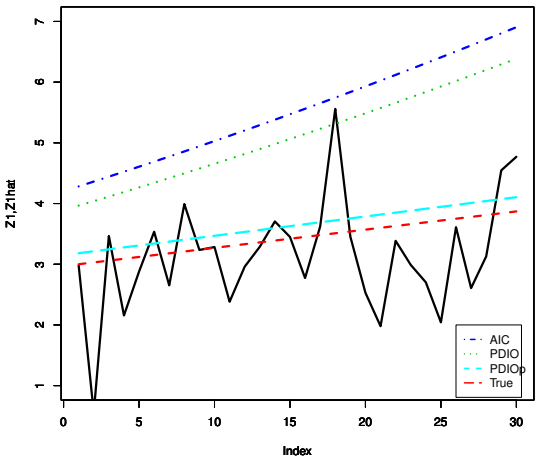

図 1　実線は真のデータ．点線は上から AIC、PDIO、PDIOp、真のパラメータによる予測

5　まとめと今後の課題

二つの新しい情報量規準を導出した．AIC の拡張、AICp はデータの一部分への当てはまりを評価できる．PDIO の拡張、PDIOp は不完全観測モデルにおいて、潜在データと観測データの一部分への当てはまりを評価することができる．いずれも、AIC,PDIO にはなかった性質を持っている．また、PDIOp がデータの一部分を評価できることをシミュレーションによって示した．なお、今回は回帰モデルでのシミュレーションであったが、欠損データのシミュレーションも進めている。

今後の課題としては、AICp、PDIOp の評価のためにさらに別の実験を行うこと、これらを実データに応用することなどが挙げられる．

参考文献

[1] Shimodaira H, *A new criterion for selecting models from partially observed data*, Selecting Models from Data: AI and Statistics IV(eds. P. Cheeseman and R. W. Oldford), Lecture Notes in Statistics, 89:21-30, 1994