回帰モデルの混合を用いた予測の性能評価

56221 松本 直久
指導教員 駒木 文保 助教授
2007 年 2 月 6 日

概要

複数の最小二乗回帰モデルを混合した推定量を考察する. モデルの混合は情報量規準によるモデル選択の不確実性を除いて, より頑健なモデルを構築するための手法である. 本論文では, パラメータとサンプルサイズについて混合推定量のリスクを漸近展開した. 結果として混合推定量がフルモデルにおける最尤推定量を改良しないことがわかった. しかし, 真のパラメータの値によってはフルモデルにおける最尤推定量よりもリスクを小さくできること, また単一のモデルを選択する推定方法に勝る場合があることを示してモデル混合の有用性を確かめた.

1 モデルの混合

回帰モデルの推定は多変量正規分布の平均の推定に帰着される. 分散共分散行列 I_n, 未知の平均ベクトル $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)^t$ の n 変量正規分布から一つのサンプル $\boldsymbol{x} = (x_1, x_2, \dots, x_n)^t$ を観測する. ここで $\boldsymbol{\theta}$ の推定量を $\hat{\boldsymbol{\theta}}$ と表すことにし, 推定量 $\hat{\boldsymbol{\theta}}$ の良さを測る指標であるリスクとして $E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$ を用いることにする. 推定量 $\hat{\boldsymbol{\theta}}$ のリスクが小さいほど, $\hat{\boldsymbol{\theta}}$ によって構築されたモデルで将来の観測値を予測する際の予測性能が良くなる.

真の構造 $\boldsymbol{\theta}$ の推定量として $n+1$ 個の最尤推定量

$$\hat{\boldsymbol{\theta}}^m = (x_1, x_2, \dots, x_m, 0, \dots, 0)^t, \quad m = 0, 1, 2, \dots, n$$

を考え, モデル m の赤池情報量規準 (AIC) を用いて混合推定量 $\hat{\boldsymbol{\theta}}$ を定数 $\beta \geq 0$ に対し

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \frac{\exp\left(-\frac{\beta}{2}\sum_{m=0}^{n} \mathrm{AIC_m}\right)\hat{\boldsymbol{\theta}}^m}{\exp\left(-\frac{\beta}{2}\sum_{m=0}^{n} \mathrm{AIC_m}\right)} \\ &= \frac{\sum_{m=0}^{n} \exp\left\{-\frac{\beta}{2}\left(\sum_{j=m+1}^{n} x_j^2 + 2m\right)\right\}\hat{\boldsymbol{\theta}}^m}{\sum_{m=0}^{n} \exp\left\{-\frac{\beta}{2}\left(\sum_{j=m+1}^{n} x_j^2 + 2m\right)\right\}}\end{aligned}$$

と定義する. この推定量は Leung and Barron [3] が提案し, $\beta = 1$ のときの重みは Akaike [1] によって "モデルの尤度" として提案されており Burnham and Anderson [2] が "Akaike weight" と名付けている.

β の値によって $\hat{\boldsymbol{\theta}}$ は次の様な意味を持つ.

- $\beta = 0$ 単純平均をとる.
- $\beta = 1$ ある事前分布を用いたベイズ法の点推定値.
- $\beta \to \infty$ AIC を用いたモデル選択をする.

2 パラメータについての漸近展開

$\boldsymbol{\theta}$ の各成分を $\theta_i = C\theta(i),\ i = 1, \dots, n$ と表し i^* を $\theta(i) \neq 0$ なる最大の i とすると, $C \to \infty$ でのリスクの極限値は次の様になる.

定理 1 $n \geq 1$, $\beta > 0$ のとき,

$$\begin{aligned}\lim_{C\to\infty} E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 &= i^* + \frac{1}{(2\pi)^{\frac{n-i^*}{2}}} \\ &\times \sum_{i=i^*+1}^{n} \int_{\boldsymbol{R}^{n-i^*}} \left(\frac{\sum_{m=i}^{n} \exp\left\{-\frac{\beta}{2}\left(\sum_{j=m+1}^{n} t_j^2 + 2m\right)\right\}}{\sum_{m=i^*}^{n} \exp\left\{-\frac{\beta}{2}\left(\sum_{j=m+1}^{n} t_j^2 + 2m\right)\right\}}\right)^2 \\ &\times t_i^2 \exp\left(-\frac{\sum_{j=i^*+1}^{n} t_j^2}{2}\right) dt_{i^*+1}^n.\end{aligned}$$

ここで, 定理 1 の右辺の積分がそれぞれ 1 より小さいことに注意すると次の系 1 が示せる.

系 1 $n \geq 1$, $\beta > 0$ とする.
(i) $i^* < n$ のとき, $i^* < \lim_{C\to\infty} E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 < n$,
(ii) $i^* = n$ のとき, $\lim_{C\to\infty} E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = n$.

さらに, $\boldsymbol{\theta}$ の各成分を $\theta(i) \neq 0$ なる関数を用いて $\theta_i = C\theta(i),\ i = 1, \dots, n$ と表して定理 1 からリスクの極限値が n になることに注意して漸近展開をすると次の様になる.

定理 2 $n \geq 1$, $\beta > 0$ で C が十分大きいとき,

$$E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 - n \sim \frac{2\beta\theta(n)^2 C^2}{(\beta+1)^{\frac{5}{2}}} \exp\left\{-\frac{\beta\theta(n)^2 C^2}{2(\beta+1)} + \beta\right\}.$$

なお, $\beta = 0$ のときは直接リスクを求められる.

定理 3 $n \geq 1$, $\beta = 0$ のとき,

$$E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \frac{n(2n+1)}{6(n+1)} + \sum_{i=1}^{n} \frac{\theta_i^2 i^2}{(n+1)^2}.$$

3 サンプルサイズについての漸近展開

先の議論を用いて, サンプルサイズを増やしたときの漸近論を考えることができる. 分散共分散行列 I_n, 平均ベクトル $\boldsymbol{\theta}$ の n 変量正規分布から N 個のサンプル $\boldsymbol{X}_i = (x_{1i}, \cdots, x_{ni})^t$, $i = 1, \ldots, N$ を観測する.

ここで, $\bar{\boldsymbol{X}} = \frac{\sum_{i=1}^{N} \boldsymbol{X}_i}{N} = (\bar{x}_1, \ldots, \bar{x}_n)^t$ とすると $n+1$ 個の最尤推定量は

$$\hat{\boldsymbol{\theta}}^m = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_m, 0, \ldots, 0)^t, \quad m = 0, 1, 2, \ldots, n$$

であり混合推定量 $\hat{\boldsymbol{\theta}}$ は

$$\hat{\boldsymbol{\theta}} = \frac{\sum_{m=0}^{n} \exp\left\{-\frac{\beta}{2}\left(N\sum_{j=m+1}^{n} \bar{x}_j^2 + 2m\right)\right\} \hat{\boldsymbol{\theta}}^m}{\sum_{m=0}^{n} \exp\left\{-\frac{\beta}{2}\left(N\sum_{j=m+1}^{n} \bar{x}_j^2 + 2m\right)\right\}}$$

となる. 一方, 正規分布の性質から

$$\bar{\boldsymbol{X}} \sim N_n(\boldsymbol{\theta}, N^{-1} I_n)$$

でありスケーリングし直せば

$$\sqrt{N}\bar{\boldsymbol{X}} \sim N_n(\sqrt{N}\boldsymbol{\theta}, I_n)$$

である.

以上から,

$$\boldsymbol{X} \longrightarrow \sqrt{N}\bar{\boldsymbol{X}}, \ \ \boldsymbol{\theta} \longrightarrow \sqrt{N}\boldsymbol{\theta}, \ \ \hat{\boldsymbol{\theta}} \longrightarrow \sqrt{N}\hat{\boldsymbol{\theta}}$$

と置き換えればサンプルサイズが 1 のときの議論が同様に成り立つことがわかる. 定理 2 において, C が十分大きいという条件を $\sqrt{N}$ が十分大きいという条件に置き換えると, 次が成り立つ.

定理 4 $n \geq 1$, $\beta > 0$ でサンプルサイズ N が十分大きいとき,

$$NE\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 - n \sim \frac{2\beta\theta_n^2 N}{(\beta+1)^{\frac{5}{2}}} \exp\left\{-\frac{\beta\theta_n^2 N}{2(\beta+1)} + \beta\right\}.$$

また, $\beta = 0$ の場合はリスクを直接求められる.

定理 5 $n \geq 1$, $\beta = 0$ のとき,

$$E\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \frac{1}{N} \cdot \frac{n\,(2n+1)}{6\,(n+1)} + \sum_{i=1}^{n} \frac{\theta_i^2 i^2}{(n+1)^2}.$$

4 数値実験

$\beta > 0$ の場合の混合推定量 $\hat{\boldsymbol{\theta}}$ に注目する. まずフルモデルにおける最尤推定量 $\hat{\boldsymbol{\theta}}^n$ との比較であるが, 定理 2 からわかるように $\hat{\boldsymbol{\theta}}$ は $\hat{\boldsymbol{\theta}}^n$ を改良してはいないが, $\|\boldsymbol{\theta}\|^2$ が小さいときは $\hat{\boldsymbol{\theta}}^n$ よりもリスクを大幅に小さくできていることがわかる. また, 真のパラメータに零成分が多く含まれている場合に広い領域で $\hat{\boldsymbol{\theta}}^n$ よりも優れていることがわかる. これらは $\hat{\boldsymbol{\theta}}^n$ がパラメータの全ての成分を最尤推定しているのに対し, $\hat{\boldsymbol{\theta}}$ は混合元のモデルとして零成分を含んだものを仮定しているので真のパラメータの零成分及び 0 に近い成分をより正確に推定できているためと考えられる.

次に AIC との比較であるが, $n = 6$ の場合には広い領域でリスクを小さくできていることがわかる. これは, 候補となるモデルが多いとモデル選択による推定値にはモデル選択の不確実性が入り込んでしまう一方, $\hat{\boldsymbol{\theta}}$ はその分だけ推定値が安定するからと考えられる.

5 結論と今後の課題

本研究では, パラメータとサンプルサイズについて混合推定量 $\hat{\boldsymbol{\theta}}$ のリスクを漸近展開した. 結果として $\hat{\boldsymbol{\theta}}$ がフルモデルにおける最尤推定量 $\hat{\boldsymbol{\theta}}^n$ を改良しないことがわかったが, 真のパラメータの値によってはフルモデルにおける最尤推定量よりもリスクを小さくできるということと, 単一のモデルを選択する推定方法にも勝る場合があることを示した.

実際に混合推定量 $\hat{\boldsymbol{\theta}}$ を用いる際には β の値を定める必要がある. β の値によっては全ての $\|\boldsymbol{\theta}\|^2$ とはいかないものの, ある範囲では他の値よりもリスクを小さくできる. したがって, β をデータの関数として表してより良い推定量を構成するということは一考に価する. また, 本研究では C 及び N が大きいときの漸近論を考えたが, C が小さいときと小標本の場合のリスクの振る舞いを明らかにすることも望まれる.

参考文献

[1] Akaike, H.: A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, **30** (1978), 9–14.

[2] Burnham, K. P. and D. R. Anderson: *Model Selection and Multimodel Inference*. Springer, New York, 2003.

[3] Leung, G. and A. R. Barron: Information Theory and Mixing Least-Squares Regressions. *IEEE Transactions on Information Theory*, **52** (2006), 3396–3410.