

Empirical and Generalized Bayes Ridge Regression Estimators with Minimaxity and Stability

久保川 達也

May 9, 2002

1 問題設定

次のような簡単な多重回帰モデルにおいて，回帰係数ベクトル $\beta \in R^p$ を推定する問題を考える。

$$y = A\beta + \epsilon \quad (1.1)$$

ただし， y は N -次元の観測ベクトル， A はランク p の $N \times p$ 行列，誤差ベクトル ϵ は $\mathcal{N}_N(0, \sigma^2 I_N)$ なる正規分布に従う。 $\hat{\beta}$ の最小 2 乗推定量 $\hat{\beta} = (A^t A)^{-1} A^t y$ の共分散行列は

$$\text{Cov}(\hat{\beta}) = \sigma^2 (A^t A)^{-1}$$

で与えられるが，説明変数の間に多重共線性が存在する場合，推定値が不安定になってしまうことが知られている。多重共線性とは， p 個の説明変数の間に高い相関関係が存在していることを意味しており，その結果 $(A^t A)^{-1}$ が不安定になってしまい，最小 2 乗推定量の精度が悪くなってしまう。

多重共線性が存在する場合の推定方法としては，リッジ回帰推定法と主成分回帰推定法が知られている。前者は，Hoerl and Kennard (1970) によって提案され

$$\begin{aligned} \hat{\beta}^R(\lambda) &= [A^t A + kI]^{-1} A^t y \\ &= \hat{\beta} - [I + \lambda A^t A]^{-1} \hat{\beta} \end{aligned} \quad (1.2)$$

なる形で与えられる。 $k > 0$ はリッジ・パラメタと呼ばれ， $k = 1/\lambda$ である。しかし，このリッジ回帰推定量には次の 2 つの欠点がある。1 つは， k の決め方に恣意性があるという点で， k を 0 に近づけるとリッジ回帰推定量は最小 2 乗推定量に近づき， $k \rightarrow \infty$ とすると 0 の方向へ縮小されてしまう。もう 1 つは，最小 2 乗推定量を一樣に改良していない点である。最小 2 乗推定量は定数リスクをもつミニマクス推定量として特徴づけられるので，リッジ回帰推定量はミニマクス性を持たないと表現することもできる。

そこで，リッジ・パラメタ k もしくは λ をデータに基づいて推定することが考えられる。ここでは，誤差分散 σ^2 が未知の場合を扱うので， $S = (y - A\hat{\beta})^t (y - A\hat{\beta})$ とおくと， λ は $\hat{\beta}$, S に基づいて $\hat{\lambda} = \hat{\lambda}(\hat{\beta}, S)$ によって推定される。これをリッジ回帰推定量 (1.2) に代入した推定量

$$\hat{\beta}^R(\hat{\lambda}) = \hat{\beta} - [I + \hat{\lambda} A^t A]^{-1} \hat{\beta} \quad (1.3)$$

は、適応型リッジ回帰推定量と呼ばれる。問題は、この形の推定量 $\hat{\beta}^R(\hat{\lambda})$ が安定性とミニマクス性を併せ持つように λ の推定量 $\hat{\lambda}$ を構成することである。

多重共線性の問題を扱いやすくするために、次のような標準系を導入する。 H を

$$H(A^t A)^{-1} H^t = D, \quad D = \text{diag}(d_1, \dots, d_p), \quad d_1 \geq \dots \geq d_p$$

なる $p \times p$ 直交行列とする。 $x = (x_1, \dots, x_p)^t = H\hat{\beta}$, $\theta = (\theta_1, \dots, \theta_p)^t = H\beta$ とおくと $x \sim \mathcal{N}_p(\theta, D)$ となり、 θ_i のリッジ回帰推定量は

$$\hat{\theta}_i(\lambda) = x_i - d_i(\lambda + d_i)^{-1}x_i, \quad i = 1, \dots, p$$

と表される。リッジ回帰推定量は、 d_i が大きく x_i の推定精度が悪いときには 0 の方向へ縮小され、 d_i が小さく x_i の精度が良いときには x_i に近づく。従って x_i のバラツキの程度に応じて縮小の程度が調整されるという性質をもつ。多重共線性は 1 つ以上の個有値 d_i が非常に大きいことを意味しており、少なくとも d_1 は極めて大きい値を示している。 d_i が非常に大きいときには x_i の信頼性が失われるので、 θ_i を事前情報や予想される値で推定する方が安定している。そこで推定量 $\hat{\theta}_i$ の安定性を次のように定義して、推定量が安定性条件をみたすか否かを検討するのが合理的であるように思える。任意に固定した未知母数に対して、ある定数 $B > 0$ と $\delta > 0$ が存在して、任意の十分大きな d_i に対して

$$P[|\hat{\theta}_i| < B] > 1 - \delta \quad (1.4)$$

なる不等式をみたすとき、推定量 $\hat{\theta}_i$ は安定であるという。リッジ回帰推定量 $\hat{\beta}^R(\lambda)$ は安定性条件 (1.4) をみたしている。

リッジ・パラメタ λ の通常の推定量は、定数 $a > 0$, $b \geq 0$ に対して

$$\hat{\lambda}_1(a, b) = \hat{\beta}^t A^t A \hat{\beta} / (aS) + b = \sum_{i=1}^p x_i^2 / (ad_i S) + b$$

なる形で与えられる。2 乗損失関数

$$L_s(\hat{\beta}, \beta, \sigma^2) = (\hat{\beta} - \beta)^t (\hat{\beta} - \beta) / \sigma^2$$

のもとで推定量の良さを評価するとき、適応型リッジ回帰推定量 $\hat{\beta}^R(\hat{\lambda}_1(a, 0))$ が最小 2 乗推定量を改良するための条件 (ミニマクス性条件) は

$$ad_1 \leq 2 \left\{ \sum_{i=1}^p (d_i/d_1)^2 - 2 \right\}$$

で与えられる。多重共線性が存在するとき、すなわち d_1 が非常に大きいときにはこのミニマクス性の条件がみたされない。

もう 1 つの代表的な λ の推定量は、Fay and Herriot (1979), Shinozaki and Chang (1993) によって与えられた経験ベイズ推定量

$$\hat{\lambda}_2 = \max(\lambda^*, 0) \quad (1.5)$$

である。ただし λ^* は方程式

$$\hat{\beta}^t \{ (A^t A)^{-1} + \lambda^* I \}^{-1} \hat{\beta} = \sum_{i=1}^p \frac{x_i^2}{d_i + \lambda^*} = \frac{p-2}{n+2} S \quad (1.6)$$

の根である。2乗損失関数 $L_s(\hat{\beta}, \beta, \sigma^2)$ に関して $\hat{\beta}^R(\hat{\lambda}_2)$ がミニマクスになるための条件は

$$\sum_{i=1}^p \frac{d_i^2}{d_1^2} - 2 \geq \frac{p-2}{2}$$

で与えられ、 d_1 が大きくなるとこの条件はみたされなくなる。

これらの結果は、多重共線性が存在する場合2乗損失関数のもとでは適応型リッジ回帰推定量のミニマクス性を正当化できないことを示している。しかし、適応型リッジ回帰推定量は実際の場合で有用であり、何らかの形で理論上の正当性を与えたい。そこで、ここでは2乗損失関数 L_s の代わりに重み付き2乗損失関数

$$L_w(\hat{\beta}, \beta, \sigma^2) = (\hat{\beta} - \beta)^t (A A^t)^2 (\hat{\beta} - \beta) / \sigma^2 = \sum_{i=1}^p (\hat{\theta}_i - \theta_i)^2 / (d_i^2 \sigma^2)$$

を用いて、適応型リッジ回帰推定量のミニマクス性を主張したい。この損失関数は Strawderman (1978) によって提案されたもので、 d_i が大きければ x_i を大きく縮小することを許している。

2 ミニマクスな適応型リッジ回帰推定量

適応型リッジ回帰推定量 (1.3) が損失関数 L_w のもとでミニマクスになるための $\hat{\lambda}$ の条件を与える。ここでは、より一般的に仮説

$$H_0: \beta = C\alpha$$

が期待される状況を考える。ただし C は $p \times q$ の既知の行列でランクが q 、 $\alpha \in R^q$ は未知のベクトルである。この仮説の方向へ縮小される適応型リッジ回帰推定量をもとめるために、 β に事前分布 $\mathcal{N}_p(C\alpha, \sigma^2 \lambda I_p)$ を仮定する。このとき $\hat{\beta}$ を与えたときの β の事後分布及び $\hat{\beta}$ の周辺分布は

$$\begin{aligned} \beta | \hat{\beta} &\sim \mathcal{N}_p(\hat{\beta}^B(\lambda, \alpha), \sigma^2(A^t A + \lambda^{-1} I)^{-1}) \\ \hat{\beta} &\sim \mathcal{N}_p(C\alpha, \sigma^2\{(A^t A)^{-1} + \lambda I\}) \end{aligned}$$

で与えられる。ここで $\hat{\beta}^B(\lambda, \alpha)$ は β のベイズ推定量であり

$$\begin{aligned} \hat{\beta}^B(\lambda, \alpha) &= (A^t A + \lambda^{-1} I)^{-1} A^t A (\hat{\beta} - C\alpha) + C\alpha \\ &= \hat{\beta} - (I + \lambda A^t A)^{-1} (\hat{\beta} - C\alpha) \end{aligned}$$

なる形をしている。 α, λ は未知母数なので、推定する必要がある。 α は、 $(\hat{\beta} - C\alpha)^t A^t A (\hat{\beta} - C\alpha)$ を最小にする重み付き最小2乗推定量

$$\hat{\alpha} = (C^t A^t A C)^{-1} C^t A^t A \hat{\beta},$$

によって推定される。また λ は S と

$$\hat{\beta} - C\hat{\alpha} = \{I - C(C^t A^t A C)^{-1} C^t A^t A\} \hat{\beta}$$

に基づいて $\hat{\lambda} = \hat{\lambda}(\hat{\beta} - C\hat{\alpha}, S)$ として推定することができる。これらをベイズ推定量 $\hat{\beta}^B(\lambda, \alpha)$ に代入すると、経験ベイズ推定量

$$\hat{\beta}^B(\hat{\lambda}, \hat{\alpha}) = \hat{\beta} - (I + \hat{\lambda}A^tA)^{-1}(\hat{\beta} - C\hat{\alpha}) \quad (2.1)$$

が得られる。これは最小 2 乗推定量を部分空間 $\{C\alpha \mid \alpha \in R^q\}$ の方向へ縮小する推定量である。

経験ベイズ推定量 $\hat{\beta}^B(\hat{\lambda}, \hat{\alpha})$ が最小 2 乗推定量を改良するための $\hat{\lambda}$ に関する条件は、次の定理で与えられる。

定理 1. $\hat{\lambda}$ が次の条件をみたすとき、経験ベイズ推定量 $\hat{\beta}^B(\hat{\lambda}, \hat{\alpha})$ は L_w 損失関数に関して最小 2 乗推定量を改良する、すなわちミニマクスになる。

- (a) 非負定数 λ_0 に対して $\hat{\lambda} \geq \lambda_0$ であり、 $\hat{\lambda}$ は $x_1, \dots, x_p, S$ に関して絶対連続である。
- (b) $i = 1, \dots, p$ に対して $(x_i - b_i^t x) \partial \hat{\lambda} / \partial x_i \geq 0$ であり、次の条件をみたす。

$$\sum_{i=1}^p \frac{(x_i - b_i^t x)}{d_i + \hat{\lambda}} \frac{\partial \hat{\lambda}}{\partial x_i} \leq 2 \quad (2.2)$$

- (c) $\partial \hat{\lambda} / \partial S \leq 0$ であり、正の定数 α, β に対して次の条件をみたす。

$$\sum_{i=1}^p \frac{(x_i - b_i^t x)^2 / S}{d_i + \hat{\lambda}} \leq \alpha, \quad - \sum_{i=1}^p \frac{(x_i - b_i^t x)^2}{(d_i + \hat{\lambda})^2} \frac{\partial \hat{\lambda}}{\partial S} \leq \beta \quad (2.3)$$

- (d) 定数 λ_0, α, β は次の不等式をみたす。

$$\sum_{i=1}^p (1 - b_{ii}) \frac{d_i - d_p}{d_i + \lambda_0} + \frac{(n-2)\alpha}{2} + 2\beta \leq p - q - 2 \quad (2.4)$$

ただし b_{ii} は $Z(Z^t D^{-1} Z)^{-1} Z^t D^{-1} = (b_1, \dots, b_p)^t$ の (i, i) 成分である。

定理 1 を用いると、ミニマクスな適応型ベイズ推定量を与えることができる。

[1] 典型的な推定量. 通常考えられる λ の推定量 $\hat{\lambda}_1(a, b)$ が定理 1 の条件をみたすように定数 a, b の条件をもとめると、

$$\begin{aligned} \hat{\lambda}_{AD} &= \frac{n+2}{\text{ch}_{\min}(A^t A)S} (\hat{\beta} - C\hat{\alpha})^t A^t A (\hat{\beta} - C\hat{\alpha}) + \lambda_0 \\ &= (n+2)d_1 \sum_{i=1}^p (x_i - b_i^t x)^2 / (d_i S) + \lambda_0 \end{aligned} \quad (2.5)$$

となる。ただし $\text{ch}_{\min}(M)$ は行列 M の最小の固有値を表しており、 λ_0 は

$$\sum_{i=1}^p (1 - b_{ii}) \frac{d_i - d_p}{d_i + \lambda_0} = p - q - 2.5 \quad (2.6)$$

の根である。このとき、適応型リッジ回帰推定量 $\hat{\beta}^{AD} = \hat{\beta}^B(\hat{\lambda}_{AD}, \hat{\alpha})$ はミニマクスになる。

しかし，この推定量 (2.5) は d_p が小さいときに次のような欠点をもつ。 d_p を小さくするにつれ $\hat{\lambda}_{AD}$ は大きくなるので，適応型リッジ回帰推定量 $\hat{\beta}^{AD}$ は最小 2 乗推定量に近づいてしまう。多重共線性が存在する場合には，最小 2 乗推定量は不安定になってしまうので，その適応型リッジ回帰推定量 $\hat{\beta}^{AD}$ は安定化されない。

そこで重みを d_i から $d_i + 1$ に代えた推定量が考えられ，具体的には次で与えられる。

[2] 打ち切り推定量 λ の打ち切り推定量を

$$\begin{aligned}\hat{\lambda}_{TR} &= \max \left\{ \frac{(n+2)(d_1+1)}{S} (\hat{\beta} - C\hat{\alpha})^t [(A^t A)^{-1} + I_p]^{-1} (\hat{\beta} - C\hat{\alpha}), \lambda_0 \right\} \\ &= \max \left\{ \frac{(n+2)(d_1+1)}{S} \sum_{i=1}^p \frac{(x_i - \mathbf{b}_i^t \mathbf{x})^2}{d_i + 1}, \lambda_0 \right\}\end{aligned}\quad (2.7)$$

で定義する。これは，上述の重みを取り替えるとともに λ_0 で打ち切ることによって $\hat{\lambda}_{AD}$ を改良している。また λ_0 を (2.6) の根とすると，適応型リッジ回帰推定量 $\hat{\beta}^B(\hat{\lambda}_{TR}, \hat{\alpha})$ はミニマクスになる。

しかし，容易に確かめられるように $\lim_{d_1 \rightarrow \infty} \hat{\theta}_1(\hat{\lambda}_{AD}) \neq 0$, $\lim_{d_1 \rightarrow \infty} \hat{\theta}_1(\hat{\lambda}_{TR}) \neq 0$ であり，安定性条件がみたされない。

[3] 経験ベイズ推定量 Shinozaki and Chang (1993) で与えられているように， λ^* を方程式

$$(\hat{\beta} - C\hat{\alpha})^t \left\{ (A^t A)^{-1} + \lambda^* I \right\}^{-1} (\hat{\beta} - C\hat{\alpha}) = \frac{p-q-2}{n+2} S \quad (2.8)$$

もしくは

$$\sum_{i=1}^p \frac{(x_i - \mathbf{b}_i^t \mathbf{x})^2 / S}{d_i + \lambda^*} = \frac{p-q-2}{n+2} S \quad (2.9)$$

の根とする。また定数 λ_0 を方程式

$$\sum_{i=1}^p (1 - b_{ii}) \frac{d_i - d_p}{d_i + \lambda_0} = (p-q-2)/2 \quad (2.10)$$

の根とするとき， λ の経験ベイズ推定量は

$$\hat{\lambda}_{EB} = \max(\lambda^*, \lambda_0) \quad (2.11)$$

で与えられる。このとき，定理 1 の条件が成り立つことを容易に確かめることができ， β の経験ベイズ推定量 $\hat{\beta}^{EB} = \hat{\beta}^B(\hat{\lambda}_{EB}, \hat{\alpha})$ のミニマクス性が示される。

また， $\lim_{d_1 \rightarrow \infty} \hat{\lambda}_{EB} < \infty$ であることから， $\lim_{d_1 \rightarrow \infty} \hat{\theta}_1(\hat{\lambda}_{EB}) = 0$ となり，安定化されている。

3 ミニマクスな一般化ベイズ・リッジ回帰推定量

ミニマクスな一般化ベイズ・リッジ回帰推定量を導出するために，モデルの標準系において (θ, η) が次のような事前分布をもつと仮定する。

$$\begin{aligned}\theta | \tau, \eta &\sim \mathcal{N}(\mathbf{0}, (\tau/\eta) \mathbf{R}) \\ (\tau, \eta) &\sim I(\tau > \tau_0) \times \eta^{-1}\end{aligned}\quad (3.1)$$

ただし $\eta = 1/\sigma^2$ であり, τ_0 は非負の定数である。

$$u \equiv u(\tau) = \prod_{j=1}^p (\tau + d_j)^{-1/p} \quad (3.2)$$

に対して

$$\phi_i = \frac{\int_{\tau_0}^{\infty} (\tau + d_i)^{-1} u^{p/2} \eta^{(n+p)/2} [\sum_{j=1}^p x_j^2 / (\tau + d_j) + S]^{-(n+p)/2-1} d\tau}{\int_{\tau_0}^{\infty} u^{p/2} \eta^{(n+p)/2} [\sum_{j=1}^p x_j^2 / (\tau + d_j) + S]^{-(n+p)/2-1} d\tau} \quad (3.3)$$

とおくとき, θ の一般化ベイズ推定量は

$$\hat{\theta}_i^{GB}(\tau_0) = x_i - d_i \phi_i x_i \quad (3.4)$$

で与えられる。

定理 2. 定数 τ_0 が条件

$$6 \frac{d_1 - d_p}{\tau_0 + d_p} + 2 \sum_{j=1}^p \frac{d_1 - d_j}{\tau_0 + d_j} + \sum_{j=1}^p \frac{d_j - d_p}{\tau_0 + d_j} \leq (n+2)/2$$

をみたすとき, 一般化ベイズ・リッジ回帰推定量 $\hat{\theta}_i^{GB}(\tau_0)$ は L_w 損失関数に関して最小 2 乗推定量を改良する, すなわちミニマクスになる。

REFERENCES

- Casella, G. (1980). Minimax ridge regression estimation. *Ann. Statist.*, **8**, 1036-1056.
- Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.*, **70**, 311-319.
- Fay, R.E. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.
- Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.
- Shinozaki, N. and Chang, Y.-T. (1993). Minimality of empirical Bayes estimators of the means of independent normal variables with unequal variances. *Commun. Statist. - Theory Method*, **22**, 2147-2169.
- Shinozaki, N. and Chang, Y.-T. (1996). Minimality of empirical Bayes estimators shrinking toward the grand mean when variances are unequal. *Commun. Statist. - Theory Method*, **25**, 183-199.
- Strawderman, W.E. (1978). Minimax adaptive generalized ridge regression estimators. *J. Amer. Statist. Assoc.*, **73**, 623-627.