数理解析研究所講究録
第1603巻 2008年 43-49

縮小型推定法における自由度の不偏推定について

東京大学経済学研究科　加藤賢悟 (Kengo Kato)
Graduate School of Economics, University of Tokyo

1　はじめに

近年，線形回帰における係数ベクトルの推定において，Lasso[5] などの縮小型推定法が注目を集めている．OLS とは異なり，Lasso は係数を exact にゼロに推定することができるため，係数ベクトルの推定と変数選択を同時に実行することが出来るという特徴を持つ．

$y=(y_1,\ldots,y_n)'$ を目的変数，$x_j=(x_{1j},\ldots,x_{nj})'$，$j=1,\ldots,p$ を p 個の線形独立な説明変数，$X=[x_1\cdots x_p]$ をデータ行列とする．このとき，次のような線形回帰モデルを考える：

$$y=X\beta+\epsilon.$$

ここで，$\beta=(\beta_1,\ldots,\beta_p)'$ は係数ベクトル，$\epsilon\sim N_n(0,\sigma^2 I_n)$ は誤差項である．

このとき，Lasso 推定量は，次の制約付き最小化問題の解で与えられる：

$$\min_{\beta}\|y-X\beta\|_2^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq t.$$

ここで $t\geq 0$ はチューニングパラメータと呼ばれる．簡単な計算により，この最小化問題は次の最小化問題に帰着することがわかる．まず，$\mathbb{R}^p$ の内積 $\langle\cdot,\cdot\rangle$ を，$\langle u,v\rangle=u'Vv$，$u,v\in\mathbb{R}^p$ で定める．ここで，$V=X'X$ であって，$\|\cdot\|=\sqrt{\langle\cdot,\cdot\rangle}$ と書く．このとき，Lasso の最小化問題は

$$\min_{\beta}\|\beta-\hat{\beta}^{\circ}\| \text{ subject to } \sum_{j=1}^{p}|\beta_j|\leq t$$

と書き直すことが出来る．ここで，$\hat{\beta}^{\circ}$ は β の OLS 推定量である．したがって，Lasso 推定量は OLS 推定量 $\hat{\beta}^{\circ}$ を制約集合 $\{\beta\in\mathbb{R}^p|\sum_{j=1}^{p}|\beta_j|\leq t\}$ に射影したものとみることが出来る．なお，制約集合の形から，係数を exact にゼロに推定するという Lasso の特徴が幾何的に理解できる．

縮小型推定法のもう一つの例として，group Lasso[7] が挙げられる．β を $\beta=(\beta_{[1]}',\ldots,\beta_{[J]}')'$ と分割する．ここで，$\beta_{[j]}$ は $p_j\times 1$ ベクトルである．V_j を $p_j\times p_j$ 正定値対称行列とする．Group Lasso 推定量は，Lasso の最小化問題において制約集合を $\{\beta\in\mathbb{R}^p\mid\sum_{j=1}^{J}(\beta_{[j]}'V_j\beta_{[j]})^{1/2}\leq t\}$ に取り替えたときの最適解で与えられ

る．Group Lassoは，説明変数をいくつかのグループに分割して，一つのグループに入っている説明変数の係数をまとめてゼロに推定するという特徴を持つ．

ところで，Lasso，group Lassoといった縮小型推定法を実行するに当たって，本質的な問題となるのがチューニングパラメータの選択である．本論文では，以下で見るように，Lasso，group Lassoを含むより一般的な推定法を考え，チューニングパラメータの選択規準を導出する統一的な方法を紹介する．

Lasso，group Lassoを一般化して，次のような最小化問題の解で定義される推定量$\hat{\beta}_K$を考える：

$$\min_{\beta} \|\beta - \hat{\beta}^{\circ}\| \text{ subject to } \beta \in K.$$

ここで，$K \subset \mathbb{R}^p$は閉凸集合である．$\hat{\beta}_K$は$\hat{\beta}^{\circ}$をKに射影したものである．実際には，与えられた閉凸集合の族$\mathcal{K}$のなかから最適な$\hat{K}$を選んで，$\hat{\beta}_{\hat{K}}$を最終的な推定量とすることが多い．たとえば，Lassoだと$\mathcal{K} = \{\beta \in \mathbb{R}^p | \sum_{j=1}^{p} |\beta_j| \leq t\}$である．本論文の目的は，$K$の選択規準を導出することである．

本論文に関連する論文として，[8]が挙げられる．[8]は，penalization formulationのもとで，Lassoの自由度(degrees of freedom，定義は2節で与える)の不偏推定量がLasso推定量のゼロでない成分の個数で与えられことを示し，チューニングパラメータの選択規準を与えている．しかしながら，彼らの導出方法はLassoのsolution path (Lasso推定値のチューニングパラメータの関数としてのパス)がpiecewise linearであることに大きく依存している．そのため，例えばgroup Lassoのように，制約集合が多面集合でない場合や，3節で扱うfused Lasso[6]のように複数のチューニングパラメータがある場合には彼らのテクニックを適用することは難しいと思われる．

なお，本論文は著者による既発表論文[2]で得られた結果を要約したものであり，詳細は上記論文を参照すること．

2 主結果

2.1 SURE法

最適なKとして，モデルの予測リスクを最小にするものを選ぶこととする．予測リスクは未知なので，推定する必要がある．

予測量$\hat{\mu}_K = \hat{\mu}_K(y) = X\hat{\beta}_K$を考える．$y^{new}$を$y$と同じ分布に従う独立なランダムベクトルとし，$\hat{\mu}_K$の予測リスクを$E(\|y^{new} - \hat{\mu}_K\|_2^2)/n$で測る．

このとき，予測リスクは

$$E(\|y^{new} - \hat{\mu}_K\|_2^2) = E(\|y - \hat{\mu}_K\|_2^2) + 2df(\hat{\mu}_K)\sigma^2 \quad (1)$$

と分解することができる．ここで，

$$df(\hat{\mu}_K) = \sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, y_i)/\sigma^2$$

は予測量 $\hat{\mu}_K$ の 自由度(degrees of freedom) と呼ばれる [1]．ところで，式 (1) の右辺第一項に関しては，$\|y - \hat{\mu}_K\|_2^2$ がその不偏推定量となる．また，自由度の推定については，次の Stein の補題 [4] が有効である．

補題 2.1 ([4])**.** 各 $\hat{\mu}_i : \mathbb{R}^n \to \mathbb{R}$ は i 番目の座標に関して絶対連続とする．$E(|\partial\hat{\mu}_i/\partial y_i|) < \infty$ なら，

$$\sum_{i=1}^{n} \mathrm{cov}(\hat{\mu}_i, y_i)/\sigma^2 = E(\mathrm{div}\,\hat{\mu})$$

が成り立つ．ここで，$\mathrm{div}\,\hat{\mu} = \sum_{i=1}^{n} \partial\hat{\mu}_i/\partial y_i$ である．

$\hat{\mu}_K$ に対して Stein の補題が適用できることは以下の補題からわかる．

補題 2.2. 各 i に対して，$\hat{\mu}_{K,i}$ は y の各座標に関して絶対連続であり，$\partial\hat{\mu}_{K,i}/\partial y$ は本質的有界である．

従って，Stein の補題を使うと，自由度の不偏推定量が

$$\widehat{df}(\hat{\mu}_K) = \mathrm{div}\,\hat{\mu}_K$$

で与えられることがわかる．このとき，予測リスクの不偏推定量として，C_p 型の規準

$$C_p(\hat{\mu}_K) = \frac{\|y - \hat{\mu}_K\|_2^2}{n} + \frac{2\widehat{df}(\hat{\mu}_K)}{n}\sigma^2$$

を得る．また，C_p と同値な規準として，AIC を

$$\mathrm{AIC}(\hat{\mu}_K) = \frac{\|y - \hat{\mu}_K\|_2^2}{n\sigma^2} + \frac{2\widehat{df}(\hat{\mu}_K)}{n}.$$

で定義することができる．

ところで，$\hat{\beta}_K$ が $\hat{\beta}^\circ$ に関して微分可能であれば，$\mathrm{div}\,\hat{\mu}_K = \mathrm{tr}(\partial\hat{\beta}_K/\partial\hat{\beta}^\circ)$ となる．したがって，$\hat{\beta}_K$ の $\hat{\beta}^\circ$ に関するダイバージェンスが計算できれば，$\hat{\mu}_K$ の自由度の不偏推定量が得られることがわかる．しかしながら，通常 $\hat{\beta}_K$ の明示的な関数形はわからないことが多い。そのため，直接ダイバージェンスを計算するのは難しいが，次節で示すように，制約集合 K の境界に区分的な滑らかさを仮定すれば，チューブ座標を導入することによって，ダイバージェンスを制約集合に関する幾何的な量で表すことが出来る．

2.2 Divergence formula

$K \subset \mathbb{R}^p$ を閉凸集合とする．$x \in \mathbb{R}^p$ に対して，x_K を $\langle \cdot, \cdot \rangle$ に関する x の K への射影とする．いま，写像 $f : \mathbb{R}^p \to \mathbb{R}^p$ を $f(x) = x_K$ で定める．f のダイバージェンスを求める．

$s \in \partial K$ に対して，s における法錘は $N(K,s) = \{y - s | y_K = s\}$ で与えられる．法錘 $N(K,s)$ の次元に応じて，境界 ∂K は $\partial K = D_1 \cup \cdots \cup D_p$ と排反に分割される．ここで，$D_m = \{s \in \partial K | \dim N(K,s) = m\}$ である．いま，$E_m = \{x \in \mathbb{R}^p \setminus K | x_K \in D_m\}$ と定めると，$\mathbb{R}^p \setminus K$ の排反な分割 $\mathbb{R}^p \setminus K = E_1 \cup \cdots \cup E_p$ を得る．次の仮定をおく [3].

仮定 2.1. D_m は $(p-m)$ 次元の C^2 級多様体であって，有限個の相対開連結成分からなる．さらに，$E_m \backslash E_m^\circ$ の *Lebesgue* 測度は 0 である．

この仮定をみたす境界を区分的に滑らかな境界と呼ぶ．

$\theta = (\theta^1, \ldots, \theta^{p-m})$ を D_m の C^2 級局所座標系とし，$s \in D_m$ を $s = s(\theta)$ と書く．D_m の s における接空間は $T_{s(\theta)} D_m = \mathrm{span}\{b_a(\theta) = \partial s / \partial \theta^a(\theta), a = 1, \ldots, p-m\}$ で与えられる．また，$T_{s(\theta)} D_m$ に直交する正規直交系 $\{n_\alpha(\theta), \alpha = 1, \ldots, m\}$ を 1 つとる．このとき，

$$(\theta, \tau) \mapsto \varphi(\theta, \tau) = (s(\theta) + \sum_{\alpha=1}^{m} \tau^\alpha n_\alpha(\theta))$$

を E_m° の C^1 級局所パラメータ付けとしてとって，f を局所座標 (θ, τ) に関して $f(\theta, \tau) = s(\theta)$ と表すことができる ([2] の Lemma 3.1 参照).

局所座標系 $\theta = (\theta^1, \ldots, \theta^{p-m})$ に付随する第 1 基本形式を $G(\theta)$，D_m の法線方向 $n_\alpha(\theta)$ に関する第 2 基本形式を $H_\alpha(\theta)$ と書く．また，$x = \varphi(\theta, \tau)$ に対して，$H(\theta, \tau) = -\sum_{\alpha=1}^{m} \tau^\alpha H_\alpha(\theta)$ とおく．これは，半正定値行列である．このとき，次の補題を得る．

補題 2.3. ダイバージェンス $\mathrm{div}\, f(x) = \sum_{j=1}^{p} \partial f_j(x) / \partial x_j$，$x \in E_m^\circ$ は

$$\mathrm{div}\, f(x) = \sum_{a=1}^{p-m} \frac{1}{1 + \kappa_a(x)}$$

で与えられる．ここで，$\kappa_a(x) = \kappa_a(\theta, \tau)$，$a = 1, \ldots p-m$ は，

$$|H(\theta, \tau) - \kappa G(\theta)| = 0 \tag{2}$$

をみたす固有値である．

2.3 Degrees of freedom

$\hat{\beta}^\circ \in E_m$ に対して，$x = \hat{\beta}^\circ$，$x_K = \hat{\beta}_K$ としたときの固有方程式 (2) の根を $\kappa_{m,a}(\hat{\beta}^\circ), a = 1, \ldots, p-m$ と書く．また，形式的に $E_0 = K$，$\kappa_{0,a} \equiv 0$，$a = 1, \ldots, p$ と定める．

定理 2.1. $K \subset \mathbb{R}^p$ を仮定 *2.1* をみたす閉凸集合とする．このとき，

$$\widehat{df}(\hat{\mu}_K) = \sum_{m=0}^{p} \sum_{a=1}^{p-m} \frac{1}{1 + \kappa_{m,a}(\hat{\beta}^\circ)} I(\hat{\beta}_K \in E_m)$$

は $\hat{\mu}_K = X\hat{\beta}_K$ の自由度 $df(\hat{\mu}_K)$ の不偏推定量を与える．

ここで，$\widehat{df}(\hat{\mu}_K)$ を計算する際に，$\hat{\beta}_K$ の関数形は必要としないことに注意する．$\widehat{df}(\hat{\mu}_K)$ を計算する際には，$\hat{\beta}^\circ$ と $\hat{\beta}_K$ の数値的な値と，制約集合 K の境界に関する第 1 基本形式と第 2 基本形式が計算できればよい．特に K が凸多面集合なら，

$$\widehat{df}(\hat{\mu}_K) = p - \sum_{m=1}^{p} m I(\hat{\beta}_K \in D_m)$$

となる．これは，$\hat{\beta}_K$ を相対内点として含むフェイスの次元に等しい．

3 例

以下，Lasso，group Lasso，fused Lasso を取り上げ，自由度の不偏推定量を与える．なお，以下では $\hat{\beta}_K$ の代わりに $\hat{\beta}(t)$ と書いている．

3.1 Lasso

制約集合：

$$K = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{p} |\beta_j| \leq t\}.$$

自由度の不偏推定量：

$$\widehat{df}(t) = \begin{cases} \#\{j \mid \hat{\beta}(t)_j \neq 0\} - 1 & \text{if } \sum_{j=1}^{p} |\hat{\beta}_j^\circ| > t, \\ p & \text{if } \sum_{j=1}^{p} |\hat{\beta}_j^\circ| \leq t. \end{cases}$$

3.2 Fused Lasso

制約集合：

$$K = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{p} |\beta_j| \leq t_1,\ \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq t_2\}.$$

ここで，$t_1 \neq t_2$ とする．いま

$$K_1 = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{p} |\beta_j| \leq t_1\},$$

$$K_2 = \{\beta \in \mathbb{R}^p \mid \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq t_2\}.$$

とおく．自由度の不偏推定量は

$$\widehat{df}(t) = \begin{cases} p - m_1(t) & \text{if } \hat{\beta}(t) \in \partial K_1 \cap K_2^\circ \text{ and } \hat{\beta}^\circ \notin K, \\ p - m_2(t) & \text{if } \hat{\beta}(t) \in K_1^\circ \cap \partial K_2 \text{ and } \hat{\beta}^\circ \notin K, \\ p - m_3(t) & \text{if } \hat{\beta}(t) \in \partial K_1 \cap \partial K_2 \text{ and } \hat{\beta}^\circ \notin K, \\ p & \text{if } \hat{\beta}^\circ \in K, \end{cases}$$

で与えられる．ここで，

$m_1(t) = \#\{j \mid \hat{\beta}(t)_j = 0\} + 1,$
$m_2(t) = \#\{j \geq 2 \mid \hat{\beta}(t)_j - \hat{\beta}(t)_{j-1} = 0\} + 1,$
$m_3(t) = \#\{j \mid \hat{\beta}(t)_j = 0\} + \#\{j \geq 2 \mid \hat{\beta}(t)_j - \hat{\beta}(t)_{j-1} = 0, \hat{\beta}(t)_{j-1}, \hat{\beta}(t)_j \neq 0\} + 2$

である．

4 Group Lasso

β を $\beta = (\beta'_{[1]}, \ldots, \beta'_{[J]})'$ と分割する．ここで，$\beta_{[j]}$ は $p_j \times 1$ ベクトルである．V_j を $p_j \times p_j$ 正定値対称行列とする．このとき，制約集合は

$$K = \{\beta \in \mathbb{R}^p \mid \sum_{j=1}^{J} (\beta'_{[j]} V_j \beta_{[j]})^{1/2} \leq t\},$$

で与えられる．ここでは，$V_j = I_{p_j}$ なるケースを考える．また，$X'X = I_p$ とする．このとき，自由度の不偏推定量は，

$$\widetilde{df}(t) = \begin{cases} \sum_{j=1}^{J} I(\|\hat{\beta}(t)\|_{[j]} > 0) + \sum_{j=1}^{J}(p_j - 1)\frac{\|\hat{\beta}(t)\|_{[j]}}{\|\hat{\beta}^{\circ}\|_{[j]}} - 1 & \text{if } \hat{\beta}^{\circ} \notin K, \\ p & \text{if } \hat{\beta}^{\circ} \in K \end{cases}$$

で与えられる．ただし，$\|\beta\|_{[j]} = (\beta'_{[j]}\beta_{[j]})^{\frac{1}{2}}$ である．

参考文献

[1] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross validation. *J. Amer. Statist. Assoc.* **99** 619-632.

[2] Kato, K. (2008). On the degrees of freedom in shrinkage estimation. Submitted for publication.

[3] Kuriki, S. and Takemura, A. (2000). Shrinkage estimation towards a closed convex set with a smooth boundary. *J. Multivariate Anal.* **75** 79-111.

[4] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135-1151.

[5] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **58** 267-288.

[6] Tibshirani, R., Saunders, M., Rosset, S., Zu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 91-108.

[7] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49-67.

[8] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the Lasso. *Ann. Statist.* **35** 2173-2192.