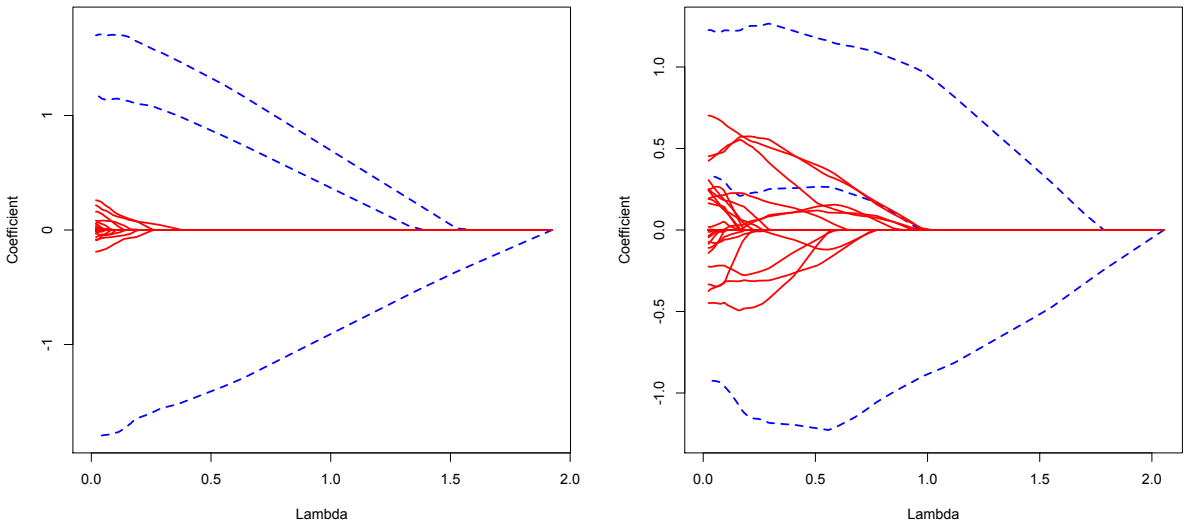
スパース正則化法によるロバスト高次元回帰

東京工業大学　片山 翔太
統計数理研究所　藤澤 洋徳

1　はじめに

近年，変数の次元が大きな場合 (高次元) の線形回帰問題が，その応用範囲の広さから重要視されている．このような場合，通常の最小二乗推定量は利用できないため，Thibshirani (1996) の提案した Lasso (least absolute shrinkage and selection operator) と呼ばれる方法が注目を浴びている．Lasso は変数の選択と推定を同時に行うことが可能で，高次元でも，結果変数に関係する説明変数の個数が少ない場合には性能がよいことが知られている (例えば Bickel et al. (2009), Meinshausen and Yu (2009), Wainwright (2009) など)．しかしながら，Lasso は通常の二乗損失に ℓ_1 スパース制約を課して推定を行っているため，外れ値が結果変数のなかに混在する場合，著しく性能が低下する．次の図は，Lasso の solution path (Lasso 推定値をチューニングパラメータの関数としてみたときのパス) を描いたものである．左図がクリーンなデータに対するもので，右図が外れ値が混入した場合である．青線 (破線) がそれぞれ，真の回帰係数が非 0 のパスを表しており，赤線 (実線) が 0 のパスを表している．この図から，外れ値が混入する場合は，非 0 のパスが 0 のパスの中に隠れてしまうことが確認できる．



通常，線形回帰を外れ値に対してロバストにするためには，二乗損失を Huber 損失や Hampel 損失 (Huber and Ronchetti (2009)) に変更して推定する，M-推定法がとられる．

しかし，これらの損失にスパース制約を課して推定を行う場合，計算コストが高くなるといった問題点が生じる．計算コストの観点からは二乗損失を利用したい．そこで She and Owen (2011) は，線形回帰モデルに新たに外れ値パラメータを追加して，外れ値パラメータと回帰係数パラメータを二乗損失を使って推定を行う方法を提案している．この方法に基づけば，計算コストを大きくせずにロバストかつスパースな回帰係数の推定値が得られる．ただし She and Owen (2011) では，得られる回帰係数の推定値に対する統計的性質を明らかにしていない．

本報告の目的は，She and Owen (2011) で提案された，外れ値を含む線形回帰モデルを基に，高次元でも利用可能な推定量を提案し，その統計的性質を明らかにすることである．

2　ロバスト高次元線形回帰

2.1　モデルとパラメータ推定

本報告では，次のモデルを考える．

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \sqrt{n}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}. \tag{2.1}$$

ここで $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ は n 次元結果変数ベクトル，$\boldsymbol{X} = (x_{ij}) = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ は $n \times p$ 説明変数行列，$\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ は p 次元回帰係数ベクトル (未知)，$\boldsymbol{\gamma}^* = (\gamma_1^*, \ldots, \gamma_n^*)^T$ は n 次元の未知パラメータベクトルで，各要素が外れ値に対応する．すなわち，非 0 の γ_i^* が外れ値の混入を意味する．また，$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ は n 次元の誤差ベクトルである．説明変数行列 $\boldsymbol{X}$ の各列の ℓ_2 ノルムは $\sqrt{n}$ に設定し，それに対応して $\boldsymbol{\gamma}^*$ の係数も $\sqrt{n}$ に設定する．高次元での推定精度を保証するために，Lasso と同様に，$\boldsymbol{\beta}^*$ にスパースネスの仮定をおく．これは，$\boldsymbol{\beta}^*$ の成分の幾つかが 0 であるという仮定である．また，$\boldsymbol{\gamma}^*$ にもスパースネスを仮定する．外れ値の個数は一般的にそれほど多くなく，これは自然な仮定である．これらのスパースネスの仮定から，次のようにパラメータ $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ の推定を行う．

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p, \boldsymbol{\gamma}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \sqrt{n}\boldsymbol{\gamma}\|_2^2 + \lambda_\beta \sum_{j=1}^{p} w_{\beta,j}|\beta_j| + \lambda_\gamma \sum_{i=1}^{n} w_{\gamma,i} P(\gamma_i). \tag{2.2}$$

ここで $(\lambda_\beta, \lambda_\gamma)$ は $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ のチューニングパラメータであり，$P(\cdot)$ はスパース制約である．外れ値へのスパース制約 $P(\cdot)$ に関しては，大きな外れ値の影響からもロバストにするために，再下降的 (redescending) な制約を含むよう，様々なものを考える．なお (2.2) においては，Zou (2006) で提案された，適応型 Lasso で推定を行っており，$w_{\beta,i}$ と $w_{\gamma,j}$ は既知の重みである．一般に，初期推定量 $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p)^T$, $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_n)^T$ が与えられたとき，重みは $w_{\beta,j} = 1/|\tilde{\beta}_j|$, $w_{\gamma,i} = 1/|\tilde{\gamma}_i|$ のように定義される．

2.2 最適化アルゴリズム

式 (2.2) における目的関数を $L(\boldsymbol{\beta}, \boldsymbol{\gamma})$ とする．目的関数は 2 つのパラメータを持っているため，次の交互最適化アルゴリズムを考える．

Algorithm 1

Step 1. Initialize $k \leftarrow 0$, $\boldsymbol{\beta}^k \leftarrow \boldsymbol{\beta}^{init}$ and $\boldsymbol{\gamma}^k \leftarrow \mathrm{argmin}_{\boldsymbol{\gamma}} L(\boldsymbol{\beta}^0, \boldsymbol{\gamma})$.

Step 2. Update $k \leftarrow k + 1$,

$$\boldsymbol{\beta}^k \leftarrow \mathrm{argmin}_{\boldsymbol{\beta}} L(\boldsymbol{\beta}, \boldsymbol{\gamma}^{k-1}), \tag{2.3}$$

$$\boldsymbol{\gamma}^k \leftarrow \mathrm{argmin}_{\boldsymbol{\gamma}} L(\boldsymbol{\beta}^k, \boldsymbol{\gamma}). \tag{2.4}$$

Step 3. If they converge, then output current $(\boldsymbol{\beta}^k, \boldsymbol{\gamma}^k)$ and stop the algorithm, otherwise return to Step 2.

更新 (2.3) は単なる Lasso の最適化問題になっており，例えば Coordinate descent アルゴリズム (Friedman et al. (2010)) などによって解かれる．また，更新 (2.4) は次のように書き直せる．

$$\underset{\boldsymbol{\gamma} \in \mathbb{R}^n}{\mathrm{argmin}} \frac{1}{2n} \sum_{i=1}^{n} \big\{ (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^k - \sqrt{n}\gamma_i)^2 + \lambda_\gamma w_{\gamma,i} P(\gamma_i) \big\}.$$

いま，最適化問題 $\mathrm{argmin}_x (z - x)^2/2 + \lambda P(x)$ の解を $\Theta(z; \lambda)$ とすると，更新 (2.4) は

$$\gamma_i^k \leftarrow \frac{1}{\sqrt{n}} \Theta(y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}^k; \lambda_\gamma w_{\gamma,i}), \quad i = 1, \ldots, n,$$

のように陽に解かれる．ここで γ_i^k は $\boldsymbol{\gamma}^k$ の第 i 成分である．この更新を $\boldsymbol{\gamma}^k \leftarrow h(\boldsymbol{\beta}^k)$ と書くと，上記のアルゴリズムは次のような $\boldsymbol{\beta}$ だけの更新式として書き直すことができる．

$$\boldsymbol{\beta}^k \leftarrow \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\mathrm{argmin}} L\big\{ \boldsymbol{\beta}, h(\boldsymbol{\beta}^{k-1}) \big\}.$$

関数 Θ を閾値関数 (thresholding function) と呼ぶことにする．閾値関数はスパース制約 $P(\cdot)$ に対応してさまざまな形をとる．例えば ℓ_1 制約に対しては $\Theta(z; \lambda) = \mathrm{sgn}(z) \max(|z| - \lambda, 0)$，$\ell_0$ 制約に対しては $\Theta(z; \lambda) = zI(|z| > \lambda)$，Fan (1997) で提案された SCAD(smoothly clipped absolute deviation) 制約に対しては

$$\Theta(z; \lambda) = \begin{cases} \mathrm{sgn}(z)(|z| - \lambda) & \text{if } |z| \le 2\lambda \\ \frac{(a-1)z - a\lambda \mathrm{sgn}(z)}{a-2} & \text{if } 2\lambda < |z| \le a\lambda \\ z & \text{if } |z| > a\lambda, \end{cases}$$

となる．ここで $a = 3.7$ が Fan (1997) で推奨されており，本報告でもその値を用いることにする．

2.3 M-推定法との対応関係

She and Owen (2011) と同様に，本報告で提案する適応型 Lasso 推定量も，M-推定法との対応関係を持っている．

命題 2.1 $\hat{\boldsymbol{\beta}}$ を Algorithm 1 の出力とし，$\psi(z;\lambda) = z - \Theta(z;\lambda)$ とする．このとき，

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}\psi(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}; \lambda_\gamma w_{\gamma,i}) + \lambda_\beta w_{\beta,j}\partial|\hat{\beta}_j| = 0 \tag{2.5}$$

が成立つ．ここで $\partial|\hat{\beta}_j|$ は $|\hat{\beta}_j|$ の劣微分，すなわち，$\hat{\beta}_j \neq 0$ のとき $\partial|\hat{\beta}_j| = \mathrm{sgn}(\hat{\beta}_j)$，それ以外で $\partial|\hat{\beta}_j| \in [-1, 1]$ となる．

命題 2.1 より，Algorithm 1 は次の M-推定法と関連していることが分かる．

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\mathrm{argmin}} \frac{1}{2n}\sum_{i=1}^{n} \Psi(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}; \lambda_\gamma w_{\gamma,i}) + \lambda_\beta \sum_{j=1}^{p} w_{\beta,j}|\beta_j|. \tag{2.6}$$

ここで $\frac{d}{dt}\Psi(t;\lambda) = \psi(t;\lambda)$ である．すなわち，Algorithm 1 の出力 $\hat{\boldsymbol{\beta}}$ と (2.6) の解は共に推定方程式 (2.5) を満たす．なお，M-推定法 (2.5) との対応関係は，$\psi(z;\lambda) = z - \Theta(z;\lambda)$ に基づいており，ℓ_1 制約と Huber 損失，ℓ_0 制約と Skipped-Mean 損失，SCAD 制約と Hampel 損失が例えば対応する．

3 統計的性質

最適化 (2.2) は一般に非凸となる．そのため，Algorithm 1 で得られる出力 $\hat{\boldsymbol{\beta}}$ は (2.2) の局所解となる可能性がある．スパース制約に基づく推定法に関する統計的性質の多くは，大域解に対してのみ導出されており，実際のアルゴリズムで得られる局所解が，その統計的性質を持つかどうかは一般に不明である．この問題に対処するため，本報告では，Algorithm 1 における第 k ステップ解 $\boldsymbol{\beta}^k$ に対する性質を直接導出することを目指す．

3.1 記号の定義

真のパラメータの台をそれぞれ，$S^* = \mathrm{supp}(\boldsymbol{\beta}^*) = \{i|\,\beta_i^* \neq 0\} \subset \{1,\ldots,p\}$, $G^* = \mathrm{supp}(\boldsymbol{\gamma}^*) \subset \{1,\ldots,n\}$ と定義し，それらの要素数をそれぞれ $s^* = |S^*|$, $g^* = |G^*|$ と定義

する．初期推定量に対しても同様にして，$\tilde{S} = \mathrm{supp}(\tilde{\boldsymbol{\beta}})$, $\tilde{G} = \mathrm{supp}(\tilde{\boldsymbol{\gamma}})$, $\tilde{s} = |\tilde{S}|$, $\tilde{g} = |\tilde{G}|$ を定義する．また制限付き最小固有値

$$\delta_{min}(u) = \inf_{\|\boldsymbol{\delta}\|_0 \leq u} \frac{\|\boldsymbol{X}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_2^2} \tag{3.1}$$

および (2 重) 制限付き最大固有値

$$\delta_{max}(u, u') = \sup_{\|\boldsymbol{\delta}\|_0 \leq u} \sup_{|G| \leq u'} \frac{\|\boldsymbol{X}_{(G)}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}\|_2^2} \tag{3.2}$$

を定義する．ここで $\boldsymbol{X}_{(G)} = \{x_{ij} |\, i \in G, 1 \leq j \leq p\}$, $G \subset \{1, \ldots, n\}$ である．

3.2 条件

出力 $\boldsymbol{\beta}^k$ に対する統計的性質を導くために，次の条件を用意する．

条件 **1** ある定数 $\sigma > 0$ に対して $\mathbb{E}(\varepsilon_i) = 0$, $\mathbb{E}\{\exp(t\varepsilon_i)\} \leq \exp(t^2\sigma^2/2)$ for all $t \in \mathbb{R}$.

条件 **2** $\Theta(x; \lambda) = 0$ if $|x| \leq \lambda$, $|\Theta(x; \lambda) - x| \leq \lambda$ for all $x \in \mathbb{R}$.

条件 **3** $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 + \|\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2 \leq \tilde{C}a_{n,1}$, $\delta_{min}(\tilde{s}) \geq \kappa$ が高い確率で成立つようなある数列 $a_{n,1} \to 0$ と定数 $\kappa > 0$ が存在する．

条件 **4** $\min\left\{ \min_{j \in S^*} |\beta_j^*|, \min_{i \in G^*} |\gamma_i^*| \right\} > \tilde{C}a_{n,1}$.

条件 1 は誤差に劣ガウス分布を仮定するもので，例えばガウス分布，2 項分布，任意の台が有界な分布が含まれる．条件 2 は閾値関数に対する仮定で，これまでに挙げた ℓ_1, ℓ_0, SCAD 制約などが含まれる．条件 3 の最初は一致性を持つ初期推定量を要求しており，残りは $\tilde{s}$ がそれほど大きくないことを要求している．実際，$\tilde{s}$ が p に近く，$n < p$ である場合は κ がほとんど 0 になってしまう．条件 4 は真のパラメータに関する条件であり，ある程度の大きさを持つことを要求する．

また，技術的な理由から，定数 $R_w > 0$ を用いて次のように重みを再定義する．

$$w_{\beta,j} = \max\left(\frac{1}{|\tilde{\beta}_j|}, \frac{1}{R_w}\right), \quad w_{\gamma,i} = \min\left(\frac{1}{|\tilde{\gamma}_i|}, R_w\right)$$

ただし $i \in \tilde{G}$, $j \in \tilde{S}$ である．この再定義により，$\min_{j \in \tilde{S}} w_{\beta,j} \geq R_w^{-1}$, $\max_{i \in \tilde{G}} w_{\gamma,i} \leq R_w$ となるが，これらの条件が以下の性質の導出に必要となる．

3.3 Algorithm 1 の出力に対する統計的性質

以上の条件と再定義した重みの下で，$\boldsymbol{\beta}^k$ の収束レートに関する次の定理が成立つ．

定理 3.1 条件 *1–4* を仮定し，定数 $C > \sqrt{2}$ に対して

$$\lambda_\beta \geq 2CR_w\sqrt{\frac{\sigma^2 \log p}{n}}, \quad \lambda_\gamma \leq \frac{Cn}{R_w \max_{j\in\tilde{S}} \sum_{i\in\tilde{G}} |x_{ij}|}\sqrt{\frac{\sigma^2 \log p}{n}}$$

とする．このとき任意のステップ数 $k \geq 1$ と任意の初期値 $\boldsymbol{\beta}^{init}$ に対して

$$\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^*\|_2 \leq \rho^k \|\boldsymbol{\beta}^{init} - \boldsymbol{\beta}^*\|_2 + 2\kappa^{-1}\sqrt{s^*}\lambda_\beta\big(R_w^{-1} + \max_{j\in S^*} w_{\beta,j}\big)\sum_{i=0}^{k-1}\rho^i, \tag{3.3}$$

が高い確率で成立つ．ここで $\rho = 2\kappa^{-1}\delta_{max}(\tilde{s}, \tilde{g})$ である．

定理 3.1 における (3.3) の右辺第一項は，アルゴリズム的な誤差と見なすことができ，第二項は統計的な誤差とみなすことができる．また，第一項は $\rho < 1$ のときに，k が増加するにしたがって指数的に減少する．定理 3.1 により，適当な条件をさらに課すと，

$$\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^*\|_2 \leq C\sqrt{s^*}\lambda_\beta, \quad \mathrm{supp}(\boldsymbol{\beta}^k) = \mathrm{supp}(\boldsymbol{\beta}^*)$$

が高い確率で成立つ．最初の性質は $\boldsymbol{\beta}^k$ が Oracle レート，すなわち，外れ値 $\sqrt{n}\boldsymbol{\gamma}^*$ をあらかじめモデルから除いて Lasso を適用した際の推定値の収束レートを達成できることを示している．もうひとつの性質は，出力 $\boldsymbol{\beta}^k$ の台が真の台に正確に一致することを示している．

4 初期推定量

適応型 Lasso 推定量 (2.2) を用いるためには，初期推定量 $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}})$ が必要である．本報告では次の初期推定量を考える．

$$\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\gamma}}) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\mathbb{R}^{np}} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\theta}\|_2^2 + \lambda_\theta\|\boldsymbol{\theta}\|_1.$$

ここで $\boldsymbol{Z} = (\boldsymbol{X}, \sqrt{n}\boldsymbol{I}_n)$ である．明らかにこの推定量は Lasso 型であり，一般的な仮定 (Bickel et al.(2009)) の下で

$$\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq C\sqrt{\frac{(s^* + g^*)\log\max(n,p)}{n}}, \tag{4.1}$$

$$|\mathrm{supp}(\tilde{\boldsymbol{\theta}})| \leq C\xi_{max}(\boldsymbol{Z}^T\boldsymbol{Z}/n)(s^* + g^*), \tag{4.2}$$

が高い確率で成立つ．ここで$\xi_{max}(\cdot)$は最大固有値である．しかし，(4.2)には$\xi_{max}(\boldsymbol{Z}^T\boldsymbol{Z}/n)$が含まれており，supp$(\tilde{\boldsymbol{\theta}})$の要素数が大きくなってしまう可能性がある．そこで$\tilde{\boldsymbol{\theta}}$の要素の中で小さいものをさらに0に縮小する推定量

$$\tilde{\theta}_j^{th} = \tilde{\theta}_j I(|\tilde{\theta}_j| > \tau_\theta \lambda_\theta), \quad j = 1, \ldots, n+p$$

も考える．この推定量は実際，収束レート(4.1)を保ちつつ，$|\text{supp}(\tilde{\boldsymbol{\theta}}^{th})| \leq C'(s^* + g^*)$となることが示される．

参考文献

[1] Bickel, P., Ritov, Y. and Tsybakov, A. (2009) Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* 37: 1705–1732.

[2] Fan, J. (1997). Comments on "wavelets in statistics: A review" by A. Antoniadis. *Journal of the American Statistical Society*. 6: 131–139.

[3] Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 33: 1–22.

[4] Huber, P. and Ronchetti, E. (2009). *Robust statistics, 2nd edition*. Wiley, New York.

[5] Meinshausen, N., and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*. 37: 246–270.

[6] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*. 58: 267–288.

[7] Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions Information Theory*. 55: 2183–2202.