数理解析研究所講究録
第1908巻 2014年 39-48

スパース推定における確率集中不等式

東京工業大学大学院 情報理工学研究科　　鈴木 大慈 (Taiji Suzuki)
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

概要

スパース推定の精度解析において有用な確率集中不等式を紹介する．また，確率集中不等式を用いて実際に精度の上界を求める．その際に，スパース推定の理論解析において標準的ないくつかのデザイン行列の条件を紹介する．L_1 正則化を用いることで，高次元推定問題においても真のパラメータの非ゼロ要素がサンプル数より十分小さければ，良い推定精度がえられることが示される．また，量子トモグラフィーへの応用として，低ランク密度行列の推定に関する理論も紹介する．

1 はじめに

モデルの次元がサンプル数より多いような高次元推定問題において，真のパラメータのスパース性を利用したスパース推定が有用である．本ノートでは，確率集中不等式を用いたスパース推定の理論を紹介する．特にスパース推定の中でも最も基本的な Lasso 推定量 (Tibshirani, 1996) に焦点を当て，その推定精度の上界を与える．

ここで用いる技法は，基本的に Bickel et al. (2009) に沿っている．彼らは Lasso の精度解析において，デザインの条件に restricted eigenvalue 条件を導入した．他のスパース推定の精度解析においても同様の条件が仮定されることが多く，この条件はその意味で標準的である．本ノートでは，restricted eigenvalue 条件およびそれより弱い compatible 条件を用いて，Lasso の推定精度を解析する．

Lasso は線形回帰において用いられるが，本ノートでは量子トモグラフィーへの応用として，密度行列の推定についてもその理論解析を与える．量子トモグラフィーにおいては，サイズが大きいが（ほぼ）低ランクであるような密度行列を推定する問題設定が現れる．そのような問題ではスパース推定の考え方が有用である．ここでは Koltchinskii (2013) による結果を紹介する．

2 問題設定

説明変数の次元を p 次元とし，n 個のサンプル $D_n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ が線形モデル $y_i = x_i^\top \beta^* + \epsilon_i$ に従って i.i.d. で生成されているとする．ここで $\beta^* \in \mathbb{R}^p$ は真の回帰係数で，$\{\epsilon_i\}_{i=1}^n$ は i.i.d. 雑音である (雑音の条件については後で述べる)．今，

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}, \ Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \ \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

とおくと，

$$Y = X\beta^* + \epsilon$$

である．我々の目的は観測されたサンプル D_n から真の回帰係数 β^* を推定することである．そこで，最小二乗法 $\hat{\beta}_{\mathrm{LS}} = (X^\top X)^\dagger X^\top Y$ を用いた場合 [1]，雑音の分散 σ^2 を用いて平均二乗誤差は

$$\frac{1}{n}\mathrm{E}_{\{\epsilon_i\}_{i=1}^n}[\|X\beta^* - X\hat{\beta}_{\mathrm{LS}}\|_2^2] = \sigma^2\frac{p}{n}$$

と評価できる．よって，もしサンプル数に対して次元が高い場合 ($p \gg n$)，最小二乗推定量によって意味のある推定は望めない．

しかし，真がスパースな場合 (β^* の非ゼロ成分の数が p に比べて十分小さい) には，スパース性を積極的に利用したスパース推定が有用である．そのもっとも基本的な方法が Lasso 推定量である:

$$\hat{\beta} := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n\|\beta\|_1.$$

ここで, $\lambda_n > 0$ かつ $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ である．$\lambda_n\|\beta\|_1$ のことを L_1 正則化項と呼ぶ．これはゼロ要素を持つ点で微分不可能であり，そのために解 $\hat{\beta}$ がスパースになりやすい (多くの成分がゼロになりやすい) という性質がある．そのため，たとえ $p \gg n$ であっても，真がスパースで実質推定する必要のあるパラメータ数が小さい場合は，上記の Lasso 推定量を用いることでそれなりに良い推定精度をえることができる．なお，λ_n は正則化定数で雑音の強さに応じて決める定数である．この定数がどのように推定精度に影響するかは後で述べる．

Lasso が実用上有用である点は凸最適化で解ける点である．真の非ゼロ要素がたとえ少なくても，非ゼロ要素の場所の候補は組み合わせの数分だけ存在する．よって，通常のモデル選択を用いることは計算量の面から実用上難しいが，Lasso は凸最適化で容易に解けるため，そのような組み合わせ的計算量を避けることができる．

3 確率集中不等式

Lasso の推定精度を解析する際に，確率集中不等式が有用である．ここでは，最も基本的な二つの不等式，Hoeffding の不等式と Bernstein の不等式，を紹介する．確率集中不等式を用いることにより，ある確率変数列の平均がどれだけその期待値へ近いかを見積もることができる．すなわち，実確率変数列 ξ_i $(i = 1, \ldots, n)$ があった時,

$$\frac{1}{n}\sum_{i=1}^n \xi_i - \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i\right]$$

の上界を与えることができる．

3.1 Hoeffding の不等式

補題 1 (Hoeffding の不等式). $\xi_i, \ldots, \xi_n \in \mathbb{R}$ を平均 0 の独立な実確率変数とし，次で定義されるサブガウシアン性を満たしているとする:

$$\mathrm{E}[e^{t\xi_i}] \le e^{\sigma_i^2 t^2/2} \quad (\forall t \in \mathbb{R},\ i = 1, 2, \ldots, n). \tag{1}$$

[1] ここで $(X^\top X)^\dagger$ は $X^\top X$ のムーア-ペンローズの擬似逆行列である．

すると，

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right| \geq \tau\right) \leq 2\exp\left(-\frac{n\tau^2}{2(\sum_{i=1}^{n}\sigma_i^2/n)}\right),$$

が成り立つ．

Proof. Chebyshev の不等式より，$t>0$ に対して次が成り立つ:

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq \tau\right) \leq \frac{\mathrm{E}[e^{\frac{1}{n}\sum_{i=1}^{n}\xi_i t}]}{e^{\tau t}} = \frac{\prod_{i=1}^{n}\mathrm{E}[e^{\frac{t}{n}\xi_i}]}{e^{\tau t}} \leq \frac{\prod_{i=1}^{n}e^{\frac{t^2\sigma_i^2}{2n^2}}}{e^{\tau t}} = e^{\frac{t^2\sum_{i=1}^{n}\sigma_i^2}{2n^2}-\tau t}.$$

これに，$t=\frac{n^2\tau}{\sum_{i=1}^{n}\sigma_i^2}$ を代入すれば，

$$P\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq \tau\right) \leq \exp\left(-\frac{n\tau^2}{2(\sum_{i=1}^{n}\sigma_i^2/n)}\right),$$

をえる．同様のことを $-\xi_i$ にも適用すれば，題意をえる． □

簡単のため $\sigma_i=\sigma$ $(\forall i)$ として Hoeffding の不等式を書きかえると，確率 $1-\delta$ 以上で

$$\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right| < \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}} \tag{2}$$

が成り立つ．ここで注意されたいのは，裾確率 δ を固定すれば右辺が $O(1/\sqrt{n})$ で落ちてゆくことである．また，δ の右辺への影響は対数オーダーである．これは，ξ_i のサブガウシアン性による．

サブガウシアン性を満たす分布としては，平均 0 の一様分布 $U([-\sigma,\sigma])$ や正規分布 $N(0,\sigma^2)$ などがある．特に，正規分布 $N(0,\sigma^2)$ の場合は

$$\mathrm{E}[e^{t\xi}] = e^{\sigma^2t^2/2}$$

であり，式 (1) が等式で成り立つ．直観的には式 (1) は正規分布より裾が軽いことを意味しており，その意味でサブガウシアン性と呼んでいる．

3.2 Bernstein の不等式

Hoeffding の不等式とはやや異なる条件のもとで成り立つ Bernstein の不等式も有用である．こちらは最初から式 (2) の形式で与える．

補題 2 (Bernstein の不等式). $\xi_i,\ldots,\xi_n \in \mathbb{R}$ を平均 0 の独立な実確率変数とし，ある正の実数 $\sigma>0$, $M>0$ に対して次の性質を満たしているとする:

$$\mathrm{E}[|\xi_i|^m] \leq \frac{m!}{2}\sigma^2M^{m-2} \ (m=2,3,\ldots). \tag{3}$$

すると，$\forall\delta>0$ で，

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i\right| \geq \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}+\frac{M\log(2/\delta)}{n}\right) \leq \delta,$$

が成り立つ．

条件 (3) は ξ が分散 σ^2 を持ち $|\xi|\leq M$ (a.s.) ならば成り立つが，それよりもかなり弱い条件である．

4 Lassoの収束レート

前の節で紹介した確率集中不等式を用いて，Lasso推定量の収束レートを導出しよう．

4.1 デザイン行列と雑音の条件

収束レートを導出するために，いくつかの条件を仮定する．

仮定 1. デザイン行列 X と雑音 ϵ_i は次の条件を満たしていると仮定する:

(i) $\max_{i,j}|X_{ij}| \leq 1$,

(ii) 雑音 ϵ_i はi.i.d. で，次のどちらかを満たす:

(a) ある $\sigma > 0$ が存在し，次が成り立つ:

$$\mathrm{E}[e^{t\epsilon_i}] \leq e^{\sigma^2 t^2/2} \quad (\forall t \in \mathbb{R}).$$

(b) ある $\sigma, M > 0$ が存在し，次が成り立つ:

$$\mathrm{E}[|\epsilon_i|^m] \leq \frac{m!}{2}\sigma^2 M^{m-2} \quad (m = 2, 3, \dots).$$

また，$J := \{j \mid |\beta_j^*| \neq 0\}$ とし，$k := |J|$ は n より十分小さいものと想定する．以下，数学的には $k \ll n$ は仮定する必要はないが，後で導出する誤差の上界が意味を持つためには $k \ll n$ を想定する必要がある．

この仮定のもと，次の補題をえる．

補題 3. 仮定1が成り立っているとする．$\forall \delta > 0$ に対して $\gamma_n = \gamma_n(\delta)$ を次のように定める:

- 条件 (ii-a) に対して:

$$\gamma_n = \sigma\sqrt{\frac{2\log(2p/\delta)}{n}}.$$

- 条件 (ii-b) に対して:

$$\gamma_n = \sigma\sqrt{\frac{2\log(2p/\delta)}{n}} + \frac{M\log(2p/\delta)}{n}.$$

すると，

$$P\left(\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty \geq \gamma_n\right) \leq \delta.$$

が成り立つ．

Proof. まず，

$$P\left(\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty \geq \gamma\right) = P\left(\max_{1\leq j\leq p}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i X_{ij}\right| \geq \gamma\right)$$

$$= P\left(\bigcup_{1\le j\le p}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{ij}\right|\ge\gamma\right\}\right)$$
$$\le \sum_{j=1}^{p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{ij}\right|\ge\gamma\right) \le p\max_{1\le j\le p} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i X_{ij}\right|\ge\gamma\right)$$

に注意する．ここで，$|X_{ij}|\le 1\ (\forall i,j)$ より，固定された j に対して $\xi_i = X_{ij}\epsilon_i$ は条件 (ii-a) もしくは条件 (ii-b) に対して，それぞれ Hoeffding の不等式および Bernstein の不等式の条件を満たす．よって，それぞれに補題 1 または補題 2 を適用し，$\delta \leftarrow \delta/p$ とすれば題意をえる． □

次に，デザイン行列の性質の良さに関する条件を定義する．$A = X^\top X/n$ とする．

定義 4 (Restricted Eigenvalue Condition (RE$(2k,3)$)).

$$\phi_{\mathrm{RE}} = \phi_{\mathrm{RE}}(2k,3) := \inf_{\substack{I\subseteq\{1,\dots,n\},v\in\mathbb{R}^p:\\ |I|\le 2k, 3\|v_I\|_1\ge\|v_{I^c}\|_1}} \frac{v^\top A v}{\|v_I\|_2^2}$$

に対し，$\phi_{\mathrm{RE}} > 0$ が成り立つ．

定義 5 (Compatibility Condition (COM$(J,3)$)).

$$\phi_{\mathrm{COM}} = \phi_{\mathrm{COM}}(J,3) := \inf_{\substack{v\in\mathbb{R}^p:\\ 3\|v_I\|_1\ge\|v_{I^c}\|_1}} k\frac{v^\top A v}{\|v_I\|_1^2}$$

に対し，$\phi_{\mathrm{RE}} > 0$ が成り立つ．

$x\in\mathbb{R}^k$ に対し $\sqrt{k}\|x\| \ge \|x\|_1$ が成り立つことに注意すると，RE$(2k,3)\Rightarrow$ COM$(J,3)$ はすぐにわかる．これらの条件は，X の列の部分集合を取ってきて部分行列を構成すれば列フルランクであり，それは他の列と強い相関を持たないことを意味している．つまり，$X\beta$ のように X を通して β を観測しても，β がほぼスパースであれば β の値がおおよそ推定できると言い換えてもよい．

4.2 収束レート

ここではLasso推定量の収束レートを導出する．これより，仮定1は成り立っているとし，ある $\delta > 0$ に対して $\gamma_n = \gamma_n(\delta)$ は補題3で定義した通りであるとする．すると，次の定理をえる．

定理 6. 任意の $0 < \delta < 1$ に対し，正則化定数を $\lambda_n = 4\gamma_n$ とする．すると，

- COM$(J,3)$ のもと，

$$\frac{1}{n}\left\|X\hat{\beta} - X\beta^*\right\|_2^2 \le \frac{C\sigma^2}{\phi_{\mathrm{COM}}^2}k\lambda_n^2, \tag{4}$$
$$\|\hat{\beta} - \beta^*\|_1^2 \le \frac{C'\sigma^2}{\phi_{\mathrm{COM}}^2}k^2\lambda_n^2, \tag{5}$$

が確率 $1-\delta$ で成り立つ．

- RE$(2k,3)$ のもと，

$$\|\hat{\beta}-\beta^*\|_2^2 \le \frac{C''\sigma^2}{\phi_{\mathrm{RE}}^2} k\lambda_n^2, \tag{6}$$

が確率 $1-\delta$ で成り立つ．

ただし，C, C', C'' は普遍定数である．

Proof. 後半 (式 (6)) から示す．$\hat{\beta}$ の最適性より，

$$\frac{1}{n}\|X\hat{\beta}-Y\|_2^2+\lambda_n\|\hat{\beta}\|_1 \le \frac{1}{n}\|X\beta^*-Y\|_2^2+\lambda_n\|\beta^*\|_1$$

が成り立つ．今，$Y=X\beta^*+\epsilon$ より，上式は次のように書きかえられる:

$$\begin{aligned} &\frac{1}{n}\|X(\hat{\beta}-\beta^*)-\epsilon\|_2^2+\lambda_n\|\hat{\beta}\|_1 \le \frac{1}{n}\|\epsilon\|_2^2+\lambda_n\|\beta^*\|_1 \\ \Rightarrow\ &\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2+\lambda_n\|\hat{\beta}\|_1 \le -\frac{2}{n}\epsilon^\top X(\beta^*-\hat{\beta})+\lambda_n\|\beta^*\|_1. \end{aligned} \tag{7}$$

ここで，補題 3 より，

$$P\left(\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty > \gamma_n\right) \le \delta,$$

が成り立つ．以下では，事象 $\{\left\|\frac{1}{n}X^\top\epsilon\right\|_\infty \le \gamma_n\}$ が起きている上で話を進める．

すると式 (7) より，

$$\begin{aligned} \frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2+\lambda_n\|\hat{\beta}\|_1 &\le \frac{2}{n}\|\epsilon^\top X\|_\infty\|\beta^*-\hat{\beta}\|_1+\lambda_n\|\beta^*\|_1 \\ &\le 2\gamma_n\|\beta^*-\hat{\beta}\|_1+\lambda_n\|\beta^*\|_1 = \frac{\lambda_n}{2}\|\beta^*-\hat{\beta}\|_1+\lambda_n\|\beta^*\|_1. \end{aligned} \tag{8}$$

ここで，インデックス集合 J^c を，

$$|\hat{\beta}_{j_1}-\beta^*_{j_1}| \ge |\hat{\beta}_{j_2}-\beta^*_{j_2}| \ge \ldots |\hat{\beta}_{j_{p-k}}-\beta^*_{j_{p-k}}|$$

となるように降順に並べ替えて，その上から k 個を F とおく:

$$F=\{j_1,\ldots,j_k\}.$$

また，$I=J\cup F$ としておく．すると，$\forall j\in I^c$ で $\beta_j^*=0$ が成り立っているので，

$$\hat{\beta}_{I^c}-\beta^*_{I^c}=\hat{\beta}_{I^c}$$

である．よって，式 (8) より，

$$\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2+\lambda_n\|\hat{\beta}_I\|_1+\lambda_n\|\hat{\beta}_{I^c}\|_1 \le \frac{\lambda_n}{2}(\|\beta^*_I-\hat{\beta}_I\|_1+\underbrace{\|\beta^*_{I^c}-\hat{\beta}_{I^c}\|_1}_{=\|\hat{\beta}_{I^c}\|_1})+\lambda_n\|\beta^*_I\|_1$$

$$\Rightarrow \quad \frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2+\frac{1}{2}\lambda_n\|\hat{\beta}_{I^c}\|_1 \le \frac{\lambda_n}{2}\|\beta_I^*-\hat{\beta}_I\|_1+\lambda_n(\|\beta_I^*\|_1-\lambda_n\|\hat{\beta}_I\|_1)$$
$$\le \frac{\lambda_n}{2}\|\beta_I^*-\hat{\beta}_I\|_1+\lambda_n\|\beta_I^*-\hat{\beta}_I\|_1=\frac{3}{2}\lambda_n\|\beta_I^*-\hat{\beta}_I\|_1, \tag{9}$$

が成り立つ．これより

$$\|\hat{\beta}_{I^c}-\beta_{I^c}^*\|_1=\|\hat{\beta}_{I^c}\|_1 \le 3\|\hat{\beta}_I-\beta_I^*\|_1 \tag{10}$$

をえる．すると，$\hat{\beta}-\beta^*$ は ϕ_{RE} の定義に現れる v の条件を満たしているので，仮定より

$$\phi_{\mathrm{RE}}\|\hat{\beta}_I-\beta_I^*\|_2^2 \le \frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2$$

である．これを式 (9) に代入すると，

$$\phi_{\mathrm{RE}}\|\hat{\beta}_I-\beta_I^*\|_2^2 \le \frac{3}{2}\lambda_n\|\hat{\beta}_I-\beta_I^*\|_1 \le \frac{3}{2}\lambda_n\sqrt{2k}\|\hat{\beta}_I-\beta_I^*\|_2$$

である．よって，

$$\|\hat{\beta}_I-\beta_I^*\|_2 \le \frac{3}{2\phi_{\mathrm{RE}}}\lambda_n\sqrt{2k}$$

をえる．また，F の定義より，

$$\|\hat{\beta}_{I^c}-\beta_{I^c}^*\|_2 \le \sqrt{\|\hat{\beta}_{I^c}-\beta_{I^c}^*\|_\infty\|\hat{\beta}_{I^c}-\beta_{I^c}^*\|_1} \quad (\because \text{Hölder の不等式})$$
$$\le \sqrt{\frac{\|\hat{\beta}_I-\beta_I^*\|_1}{k}\|\hat{\beta}_{I^c}-\beta_{I^c}^*\|_1} \quad (\because F \text{の取り方より})$$
$$\le \sqrt{\frac{3}{k}}\|\hat{\beta}_I-\beta_I^*\|_1 \le \sqrt{6}\|\hat{\beta}_I-\beta_I^*\|_2$$

である．すると，

$$\|\hat{\beta}-\beta^*\|_2 \le \|\hat{\beta}_I-\beta_I^*\|_2+\|\hat{\beta}_{I^c}-\beta_{I^c}^*\|_2$$
$$\le (1+\sqrt{6})\|\hat{\beta}_I-\beta_I^*\|_2$$
$$\le (1+\sqrt{6})\frac{3}{2\phi_{\mathrm{RE}}}\lambda_n\sqrt{2k}.$$

よって題意をえる．

前半の式 (5) は，次のようにして示す．上と同様の議論を $I=J$ として行い，式 (9) と式 (10) をえる．すると，式 (9) から，

$$\frac{\phi_{\mathrm{COM}}}{k}\|\hat{\beta}_I-\beta_I^*\|_1^2 \le \frac{3}{2}\lambda_n\|\hat{\beta}_I-\beta_I^*\|_1$$
$$\Rightarrow \|\hat{\beta}-\beta^*\|_1=\|\hat{\beta}_I-\beta_I^*\|_1+\|\hat{\beta}_{I^c}-\beta_{I^c}^*\|_1 \le 4\|\hat{\beta}_I-\beta_I^*\|_1 \le \frac{6}{\phi_{\mathrm{COM}}}k\lambda_n.$$

再度式 (9) を用いると，

$$\frac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2 \le \frac{C}{\phi_{\mathrm{COM}}}k\lambda_n^2$$

をえる．□

γ_n の定義より，n が十分大きい時には $\lambda_n^2 \simeq \frac{\log(p/\delta)}{n}$ である．これより，上記の結果をまとめて図式化すると以下のようになる．

$$\begin{array}{lcl} \mathrm{RE}(2k,3) & \Rightarrow & \|\hat{\beta}-\beta^*\|_2^2 \le C\dfrac{k\log(p/\delta)}{n}. \\ \Downarrow & & \\ \mathrm{COM}(J,3) & \Rightarrow & \dfrac{1}{n}\|X(\hat{\beta}-\beta^*)\|_2^2 = C\dfrac{k\log(p/\delta)}{n},\ \|\hat{\beta}-\beta^*\|_1^2 = C\dfrac{k^2\log(p/\delta)}{n}. \end{array}$$

これらがある意味でミニマックス最適であることは (Raskutti et al., 2011) で示されている．

本ノートでは線形回帰の場合のみを扱ったが，より一般の線形モデルにおける誤差解析については，Bühlmann and van de Geer (2011) に網羅されている．また，Bühlmann and van de Geer (2011) には他のデザイン行列の条件も紹介されており，それらと本ノートで紹介した条件との関係についてもその詳細が記されている．

5 密度行列推定への拡張

上では高次元ベクトル推定を扱ってきたが，大きなサイズの密度行列の推定理論についても紹介しよう．ここでは，Koltchinskii (2013) の結果を中心に紹介する．ベクトルの場合は真の回帰係数の非ゼロ要素の数が少ないと仮定したが，密度行列の場合は代わりに低ランク性を仮定する．そのように仮定すると，実質的な未知パラメータ数が小さく，次元に比べてサンプル数が十分多くなくても精度よく推定することが可能になる．

5.1 行列版 Bernstein の不等式

高次元行列推定の理論解析で有用な，行列版の Bernstein の不等式を紹介しよう．$m \times m$ エルミート行列全体の集合を $\mathbb{H}_m$ と書く．以下，$X_i \in \mathbb{H}_m$ $(i = 1, \ldots, n)$ を独立（同一とは限らない）なエルミート行列に値を取る確率変数とする．

定理 7 (Tropp (2012)). $\|X_i\|_\infty \le M$, $\mathrm{E}[X_i] = 0$ を仮定する (ただし，$\|\cdot\|_\infty$ は行列のスペクトルノルムである)．ここで，

$$\sigma_n^2 := \frac{1}{n}\|\mathrm{E}[X_1^2 + \cdots + X_n^2]\|_\infty, \tag{11}$$

とすると，次の不等式が成り立つ．

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_\infty \ge t\right) \le 2m\exp\left(-\frac{nt^2}{2(\sigma_n^2 + Mt/3)}\right).$$

先の実確率変数の Bernstein の不等式から導かれた補題 3 と見比べると，p と m が対応していることがわかる．

さらに Koltchinskii (2013) はこれを拡張した不等式を導出している．準備として Orlicz ノルムを定義する．

定義 8. $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ を単調非減少な凸関数で $\psi(0) = 0$ を満たすものとする．すると，実確率変数 ξ の ψ に付随した Orlicz ノルム $\|\xi\|_\psi$ は，

$$\|\xi\|_\psi := \inf\{C > 0 \mid \mathrm{E}[\psi(|\xi|/C)] \leq 1\},$$

と定義される．

たとえば，$\psi(u) = u^p$ とすると，対応する Orlicz ノルムは L_p ノルム $\|\xi\|_\psi = \mathrm{E}[|\xi|^p]^{\frac{1}{p}}$ になる．

上で紹介した行列版 Bernstein の不等式では行列のスペクトルノルムを M で抑えていたが，次の定理はそれを Orlicz ノルムで特徴付けられる量へ拡張したものである．

定理 9 (Koltchinskii (2013))**.** 凸関数 $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ を，単調非減少な凸関数で $\psi(0) = 0$ を満たし，

$$\psi(u) \geq e^u - 1 - u \ (\forall u \geq 1),\ \psi(u) \geq u^p \ (\exists p \geq 1,\ \forall u \geq 0)$$

が成り立っているものとする．$U > 0$ を $\|\|X_j\|_\infty\|_{\psi^2} \leq U$ $(\forall j = 1, \ldots, n)$ なる正の実数とし，ある $\delta \in (0, 2/\psi(1))$ に対し $M := U\psi^{-1}\left(\frac{2U^2}{\delta\sigma_n^2}\right)$ とおく．すると，次が成り立つ:

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^n X_i\right\|_\infty \geq t\right) \leq 2m\exp\left\{-\frac{nt^2}{\max\{2[(1+\delta)\sigma_n^2 + Mt/3], (e-1)M\}}\right\}.$$

5.2 量子トモグラフィーへの応用

Koltchinskii (2013) は拡張された行列版 Bernstein の不等式 (定理 9) を，量子トモグラフィーにおける大規模密度行列推定問題へ応用している．

密度行列の集合を

$$\mathcal{S} := \{S \mid S \in \mathbb{H}_m, S \succeq O, \mathrm{Tr}[S] = 1\}$$

とする．また，$X, Y \in \mathbb{H}_m$ に対し $\langle X, Y\rangle := \mathrm{Tr}[XY]$ とし，$\|X\|_2^2 := \langle X, X\rangle$ $(X \in \mathbb{H}_m)$ とする．今，エルミート行列に値を取る確率変数 $X_1, \ldots, X_n$ は，$\exists U_X > 0$, $\|X_i\|_\infty \leq U_X$ $(i = 1, \ldots, n)$, $\mathrm{E}[X_i] = 0$ かつ

$$\frac{1}{n}\sum_{i=1}^n \mathrm{E}\left[X_i\langle X_i, \rho\rangle\right] = \rho \quad (\forall \rho \in \mathcal{S})$$

を満たすと仮定する．また，$\sigma_X^2 := \max_{1\leq i\leq n}\|\mathrm{E}X_i^2\|_\infty$ と定義する．

真の密度行列を $\rho^* \in \mathcal{S}$ として，独立な n サンプル $\{(X_i, Y_i)\}_{i=1}^n$ が

$$Y_i = \langle X_i, \rho^*\rangle + W_i.$$

に従って観測されているとする．ただし，W_i は観測雑音である．今，ρ^* が閉凸部分集合 $\mathbb{D} \subseteq \mathcal{S}$ に含まれているとし $(\rho^* \in \mathbb{D})$，ρ^* は低ランクであると想定する．

ρ^* の推定量として，次の最適化問題の解 $\hat{\rho}$ を考える:

$$\hat{\rho} = \arg\min_{\rho\in\mathbb{D}}\left\{\|\rho\|_2^2 - 2\left\langle\frac{1}{n}\sum_{i=1}^n Y_iX_i, \rho\right\rangle\right\}. \tag{12}$$

すると，$\hat{\rho}$ は次の性質を有する．

定理 10 (Koltchinskii (2013)). ある $\alpha \geq 1$ を用いて $\psi_\alpha(u) = e^{u^\alpha} - 1$ $(u \geq 0)$ とする．雑音 W_i が，$\mathrm{E}W_i = 0$, $\mathrm{E}W_i^2 \leq \sigma_W^2$ $(\forall i = 1, \ldots, n)$ かつ，

$$2^{1/\alpha}\|W_i\|_{\psi_\alpha} \leq M_W^{(\alpha)} \ (i = 1, \ldots, n)$$

を満たしているとする．すると，$\sigma_W, M_W^{(\alpha)}, \sigma_X, U_X$ に依存した定数 C が存在して，確率 $1 - e^{-t}$ で，

$$\|\hat{\rho} - \rho^*\|_2^2 \leq C\left(\frac{\log(m) + t}{n} + \frac{(\log(m) + t)^2}{n^2}\right)\mathrm{rank}(\rho^*),$$

が成り立つ．

上の定理より，ρ^* が低ランクであれば，$m \gg n$ であっても，n が $\mathrm{rank}(\rho^*)$ や $\log(m)$ より十分大きければ，$\hat{\rho}$ は ρ^* を十分精度よく推定できることがわかる．

6 まとめ

スパース推定の理論を，Bickel et al. (2009) に沿って紹介した．そこでは，Hoeffding の不等式や Bernstein の不等式といった確率集中不等式が有用であった．結果として，適当なデザイン行列の条件のもと，収束レートに次元は対数オーダーでしか効いてこず，真の非ゼロ要素の数が直接的に影響することが示された．スパース推定の理論を体系的にまとめてある文献としては，Bühlmann and van de Geer (2011) がある．

また，スパース推定の量子トモグラフィーへの応用として，低ランク密度行列の推定問題に関する理論 (Koltchinskii, 2013) を紹介した．ここでも，実際の次元よりも真の密度行列のランクが直接的に収束レートを支配することが見てとれた．

References

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, *37*, 1705–1732.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data.* Springer.

Koltchinskii, V. (2013). A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. *9*, 213–226.

Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, *57*, 6976–6994.

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

Tropp, J. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, *12*, 389–434.