高次元平均ベクトルのスパース推定

大阪大学大学院基礎工学研究科　片山 翔太[1]

2つの p 変量正規母集団 $N_p(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}^*)$ $(i=1,2)$ からそれぞれ独立に n_i $(i=1,2)$ 個の独立サンプル $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ $(i=1,2; j=1,\ldots,n_i)$ が得られたとする．本研究では，高次元データ (変数の次元 p が標本サイズ n_i と同じまたは大きなデータ) における，2つの母平均ベクトルの差 $\boldsymbol{\delta}^* = \boldsymbol{\mu}_1^* - \boldsymbol{\mu}_2^*$ に関する統計的推測問題を扱う．高次元データにおける $\boldsymbol{\delta}^*$ の仮説検定問題 $H_0 : \boldsymbol{\delta}^* = \mathbf{0}$ v.s. $H_1 : \boldsymbol{\delta}^* \neq \mathbf{0}$ に対しては，Bai and Saranadasa (1996), Srivastava and Du (2008), Chen and Qin (2010) などによって，高次元データに対しても検出力の高い検定方法がこれまでに様々提案されている．そこで本報告では，帰無仮説 H_0 が棄却された後の問題，すなわち，$\boldsymbol{\delta}^*$ のどの要素が 0 ではなくかつその値はどの程度かを推定する問題を考える．

近年，回帰分析モデルにおいて Lasso (Tibshirani, 1996) とよばれる，変数選択と回帰係数の推定を同時に行うことが可能な方法が提案されている．この方法に基づくと，母共分散行列 $\boldsymbol{\Sigma}^*$ が既知ならば，$\boldsymbol{\delta}^*$ のひとつの推定量は次の目的関数を $\boldsymbol{\delta}$ に関して最小化することによって得られる:

$$Q(\boldsymbol{\delta}; \boldsymbol{\Sigma}^*) = \boldsymbol{\delta}^T \boldsymbol{\Sigma}^{*-1} \boldsymbol{\delta} - 2\boldsymbol{\delta}^T \boldsymbol{\Sigma}^{*-1} \bar{\boldsymbol{\delta}} + 2\tau_{n,p} \|\boldsymbol{\delta}\|_1.$$

ここに，$\bar{\boldsymbol{\delta}} = \bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$ ($\bar{\boldsymbol{X}}_i$ は標本平均) であり，$\|\cdot\|_1$ は ℓ_1 ノルム，$\tau_{n,p}$ はあるチューニングパラメータである．すなわち，$\tau_{n,p}$ が大きければ推定量は 0 に縮小され，$\tau_{n,p}$ が小さければ推定量は $\bar{\boldsymbol{\delta}}$ とほぼ同じ挙動を示す．しかしながら，ほとんどの場合 $\boldsymbol{\Sigma}^*$ は未知であり，そのうえ高次元データの場合，$\boldsymbol{\Sigma}^*$ の一般的な推定量である標本共分散行列は，逆行列が存在しないため用いることができない．そこで本報告では，近年 Cai and Liu (2011) によって提案されている，高次元データに対しても非常に良い性質を持つ $\boldsymbol{\Sigma}^*$ の推定量 $\hat{\boldsymbol{\Sigma}}_T$ を $Q(\boldsymbol{\delta}; \boldsymbol{\Sigma}^*)$ に Plug-In して得られる次の推定量を提案する:

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta} \in \mathbb{R}^p}{\operatorname{argmin}} \, Q(\boldsymbol{\delta}; \hat{\boldsymbol{\Sigma}}_T).$$

このようにして得られる推定量 $\hat{\boldsymbol{\delta}}$ が，真に 0 である要素を正確に 0 と推定しているか (Sign Recovery) および推定精度はどの程度か (Mean Squared Error) については当日報告する．

参考文献

[1] Bai, Z and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 6: 311–329.

[2] Cai, T and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Society.* 106: 672–684.

[3] Chen, S. X. and Qin, Y. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* 38: 808–835.

[4] Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Jounal of Multivariate Analysis.* 99: 386–402.

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

[1]日本学術振興会特別研究員 (DC1)