高欠測データにおけるスパースモデリング

株式会社東芝　　髙田　正彬
統計数理研究所　　藤澤　洋徳
株式会社東芝　　西川　武一郎

はじめに

Lasso をはじめとするスパース回帰は、高次元データに対する非常に有効な手法である。しかしながら、実際の高次元データにはたびたび欠測値が多く含まれる。Convex Conditioned Lasso (CoCoLasso)[1]は、欠測データを扱うスパース回帰手法として提案されたが、欠測が多い場合には推定値が大幅に悪化してしまう。そこで、高次元・高欠測データのための新しいスパース回帰手法を提案する。

提案手法

回帰モデル $y = X\beta + \varepsilon$ $(X \in \mathbb{R}^{n\times p}, y \in \mathbb{R}^n, \beta \in \mathbb{R}^p, \varepsilon \in \mathbb{R}^n)$を考え、$X$に多くの欠測が存在するとする。CoCoLasso では、以下の問題を解く。

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2}\beta^T \tilde{\Sigma}\beta - \tilde{\rho}^T\beta + \lambda\|\beta\|_1, \qquad \tilde{\Sigma} = \underset{\Sigma\geq 0}{\operatorname{argmin}} \left\|\Sigma - S^{\mathrm{pair}}\right\|_{\max}$$

ここで、$S^{\mathrm{pair}}, \tilde{\rho}$ はそれぞれ、Xのペアワイズ共分散行列、Xとyのペアワイズ共分散ベクトルであり、λは正則化パラメータである。ところが、この定式化では、欠測率の高い変数がデータに 1 つでも含まれていると、共分散の推定が非効率になってしまう。そこで、我々は、共分散行列の推定の際に、以下のように重みづけた距離を最小化する方法を提案する。

$$\tilde{\Sigma} = \underset{\Sigma\geq 0}{\operatorname{argmin}} \left\|R \odot (\Sigma - S^{\mathrm{pair}})\right\|_{\mathrm{F}}^2$$

ここで、$R \in \mathbb{R}^{p\times p}$はペアワイズ実測率行列($R_{jk}$は$j$列目と$k$列目のペアワイズ実測率)である。これにより、欠測によるペアワイズ共分散のばらつきをバランスさせることができ、高欠測データに対して安定的にモデルを推定することができる。

結果

理論解析では、共分散行列の非漸近限界を導出し、ペアワイズ実測率行列で重みづけることに最適性があり、CoCoLasso に対して優位性があることを示した。数値実験では、他の手法と比較して、推定誤差を大幅に削減できることを確認した。

[1] Datta, A., & Zou, H. (2017). CoCoLasso for high-dimensional error-in-variables regression. The Annals of Statistics, 45(6), 2400-2426.