2019年6月5日　統計数理研究所　オープンハウス

スパースモデリング・ロバスト統計・欠測データ解析

藤澤　洋徳　　数理・推論研究系　教授　（ものづくりデータ科学研究センター副センター長・
理化学研究所AIP客員研究員・名古屋大学大学院医学系研究科客員教授）

はじめに

本ポスター：　スパース性・外れ値・欠測に同時に対処する手法に関して紹介しています.

研究などの具体的な情報は藤澤のホームページや arXiv をご覧ください.

研究テーマ: ロバスト統計. ダイバージェンス. スパース･モデリング. グラフィカル･モデリング. 非対称分布. 遺伝子発現データ. モデル選択. 混合効果モデル. 経時データ. 欠測データ. 多重検定. 多重代入. クラスタリング. 高欠測データ. 異常検知. など

共同研究やコンサルテーション：　製薬企業・製造業など

大学院生：　思考力とやる気のある学生を歓迎します.

高欠測データに対するLasso（with 東芝）

背景：　ビッグデータでは，たびたび欠測値が非常に多い.

目的：　高欠測データに対して欠測補完なしにLassoを行う.

Lasso　（データは事前に標準化）

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2n}\|y - X\beta\|^2 + \lambda\|\beta\|_1$$
$$= \arg\min_{\beta} \frac{1}{2n}\beta^\top X^\top X\beta - y^\top X\beta + \frac{1}{2n}y^\top y + \lambda\|\beta\|_1$$
$$= \arg\min_{\beta} \frac{1}{2}\beta^\top S_{xx}\beta - S_{yx}\beta + \frac{1}{2}S_{yy} + \lambda\|\beta\|_1$$

ペア毎に共分散を推定：$S_{xx}^{\text{pair}} = (s_{jk}^{\text{pair}})$.　$S_{yx}^{\text{pair}} = (s_{yk}^{\text{pair}})$.

$$s_{jk}^{\text{pair}} = \frac{1}{\#I_{jk}} \sum_{i \in I_{jk}} x_{ij}x_{ik} \qquad I_{jk} = \{i;\ x_{ij} \text{と} x_{ik} \text{が同時に観測}\}$$

ペア毎なら標本サイズが大きいことは期待できる.
SをS^{pair}で置き換えればよいのでは？ 問題点：$S_{xx}^{\text{pair}} > 0$？

CoCoLasso (2017): S_{xx}^{pair}を正定値行列で近似 & 理論解析

$$\tilde{S}_{xx} = \arg\min_{\Sigma_{xx} \geq 0} \|\Sigma_{xx} - S_{xx}^{\text{pair}}\|_{\max}$$

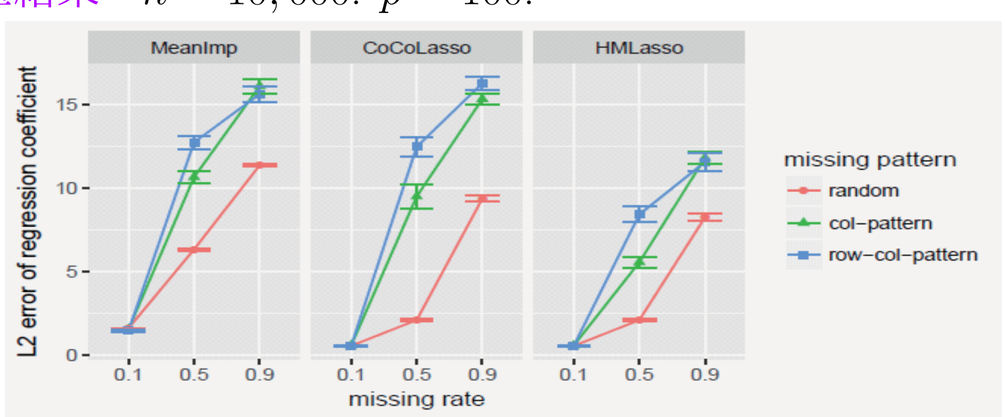
HMLasso (Lasso with High Missing Rate)：提案手法

$$\tilde{S}_{xx} = \arg\min_{\Sigma_{xx} \geq 0} \|R \odot \left(\Sigma_{xx} - S_{xx}^{\text{pair}}\right)\|_F^2 \qquad R = (\#I_{jk}/n)$$

特徴

観測数の違いによる推定量S_{jk}^{pair}の不安定性の違いを考慮
提案した重みは非漸近理論に基づいて最適性を示せる
ADMMにより最適化が可能

数値結果　$n = 10,000.\ p = 100.$



スパース性とロバスト性をもつ回帰手法

背景：　ロバスト性とスパース性を併せもつ手法は推定アルゴリズムが効率的ではない.

目的：　ロバスト性とスパース性を同時に併せ持ち推定精度も良い効率的な推定アルゴリズムの提案

回帰用γ-交差エントロピー　$\gamma > 0$

$$d_\gamma(p_{y|x}, q_{y|x}; p_x) = -\frac{1}{\gamma}\log \int \frac{\int p_{y|x}(y|x)q_{y|x}(y|x)^\gamma dy}{\int q_{y|x}(y|x)^{1+\gamma}dy} p_x(x)dx$$

提案手法　$q_{y|x}(y|x) = f(y|x;\theta) = \phi(y;\beta^\top x, \sigma)$

$$\hat{\theta} = -\frac{1}{\gamma}\log\frac{1}{n}\sum_{i=1}^{n} \frac{f_{y|x}(y_i|x_i;\theta)^\gamma}{\int f_{y|x}(y|x_i;\theta)^{1+\gamma}dy} + \lambda\|\beta\|_1$$

特徴

MMアルゴリズムと座標降下法に基づく推定アルゴリズム
一般的な回帰モデルへ容易に拡張可能
確率的最適化に基づく計算量削減も可能

数値結果

遺伝子発現データ解析
$n = 59.\ p = 22,283.$

Method	RTMSPE	$\hat{p}$
Lasso	1.058	52
RLARS	0.936	18
sLTS	0.721	33
$\gamma = 0.1$	0.679	29
$\gamma = 0.5$	0.700	30

関連研究

上記はGLMへ拡張可能
ロバストでスパースなガウシアン･グラフィカル･モデリング

そのほかの最近の研究

外れ値が変量毎に現れる場合の効率的なガウシアン・グラフィカル・モデリング（この研究は外れ値も欠測もある場合に対応できる.）
主成分回帰モデリング
特徴量の相関構造を積極的に利用したスパース・モデリング
最適な半教師付き学習
多重代入法のバイアス補正
歪正規分布に対するEMアルゴリズム

R Package

汎用性が高い手法はソフトウェア化して頂いています.

gamreg: Robust Regression Modeling with Sparcity
rsggm: Robust Spase Gaussian Graphical Modeling
iilasso: Independently Interpretable Lasso
spcr: Sparse Principal Component Regression
snem: EM Algorithm for Skew-Normal Distribution