クラスタリングと Fused Lasso を用いたグラフィカル Lasso の改良版

大阪大学・基礎工学研究科　　◯張 秉元
大阪大学・基礎工学研究科　　鈴木 譲

1.Graphical Lasso

本研究では，多変量のデータから，変数間の条件付き独立性を見出す問題を検討する．確率変数 $X = (X_1, X_2, \ldots, X_p)^T$ が p 次元の正規分布に従うとすると，X_i と $X_j (i \neq j)$ が条件付き独立であることと，精度行列 $\Theta = \Sigma^{-1}$ の ij 成分が 0 であることが同値になる．その性質を用いると，スパースな（成分に 0 が多い）精度行列を推定することによって，条件付き独立性を調べることができる．そのための手段として，グラフィカル Lasso(glasso) がよく知られている (Friedman et al.， 2008)．

2. 提案手法

Graphical Lasso は，データが一つの正規分布から発生することを仮定する．しかし，現実にはデータが混合分布から発生することが多い．その場合に，同一の分布として取り扱う Graphical Lasso を用いると，条件付き独立性を正しく推定できないことがわかる．この問題を解決するため，本研究では，以下の定式化を考える．

データ $x_i \in \mathbb{R}^p (i = 1, \ldots, n)$ が混合正規分布 $\sum_{k=1}^{K} \pi_k f_k(x_i \mid \mu_k, \Sigma_k)$ から発生するとする．π_k は混合比，$\sum_{k=1}^{K} \pi_k = 1$．各正規分布において，変数の条件付き独立性が類似する場合を考える．その時に，各正規分布の精度行列がスパースになるように推定すると同時に，類似になるように推定したい．

そこで，本研究では，次の式を最大にする問題を考える．

$$\text{maximize}_{\{\Theta\}} \left(\sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k f_k(x_i \mid \mu_k, \Sigma_k) \right] - P(\{\Theta\}) \right) \tag{1}$$

特に，$P(\{\Theta\})$ として，Fused Lasso の罰則項を用いている．

$$P(\{\Theta\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\theta_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\theta_{ij}^{(k)} - \theta_{ij}^{(k')}| \tag{2}$$

上の λ_1 を大きくすると，推定した精度行列 $\hat{\Theta}^{(1)}, \hat{\Theta}^{(2)}, \ldots, \hat{\Theta}^{(K)}$ の 0 成分が多くなる．また，λ_2 を大きくすると，各精度行列の成分が近くなるように推定できる．解を求めるために，$\{(\pi_j, \mu_j, \Sigma_j), j = 1, \ldots, K\}$ を推定することに EM アルゴリズムを用いる．

提案法を用いると，混合分布の場合において，変数同士の条件付き独立性を精度よく推定できることがわかった．また，真の精度行列がスパースになるときに，精度行列の正しく推定することによって，クラスタリングの結果も良くなることが期待できる．数値実験の結果は当日報告する．