ロバストでスパースなグラフィカル・モデリングと遺伝子ネットワーク同定への応用

藤澤　洋徳　数理・推論研究系　教授

はじめに

グラフィカル・モデルを同定する際に，スパース性を利用したパラメータ推定値に基づく手法がよく使われている (graphical lasso; glasso)．その手法をロバスト化して外れ値に強くする (γ-lasso)．外れ値が含まれているデータに γ-lasso を適用すると，外れ値を事前にうまく取り除いたデータに glasso を適用した場合と，同じような結果を提示できる．提案手法 γ-lasso を遺伝子ネットワーク同定へも応用する．

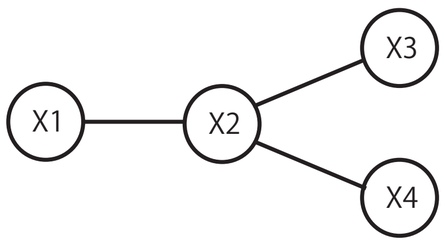
グラフィカル・モデル

分散行列と逆行列

$$\mathrm{Var}[\boldsymbol{X}] = \Sigma. \qquad \Omega = \Sigma^{-1} = (\omega_{jk}).$$

条件付き独立

$$\omega_{jk} = 0 \quad \Leftrightarrow \quad X_j \text{と} X_k \text{は条件付き独立 given } \boldsymbol{X}_{-j,-k}$$



スパースなグラフィカル・モデリング

スパース罰則をもつ罰則付き最尤推定

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}=(\boldsymbol{\mu},\Omega)} \left\{ \frac{1}{n}\sum_{i=1}^{n} -\log\phi(\boldsymbol{x}_i;\boldsymbol{\mu},\Omega^{-1}) + \lambda \sum_{j<k} |\omega_{jk}| \right\}$$

$$\hat{\omega}_{jk} = 0 \quad \Rightarrow \quad X_j \text{と} X_k \text{は条件付き独立 given } \boldsymbol{X}_{-j,-k}$$

パラメータ推定アルゴリズム： glasso (Friedman+2007)

ロバスト化 (γ-lasso)

ロバスト化 (Hirose, Fujisawa, Sese, 2016)

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \arg\min_{\boldsymbol{\theta}} \left\{ \frac{1}{n}\sum_{i=1}^{n} -\log\phi(\boldsymbol{x}_i;\boldsymbol{\mu},\Omega^{-1}) + \lambda \sum_{j<k} |\omega_{jk}| \right\} \\ &= \arg\min_{\boldsymbol{\theta}} \left\{ d_{\mathrm{KL}}(\bar{g},\phi_{\boldsymbol{\theta}}) + \lambda \sum_{j<k} |\omega_{jk}| \right\}\end{aligned}$$

$d_{\mathrm{KL}}(g,f)$ ： 密度関数 g と f のKL相互エントロピー
$\bar{g}$ ： データ $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$ から得られる経験密度関数

ロバスト化

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ d_{\gamma}(\bar{g},\phi_{\boldsymbol{\theta}}) + \lambda \sum_{j<k} |\omega_{jk}| \right\}$$

γ-相互エントロピー (Fujisawa and Eguchi, 2008)

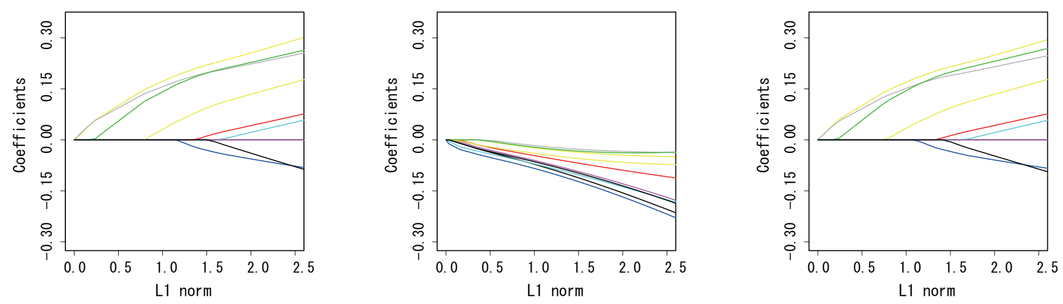
$$d_{\gamma}(g,f) = -\frac{1}{\gamma}\log\int g(x)f(x)^{\gamma}dx + \frac{1}{1+\gamma}\log\int f(x)^{1+\gamma}dx$$

パラメータ推定アルゴリズム： MMアルゴリズム ＋ glasso
R package： rsggm

Solution Path

$n = 200.$　$g = 0.9N(\boldsymbol{0},\Omega^{-1}) + 0.1N(5\boldsymbol{1}_5, I).$

$$\Omega = \begin{pmatrix} 1.0 & 0.3 & 0.3 & 0.0 & 0.0 \\ 0.3 & 1.0 & 0.0 & 0.0 & 0.3 \\ 0.3 & 0.0 & 1.0 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.0 & 1.0 \end{pmatrix}$$



外れ値なしの glasso　外れ値ありの glasso　外れ値ありのγ-lasso

ポイント！： 外れ値がある場合の γ-lasso の結果は，外れ値のない場合の glasso の結果の振る舞いに近い．つまり，γ-lasso は，外れ値はなかったかのように振る舞っている．

遺伝子ネットワーク

$n = 155.$ $p = 11.$
（2次元のPCAから10%程度以上の外れ値が含まれていると見れた.）

左側の6遺伝子と右側の5遺伝子を分離できれば良い．
γ-lasso だけが成功している．