Graphical Models, UWr March 2020
PRACTICAL SELECTION OF THE BEST GRAPHICAL GAUSSIAN MODEL

Let $X = (X_1, \ldots, X_p)^T$ be a Gaussian random vector $N(\xi, \Sigma)$ on $\mathbb{R}^p$ with unknown mean ξ and covariance Σ

We have a sample $X^{(1)}, X^{(2)}, ..., X^{(n)}$ of size n of X.
We want to **do model selection** among all Gaussian graphical models $\mathcal{G} = (V, E)$ with $|V| = p$.
Which graphical model $\mathcal{G} = (V, E)$ with $|V| = p$ fits the best the sample $X^{(1)}, X^{(2)}, ..., X^{(n)}$?

Equivalently,
where to put zeros in the precision matrix $K = \Sigma^{-1}$?

METHOD 1. CASE $n > p$: COMPUTATION OF EMPIRICAL SCALED PRECISION MATRIX $\tilde{K}_{\text{emp}}$

1.1. SAMPLE (EMPIRICAL) COVARIANCE MATRIX:

$$\Sigma_{\text{emp}} = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T \in Sym^{>0}(p \times p)$$

Σ_{emp} is the Max Likelihood Estimator of Σ

1.2. SAMPLE (EMPIRICAL) PRECISION MATRIX: $K_{\text{emp}} = \Sigma_{\text{emp}}^{-1}$

1.3. SAMPLE (EMPIRICAL) SCALED PRECISION MATRIX: $\tilde{K}_{\text{emp}}$, $\tilde{k}_{lm} = \frac{k_{lm}}{\sqrt{k_{ll}}\sqrt{k_{mm}}} = -\rho_{lm|V\setminus\{l,m\}}$.

When $\tilde{k}_{lm} \approx 0$,
we decide $X_l \perp\!\!\!\perp X_m \,|X_{V\setminus\{l,m\}}$ and $k_{lm} = 0$.

2. BIG DATA CASE $\boxed{n < p}$

GRAPHICAL LASSO METHODS

(also possible in the case $n \geq p$)

Big problem when $n < p$: Σ_{emp}^{-1} does not exists, $K_{\text{emp}} = \Sigma_{\text{emp}}^{-1}$ makes no sense

2.0. Shortly on LASSO
(in programme of Big Data Statistics, Master)

Classical Linear Regression problem

$$Y = \mathbf{X}\beta + \varepsilon \qquad (\varepsilon = \text{ noise})$$
$$\widehat{\beta} = \arg\min_{\beta} \|Y - \mathbf{X}\beta\|_2^2$$

– has a unique solution when $n > p$ (classical case)
– has infinity of solutions when $n \leq p$ (Big Data case)

Genius idea of LASSO:
one introduces a penalty $\lambda \sum_{i=1}^{p} |\beta_i| = \lambda\|\beta\|_1$, $\lambda > 0$

$$Y = \mathbf{X}\beta + \varepsilon \qquad (\varepsilon = \text{ noise})$$
$$\widehat{\beta} = \arg\min_{\beta}(\|Y - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1), \qquad (\lambda > 0).$$

Regression LASSO method generates **sparsity**, i.e. a lot of zero coefficients β_i of the vector β in the regression problem.

If λ is bigger, we get more sparsity (more $\beta_i = 0$)

R package: $glmnet(X, Y, alpha = 1)$

Graphical Lasso = G-Lasso

In graphical models there is, in principle, no response variable Y to X (unsupervised learning).

We seek to have zeros in the precision matrix K.

2 methods of Graphical Lasso exist:

- **by Penalized Log-Likelihood (Friedman 2008)**

- **by Regression LASSO for each X_i as response (Meinshausen, Bühlmann 2006)**

2.1. Graphical Lasso via Penalized Log-Likelihood

(d'Aspremont, Banerjee, Ghaoui 2008,
Friedman, Hastie, Tibshirani 2008)

Regression LASSO has an equivalent formulation via maximization of the L^1-Penalized Log-Likelihood. One exploits such formulation for a method of Graphical Lasso.

The likelihood (density) function of the sample
$X^{(1)}, \ldots, X^{(n)}$:
$f(x^{(1)}, \ldots, x^{(n)}; K) = (2\pi)^{-pn/2} (\det K)^{n/2} \exp(-\frac{n}{2} \langle \Sigma_{\text{emp}}, K \rangle)$
where $\Sigma_{\text{emp}} = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T$
(this will be proved in a further lecture)

The log-likelihood function

$$\log f(x^{(1)}, \ldots, x^{(n)}; K) = c + \frac{n}{2} \log \det K - \frac{n}{2} \langle \Sigma_{\text{emp}}, K \rangle)$$

Graphical Lasso via Penalized Log-Likelihood:

$$\hat{K} = \arg\max_{K \in Sym^+(p)} [\log \det K - \langle \Sigma_{\text{emp}}, K \rangle - \lambda \sum_{l \neq m} |k_{lm}|]$$

where $\lambda > 0$, Σ_{emp}= sample covariance matrix.

The penalty is proportional to the L^1-norm of the off-diagonal entries of the precision matrix K.

Fact. The resulting optimal precision matrix $\hat{K}$ has sparsity in off-diagonal terms k_{lm}.

R package: *glasso*

2.2 Regression LASSO for each X_i as response variable to all other $X_{\hat{i}}$ *("'Neighborhood-Based Likelihood")*

(Meinshausen, Bühlmann 2006)

Main Idea. In the linear regression $X_i = \sum_{j \neq i} \beta_{ij} X_j + \varepsilon_i$ we estimate the coefficients β_{ij} by

$$\beta_{ij} = \frac{Cov(X_i, X_j | X_{V \setminus \{i,j\}})}{Var(X_j | X_{V \setminus \{i,j\}})} = \frac{-\kappa_{ij}}{\kappa_{ii}},$$

(Choose X_i, X_j, treat all other variables as fixed,
use $\Sigma_{X_i, X_j | X_{V \setminus \{i,j\}}} = K^{-1}_{\{i,j\}} = \frac{1}{\det K_{\{i,j\}}} \begin{pmatrix} \kappa_{jj} & -\kappa_{ij} \\ -\kappa_{ij} & \kappa_{ii} \end{pmatrix}$.)

Conclusion: $\beta_{ij} = 0$ iff $\kappa_{ij} = 0$.

Method of Meinshausen, Bühlmann:

(i) Apply LASSO to each X_i in turn as the response (apply usual LASSO p times)

(ii) Decide $i \not\sim j$ in the graph $\mathcal{G}$ if both $\beta_{ij} = 0 = \beta_{ji}$.

COMPUTER PROBLEM
5-9 March, 2020

Apply 3 Methods (Method $\tilde{K}_{\text{emp}}$ and 2 methods of graphical Lasso) for the famous Frets' Heads data (1921):

The head dimensions:
length L_i and breadth B_i, $i = 1, 2$
of 25 pairs of first and second sons were measured.

Thus we have $n = 25$ and $p = 4$.

Frets' Heads Data is available in R:
$library(boot)$
$frets$

Table 5.1.1 The measurements on the first and second adult sons in a sample of 25 families. (Data from Frets, 1921.)

First son		Second son	
Head length	Head breadth	Head length	Head breadth
191	155	179	145
195	149	201	152
181	148	185	149
183	153	188	149
176	144	171	142
208	157	192	152
189	150	190	149
197	159	189	152
188	152	197	159
192	150	187	151
179	158	186	148
183	147	174	147
174	150	185	152
190	159	195	157
188	151	187	158
163	137	161	130
195	155	183	158
186	153	173	148
181	145	182	146
175	140	165	137
192	154	185	152
174	143	178	147
176	139	176	143
197	167	200	158
190	163	187	150