MATH 829: Introduction to Data Mining and Analysis
Graphical Models III - Gaussian Graphical Models (cont.)

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

May 6, 2016

Estimating the conditional independence structure of a GGM

During the last lecture, we have shown that when $X \sim N(\mu, \Sigma)$,

1. $X_i \perp\!\!\!\perp X_j$ iff $\Sigma_{ij} = 0$.
2. $X_i \perp\!\!\!\perp X_j \mid \text{rest}$ iff $(\Sigma^{-1})_{ij} = 0$.

- To discover the conditional structure of X, we need to estimate the **structure of zeros** of the precision matrix $\Omega = \Sigma^{-1}$.
- We will proceed in a way that is similar to the lasso.
- Suppose $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$ are iid observations of X. The associated **log-likelihood** of (μ, Σ) is given by

$$l(\mu, \Sigma) := -\frac{n}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^n (x^{(i)}-\mu)^T\Sigma^{-1}(x^{(i)}-\mu) - \frac{np}{2}\log(2\pi).$$

Classical result: the MLE of (μ, Σ) is given by

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^n x^{(i)}, \qquad S := \frac{1}{n}\sum_{i=1}^n (x^{(i)} - \hat{\mu})(x^{(i)} - \hat{\mu})^T.$$

Estimating the CI structure of a GGM (cont.)

- Using $\hat{\mu}$ and $\widehat{\Sigma}$, we can conveniently rewrite the log-likelihood as:

$$l(\mu, \Sigma) = -\frac{n}{2}\log\det\Sigma - \frac{n}{2}\operatorname{Tr}(S\Sigma^{-1}) - \frac{np}{2}\log(2\pi) - \frac{n}{2}\operatorname{Tr}(\Sigma^{-1}(\hat{\mu}-\mu)(\hat{\mu}-\mu)^T).$$

(use the identity $x^T A x = \operatorname{Tr}(Axx^T)$).

- Note that the last term is minimized when $\mu = \hat{\mu}$ (independently of Σ) since

$$\operatorname{Tr}(\Sigma^{-1}(\hat{\mu}-\mu)(\hat{\mu}-\mu)^T) = (\hat{\mu}-\mu)^T\Sigma^{-1}(\hat{\mu}-\mu) \geq 0.$$

(The last inequality holds since Σ^{-1} is positive definite.)

- Therefore the log-likelihood of $\Omega := \Sigma^{-1}$ is

$$l(\Omega) \propto \log\det\Omega - \operatorname{Tr}(S\Omega) \qquad \text{(up to a constant)}.$$

The Graphical Lasso

The Graphical Lasso (glasso) algorithm (Friedman, Hastie, Tibshirani, 2007), Banerjee et al. (2007), solves the **penalized likelihood** problem:

$$\hat{\Omega}_\rho = \underset{\Omega \text{ psd}}{\operatorname{argmax}} \left[\log\det\Omega - \operatorname{Tr}(S\Omega) - \rho\sum_{i,j=1}^p \|\Omega\|_1\right],$$

where $\|\Omega\|_1 := \sum_{i,j=1}^p |\Omega_{ij}|$, and $\rho > 0$ is a fixed regularization parameter.

- Idea: Make a trade-off between maximizing the likelihood and having a sparse Ω.
- Just like in the lasso problem, using a 1-norm tends to introduce many zeros into Ω.
- The regularization parameter ρ can be chosen by cross-validation.
- The above problem can be efficiently solved for problems with up to a few thousand variables (see e.g. ESL, Algorithm 17.2).

The Graphical Lasso (cont.)

- We need to maximize

$$F(\Omega) := \log\det\Omega - \mathrm{Tr}(S\Omega) - \rho \sum_{i,j=1}^{p} \|\Omega\|_1.$$

- Since F is concave, we can use the *sub-gradient* to identify optimal points of F (to be really rigorous, we should be working with $-F$ in order to use the sub-gradient, but the derivation is the same).
- We have

$$\frac{\partial}{\partial\Omega}\log\det\Omega = \Omega^{-1}, \qquad \frac{\partial}{\partial\Omega}\mathrm{Tr}(S\Omega) = S.$$

Also,

$$\partial \sum_{i,j=1}^{p} |\Omega_{ij}| = \mathrm{Sign}(\Omega)$$

where

$$\mathrm{Sign}(\Omega)_{ij} = \begin{cases} 1 & \text{if } \Omega_{ij} > 0 \\ -1 & \text{if } \Omega_{ij} < 0 \\ [-1,1] & \text{if } \Omega_{ij} = 0 \end{cases}.$$

The Graphical Lasso (cont.)

- Putting everything together, we get

$$\partial F = \Omega^{-1} - S - \rho\cdot\mathrm{Sign}(\Omega).$$

- Just like for the lasso problem, we will derive a **coordinate-wise** approach to solve the glasso problem.
- Let $W = \Omega^{-1}$. Write W and Ω in **block form**

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, \qquad \Omega = \begin{pmatrix} \Omega_{11} & \omega_{12} \\ \omega_{12}^T & \omega_{22} \end{pmatrix},$$

where $W_{11}, \Omega_{11} \in \mathbb{R}^{(p-1)\times(p-1)}$.
- We will cyclically optimize F, one column/row at a time.
- Note that since $W\Omega = I$, we have

$$\begin{pmatrix} W_{11}\Omega_{11} + w_{12}\omega_{12}^T & W_{11}\omega_{12} + w_{12}\omega_{22} \\ w_{12}^T\Omega_{11} + w_{22}\omega_{12}^T & w_{12}^T\omega_{12} + w_{22}\omega_{22} \end{pmatrix} = \begin{pmatrix} I_{(p-1)\times(p-1)} & \mathbf{0}_{(p-1)\times 1} \\ \mathbf{0}_{1\times(p-1)} & 0 \end{pmatrix}.$$

The Graphical Lasso (cont.)

- In particular, we have $W_{11}\omega_{12} + w_{12}\omega_{22} = 0$, i.e.,

$$w_{12} = -W_{11}\frac{\omega_{12}}{\omega_{22}} = W_{11}\beta,$$

where $\beta := -\omega_{12}/\omega_{22}$.
- Now, the upper right block of $\Omega^{-1} - S - \rho\cdot\mathrm{Sign}(\Omega)$ is equal to

$$w_{12} - s_{12} + \rho\cdot\mathrm{Sign}(\beta)$$

since $\omega_{22} > 0$.
- We need to choose w_{12} such that

$$0 \in w_{12} - s_{12} + \rho\cdot\mathrm{Sign}(\beta) \Leftrightarrow 0 \in W_{11}\beta - s_{12} + \rho\cdot\mathrm{Sign}(\beta).$$

Observation: in the lasso problem $\min_\beta \frac{1}{2}\|y - Z\beta\|^2 + \rho\|\beta\|_1$, we have

$$\partial\left(\frac{1}{2}\|y - Z\beta\|^2 + \rho\|\beta\|_1\right) = Z^TZ\beta - Z^Ty + \rho\cdot\mathrm{Sign}(\beta).$$

The Graphical Lasso (cont.)

So, we have the two optimality conditions:

- **Glasso update**: $0 \in W_{11}\beta - s_{12} + \rho\cdot\mathrm{Sign}(\beta)$
- **Lasso problem**: $0 \in Z^TZ\beta - Z^Ty + \rho\cdot\mathrm{Sign}(\beta)$

Now, let $Z := W_{11}^{1/2}$, and $y := W_{11}^{-1/2}s_{12}$.
- The glasso update is thus equivalent to solving the lasso problem:

$$\min_\beta \frac{1}{2}\|W_{11}^{-1/2}s_{12} - W_{11}^{1/2}\beta\|_2^2 + \rho\|\beta\|_1.$$

We can therefore solve the glasso problem by cycling through the row/columns of W, and updating them by solving a lasso problem!

The Graphical Lasso (cont.)

We therefore have the following algorithm to solve the glasso problem.

Algorithm 17.2 *Graphical Lasso.*

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains unchanged in what follows.
2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.
   (b) Solve the estimating equations $\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \mathrm{Sign}(\beta) = 0$ using the cyclical coordinate-descent algorithm (17.26) for the modified lasso.
   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$
3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.

ESL, Algorithm 17.2.

MLE estimation of a GGM

- From the glasso solution, one infers a **conditional independence graph** for $X = (X_1, \ldots, X_p)$ by examining the zeros in the estimated inverse covariance matrix.
- Given a graph $G = (V, E)$ with p vertices, let

$$\mathbb{P}_G := \{A \in \mathbb{P}_p : A_{ij} = 0 \text{ if } (i,j) \notin E\}.$$

- We can now estimate the *optimal* covariance matrix with the given graph structure by solving:

$$\hat{\Sigma}_G := \underset{\Sigma\,:\,\Omega=\Sigma^{-1}\in\mathbb{P}_G}{\operatorname{argmax}} \; l(\Sigma),$$

where $l(\Sigma)$ denotes the log-likelihood of Σ.
- Note: Instead of maximizing the log-likelihood over all possible psd matrices as in the classical case, we restrict ourselves to the matrices having the right conditional independence structure.
- The "graphical MLE" problem can be solved efficiently for up to a few thousand variables (see e.g. ESL, Algorithm 17.1).

MLE estimation of a GGM (cont.)

Computing the Gaussian MLE of a multivariate normal random vector with known conditional independence graph G:

Algorithm 17.1 *A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure.*

1. Initialize $\mathbf{W} = \mathbf{S}$.
2. Repeat for $j = 1, 2, \ldots, p$ until convergence:
   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.
   (b) Solve $\mathbf{W}_{11}^*\beta^* - s_{12}^* = 0$ for the unconstrained edge parameters β^*, using the reduced system of equations as in (17.19). Obtain $\hat{\beta}$ by padding $\hat{\beta}^*$ with zeros in the appropriate positions.
   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$
3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = s_{22} - w_{12}^T\hat{\beta}$.

ESL, Algorithm 17.1.

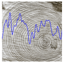
The derivation of the algorithm is similar to the derivation of the glasso algorithm (see ESL, Section 17.3.1).

Application

Example: Estimating the conditional independencies in temperature fields (Guillot et al., 2015)

FIG. 3. *Example of estimated graphical structure of a temperature field (HadCRUT3v).*

Reconstructing climate fields using paleoclimate proxies:



- Estimate conditional independence graph on instrumental period.
- Use an EM algorithm with an embedded graphical model.
- The resulting algorithm is called **GraphEM**.

See Guillot et al.(2015) for more details.