MATH 829: Introduction to Data Mining and Analysis
Graphical Models II - Gaussian Graphical Models

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

May 4, 2016

Recall

- An undirected graphical model is a set of random variables $\{X_1, \dots, X_p\}$ satisfying a Markov property.
- Let $G = (V, E)$ be a graph on $\{1, \dots, p\}$.
- The pairwise Markov property: $X_i \perp\!\!\!\perp X_j \mid \text{rest}$ whenever $(i,j) \notin E$.
- If the density of $X = (X_1, \dots, X_p)$ is continuous and positive, then

$$\text{pairwise} \Leftrightarrow \text{local} \Leftrightarrow \text{global}.$$

- The Hammersley–Clifford theorem provides a necessary and sufficient condition for a random vector to have a Markov random field structure with respect to a given graph G.

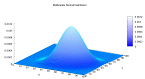
We will now turn our attention to the special case of a random vector with a multivariate **Gaussian** distribution.

Recall: Multivariate Gaussian/normal distribution

Recall: $X = (X_1, \dots, X_p) \sim N(\mu, \Sigma)$ where $\mu \in \mathbb{R}^p$ and $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ is positive definite if

$$P(X \in A) = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \int_A e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}\, dx_1 \dots dx_p.$$

Bivariate case:



We have

$$E(X) = \mu, \quad \operatorname{Cov}(X_i, X_j) = \sigma_{ij}.$$

If $Y = c + BX$, where $c \in \mathbb{R}^p$ and $B \in \mathbb{R}^{m \times p}$, then

$$Y \sim N(c + B\mu, B\Sigma B^T).$$

Note: $\Omega := \Sigma^{-1}$ is called the *precision* matrix or the *concentration* matrix of the distribution.

The Schur complement

Let

$$M := \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

where $A = A_{m\times m}$, $B = B_{m \times n}$, $C = C_{n \times m}$, and $D = D_{n \times n}$.
Assuming D is invertible, the *Schur complement* of D in M is

$$M/D := A - BD^{-1}C.$$

Important properties:

1. $\det M = \det D \cdot \det(M/D)$.
2. $M \in \mathbb{P}_{n+m}$ if and only if $D \in \mathbb{P}_n$ and $M/D \in \mathbb{P}_m$.
   where $\mathbb{P}_k =$ denotes the cone of $k \times k$ real symmetric positive semidefinite matrices.

Proof:

$$M = \begin{pmatrix} I_m & BD^{-1} \\ 0 & I_n \end{pmatrix} \begin{pmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I_m & 0 \\ D^{-1}C & I_n \end{pmatrix}.$$

Multivariate Gaussian/normal distribution (cont.)

- Conditional distribution: if $A \cup B$ is a partition of $\{1, \dots, p\}$, then
$$X_A | X_B = x_B \sim N(\mu_{A|B}, \Sigma_{A|B}),$$
with
$$\mu_{A|B} := \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B),$$
and
$$\Sigma_{A|B} := \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}.$$
- Marginals: to obtain the joint distribution of (X_i, X_j), note that
$$(X_i, X_j)^T = B(X_1, \dots, X_p)^T$$
where
$$B = \begin{pmatrix} I_{2\times 2} & \mathbf{0}_{2\times(p-2)} \end{pmatrix} \in \mathbb{R}^{2\times p}.$$
Therefore
$$(X_i, X_j)^T \sim N(B\mu, B\Sigma B^T),$$
and
$$B\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad B\Sigma B^T = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Multivariate Gaussian/normal distribution (cont.)

Now, suppose
$$X \sim N(\mu, \Sigma)$$
with $\mu \in \mathbb{R}^p$ and $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$ psd.

Claim:

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

Proof of (1):
$$X_i \perp\!\!\!\perp X_j \Leftrightarrow X_i | X_j = x_j \overset{\mathcal{L}}{=} X_i \quad \forall x_j.$$

Now
$$X_i | X_j = x_j \sim N\left(\mu_i + \frac{\sigma_{ii}}{\sigma_{jj}}\rho(x_j - \mu_j), (1-\rho^2)\sigma_{ii}^2\right),$$
where $\rho = \frac{\sigma_{ij}}{\sigma_{ii}\sigma_{jj}}$ is the correlation coefficient between X_i and X_j.
Therefore $X_i \perp\!\!\!\perp X_j$ iff $\rho = 0$ iff $\sigma_{ij} = 0$.

Multivariate Gaussian/normal distribution (cont.)

Proof of (2): Without loss of generality, assume $(i, j) = (1, 2)$. Write μ, Σ in block form according to the partition $A = \{1, 2\}, B = \{3, \dots, p\}$:
$$\mu = (\mu_A, \mu_B)^T, \qquad \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}.$$

Now
$$(X_1, X_2)^T \mid \text{rest} = x_B \sim N(\mu_{A|B}, \Sigma_{A|B}),$$
where
$$\mu_{A|B} := \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B),$$
and
$$\Sigma_{A|B} := \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$$
By part (1), $X_1 \perp\!\!\!\perp X_2 \mid \text{rest}$ iff $(\Sigma_{A|B})_{12} = 0$.

The inverse of a block matrix

Computing the inverse of a block matrix:

9.1.3 The Inverse

The inverse can be expressed as by the use of
$$\mathbf{C}_1 = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} \tag{399}$$
$$\mathbf{C}_2 = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} \tag{400}$$
as
$$\left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right]^{-1} = \left[\begin{array}{c|c} \mathbf{C}_1^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C}_2^{-1} \\ \hline -\mathbf{C}_2^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & \mathbf{C}_2^{-1} \end{array}\right]$$
$$= \left[\begin{array}{c|c} \mathbf{A}_{11}^{-1} + \mathbf{A}_{11}^{-1}\mathbf{A}_{12}\mathbf{C}_2^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & -\mathbf{C}_1^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \hline -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{C}_1^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{C}_1^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{array}\right]$$

Ref.: Petersen and Pedersen, *The matrix cookbook*.

It follows that
$$\Sigma_{A|B}^{-1} = (\Sigma^{-1})_{1:2,1:2}$$

Multivariate Gaussian/normal distribution (cont.)

We have shown

$$\Sigma_{A|B}^{-1} = (\Sigma^{-1})_{1:2,1:2}.$$

Also, we have

$$\begin{pmatrix} a & b \\ b & c \end{pmatrix}^{-1} = \frac{1}{ac-b^2}\begin{pmatrix} c & -b \\ -b & a \end{pmatrix}.$$

Finally,

$$(\Sigma_{A|B})_{12} = 0 \Leftrightarrow (\Sigma_{A|B}^{-1})_{12} = 0 \Leftrightarrow (\Sigma^{-1})_{12} = 0.$$

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

Estimating the conditional independence structure of a GGM

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.
- To discover the conditional structure of X, we need to estimate the **structure of zeros** of the precision matrix $\Omega = \Sigma^{-1}$.
- 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:

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

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

- 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).

MLE estimation of a GGM

- From the glasso solution, one infers a **conditional independence graph** for $X = (X_1, \dots, X_p)$.
- 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).