MODELE GRAFICZNE

Piotr GRACZYK

5. MAXIMUM LIKELIHOOD ESTIMATION

Let X be a Gaussian random vector $N(\xi, \Sigma)$ on $\mathbb{R}^p$

(we consider p variables $X_1, \ldots, X_p$)

with unknown mean ξ and covariance Σ

We dispose of a sample $X^{(1)}, X^{(2)}, ..., X^{(n)}$ of X.

We want to **estimate**:
the unknown mean ξ
the unknown covariance Σ.

CLASSICAL CASE that you know after a course in multivariate statistics: *no information on conditional independence between X_i's.*
(saturated graphical model, complete graph $\mathcal{G}$)

The maximum likelihood estimators are well known:

for the mean ξ, the empirical mean $\widehat{\xi} = \bar{X}$

for the covariance Σ, the empirical covariance

$$\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T$$

These maximum likelihood estimators exist if and only if the matrix $\tilde{\Sigma}$ is **strictly** positive definite. This happens with probability 1 if $n > p$ and never if $n \leq p$.

$\tilde{\Sigma}$ has a Wishart law on the matrix cone $\text{Sym}^+(p, \mathbb{R})$.

This is a matrix analog of KHI^2 law χ^2_{n-1} sur $\mathbb{R}^+$ for $p = 1$.

(C is a cone if $x \in C \;\Rightarrow\; \forall t > 0 \quad tx \in C$)

GAUSSIAN GRAPHICAL MODEL CASE

Estimation under conditional independence between X_i's.

(graphical model with non-complete graph $\mathcal{G}$)

Let $V = \{1, \ldots, p\}$ and let $\mathcal{G} = (V, E)$ be an undirected graph.

Let $\boxed{\mathcal{S}(\mathcal{G}) = \{Z \in Sym(p \times p) |\ i \not\sim j \ \Rightarrow \ Z_{ij} = 0\}}$

$\mathcal{S}(\mathcal{G})$ is the space of symmetric $p \times p$ matrices with **obligatory zero terms** $Z_{ij} = 0$ for $i \not\sim j$

Let $\mathcal{S}^+(\mathcal{G}) = Sym^+(p, \mathbb{R}) \cap \mathcal{S}(\mathcal{G})$ be the open cone of positive definite matrices with obligatory zero terms $Z_{ij} = 0$ for $i \not\sim j$.

Example 1. (Simpson paradox) $X_1 \perp\!\!\!\perp X_2 \mid X_3$
X_1 and X_2 are conditionally independent knowing X_3

Graphe $\mathcal{G}$: 1———3———2

The precision matrix $K = \Sigma^{-1}$ has **obligatory zeros**
$\kappa_{12} = \kappa_{21} = 0$

$$K \in \left\{ \begin{pmatrix} x_{11} & 0 & x_{31} \\ 0 & x_{22} & x_{32} \\ x_{31} & x_{32} & x_{33} \end{pmatrix} \mid x_{11}, x_{22}, x_{31}, x_{32}, x_{33} \in \mathbb{R} \right\} \cap Sym^{+}(3)$$

$\boxed{K \in \mathcal{S}^{+}(\mathcal{G})}$ is a supplementary restriction to the MLE problem

Example 2. Nearest neighbours interaction graph A_4

Graphe $\mathcal{G}$: 1———2———3———4

$$K \in \left\{ \begin{pmatrix} x_{11} & x_{21} & 0 & 0 \\ x_{21} & x_{22} & x_{32} & 0 \\ 0 & x_{32} & x_{33} & x_{43} \\ 0 & 0 & x_{43} & x_{44} \end{pmatrix} \mid x_{11}, \ldots, x_{44} \in \mathbb{R} \right\} \cap Sym^{+}(4)$$

$\boxed{K \in \mathcal{S}^{+}(\mathcal{G})}$ is a supplementary restriction to the MLE problem

GAUSSIAN GRAPHICAL MODEL $\mathcal{G}$

Conditional independence case

n-sample of X $\Rightarrow$ estimation of parameters ξ, Σ of X

In order to formulate the MLE formula, we need the natural **projection** $\boxed{\pi_{\mathcal{G}} : Sym \to \mathcal{S}(\mathcal{G})}$

This projection puts 0 instead of x_{ij} when $i \not\sim j$ in $\mathcal{G}$.

Example 1.(Simpson paradox) $\mathcal{G}$: 1———3———2

$$\pi_{\mathcal{G}}(\begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{21} & x_{22} & x_{32} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}) = \begin{pmatrix} x_{11} & 0 & x_{31} \\ 0 & x_{22} & x_{32} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}$$

Sample $X^{(1)}, \ldots, X^{(n)}$;   each $X^{(i)} \in \mathbb{R}^p$

A natural candidate to estimate Σ is (when $n > p$)

$$\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T$$

but it **does not take into account the restriction** $K = \Sigma^{-1} \in \mathcal{S}^+(\mathcal{G})$

MLE Theorem. Let the graph $\mathcal{G} = (V, E)$ govern the Gaussian graphical model $X = (X_v)_{v \in V} \sim N_p(\xi, \Sigma)$, with precision matrix $K = \Sigma^{-1} \in \mathcal{S}^+(\mathcal{G})$. Consider an n-sample $X^{(1)}, \ldots, X^{(n)}$ of $X \in \mathbb{R}^p$ with $n > p = |V|$. The MLE of the mean is $\hat{\xi} = \bar{X}$.

The MLE $\hat{K} \in \mathcal{S}^+(\mathcal{G})$ of the precision matrix is the unique solution of the equation

$$\pi_{\mathcal{G}}(\hat{K}^{-1}) = \pi_{\mathcal{G}}(\tilde{\Sigma}), \tag{1}$$

where $\tilde{\Sigma}$ is the sample covariance:

$$\tilde{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T$$

The MLE $\hat{\Sigma}$ of Σ is given by $\hat{\Sigma} = \hat{K}^{-1}$.

Proof. Simplified case: known zero mean $\xi = 0$.

$X = (X_1, \ldots, X_p)^T$: random vector obeying $N(0, \Sigma)$

with $\boxed{\textbf{unknown covariance matrix } \Sigma \in Sym^+(p)}$

such that $\boxed{K = \Sigma^{-1} \in \mathcal{S}^+(\mathcal{G})}$

The **likelihood (density) function** of the sample $X^{(1)}, \ldots, X^{(n)}$ equals:

$$f(x^{(1)}, \ldots, x^{(n)}; K) =$$
$$= \prod_{k=1}^{n} \{(2\pi)^{-p/2} (\det K)^{1/2} \exp(-{x^{(k)}}^T K x^{(k)}/2)\}$$
$$= (2\pi)^{-pn/2} (\det K)^{n/2} \exp(-\textstyle\sum_{k=1}^{n} {x^{(k)}}^T K x^{(k)}/2)$$

Note that the real number in the exponent equals its trace. We use the formula $\mathrm{tr}(A_{l\times m} B_{m\times l}) = \mathrm{tr}(B_{m\times l} A_{l\times m})$:

$$\sum_{k=1}^{n} {x^{(k)}}^T K x^{(k)} = \mathrm{tr}\,(\sum_{k=1}^{n} x^{(k)} {x^{(k)}}^T) K = \left\langle n\tilde{\Sigma}, K \right\rangle$$

where $< R, S >$ is the usual scalar product of two symmetric matrices $< R, S >= \sum_{i,j} r_{ij} s_{ij}$.

We explain it on an example 2×2:

$$\left\langle \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \begin{pmatrix} A & B \\ B & C \end{pmatrix} \right\rangle = aA + bB + bB + cC$$

$$trace \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} A & B \\ B & C \end{pmatrix} = (aA + bB) + (bB + cC)$$

$$f(x^{(1)}, \ldots, x^{(n)}; K) = (2\pi)^{-\frac{pn}{2}} (\det K)^{\frac{n}{2}} \exp(-\frac{1}{2} \left\langle n\tilde{\Sigma}, K \right\rangle)$$

Because of $K \in \mathcal{S}^+(\mathcal{G})$, $\left\langle n\tilde{\Sigma}, K \right\rangle = \left\langle \pi_{\mathcal{G}}(n\tilde{\Sigma}), K \right\rangle$.

(recall that K has obligatory zeros when $i \not\sim j$
and $\pi_{\mathcal{G}}$ = projection on $\mathcal{S}(\mathcal{G})$)

We explain it on the example 3×3 of Simpson paradox

$$\left\langle \begin{pmatrix} x_{11} & x_{21} & x_{31} \\ x_{21} & x_{22} & x_{32} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}, \begin{pmatrix} \kappa_{11} & 0 & \kappa_{31} \\ 0 & \kappa_{22} & \kappa_{32} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{pmatrix} \right\rangle =$$

$$\left\langle \begin{pmatrix} x_{11} & 0 & x_{31} \\ 0 & x_{22} & x_{32} \\ x_{31} & x_{32} & x_{33} \end{pmatrix}, \begin{pmatrix} \kappa_{11} & 0 & \kappa_{31} \\ 0 & \kappa_{22} & \kappa_{32} \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{pmatrix} \right\rangle$$

Which $K \in \mathcal{S}^{+}(\mathcal{G})$ is **most likely?**

Maximum Likelihood Estimation $\Rightarrow$
it is $K = \hat{K}$ for which $f(x^{(1)}, \ldots, x^{(n)}; \hat{K})$ is maximum

$\Longleftrightarrow$ $\log f(x^{(1)}, \ldots, x^{(n)}; \hat{K})$ is maximum

$\Longleftrightarrow$ $grad_K \log f(x^{(1)}, \ldots, x^{(n)}; \hat{K}) = 0.$

We study as a function of $K \in \mathcal{S}^+(\mathcal{G})$

$$\log f(x^{(1)}, \ldots, x^{(n)}; K) = c + \frac{n}{2} \log \det K - \frac{n}{2} \left\langle \pi_{\mathcal{G}}(\tilde{\Sigma}), K \right\rangle$$

For M invertible $p \times p$ real matrix we have

$$\boxed{\operatorname{grad} \log \det M = M^{-1}}$$

(EXERCISE: prove this derivation formula)

$K \in \mathcal{S}^+(\mathcal{G})$, so grad_K does not contain $\frac{\partial}{\partial \kappa_{ij}}$ for $i \not\sim j$

$$0 = \operatorname{grad}_K \log f(x^{(1)}, \ldots, x^{(n)}; K) = \frac{n}{2}(\pi_{\mathcal{G}}(K^{-1}) - \pi_{\mathcal{G}}(\tilde{\Sigma}))$$

Equation (1) is obtained: $\boxed{\pi_{\mathcal{G}}(\hat{K}^{-1}) = \pi_{\mathcal{G}}(\tilde{\Sigma})}$.

The existence and unicity of a solution $\hat{K}$ are ensured for $n \geq p$ (when $\mathbf{E}X$ is not given, for $n > p$) by a convexity argument (omitted). $\square$

Example 1.(Simpson paradox) $\mathcal{G}$: 1 —— 3 —— 2
The graph $\mathcal{G}$ governs the model.

Suppose that $n > 3$ and the sample covariance matrix equals $\tilde{\Sigma} = \begin{pmatrix} 1 & 0.5 & 1 \\ 0.5 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$. (check that $\tilde{\Sigma} >> 0$)
We have $(\tilde{\Sigma}^{-1})_{12} = -0.5 \times (-0.5) = 0.25$
so $\tilde{\Sigma}^{-1} \notin \mathcal{S}(\mathcal{G})$ (terms$_{12}$ should be 0 for matrices in $\mathcal{S}(\mathcal{G})$.). Thus $\tilde{\Sigma} \neq \hat{\Sigma}$.

We apply the MLE Theorem.

$$\pi_{\mathcal{G}}(\tilde{\Sigma}) = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}.$$ In order to find $\hat{\Sigma}$, we need to find x such that $$\Sigma_x = \begin{pmatrix} 1 & x & 1 \\ x & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix} \in Sym^+$$ and $\Sigma_x^{-1} \in \mathcal{S}(\mathcal{G})$.

PLEASE DO IT NOW!

$\Sigma_x \in Sym^+ \Leftrightarrow 2 > x^2$ and $\det \Sigma_x = 4x - 3x^2 > 0 \Leftrightarrow 0 < x < \frac{4}{3}$.

The condition $\Sigma_x^{-1} \in \mathcal{S}(\mathcal{G})$ (terms$_{12}$ should be 0) gives $\det \begin{pmatrix} x & 1 \\ 2 & 3 \end{pmatrix} = 0$, so $x = \frac{2}{3}$. By MLE Theorem

$$\widehat{\Sigma} = \Sigma_{\frac{2}{3}} = \begin{pmatrix} 1 & \frac{2}{3} & 1 \\ \frac{2}{3} & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}$$

In practice, when $n > p$, we proceed as follows:

1. We compute the empirical covariance $\tilde{\Sigma}$ from the sample $X^{(1)}, \dots, X^{(n)}$.
We do the projection $\pi_{\mathcal{G}}(\tilde{\Sigma})$.

2. We must find $\hat{K} \in \mathcal{S}^+(\mathcal{G})$ such that $\boxed{\pi_{\mathcal{G}}(\hat{K}^{-1}) = \pi_{\mathcal{G}}(\tilde{\Sigma}).}$

This is a highly non-trivial step. The Theorem says that a **unique solution exists**, but does not say how to find it.

This question is trivial only when $\mathcal{G}$=complete graph.
(Then $\pi_{\mathcal{G}} = id$ and $\hat{K} = \tilde{\Sigma}^{-1}$)

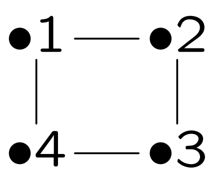
3. Once 2. solved, we compute $\hat{\Sigma} := \hat{K}^{-1}$.
(For $\mathcal{G}$ complete we find the well known MLE $\hat{\Sigma} = \tilde{\Sigma}$)

• An **explicit solution** of the Likelihood Equation (1) $\pi_{\mathcal{G}}(K^{-1}) = \pi_{\mathcal{G}}(\tilde{\Sigma})$ is known on **decomposable (***also called* **chordal** *or* **triangulated)** graphs.
It is expressed by the Lauritzen map.

• On any graphical model, in order to find approximatively a solution of (1), one can perform the **Iterative Proportional Scaling (IPS)** algorithm, which is infinite on non-decomposable graphs.

Decomposable graphs roughly means *decomposable into complete subgraphs connected by complete separators*.

The smallest non-decomposable graph is the square



The Likelihood Equation $\pi_{\mathcal{G}}(K^{-1}) = \pi_{\mathcal{G}}(\tilde{\Sigma})$ is in 2 variables and it leads to a fifth degree equation in x which would be solvable for particular values of $\pi_{\mathcal{G}}(\tilde{\Sigma})$ only.

**TOWARDS BAYESIAN METHODS

In Bayesian statistics, we need to propose a **prior law** on the precision matrix K. The law of MLE may be naturally proposed as a prior law.

• the random matrix $\pi(\tilde{\Sigma}) \in \pi_{\mathcal{G}}(Sym^+(p))$ obeys Wishart law on the cone $\pi_{\mathcal{G}}(Sym^+(p))$.

• the random matrix $K \in \mathcal{S}^+(\mathcal{G})$ such that the Likelihood Equation $\pi_{\mathcal{G}}(K^{-1}) = \pi_{\mathcal{G}}(\tilde{\Sigma})$ holds obeys Wishart law on the cone $\mathcal{S}^+(\mathcal{G})$.

Harmonic (Laplace) analysis on the convex cones is needed to study these Wishart laws (e.g. the density)

The formula for sample density
$f(x^{(1)}, \ldots, x^{(n)}; K) = (2\pi)^{-\frac{pn}{2}} (\det K)^{\frac{n}{2}} \exp(-\frac{1}{2} \left\langle n\tilde{\Sigma}, K \right\rangle)$

suggests using as a prior distribution of K the law with density

$$K \to C (\det K)^{\frac{s}{2}} e^{-\frac{1}{2}\text{tr}(K\theta)}, \quad K \in \mathcal{S}^+(\mathcal{G})$$

where $\theta \in \pi_{\mathcal{G}}(Sym^+(p))$, i.e. only the terms $(\theta_{ij})_{i\sim j}$ are essential. This is a Diaconis-Ylvisaker prior for K.

The computation of the normalizing constant C is crucial for Bayes methods (and uneasy!)