Undirected Graphical Gaussian models: introduction

Elena Stanghellini-Università di Perugia

Department of Statistics - SIMSAM project

Umeå- November, 2012
Part II
Summary of Part II

- Multivariate Gaussian distribution: well-known facts
- Undirected Graphical Gaussian models
- Inference
References

References

   but also
Multivariate normal distribution

Let $\mathbf{X} = (X_1, \ldots, X_p)^T$ a vector of random variables and $\mathbf{x} = (x_1, \ldots, x_p)^T$ be a value. Let $V = \{1, \ldots, p\}$ be the set of nodes associated with $\mathbf{X}$.

$\mathbf{X}$ is jointly normally distributed with $E(\mathbf{X}) = \mu$ e $\text{Var}(\mathbf{X}) = \Sigma$, if the density function of $\mathbf{X}$ is:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} | \Sigma |^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\}$$

for all $\mathbf{x} \in \mathbb{R}^p$.

$E(\mathbf{X}) = \mu$ e $\text{Var}(\mathbf{X}) = \Sigma$. 
The canonical parametrization

- Let $\Omega = \Sigma^{-1}$ and $\beta = \Sigma^{-1} \mu$.
- The density function can then be rewritten:

$$f(x) = \exp\left\{\alpha + \beta^T x - x^T \Omega x / 2\right\}$$

with $\alpha$ a normalizing constant. The $\Omega$ and $\beta$ are the canonical parameters, as:

$$f(x) = \exp\left\{\alpha + \sum_{i} \beta_i x_i - \sum_{i} \sum_{j} x_i x_j \omega_{ij} / 2\right\}.$$

- From the factorization criterion

$$X_i \perp X_j \mid X_{V \setminus \{ij\}} \iff \omega_{ij} = 0.$$
It is possible to show that:

- Each element $a$ of the diagonal of $\Omega = \Sigma^{-1}$ is the inverse of $\text{Var}(X_a \mid X_{V\setminus a})$.
- The correlation coefficients between $X_i$ and $X_j$ after conditioning on $X_{V\setminus \{i,j\}}$ is:

$$-\omega_{ij} / \sqrt{\omega_{ii}\omega_{jj}} = \rho_{ij} \mid V \setminus \{i,j\}.$$  

The last result confirms that

$$\omega_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j \mid X_{V\setminus \{i,j\}}.$$
A graphical Gaussian Model $\mathcal{G} = (V, E)$ is such that

$$\{i, j\} \notin E \iff \omega_{ij} = 0.$$ 

If $\Sigma$ is positive definite, then the equivalence between Markov properties ensures that

$$\mathbf{X}_A \perp \mathbf{X}_B \mid \mathbf{X}_S$$

whenever $S$ separates $A$ and $B$ in $\mathcal{G}$. 

A simple example

We have three r.v.’s $X_A$, $X_S$, $X_B$ with observed covariance matrix $S$:

$$S = \begin{pmatrix}
3.023 & 1.258 & 1.004 \\
1.258 & 1.709 & 0.842 \\
1.004 & 0.842 & 1.116
\end{pmatrix}$$

We want to fit the conditional independence model:
A simple example-continued

We have to estimate $\Sigma$ such that

$$\hat{\Sigma}^{-1} = \begin{pmatrix} * & * & 0 \\ * & * & * \\ 0 & * & * \end{pmatrix}$$

Moreover: since the model is collapsible onto $(A, S)$ and $(S, B)$ standard MLE theory suggests:

- that blocks $\{A, S\}$ and $\{S, B\}$ of $\hat{\Sigma}$ equals the same blocks of $\Sigma$. 

Maximum Likelihood Estimation

Let \( x^i \) be the \( i \)-th observation of a \( N \) dimensional random sample. The loglikelihood is then

\[
l(\mu, \Sigma) = -Np \log(2\pi)/2 - N \log |\Sigma|/2 - \sum_i (x^i - \mu)^T \Omega (x^i - \mu)/2
\]

which simplifies into (exercise):

\[
l(\mu, \Sigma) = -Np \log(2\pi)/2 - N \log |\Sigma|/2 - Ntr(\Omega S)/2 - N(\bar{x} - \mu)^T \Omega (\bar{x} - \mu)/2
\]

with \( S = \frac{1}{N} \sum_i (x^i - \bar{x})(x^i - \bar{x})^T \).
Maximum likelihood estimators

- If the model is saturated, then $\hat{\Sigma} = S$ and $\hat{\mu} = \bar{x}$.

- In a graphical model with missing edges, still $\hat{\mu} = \bar{x}$, but the estimated covariance matrix has to satisfy the constraints

  $$\hat{\omega}_{ij} = 0 \iff (i,j) \notin E$$

  and

  $$\hat{\Sigma}_C = S_C$$

  for all cliques $C$ in $G$.

- If the graph is decomposable, then we may have explicit solutions. Otherwise iterative methods are used, such as **Iterative Proportional Scaling**.
Decomposable graphs

Let $A, B, S$ form a decomposition of $G$. Then:

$$f_V = \frac{f_{AUS} f_{BUS}}{f_S}$$

and therefore:

$$l(\mu, \Sigma) = l(\mu_{AUS}, \Sigma_{AUS,AUS}) + l(\mu_{BUS}, \Sigma_{BUS,BUS}) - l(\mu_S, \Sigma_{SS})$$

Let’s assume for now that $G_{AUS}$ and $G_{BUS}$ are complete. Then the estimated covariance matrix is such that:

- $(\hat{\Omega}_{AUS,AUS})^{-1} = S_{AUS,AUS}$
- $(\hat{\Omega}_{BUS,BUS})^{-1} = S_{BUS,BUS}$
- $(\hat{\Omega}_{SS})^{-1} = S_{SS}$
- $\Omega^{AB} = 0$. 
A simple example -continued

Then the ML estimate $\hat{\Sigma}$ is

$$\hat{\Sigma} = \begin{pmatrix} 3.023 & 1.258 & 0.620 \\ 1.258 & 1.709 & 0.842 \\ 0.620 & 0.842 & 1.116 \end{pmatrix}$$

$$\hat{\Sigma}^{-1} = \begin{pmatrix} 0.477 & -0.351 & 0.000 \\ -0.351 & 1.190 & -0.703 \\ 0.000 & -0.703 & 1.426 \end{pmatrix}$$

To be confronted with:

$$S = \begin{pmatrix} 3.023 & 1.258 & 1.004 \\ 1.258 & 1.709 & 0.842 \\ 1.004 & 0.842 & 1.116 \end{pmatrix}$$

$$S^{-1} = \begin{pmatrix} 0.182 & -0.092 & -0.102 \\ -0.092 & 0.537 & -0.396 \\ -0.102 & -0.396 & 0.803 \end{pmatrix}$$

Note that the fact that the block $\{AS\}$ and $\{BS\}$ coincide in a matrix does not imply that the same block coincide in its inverse.
More generally - Iterative Proportional Fitting

The IPF algorithm is such that, for each clique \( A \) in the graph, let \( \hat{\Omega} \) be the estimated covariance at a given iteration:

\[
\begin{pmatrix}
S_{AA} & * \\
* & *
\end{pmatrix}
= \begin{pmatrix}
\hat{\Omega}^{AA} + R & \hat{\Omega}^{AB} \\
\hat{\Omega}^{BA} & \hat{\Omega}^{BB}
\end{pmatrix}^{-1}
= \begin{pmatrix}
\hat{\Sigma}_{AA} & \hat{\Sigma}_{AB} \\
\hat{\Sigma}_{BA} & \hat{\Sigma}_{BB}
\end{pmatrix}
\]

Note that \( S_{AA} \) and \( \hat{\Sigma}_{AA} \) should coincide. Solving (we use results on the inverse or a partitioned matrix)

\[
(S_{AA})^{-1} = \hat{\Omega}^{AA} + R - \hat{\Omega}^{AB} (\hat{\Omega}^{BB})^{-1} \hat{\Omega}^{BA}
\]

and therefore

\[
R = (S_{AA})^{-1} - (\hat{\Omega}^{AA} - \hat{\Omega}^{AB} (\hat{\Omega}^{BB})^{-1} \hat{\Omega}^{BA})
\]

after noting that the last term is \( \Sigma_{AA}^{-1} \)

\[
R = S_{AA}^{-1} - (\Sigma_{AA}^{-1})
\]

The algorithm converges to the unique maximum.
Deviance test

Let $M_0$ be a model, and $\hat{\Sigma}_0$ be the estimated covariance matrix. Under the saturated model $\hat{\Sigma} = S$ and

$$l_s = \text{const} - N \log |\hat{S}|/2 - Np/2 \quad \text{function of } \bar{x}$$

so

$$G^2 = -2(l_0 - l_s) = N \log(|\hat{\Sigma}|/|S|) \sim \chi^2_g$$

where $g$ is the number of edges deleted from the saturated graph.

Other tests makes use of the estimated standard errors (see Lauritzen, 1996).
Deviance test continued

We are often interested in comparisons between $M_0$ and $M_1$, with $M_1$ NOT the saturated model.

For $M_1$ any model (not just the saturated model), expression of $G^2$ becomes:

$$G^2 = N \log(|\hat{\Sigma}_0|/|\hat{\Sigma}_1|) \sim \chi^2_g$$

where $g$ is the number of edges deleted from the graph $G_0$ of $M_0$ and $G_1$ of $M_1$.

It is possible to see that this is the difference of the deviance of both models against the saturated model.
Single edge deletion

A special case of interest: the single edge deletion. Let $M_0$ and $M_1$ differ only by deletion of the $(a, b)$ edge. Then, (exercise):

$$G^2 = -N \log [1 - (\rho_{ab \setminus \{a,b\}})^2].$$
Notes

1. The inverse of a partitioned matrix (also known as inverse variance Lemma) plays a crucial role in understanding all implications of graphical model.

2. A very useful exercise is to generate a sample of observations drawn multiple Gaussian variable with a covariance matrix with zeroes in the inverse. Import it in MIM and ...see what happens.