COMPUTATIONAL STATISTICS
Module 4

4 Gaussian graphical models

by Marcelo Ruiz

4.1 Preliminaries

Graphs: A *graph* is a pair $\mathcal{G} = (V, E)$ where $V \neq \emptyset$ is the set of nodes and $E \subseteq V \times V$ the set of edges. Two nodes i and j are are adjacent or *neighbors* if $(i, j) \in E$. A graph is *undirected* if it satisfies the condition

$$\forall (i,j) \in V^2 : (i,j) \in E \text{ if and only if } (j,i) \in E.$$

For every node $i \in V$ let

$$\mathcal{A}_i = \{j \in \mathrm{V} \setminus \{i\} : (i,j) \in E\} \tag{1}$$

be its *neighborhood.*

A *path* from a node i to a node j is a sequence of nodes $\{i_1, \dots, i_l\}$ such that $i_1 = i$, $i_k = j$ and $(i_k, i_{k+1}) \in E$ for all $k = 1, \dots, l-1$. Given A, B, C, disjoint subsets of V, we say that C *separates* A and B if for any $i \in A$ and $j \in B$, if there exits a path from i to j, it intersects C.

Conditional Independence: if X, Y and Z are discrete random variables we say that X, Y are independent given Z, $X \perp\!\!\!\perp X \mid Z$, if and only if (iff)

$$P(X = x, Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)$$

$\forall x, y, z$ such that $P(Z = z) > 0$. If (X, Y, Z) admits a joint density then

$$X \perp\!\!\!\perp Y \mid Z \text{ iff } f_{XY|Z}(x, y|z) = f_{X|Z}(x|z) f_{Y|Z}(y|z).$$

$\forall x, y, z$ such that $f_Z(z) > 0$ provided all the densities are continuous.

Remark 1 *Note that:*

i) This definition can be easily extended to (vectors) $X_A \perp\!\!\!\perp X_B \mid X_C$ where A, B, C are set of indexes.

ii) Here functions on discrete spaces are considered continuous.

iii) A more general and rigorous definition of conditional independence (requiring measure theory) is beyond the scope of this course.

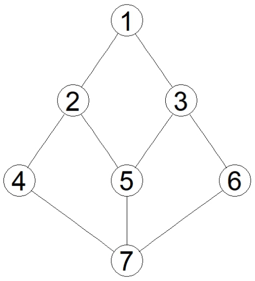

Figure 1: $\mathcal{G}_1 = (V_1, E_1)$.

Some properties of the conditional independence are listed below (their proofs are good exercises).

Lemma 2 *The following properties are valid:*

C.1) if $X \perp\!\!\!\perp Y \mid Z$ *then* $Y \perp\!\!\!\perp X \mid Z$

C.2) If $X \perp\!\!\!\perp Y \mid Z$ *and* $U = h(X)$ *then* $U \perp\!\!\!\perp Y \mid Z$ *where* h *is a (measurable) function (on the sample space of the r.v.* X*).*

C.3) If $X \perp\!\!\!\perp Y \mid Z$ *and* $U = h(X)$ *then* $X \perp\!\!\!\perp Y \mid (Z, U)$*, with* h *as in (C.2).*

C.4) if $X \perp\!\!\!\perp Y \mid Z$ *and* $X \perp\!\!\!\perp W \mid (Y, Z)$ *then* $X \perp\!\!\!\perp (W, Y) \mid Z$*.*

Proof. Exercise. ■

Graphical model: Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ be a random vector with distribution P. A *Graphical Model* (GM) is the pair $(\mathcal{G}, P)$ such that $V = \{1, \ldots, p\}$ and E is defined by

$$(i, j) \notin E \text{ if and only if } X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}. \tag{2}$$

Example 3 *Let* $(X_1, X_2, \ldots, X_7)$ *be a random vector with distribution* P *and let* $\mathcal{G}_1 = (V_1, E_1)$ *be its graphical model represented in Figure 1.*

a) Note that $\mathcal{A}_1 = \{2, 3\}$ *and* $\mathcal{A}_5 = \{2, 3, 7\}$*,* $X_1 \perp\!\!\!\perp X_7 \mid X_{V \setminus \{1,7\}}$ *and "given the remaining variables"* X_1 *and* X_2 *are dependent.*

b) But, can we conclude that $X_1 \perp\!\!\!\perp X_7 \mid X_{\{3,4,5\}}$*?*

Markov properties: one of the main purposes of graphical modeling consists in representing the dependence structure of a distribution.

Let $\boldsymbol{X} = (X_1, \ldots, X_p) \sim P$, $V = \{1, \ldots, p\}$ and let $\mathcal{G} = (V, E)$ be a graph. We say P satisfies (with respect to $\mathcal{G}$) the

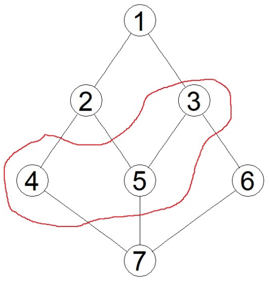

Figure 2: $\mathcal{G}_1 = (V_1, E_1)$.

- *pairwise property* (P) if for every $(i, j) \notin E$:

$$X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}}.$$

- *global* (G) property if for every disjoints sets A, B, C such that C separates A and B it is satisfied that

$$X_A \perp\!\!\!\perp X_B \mid X_C.$$

- *local* (L) property if for every node i

$$X_i \perp\!\!\!\perp X_{V \setminus \mathrm{cl}(i)} \mid X_{\mathcal{A}_i},$$

where $\mathrm{cl}(i) = \mathcal{A}_i \cup \{i\}$ denotes its closure.

Example 4 *Let $(X_1, X_2, \ldots, X_7) \sim P$ and $\mathcal{G}_1 = (V_1, E_1)$ as in the previous example and represented by Figure 2. Assume also that P satisfies, with respect to $\mathcal{G}_1$, the three Markov properties.*

Some conclusions:

a) $X_1 \perp\!\!\!\perp X_7 \mid X_{V \setminus \{1,7\}}$ *(as before)*

b) $X_1 \perp\!\!\!\perp X_7 \mid X_{\{3,4,5\}}$ *because $C = \{3, 4, 5\}$ separates the sets $A = \{1\}$ and $B = \{7\}$, and*

c) $X_1 \perp\!\!\!\perp X_{\{4,5,6,7\}} \mid X_{\{2,3\}}$.

Proposition 5 *Let $\boldsymbol{X} \sim P$ and $\mathcal{G} = (V, E)$ be as in the previous definition. Then* *a) for any distribution P*

$$\textit{Global} \Longrightarrow \textit{Local} \Longrightarrow \textit{Pairwise} \ ;$$

b) if the distribution P has a positive and continuous density (w.r.l.m.), then the Markov properties are equivalent.

Proof.

a) Assume that (G) property hold and let $i \in V$. As $\mathcal{A}_i$ separates $\{i\}$ and $V \setminus \mathrm{cl}(i)$ then, by (G), (L) follows; i.e. property $X_i \perp\!\!\!\perp X_{V\setminus\mathrm{cl}(i)} \mid X_{\mathcal{A}_i}$.

Assume now that (L) holds and consider two nodes i and j such that $(i,j) \notin E$ and let

$$Y = X_{\{i\}},\ X = X_{V\setminus\mathrm{cl}(i)},\ Z = X_{\mathcal{A}_i} \text{ and } U = h(X) = X_{(V\setminus\mathrm{cl}(i))\setminus\{j\}}$$

(note that as $j \notin \mathcal{A}_i$ h is well defined and consists in the projection of the coordinates $X_{(V\setminus\mathrm{cl}(i))}$ on $X_{(V\setminus\mathrm{cl}(i))\setminus\{j\}}$).

By the (L) property $X \perp\!\!\!\perp Y|Z$ holds. Using this fact and the (C.3) property of the conditional independence it follows that

$$X \perp\!\!\!\perp Y|(Z,U) \Leftrightarrow X_{\{i\}} \perp\!\!\!\perp X_{V\setminus\mathrm{cl}(i)}|(X_{\mathcal{A}_i}, X_{(V\setminus\mathrm{cl}(i))\setminus\{j\}})$$

Note now that, as $(i,j) \notin E$ (and then $j \notin \mathcal{A}_i$)

$$\mathcal{A}_i \cup ((V \setminus \mathrm{cl}(i)) \setminus \{j\}) = V \setminus \{i,j\};$$

and so

$$X_{\{i\}} \perp\!\!\!\perp X_{V\setminus\mathrm{cl}(i)}|X_{V\setminus\{i,j\}}.$$

Applying (C.2) with X, Y and Z as before but $U = h(X) = X_{\{j\}}$ (the projection of $X_{V\setminus\mathrm{cl}(i)}$ on the $X_{\{j\}}$ coordinate)

$$X_i \perp\!\!\!\perp X_j \perp\!\!\!\perp X_{V\setminus\{i,j\}}$$

is established.

b) See Lauritzen (1996, pp. 34–45). Schematically, to prove b) first it is shown that $(G) \Longleftrightarrow (L) \Longleftrightarrow (P)$ if and only if $\forall A, B, C, D$ disjoint subsets if $A \perp\!\!\!\perp B|(C \cup D)$ and $A \perp\!\!\!\perp C|(B \cup D)$ then $A \perp\!\!\!\perp (B \cup C)|D$. Then, the factorization concept is used and the theorem of Hammersley and Clifford close the proof. See details in in Lauritzen (1996, pp. 34–45).

■

Gaussian Graphical Model

A *Gaussian Graphical Model* (GGM) is a GM such that P is the multivariate Gaussian distribution; i.e. $\mathbf{X} = (X_1, \ldots, X_p)^\top \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Remark 6 *A GGM satisfies the three Markov Properties*

The following proposition (easy to prove) characterizes a neighborhood in terms of the conditional independence.

Proposition 7 *In a GGM, for every node i, if $\mathcal{A}_i \neq \emptyset$ then*

$$\mathcal{A}_i = \bigcap_{A \in \mathcal{E}_i} A \text{ with } \mathcal{E}_i = \{A \subset V \setminus \{i\} : X_i \perp\!\!\!\perp X_{V\setminus(A\cup\{i\})} \mid X_A\};$$

i.e. $\mathcal{A}_i$ is the smallest subset of $V\setminus\{i\}$ such that X_i is conditionally independent of the remaining variables, given $X_{\mathcal{A}_i}$.

Proof. Denote $M_i = \bigcap_{A\in\mathcal{E}_i} A$ and let us prove that $M_i = \mathcal{A}_i$. By the local Markov property, $\mathcal{A}_i \in \mathcal{E}_i$ and so

$$\mathcal{E}_i \subseteq \mathcal{A}_i.$$

Let $j \in \mathcal{A}_i$ and assume for the sake of contradiction that $j \notin M_i$. Hence, by definition of M_i, there exists $A_o \in \ \mathcal{E}_i$ such that $j \notin A_o$.

So, $X_i \perp\!\!\!\perp X_{V\setminus(A_o\cup\{i\})} \mid X_{A_o}$ and by the global property A_o separates $\{i\}$ and $V \setminus (A_o \cup \{i\})$. As $(i, j) \in E$ then (i, j) is a path from a i to a j; in consequence this path have to intersect A_o. So $i \in A_o$ or $j \in A_o$, and this is a contradiction. Hence

$$\mathcal{A}_i \subseteq \mathcal{E}_i,$$

and the proof is complete. ■

4.2 Conditional dependence in a GGM

Hereafter we assume that

$$\mathbf{X} = (X_1, \ldots, X_p)^\top \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

with a positive definite covariance matrix $\boldsymbol{\Sigma}$.

Notation: If $\mathcal{A}$ is a subset of $V = \{1, \ldots, p\}$, $\mathbf{X}_\mathcal{A}$ denotes the vector of variables with subscripts in $\mathcal{A}$ in increasing order and, for every fixed pair of nodes (i, l), set $\mathbf{X}_1^\top = (X_i, X_l)$, $\mathbf{X}_2^\top = \mathbf{X}_{V\setminus\{i,l\}}$ and $\mathbf{X} = \left(\mathbf{X}_1^\top, \mathbf{X}_2^\top\right)^\top$.

Parametrization of E: the following definitions and equalities give different representation of the set of edges.

P.1) The *conditional correlation* of $(X_i, X_l)|\mathbf{X}_{V\setminus\{i,l\}}$, is defined as

$$\text{CORR}\left(X_i, X_l|\mathbf{X}_{V\setminus\{i,l\}}\right) = \frac{\text{COV}\left(X_i, X_l \mid \mathbf{X}_{V\setminus\{i,l\}}\right)}{\sqrt{\text{VAR}\left(X_i \mid \mathbf{X}_{V\setminus\{i,l\}}\right)\text{VAR}\left(X_l \mid \mathbf{X}_{V\setminus\{i,l\}}\right)}}. \quad (3)$$

P.2) If $\mathbf{\Omega} = (\omega_{ij})_{i,j=1\ldots,p}$ denotes the precision matrix $\mathbf{\Sigma}^{-1}$, then

$$\text{CORR}\left(X_i, X_l|\mathbf{X}_{V\setminus\{i,l\}}\right) = -\frac{\omega_{il}}{\sqrt{\omega_{ii}\omega_{ll}}}. \quad (4)$$

P.3) Consider the regression error of $\mathbf{X}_1$ on $\mathbf{X}_2$ defined by $\boldsymbol{\varepsilon} = \mathbf{X}_1 - \widehat{\mathbf{X}}_1 = \mathbf{X}_1 - \boldsymbol{\beta}^\top\mathbf{X}_2$ and let ε_i and ε_l denote the entries of $\boldsymbol{\varepsilon}$ (i.e. $\boldsymbol{\varepsilon}^\top = (\varepsilon_i, \varepsilon_l)$). Usually, ε_i and ε_l are called the residuals associated to X_i and X_l in the regression of $\mathbf{X}_1$ on $\mathbf{X}_2$.

The *partial correlation coefficient* between X_i and X_l, denoted by $\rho_{il\cdot V\setminus\{i,l\}}$, is defined as the Pearson correlation coefficient of the residuals; that is,

$$\rho_{il\cdot V\setminus\{i,l\}} = \frac{\text{COV}\left(\varepsilon_i, \varepsilon_l\right)}{\sqrt{\text{VAR}\left(\varepsilon_i\right)\text{VAR}\left(\varepsilon_l\right)}} \quad (5)$$

and both, correlation and conditional coefficients are equal.

The following proposition establishes differents ways to represent (parametrize) E:

Proposition 8 *For a GGM, the set of edges E satisfies*

$$\begin{aligned} E &= \{i, l \in V : \text{CORR}\left(X_i, X_l|\mathbf{X}_{V\setminus\{i,l\}}\right) \neq 0\} \quad (6)\\ &= \{i, l \in V : \omega_{i,l} \in \Omega \textit{ and } \neq 0\}\\ &= \{i, l \in V : \rho_{il\cdot V\setminus\{i,l\}} \neq 0\}. \end{aligned}$$

Proof. The first equality is immediate. Second and third equalities follow if we show (4) and that the conditional and partial correlation coefficients are equal.

To prove (4) first note that $\left(\mathbf{X}_1^\top, \mathbf{X}_2^\top\right)^\top$ has multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \quad (7)$$

such that Σ_{11} has dimension 2×2, Σ_{12} has dimension $2 \times (p-2)$ and so on. The matrix in (7) is a partition of a permutation of the original covariance matrix Σ, according into blocks $\Sigma_{u,j}$, $u, j = 1, 2$.

Moreover, the conditional distribution of $\mathbf{X}_1|\mathbf{X}_2$ is normal and satisfies that

$$\text{E}\left(\mathbf{X}_1|\mathbf{X}_2\right) = \boldsymbol{\beta}^\top\mathbf{X}_2 \text{ and } \text{COV}\left(\mathbf{X}_1|\mathbf{X}_2\right) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{\Sigma}_{21}, \quad (8)$$

where $\boldsymbol{\beta}^{\top}=\Sigma_{12}\Sigma_{22}^{-1}$ is denominated the matrix of regression coefficients of $\mathbf{X}_1$ on $\mathbf{X}_2$, and $\widehat{\mathbf{X}}_1=\boldsymbol{\beta}^{\top}\mathbf{X}_2$ is the optimal predictor of $\mathbf{X}_1$.

If we set

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$$

then, the blocks $\Omega_{i,j}$ can be written explicitly in terms of $\Sigma_{i,j}$ or $\Sigma_{i,j}^{-1}$ and, in particular $\Omega_{11}=\left(\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)^{-1}$ where

$$\Omega_{11}=\begin{pmatrix} \omega_{ii} & \omega_{il} \\ \omega_{li} & \omega_{ll} \end{pmatrix}$$

is the submatrix of Ω (with rows i and l and columns i and l). Hence, by (8),

$$\begin{aligned} \text{COV}\left(\mathbf{X}_1|\mathbf{X}_2\right) &= \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \qquad (9)\\ &= \Omega_{11}^{-1} \\ &= \frac{1}{\omega_{ii}\omega_{ll}-\omega_{il}\omega_{li}}\begin{pmatrix} \omega_{ll} & -\omega_{il} \\ -\omega_{li} & \omega_{ii} \end{pmatrix} \end{aligned}$$

and, in consequence, (3) can be expressed as in (4).

Now consider the regression error $\boldsymbol{\varepsilon}$ of $\mathbf{X}_1$ on $\mathbf{X}_2$ as before. $\boldsymbol{\varepsilon}$ is independent of $\widehat{\mathbf{X}}_1$ and has normal distribution with mean $\mathbf{0}$ and matrix covariance Ψ_{11} with elements denoted by

$$\Psi_{11}=\begin{pmatrix} \psi_{ii} & \psi_{il} \\ \psi_{li} & \psi_{ll} \end{pmatrix}. \qquad (10)$$

A straightforward calculation shows that

$$\Psi_{11}=\text{COV}\left(\mathbf{X}_1\right)+\text{COV}\left(\widehat{\mathbf{X}}_1\right)-2\text{COV}\left(\mathbf{X}_1,\widehat{\mathbf{X}}_1\right)$$

$$=\Sigma_{11}+\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21}-2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

$$=\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}=\Omega_{11}^{-1}.$$

Therefore, by this equality, (9) and (4), the partial correlation coefficient and the conditional correlation are equal

$$\begin{aligned} \rho_{il\cdot V\setminus\{i,l\}} &= \text{CORR}\left(X_i,X_l|\mathbf{X}_{V\setminus\{i,l\}}\right) \qquad (11)\\ &= \frac{\psi_{il}}{\sqrt{\psi_{ii}\psi_{ll}}}. \end{aligned}$$

4.3 Covariance selection

Note that even when p is moderate the number of different graphs with p nodes can be big enough: $2^{p(p-1)/2}!$. Indeed, the total number of edges of a subset of p nodes are $\begin{pmatrix} p \\ 2 \end{pmatrix}$, then the total number of graphs is the total number of subsets of the set $\left\{1, \ldots, \binom{p}{2}\right\}$ that is equal to $2^{\binom{p}{2}} = 2^{p(p-1)/2}$.

Let $(X_{1j}, \ldots, X_{pj})^\top$, $j = 1, \ldots, n$ be a random sample from $\mathbf{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma})$. Covariance selection (Dempster, 1970) consists in finding procedures that, based on the sample, determine the set of edges of the associated GGM or, equivalently, the conditionally dependent pairs of variables (given the rest) or, the non-zero elements of the precision matrix.

Underlying to the covariance selection is the principle of parsimony in parametric model fitting “parameters should be introduced sparingly and only when data indicate they are required” (Dempster, 1970).

Covariance selection is not a selection variables problem. It is a problem of (conditional dependence) “relationships”selection between variables.

Moreover, note that

1. In a high-dimensional framework, when $p > n$, the sample covariance matrix S is not invertible and the maximum likehood estimate (MLE) of $\mathbf{\Sigma}$ does not exist.

2. When $p/n \leq 1$, but close to 1, S is invertible but ill-conditioned, dramatically increasing the estimation error.

To deal with these problems several covariance selection procedures have been proposed based on the assumption that the inverse of the covariance matrix, the precision matrix, is sparse.

4.4 Covariance selection using Lasso

Instead of assuming that there exists a fixed true underlying fixed model Meinshausen and Bühlmann (2006) assume a more flexible approach: the number of nodes $p = p(n) = |V(n)|$ and the covariance matrix $\Sigma = \Sigma(n)$ “depends” on the number of observations. The mentioned authours propose to estimate the graph (covariance selection) using Lasso. In this subsection, we give some aspects of their work.

OLS estimates: let $\boldsymbol{\beta}^i \in \mathbb{R}^p$ be the (population) coefficient of the regression of X_i on the remaining variables $\mathbf{X}_{V \setminus \{i\}}$; that is

$$\boldsymbol{\beta}^i = \underset{\boldsymbol{\beta}:\beta_i = 0}{\operatorname{argmin}} E \left(X_i - \sum_{k \in V} \beta_k X_k \right)^2 . \tag{12}$$

If β^i_j denotes the j–th coefficient of β^i, then

$$\beta^i_j = -\frac{\omega_{ij}}{\omega_{ii}} \text{ and } \{j \in \backslash\{i\} : \beta^i_j \neq 0\} = \{j \in \backslash\{i\} : \omega_{ij} \neq 0\}$$

that implies

$$\mathcal{A}_i = \left\{j \in \backslash\{i\} : \beta^i_j \neq 0\right\}. \tag{13}$$

Let X be the matrix of dimension $n \times p$ containing n independent observations of X, such that the columns $\mathbf{X}_i$ corresponds to the vector of n independent observations of X_i. Sea $\langle \cdot, \cdot \rangle$ the (usual) inner product in $\mathbb{R}^n$ and $\|\cdot\|_2$ its corresponding norm.

So the least square estimates $\hat{\boldsymbol{\beta}}^{i,ls}$ of $\boldsymbol{\beta}^i$ is

$$\widehat{\boldsymbol{\beta}}^{i,ls} = \underset{\boldsymbol{\beta}:\beta_i=0}{\operatorname{argmin}}\, n^{-1} \left\|\mathbf{X}_i - \mathbf{X}\boldsymbol{\beta}\right\|_2^2. \tag{14}$$

Under sparsity it is necessary to change the estimation strategy, considering Lasso estimation.

Lasso estimates $\hat{\boldsymbol{\beta}}^{i,\lambda}$ of $\boldsymbol{\beta}^i$ is given by

$$\hat{\boldsymbol{\beta}}^{i,\lambda} = \underset{\boldsymbol{\beta}:\beta_i=0}{\operatorname{argmin}} \left(n^{-1} \left\|\mathbf{X}_i - \mathbf{X}\boldsymbol{\beta}\right\|_2^2 + \lambda \left\|\boldsymbol{\beta}\right\|_1\right),$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{k\in V} |\beta_k|$ is the ℓ_1 norm and $\lambda \in [0, \infty)$.

The estimated (by lasso) neighborhood, $\widehat{\mathcal{A}}^\lambda_i$ is given by

$$\widehat{\mathcal{A}}^\lambda_i = \left\{j \in : \hat{\beta}^{i,\lambda}_j \neq 0\right\}.$$

Consistent neighborhood estimation.

It is possible to control the Type I and II errors imposing certain rate to the penalization parameter. More specifically, under certain conditions and assuming that

$$\lambda_n \sim dn^{-(1-\epsilon)/2}$$

for some $\kappa < \epsilon < \xi$ and $d > 0$ then there exists some $c > 0$ such that for $i \in\ V$,

$$\begin{array}{rcl} P\left(\hat{\mathcal{A}}^\lambda_i \subseteq \mathcal{A}_i\right) &=& 1 - O(exp(-cn^\epsilon)) \\ P\left(\mathcal{A}_i \subseteq \hat{\mathcal{A}}^\lambda_i\right) &=& 1 - O(exp(-cn^\epsilon)) \end{array}$$

as $n \to \infty$.

What kind of assumptions? We only mention two:

- ***High dimensionality.*** The number of variables can grow as a power of the number of observations: $\exists \gamma > 0$, such that

$$p(n) = O(n^{\gamma}) \qquad \text{para } n \to \infty. \tag{15}$$

- ***Sparsity.*** Restricts the size of the neighborhood: $\exists 0 \leq \kappa < 1$ such that

$$\max_{i \in V} |\mathcal{A}_i| = O(n^{\kappa}) \qquad \text{for } n \to \infty. \tag{16}$$

Lasso Estimation of E: for each variable fit a lasso model using the others as predictors and, considering that

$$E = \{(i,j) : i \in \mathcal{A}_j \text{ and } j \in \mathcal{A}_i\}$$

then define the estimated set of edges as

$$\begin{aligned} \widehat{E}^{\lambda,\wedge} &= \{(i,j) : i \in \widehat{\mathcal{A}}_j^{\lambda} \text{ and } j \in \widehat{\mathcal{A}}_i^{\lambda}\}, \text{ or} \\ \widehat{E}^{\lambda,\vee} &= \{(i,j) : i \in \widehat{\mathcal{A}}_j^{\lambda} \text{ or } j \in \widehat{\mathcal{A}}_i^{\lambda}\}. \end{aligned}$$

Both $\widehat{E}^{\lambda,\wedge}$ and $\widehat{E}^{\lambda,\vee}$ consistently estimates the set of non-zero elements of the set of edges E.

The penalty term λ is chosen such that, for a level $\alpha = \alpha(\lambda)$ the probability of falsely joining two distinct connectivity components with the edge set is bounded by that level (see detalis in Meinshausen and Bühlmann, 2006).

4.5 A Graphical Stepwise Approach to Covariance Selection

According to (11) and the discussion therein, the partial correlation can be also used to perform covariance selection. For every node $i \in V$, let $\mathcal{A}_i$ be its neighborhood as it was defined previously. As we emphasize before, conditionally on its neighbors, X_i is independent of all the other variables; i.e.,

$$\forall i \in V : \mathcal{A}_i \neq \emptyset : \forall l \notin \mathcal{A}_i \, (l \neq i) : X_i \perp\!\!\!\perp X_l | \mathbf{X}_{\mathcal{A}_i}. \tag{17}$$

In consequence, if a "tentative" graphical model is given by the system of neighborhoods $\{\mathcal{A}_i\}_{i=1}^{p}$ and $l \notin \mathcal{A}_i$ (and therefore $i \notin \mathcal{A}_l$), then the partial correlation between X_i and X_l can be obtained by the following procedure:

1. regress X_i on $\mathbf{X}_{\mathcal{A}_i}$ and form the regression residual ε_i; regress X_l on $\mathbf{X}_{\mathcal{A}_l}$ and form the regression residual ε_l and then
2. calculate the Pearson correlation between ε_i and ε_l.

Hence, this procedure suggest the following:

Graphical Stepwise Algorithm

Input: the (centered) data $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$, and the forward and backward thresholds α_f and α_b.

Initialization. Set $\mathcal{A}_1^0 = \mathcal{A}_2^0 = \cdots = \mathcal{A}_p^0 = \phi$
Given $\mathcal{A}_1^k, \mathcal{A}_2^k, ..., \mathcal{A}_p^k$ we compute $\mathcal{A}_1^{k+1}, \mathcal{A}_2^{k+1}, ..., \mathcal{A}_p^{k+1}$ as follows.

Forward Step. For each $i = 1, ..., p$ do the following.
For each $l \notin \mathcal{A}_i^k$ calculate f_{il}^k as follows.

(a) Regress the i^{th} variables on the variables with subscript in the set $\mathcal{A}_i^k$ and compute the regression residuals $\mathbf{e}_i^k = \left(e_{i1}^k, e_{i2}^k, ..., e_{in}^k\right)$.

(b) Regress the l^{th} variables on the variables with subscript in the set $\mathcal{A}_l^k$ and compute the regression residuals $\mathbf{e}_l^k = \left(e_{l1}^k, e_{l2}^k, ..., e_{ln}^k\right)$.

(c) Compute the Pearson correlation f_{il}^k between $\mathbf{e}_i^k$ and $\mathbf{e}_l^k$.

If

$$\max_{l \notin \mathcal{A}_i^k, i \in V} \left| f_{il}^k \right| = \left| f_{i_0 l_0}^k \right| \geq \alpha_f$$

set $\mathcal{A}_{i_0}^{k+1} = \mathcal{A}_{i_0}^k \cup \{l_0\}$, $\mathcal{A}_{l_0}^{k+1} = \mathcal{A}_{l_0}^k \cup \{i_0\}$, $\mathcal{A}_l^{k+1} = \mathcal{A}_l^k$ for $l \neq i_0, l_0$
If

$$\max \left| f_{il}^k \right| = \left| f_{i_0 l_0}^k \right| < \alpha_f$$

Stop.

Backward Step. For each $i = 1, ..., p$ do the following.
For each $l \in \mathcal{A}_i^{k+1}$ calculate b_{il}^k as follows.

(a) Regress the i^{th} variables on the variables with subscript in the set $\mathcal{A}_i^{k+1} \backslash \{l\}$ and compute the regression residuals $\mathbf{r}_i^k = \left(r_{i1}^k, r_{i2}^k, ..., r_{in}^k\right)$.

(b) Regress the l^{th} variables on the variables with subscript in the set $\mathcal{A}_l^{k+1} \backslash \{i\}$ and compute the regression residuals $\mathbf{r}_l^k = \left(r_{l1}^k, r_{l2}^k, ..., r_{ln}^k\right)$.

(c) Compute the Pearson correlation b_{il}^k between $\mathbf{r}_i^k$ and $\mathbf{r}_l^k$.

If

$$\min_{l \in \mathcal{A}_i^k, i \in V} \left| b_{il}^k \right| = \left| b_{i_0 l_0}^k \right| \leq \alpha_b$$

set $\mathcal{A}_{i_0}^{k+1} \to \mathcal{A}_{i_0}^{k+1} \backslash \{l_0\}$, $\mathcal{A}_{l_0}^{k+1} \to \mathcal{A}_{l_0}^{k+1} \backslash \{i_0\}$, $\mathcal{A}_l^{k+1} \to \mathcal{A}_l^{k+1}$ for $l \neq i_0, l_0$

Summarizing: the Graphical Stepwise Algorithm (GSA) begins with the family of empty neighborhoods, $\hat{\mathcal{A}}_i^{(0)} = \emptyset$ for each $i \in V$ (i.e., the initial set of edges is empty). There are two basic steps, the forward and the backward:

- in the forward step, the algorithm finds a new edge (i_0, l_0) provided the largest absolute value of the empirical Pearson correlation between residuals corresponding to the variables X_{i_0}, X_{l_0} is big enough (no less than the threshold α_f) and, otherwise the algorithm stop;

- next, in the backward step the algorithm estimates the empirical Pearson correlation between residuals under the presence of the lately added edge; if the minimum of the absolute value of the empirical Pearson correlation between residuals is not big enough (less or equal than the threshold α_b) then the corresponding edge is eliminated. And so on.

The ouput of the GSA is a collection of estimated neighbors $\widehat{\mathcal{A}}_i^{GS}$ and the set of estimated edges $\widehat{E}^{GS} = \left\{(i,j) \in V^2 : j \in \widehat{\mathcal{A}}_i^{GS}\right\}$. To simplify we omit the superscript GS.

Selection of thresholds by cross-validation

Let X be a $p \times n$ matrix with columns $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})^\top$, $j = 1, \ldots, n$, corresponding to n independent and identically distributed observations.

Assume that the graphical model is given by the system of neighborhoods $\{\mathcal{A}_i\}_{i \in V_0}$ where for every $i \in V_0$, $\mathcal{A}_i \neq \emptyset$. Without loss of generality we suppose that $V_0 = \{1, \ldots, I\}$, with $1 \leq I \leq p$ and, let $\left\{\widehat{\mathcal{A}}_i\right\}_{i \in V_0}$ be the collection of estimated neighborhoods by the GSA.

For every node $i \in V_0$, denote $\mathcal{A}_i = \left\{X_{i_1}, \ldots X_{i_{p_i}}\right\}$ with $p_i = |\mathcal{A}_i|$ and $i_1 < \ldots i_{p_i}$ and, consider the regression of X_i on $\mathcal{A}_i$

$$X_i = \boldsymbol{\beta}_{\mathcal{A}_i}^\top \mathbf{X}_{\mathcal{A}_i} + \varepsilon_i \tag{18}$$

where now $\mathbf{X}_{\mathcal{A}_i}$ is the column vector $\left(X_{i_1}, \ldots X_{i_{p_i}}\right)^\top$.

For every $i = 1, \ldots, n$ let $\mathbf{X}_i = (x_{i1}, \ldots, x_{in})$ denote the ith–row of the matrix $\mathbf{X}$. So, for the model given by (18) after the GSA (based on a pair of thresholds (α_f, α_b)) is run the regression residuals vector, $\mathbf{e}_i = (e_1, \ldots, e_n)^\top$, can be written as

$$\mathbf{e}_i^\top = \mathbf{X}_i - \widehat{\boldsymbol{\beta}}_{i\widehat{\mathcal{A}}_i}^\top \mathbf{X}_{\widehat{\mathcal{A}}_i} \tag{19}$$

where $\widehat{\boldsymbol{\beta}}_{i\widehat{\mathcal{A}}_i}^\top$ is the estimated regression coefficients and

$$\mathbf{X}_{\widehat{\mathcal{A}}_i} = \begin{pmatrix} x_{i_1 1} & \cdots & x_{i_1 n} \\ \vdots & \ldots & \vdots \\ x_{i_{\widehat{p}_i} 1} & \cdots & x_{i_{\widehat{p}_i} n} \end{pmatrix} \tag{20}$$

is the matrix of the sample observations corresponding to the selected variables $X_{i_1}, \ldots, X_{i_{\widehat{p}_i}}$ and $\widehat{p}_i$ is the size of the estimated neighborhood $\widehat{\mathcal{A}}_i$. Note that $\widehat{p}_i$, $\widehat{\mathcal{A}}_i$ and $\widehat{\boldsymbol{\beta}}_{i\widehat{\mathcal{A}}_i}$ depend on the thresholds (α_f, α_b) and that not necessarily p_i and $\widehat{p}_i$ are equal.

Partition the data set $\{\mathbf{x}_j\}_{1 \leq j \leq n}$ at random into K disjoint subsets of approximately equal size, the j^{th} subset having size $n_j \geq 2$, $\sum_{j=1}^{K} n_j = n$.

For every j, let $\{\mathbf{x}_i^{(j)}\}_{1 \leq i \leq n_j}$ be the j^{th} subset, the *validation set*, and its complement $\{\widetilde{\mathbf{x}}_i^{(j)}\}_{1 \leq i \leq n - n_j}$, the *training set*.

Fix α_f, α_b a pair of forward and backward thresholds, respectively. Now, for every j let (see the notation introduced in (19) and (20))

$$\widehat{\beta}^{(j)^\top}_{1\widehat{\mathcal{A}}_1^{(j)}}, \ldots, \widehat{\beta}^{(j)^\top}_{I_j\widehat{\mathcal{A}}_{I_j}^{(j)}}$$

be the estimated regression coefficients computed by the GSA based on the observations in the training set $\{\widetilde{\mathbf{x}}_l^{(j)}\}_{1\leq l\leq n-n_j}$ and on the chosen thresholds (α_f, α_b).

If $\mathbf{x}_k^{(j)} = (x_{1k}^{(j)}, \ldots, x_{pk}^{(j)})$, $k = 1, \ldots, n_j$ is the validation set, then for every node i $\mathbf{X}_i^{(j)} = \left(x_{i1}^{(j)}, \ldots, x_{in_j}^{(j)}\right)$ is the sample of size n_j corresponding to the variable X_i and let

$$\widehat{\mathbf{X}}_i^{(j)} = \widehat{\beta}^{(j)^\top}_{i\mathcal{A}_i^{(j)}} \mathbf{X}_{\widehat{\mathcal{A}}_i^j}, \tag{21}$$

where

$$\mathbf{X}_{\widehat{\mathcal{A}}_i^j} = \begin{pmatrix} x_{i_1 1}^{(j)} & \cdots & x_{i_1 n_j}^{(j)} \\ \vdots & \cdots & \vdots \\ x_{i_{p_i} 1}^{(j)} & \cdots & x_{i_{\widehat{p}_i} n_j}^{(j)} \end{pmatrix}.$$

If we write $\widehat{\mathbf{X}}_i^{(j)} = \widehat{\mathbf{X}}_i^{(j)}(\alpha_f, \alpha_b)$ to emphasize the dependence of $\widehat{\mathbf{X}}_i^{(j)}$ on the thresholds, the K–fold cross–validation function is defined as

$$CV(\alpha_f, \alpha_b) = \frac{1}{n} \sum_{j=1}^{K} \sum_{i=1}^{p} \left\| \mathbf{X}_i^{(j)} - \widehat{\mathbf{X}}_i^{(j)}(\alpha_f, \alpha_b) \right\|^2 \tag{22}$$

where $\|\cdot\|$ the L2-norm or euclidean distance in $\mathbb{R}^p$.

The K–fold cross–validation forward–backward thresholds $\widehat{\alpha}_f$, $\widehat{\alpha}_b$ are defined as

$$(\widehat{\alpha}_f, \widehat{\alpha}_b) = \underset{(\alpha_f, \alpha_b)\in\mathcal{H}}{\operatorname{argmin}} CV(\alpha_f, \alpha_b) \tag{23}$$

where $\mathcal{H}$ is a grid of ordered pairs (α_f, α_b) in $[0, 1] \times [0, 1]$ over which we perform the search.

Bibliography

Dempster, A. (1970). Covariance selection. *Biometrics.* **28**, 157–175.

Laffit, G., Nogales, J., Ruiz, M., Zamar, R. (2018). *A Forward-Backward Approach for High-Dimensional Gaussian Graphical Models.* To appear.

Lauritzen, S. (1996). *Graphical Models.* Clarendon Press, Oxford.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics.* **3**, 1436–1462