Estimating partial correlations and recovering network hubs with PCGLASSO

Adam Chojecki, Ivan Hejny, Bartosz Kolodziejek,
Jonas Wallin, *Malgorzata Bogdan*

Warsaw University of Technology, University of Wroclaw, Lund University

30/06/2025

Outline

- Introduction and Motivation
- Graphical LASSO (GLASSO)
- Partial Correlation LASSO (PCGLASSO)
  - Algorithm
  - Convexity issues
  - Asymptotic distribution
  - Irrepresentability condition
  - Identifying the hub structure

Covariance and Precision Matrices

$(X_1, \ldots, X_p)$ - p dimensional random vector

Covariance and Precision Matrices

$(X_1, \ldots, X_p)$ - p dimensional random vector

$\Sigma_{p \times p}$ - covariance matrix

Covariance and Precision Matrices

$(X_1, \ldots, X_p)$ - p dimensional random vector

$\Sigma_{p \times p}$ - covariance matrix

$$\Sigma_{i,j} = Cov(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j)$$

Covariance and Precision Matrices

$(X_1, \ldots, X_p)$ - p dimensional random vector

$\Sigma_{p \times p}$ - covariance matrix

$$\Sigma_{i,j} = Cov(X_i, X_j) = E(X_i - \mu_i)(X_j - \mu_j)$$

$\Omega = \Sigma^{-1}$ - precision matrix

Motivation: Global Minimum Variance Portfolio

$R = [R_1 \ \cdots \ R_K]'$ - a random vector of returns

$$\mu = [\mathbb{E}[R_1] \ \cdots \ \mathbb{E}[R_K]]'$$

$$\Sigma = \mathbb{E}\left[(R-\mu)(R-\mu)'\right].$$

$w \in R^K$ - portfolio weights

$$Var(w'R) = w'\Sigma w$$

$$w^\star = \arg\min_{w \in \mathbb{R}^K} w'\Sigma w \ \text{ subject to } w'\mathbf{1}_K = 1,$$

$$w^\star = \frac{\Sigma^{-1}\mathbf{1}_K}{\mathbf{1}_K'\Sigma^{-1}\mathbf{1}_K}$$

Graphical Model

$X = (X_1, \ldots, X_p)$ - a random vector

Graphical Model

$X = (X_1, \ldots, X_p)$ - a random vector
Graph: $G = (V, E)$
Vertices (V): indices (names) of components of X: $V = \{1, \ldots, p\}$
Edges (E): $(u, v) \notin V \iff X_u \perp\!\!\!\perp X_v | X_{V \setminus \{u,v\}}$

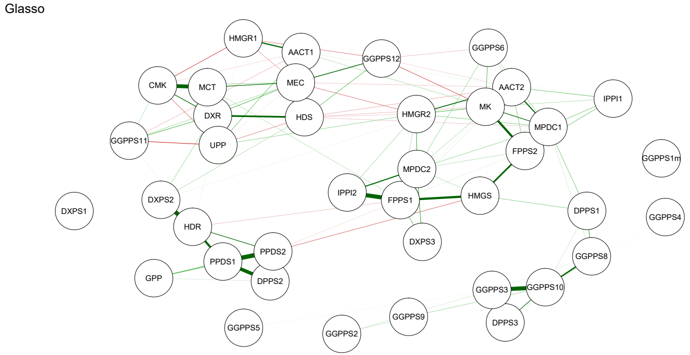
Graphical Model

$X = (X_1, \ldots, X_p)$ - a random vector

Graph: $G = (V, E)$

Vertices (V): indices (names) of components of X: $V = \{1, \ldots, p\}$

Edges (E): $(u, v) \notin V \iff X_u \perp\!\!\!\perp X_v | X_{V \setminus \{u,v\}}$



Gaussian Graphical Model

$$X = (X_1, \ldots, X_p) \sim N(0, \Sigma), \quad \Omega = \Sigma^{-1} \quad \text{- precision matrix}$$

Gaussian Graphical Model

$$X = (X_1, \ldots, X_p) \sim N(0, \Sigma), \quad \Omega = \Sigma^{-1} \quad \text{- precision matrix}$$

$$X_i \perp\!\!\!\perp X_j | X_{V \setminus \{u,v\}} \Longleftrightarrow \Omega_{ij} = 0$$

Gaussian Graphical Model

$$X = (X_1, \ldots, X_p) \sim N(0, \Sigma), \;\; \Omega = \Sigma^{-1} \;\; \text{- precision matrix}$$

$$X_i \perp\!\!\!\perp X_j | X_{V \setminus \{u,v\}} \Longleftrightarrow \Omega_{ij} = 0$$

Example: for $j = 2, \ldots, p$ $X_j = X_1 + \varepsilon_j$, where $Var(X_1) = Var(\varepsilon_j) = 2$

Gaussian Graphical Model

$$X = (X_1, \ldots, X_p) \sim N(0, \Sigma), \quad \Omega = \Sigma^{-1} \quad \text{- precision matrix}$$

$$X_i \perp\!\!\!\perp X_j | X_{V \setminus \{u,v\}} \iff \Omega_{ij} = 0$$

Example: for $j = 2, \ldots, p$ $X_j = X_1 + \varepsilon_j$, where
$Var(X_1) = Var(\varepsilon_j) = 2$
for $i \neq j$, $Cov(X_i, X_j) = Var(X_1) = 2$

Gaussian Graphical Model

$$X = (X_1, \ldots, X_p) \sim N(0, \Sigma), \quad \Omega = \Sigma^{-1} \quad \text{- precision matrix}$$

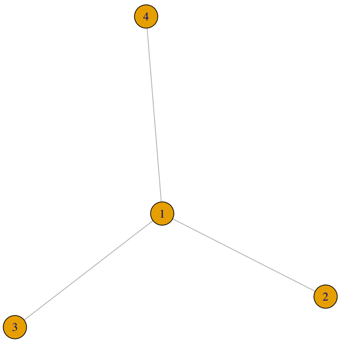
$$X_i \perp\!\!\!\perp X_j | X_{V \setminus \{u,v\}} \iff \Omega_{ij} = 0$$

Example: for $j = 2, \ldots, p$ $X_j = X_1 + \varepsilon_j$, where
$Var(X_1) = Var(\varepsilon_j) = 2$
for $i \neq j$, $Cov(X_i, X_j) = Var(X_1) = 2$
for $j \geq 2$, $Var(X_j) = 4$

Hub

$$\Sigma = \begin{bmatrix} 2 & 2 & 2 & 2 \\ 2 & 4 & 2 & 2 \\ 2 & 2 & 4 & 2 \\ 2 & 2 & 2 & 4 \end{bmatrix}, \Omega = \begin{bmatrix} 2 & -0.5 & -0.5. & -0.5 \\ -0.5 & 0.5 & 0 & 0 \\ -0.5 & 0 & 0.5 & 0 \\ -0.5 & 0 & 0 & 0.5 \end{bmatrix}$$

Example 1



Method of moments

$X_{n \times p}$ - n independent realizations of the p-dimensional random vector
$S = \frac{1}{n} X^T X$ - sample covariance matrix

Method of moments

$X_{n\times p}$ - n independent realizations of the p-dimensional random vector
$S = \frac{1}{n}X^TX$ - sample covariance matrix

$$ES = \Sigma$$

Method of moments

$X_{n\times p}$ - n independent realizations of the p-dimensional random vector
$S = \frac{1}{n}X^TX$ - sample covariance matrix

$$ES = \Sigma$$

$$\hat{\Omega}_1 = S^{-1}$$

Method of moments

$X_{n\times p}$ - n independent realizations of the p-dimensional random vector
$S = \frac{1}{n}X^T X$ - sample covariance matrix

$$ES = \Sigma$$

$$\hat{\Omega}_1 = S^{-1}$$

Gaussian cross-entropy function: $L(\Omega, X) C + \dfrac{n}{2}\log\det\Omega - \dfrac{n}{2}tr(S\Omega).$

Large p scenario

When S is invertible, $\hat{\Omega}_{MLE} = S^{-1}$

Large p scenario

When S is invertible, $\hat{\Omega}_{MLE} = S^{-1}$

$$ES^{-1} = \frac{n}{n-p-1}\Omega$$

Large p scenario

When S is invertible, $\hat{\Omega}_{MLE} = S^{-1}$

$$ES^{-1} = \frac{n}{n-p-1}\Omega$$

S is not invertible when $p > n$.

Large p scenario

When S is invertible, $\hat{\Omega}_{MLE} = S^{-1}$

$$ES^{-1} = \frac{n}{n-p-1}\Omega$$

S is not invertible when $p > n$.

$\hat{\Omega}_{MLE}$ is dense (inverse Wishart distribution)

Large p scenario

When S is invertible, $\hat{\Omega}_{MLE} = S^{-1}$

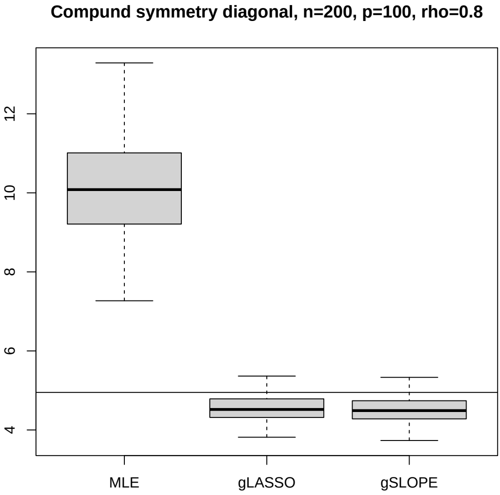
$$ES^{-1} = \frac{n}{n-p-1}\Omega$$

S is not invertible when $p > n$.

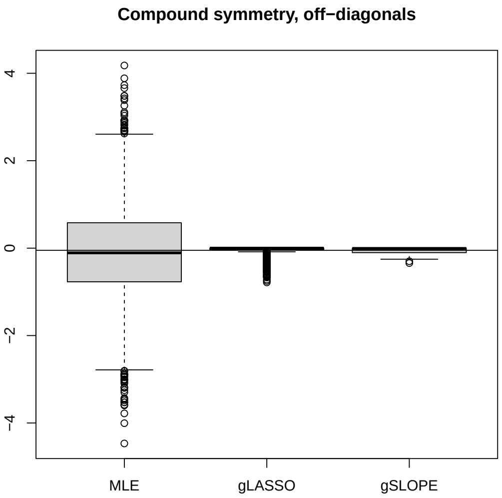
$\hat{\Omega}_{MLE}$ is dense (inverse Wishart distribution)

For large p, $MSE = E||\hat{\Omega} - \Omega||_F^2$ is large

Simulation example (1)



Simulation example (2)



Graphical LASSO, Friedman et al. (2008)

$$L(\Omega, X) = C + \frac{n}{2} \log \det \Omega - \frac{n}{2} tr(S\Omega).$$

Graphical LASSO, Friedman et al. (2008)

$$L(\Omega, X) = C + \frac{n}{2}\log\det\Omega - \frac{n}{2}tr(S\Omega).$$

$$\text{gLASSO:}\widehat{\Omega}_L = \arg\max_{\Omega \in Sym_+^p}\left[\log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1\right] \quad ,$$

$$||\Omega||_1 = \sum_{i<j} |\omega_{ij}| \quad .$$

Example: $n = 50$, $p = 30$, Σ - block diagonal with 3 blocks of dimension 10×10, correlation within blocks $\rho = 0.8$

Graphical LASSO, Friedman et al. (2008)

$$L(\Omega, X) = C + \frac{n}{2}\log\det\Omega - \frac{n}{2}tr(S\Omega).$$

gLASSO:$\widehat{\Omega}_L = \arg\max_{\Omega\in Sym_+^p}\left[\log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1\right]$,

$$||\Omega||_1 = \sum_{i<j}|\omega_{ij}| \quad .$$

Example:$n = 50$, $p = 30$, Σ - block diagonal with 3 blocks of dimension 10×10, correlation within blocks $\rho = 0.8$
$||\Omega - \hat{\Omega}_{MLE}||_F = 10800$

Graphical LASSO, Friedman et al. (2008)

$$L(\Omega, X) = C + \frac{n}{2}\log\det\Omega - \frac{n}{2}tr(S\Omega).$$

gLASSO:$\widehat{\Omega}_L = \arg\max_{\Omega \in Sym_+^p}\left[\log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1\right]$,

$$||\Omega||_1 = \sum_{i<j}|\omega_{ij}| \quad .$$

Example:$n = 50,\ p = 30$, Σ - block diagonal with 3 blocks of dimension 10×10, correlation within blocks $\rho = 0.8$
$||\Omega - \hat{\Omega}_{MLE}||_F = 10800$
Banerjee and d'Aspremont (2008), FWER control for block diagonal matrices with standardized entries:

$$\lambda_\alpha^{Banerjee} = \frac{t_{n-2}\left(1 - \frac{\alpha}{2p^2}\right)}{\sqrt{n - 2 + t_{n-2}^2\left(1 - \frac{\alpha}{2p^2}\right)}} \quad , \tag{1}$$

Graphical LASSO, Friedman et al. (2008)

$$L(\Omega, X) = C + \frac{n}{2}\log\det\Omega - \frac{n}{2}tr(S\Omega).$$

gLASSO:$\widehat{\Omega}_L = \arg\max_{\Omega \in Sym_+^p}\left[\log\det\Omega - tr(S\Omega) - \lambda||\Omega||_1\right]$,

$$||\Omega||_1 = \sum_{i<j}|\omega_{ij}| \quad .$$

Example:$n = 50$, $p = 30$, Σ - block diagonal with 3 blocks of dimension 10×10, correlation within blocks $\rho = 0.8$
$||\Omega - \hat{\Omega}_{MLE}||_F = 10800$
Banerjee and d'Aspremont (2008), FWER control for block diagonal matrices with standardized entries:

$$\lambda_\alpha^{Banerjee} = \frac{t_{n-2}\left(1 - \frac{\alpha}{2p^2}\right)}{\sqrt{n - 2 + t_{n-2}^2\left(1 - \frac{\alpha}{2p^2}\right)}} \quad , \qquad (1)$$

Perfect graph discovery, $||\Omega - \hat{\Omega}_{gLASSO}||_F = 468$

Dependence of scaling

Precision matrix depends on scaling factors:
$Cov(aX_i, X_j) = aCov(X_i, X_j)$

Dependence of scaling

Precision matrix depends on scaling factors:
$Cov(aX_i, X_j) = aCov(X_i, X_j)$

The sequence of models along the GLASSO paths depends on the scaling of individual variables

Dependence of scaling

Precision matrix depends on scaling factors:
$Cov(aX_i, X_j) = aCov(X_i, X_j)$

The sequence of models along the GLASSO paths depends on the scaling of individual variables

Standard solution - standardize the variables to the unit variance, i.e. estimate the inverse of the correlation matrix and than rescale

Dependence of scaling

Precision matrix depends on scaling factors:
$Cov(aX_i, X_j) = aCov(X_i, X_j)$

The sequence of models along the GLASSO paths depends on the scaling of individual variables

Standard solution - standardize the variables to the unit variance, i.e. estimate the inverse of the correlation matrix and than rescale

More reasonable solution - penalize the partial correlation matrix instead of the precision

Dependence of scaling

Precision matrix depends on scaling factors:
$Cov(aX_i, X_j) = aCov(X_i, X_j)$

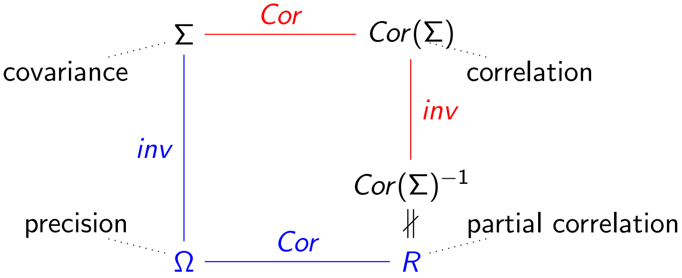
The sequence of models along the GLASSO paths depends on the scaling of individual variables

Standard solution - standardize the variables to the unit variance, i.e. estimate the inverse of the correlation matrix and than rescale

More reasonable solution - penalize the partial correlation matrix instead of the precision

Partial correlation : $\rho_{i,j} = -\frac{\Omega_{i,j}}{\sqrt{\Omega_{i,i}\Omega_{j,j}}} = -R_{i,j}$

Precision vs Partial correlation



Rysunek

Partial correlation vs inverse of the correlation

Typically, partial correlation R has a different ordering of off-diagonal entries than the precision of the standardised data $Cor(\Sigma)^{-1}$.

$$\Sigma = \begin{bmatrix} 2 & -1 & -2 & 2 \\ -1 & 1 & 1 & -1 \\ -2 & 1 & 3 & -3 \\ 2 & -1 & -3 & 3.5 \end{bmatrix} \quad Cor(\Sigma)^{-1} = \begin{bmatrix} 4 & 1.414 & 2.449 & 0 \\ 1.414 & 2 & 0 & 0 \\ 2.449 & 0 & 9 & 6.48 \\ 0 & 0 & 6.48 & 7 \end{bmatrix}$$

$$\Omega = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 3 & 2 \\ 0 & 0 & 2 & 2 \end{bmatrix} \quad R = \begin{bmatrix} 1 & 0.5 & 0.408 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0.408 & 0 & 1 & 0.816 \\ 0 & 0 & 0.816 & 1 \end{bmatrix}$$

Partial correlation LASSO

A. Chojecki, I. Hejny, B. Kołodziejek, J. Wallin, MB

J.S.Carter, D. Rossell, J. Q. Smith, SJS, 2023

We represent $\Omega = DRD$, where $D^2 = diag(\Omega)$, and define its estimator as

$$\hat{\Omega}_{\mathrm{pcg}} = \hat{D}\hat{R}\hat{D},$$

where

$$(\hat{R}, \hat{D}) \in Argmax_{R,D} \{\log\det(DRD) - SDRD - \lambda\|R\|_{1,\mathrm{off}} - 2\alpha\log\det(D) \tag{2}$$

and $\lambda \geq 0$ and $\alpha < 1$ are tuning parameters.

Convexity issues

The objective function is non-concave and there might exist more than one global maxima

Convexity issues

The objective function is non-concave and there might exist more than one global maxima

Let $\hat{C}$ denote the sample correlation matrix.

Theorem

(i) *If* $\|\hat{C} - I_p\|_\infty \leq (2(1-\alpha)p^3)^{-1/2}$, *then for any* $\lambda \geq 0$, *PCGLASSO admits a unique local maximum.*

(ii) *There exist* $\lambda_0 > 0$ *and* $\alpha_0 > 0$ *such that, for every* $\hat{C}$, $\lambda \in (0, \lambda_0)$ *and* $\alpha \in (-\infty, \alpha_0)$, *PGLASSO admits a unique local maximum.*

Algorithm

The function is biconcave, i.e. for fixed D it is concave as a function of R and vice versa

Algorithm

The function is biconcave, i.e. for fixed D it is concave as a function of R and vice versa

We iteratively optimize with respect to D and R

Algorithm

The function is biconcave, i.e. for fixed D it is concave as a function of R and vice versa

We iteratively optimize with respect to D and R

For fixed D we optimize R using an adaptation of the block coordinate descent method for GLASSO

Algorithm

The function is biconcave, i.e. for fixed D it is concave as a function of R and vice versa

We iteratively optimize with respect to D and R

For fixed D we optimize R using an adaptation of the block coordinate descent method for GLASSO

For fixed R we optimize D using a modified Newton-Raphson method with backtracking line search.

Asymptotic distribution

We minimize

$$-logdet(\Omega) + tr(\Omega S) + n^{-1/2} Pen(\Omega)$$

We want to compare penalization of the precision matrix Ω and of the partial correlation matrix
$R = g(\Omega) = Cor(\Omega) = d(\Omega)^{-1/2} \Omega d(\Omega)^{-1/2}$:

$$Pen(\Omega) = \begin{cases} f(\Omega) + h(\Omega) \ldots \text{ precision penalization} \\ f(g(\Omega)) + h(\Omega) \ldots \text{ partial correlation penalization} \end{cases},$$

where $h(\Omega) = \alpha log(det(d(\Omega)))$ is a seperate penalty on the diagonal of Ω.

Low dimensional asymptotics

Theorem

Assume x follows a centered elliptical distribution with $Cov(x) = \Sigma = \Omega_0^{-1}$. Then $\sqrt{n}(\widehat{\Omega}_n - \Omega_0)$ converges weakly and in pattern to the minimizer of $V : \mathbb{R}^{p^2} \to \mathbb{R}$;

$$\frac{1}{2} vec(U)^T C vec(U) - vec(U)^T W + Pen'(\Omega_0; U), \qquad (3)$$

where $W \sim \mathcal{N}(0, C_\triangle)$,

$$C = \frac{1}{2}(\Omega_0^{-1} \otimes \Omega_0^{-1}),$$
$$C_\triangle = Cov(vec(x^T x))/4$$

and $Pen'(\theta_0; u)$ denotes the directional derivative of the penalty Pen at θ_0 in direction u.

Pattern recovery

Theorem

Let $\Omega_0 = DRD$. The irrepresentability condition is

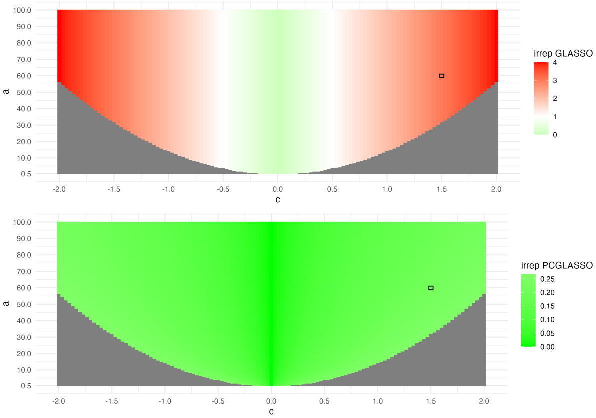
$$||\tilde{\Gamma}_{S^cS}(\tilde{\Gamma}_{SS})^+ vec(sign(od(R)))||_\infty < 1$$

where $\tilde{\Gamma} := M_R^+(R^{-1} \otimes R^{-1})$, $M_R = I - (1/2)P_d(I \otimes R + R \otimes I)$, P_d a projection satisfying $P_d vec(U) = vec(d(U))$ and M_R^+ the Moore-Penrose inverse. If the above condition is satisfied, then PCGLasso is model consistent in the sense

$$\lim_{n\to\infty} \mathbb{P}(sign(\hat{\Omega}_n) = sign(\Omega_0)) > 1 - e^{c\gamma},$$

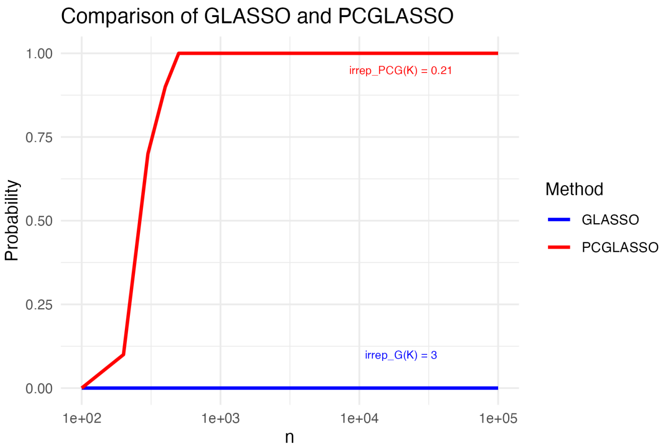
for some $c > 0$ and arbitrary $\gamma \geq 0$, where $\hat{\Omega}_n$ minimizes -logdet$(\Omega) + tr(\Omega S) + n^{-1/2}\gamma Pen(\Omega)$.

Irrepresentability condition (1)

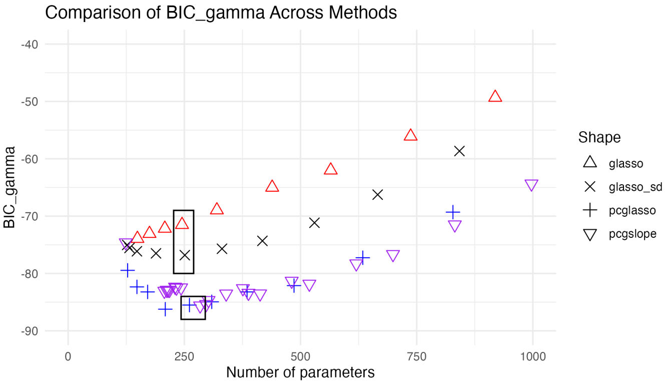


Rysunek: Heatmap representing the value of the IRR formula for the hub graphs such that $K[1,1] = a$; $K[i,i] = 1$; $K[i,1] = K[1,i] = c$ and $K[i,j] = 0$; size $p = 15$. Green area marks the region where IR is satisfied, grey area marks the region where K is not positive semidefinite (namely, $a \leq (p-1)c^2$).

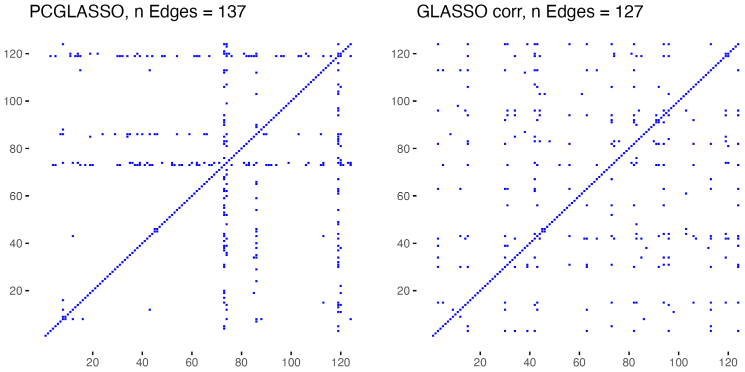
Irrepresentability condition (2)



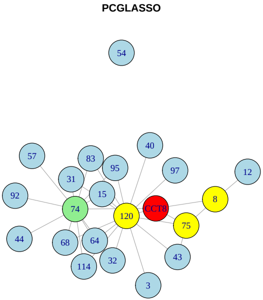
Analysis of gene expression data

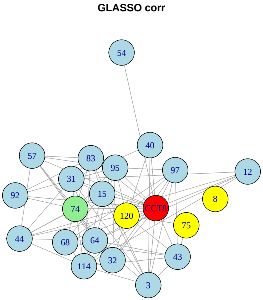


Analysis of gene expression data (2)

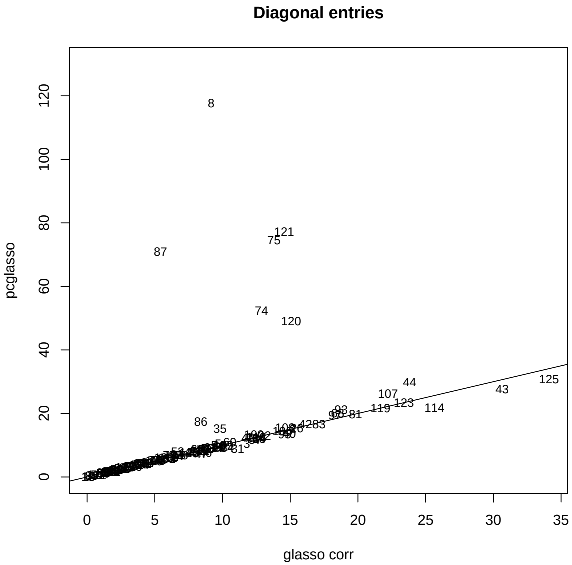


Analysis of gene expression data (3)





Analysis of gene expression data(4)



Analysis of gene expression data (5)