Precision matrix estimation in Gaussian graphical models

Michał Makowski

Faculty of Mathematics and Computer Science
University of Wrocław

michalmakowski@outlook.com

February 15, 2019

Overview

Graphical models

- Each vertex represents a random variable.
- Useful for either unsupervised or supervised learning.
- Directed or undirected.
- Represents joint distribution.

Undirected graphical models

The absence of an edge between two vertices has a special meaning: the corresponding random variables are conditionally independent, given the other variables.

Example 1/2

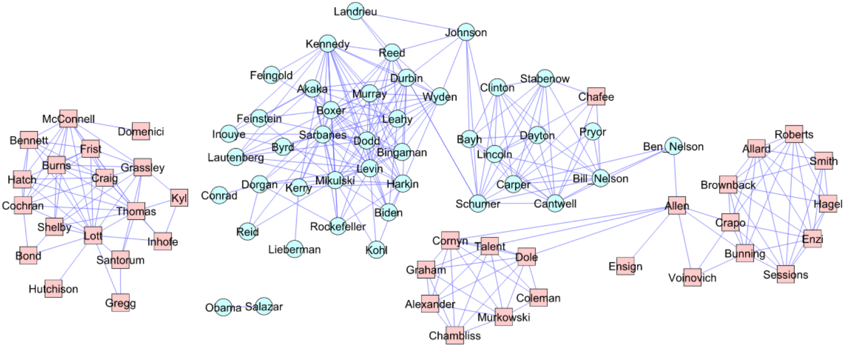


Figure 16: US Senate, 109th Congress (2004-2006). The graph displays the solution to (12) obtained using the log determinant relaxation to the log partition function of Wainwright and Jordan (2006). Democratic senators are represented by round nodes and Republican senators are represented by square nodes.

Example 2/2

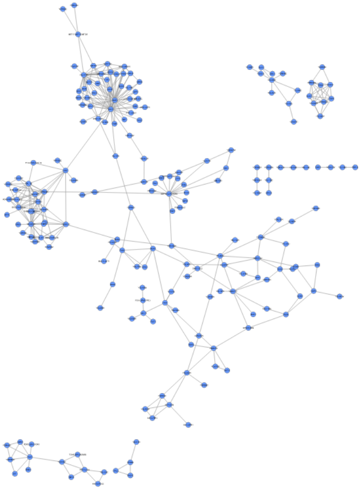


Figure 13: Application to Hughes compendium. The above graph results from solving (1) for this data set with a penalty parameter of $\lambda = 0.0313$.

Factorization

Any multivariate normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ can reparametrized into canonical parameters of the form

$$\gamma = \mathbf{\Sigma}^{-1}\mu \quad \text{and} \quad \mathbf{\Theta} = \mathbf{\Sigma}^{-1}.$$

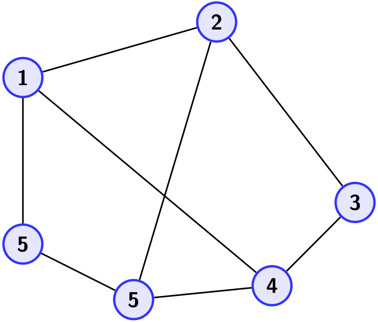
Factorization

Any multivariate normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ can reparametrized into canonical parameters of the form

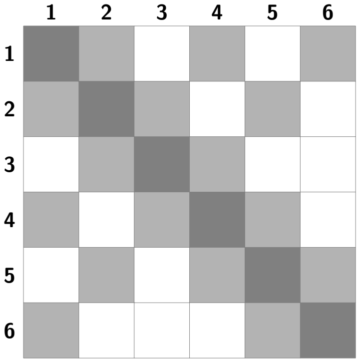
$$\gamma = \mathbf{\Sigma}^{-1}\mu \quad \text{and} \quad \mathbf{\Theta} = \mathbf{\Sigma}^{-1}.$$

If $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ factorizes according to some graph G, $\theta_{st} = 0$ for any pair $(s, t) \notin E$, which sets up correspondence between the zero pattern of the matrix $\mathbf{\Theta}$ and pattern of the underlying graph. In particular, if the $\theta_{st} = 0$, then variables s and t are conditionally independent, given the other variables.

Graph and matrix correspondence



(a) The undirected graph G on six vertices.



(b) The associated sparsity pattern of the precision matrix $\mathbf{\Theta}$. White squares correspond to zero entries.

Maximum likelihood estimator...

MLE

$$\widehat{\mathbf{\Theta}}_{ML} \in \underset{\Theta \in S_{+}^{p}}{\arg\max} \left\{ \log \det \mathbf{\Theta} - \operatorname{tr}\left(\mathbf{S}\,\mathbf{\Theta}\right) \right\}$$

Maximum likelihood estimator...

MLE

$$\widehat{\mathbf{\Theta}}_{ML} \in \underset{\Theta \in S_+^p}{\arg\max} \left\{ \log\det \mathbf{\Theta} - \operatorname{tr}\left(\mathbf{S}\,\mathbf{\Theta}\right)\right\}$$

When the maximum is attained the solution is given by

$$\mathbf{S}^{-1} = \widehat{\mathbf{\Theta}},$$

or its truncated version

...and its problems

In case when the number of nodes p is comparable to, or larger than, the sample size N, the sample covariance $\mathbf{S}$ is singular (so $\mathbf{S}^{-1}$ does not exist), so the MLE. Moreover, sometimes we are looking for *sparse* solutions.

Regularization

We can control the number of edges, which can be measured by ℓ_0-based quantity

$$\rho_0(\mathbf{\Theta}) = \sum_{s \neq t} \mathbb{I}[\theta_{st} \neq 0].$$

Note that $\rho_0(\mathbf{\Theta}) = 2|E(G)|$ for a given graph G.

Regularization

We can control the number of edges, which can be measured by ℓ_0-based quantity

$$\rho_0(\mathbf{\Theta}) = \sum_{s \neq t} \mathbb{I}[\theta_{st} \neq 0].$$

Note that $\rho_0(\mathbf{\Theta}) = 2|E(G)|$ for a given graph G.

ℓ_0-based problem

$$\widehat{\Theta} \in \underset{\substack{\Theta \in S_+^p \\ \rho_0(\mathbf{\Theta}) \leq k}}{\arg\max} \{\log\det \mathbf{\Theta} - \mathrm{tr}\,(\mathbf{S}\,\mathbf{\Theta})\}$$

Unfortunately, the ℓ_0-based constrained defines a highly nonconvex constraint set.

Graphical Lasso

Convex relaxation of ℓ_0-based constrain leads to

$$\mathbb{L}_\lambda(\mathbf{\Theta}, \mathbf{X}) = \log\det\mathbf{\Theta} - \mathrm{tr}\,(\mathbf{S}\,\mathbf{\Theta}) - \lambda\|\,\mathbf{\Theta}\,\|_1.$$

where $\|\cdot\|_1$ states for entrywise off-diagonal ℓ_1-norm $\|A\|_1 = \sum_{i\neq j}|a_{ij}|$.

Graphical Lasso

Convex relaxation of ℓ_0-based constrain leads to

$$\mathbb{L}_\lambda(\mathbf{\Theta}, \mathbf{X}) = \log\det\mathbf{\Theta} - \operatorname{tr}(\mathbf{S}\,\mathbf{\Theta}) - \lambda\|\,\mathbf{\Theta}\,\|_1.$$

where $\|\cdot\|_1$ states for entrywise off-diagonal ℓ_1-norm $\|A\|_1 = \sum_{i\neq j}|a_{ij}|$.

Graphical Lasso problem

$$\widehat{\mathbf{\Theta}} \in \underset{\Theta\in S_+^p}{\arg\max}\left\{\log\det\mathbf{\Theta} - \operatorname{tr}(\mathbf{S}\,\mathbf{\Theta}) - \lambda\|\,\mathbf{\Theta}\,\|_1\right\}.$$

Graphical Lasso parameter choice

Banerjee lambda for Graphical Lasso

$$\lambda^{\text{Banerjee}}(\alpha) = \max_{i<j}(s_{ii}, s_{jj}) \frac{\mathrm{qt}_{n-2}(1 - \frac{\alpha}{2p^2})}{\sqrt{n - 2 + \mathrm{qt}^2_{n-2}(1 - \frac{\alpha}{2p^2})}} \tag{1}$$

The following theorem was formulated by Banerjee et al.

Theorem

Using (1) as the penalty parameter in Graphical Lasso problem, for any fixed level α we obtain

$$\mathbb{P}(\text{False Discovery}) \leq \alpha,$$

where **False Discovery** *means there is a nonzero coefficient of the estimated precision matrix, which is zero in the real precision matrix.*

Graphical SLOPE

Instead of ordinary ℓ_1 norm we want to use OL1 norm

OL1

$$\mathsf{J}_{\lambda}(\boldsymbol{\Theta}) = \sum_i \lambda_i |\theta|_{(i)}$$

Graphical SLOPE

Instead of ordinary ℓ_1 norm we want to use OL1 norm

OL1

$$\mathrm{J}_\lambda(\mathbf{\Theta}) = \sum_i \lambda_i |\theta|_{(i)}$$

Thus, we maximize

$$\mathbb{L}_\lambda(\mathbf{\Theta}, \mathbf{X}) = \log\det \mathbf{\Theta} - \mathrm{tr}\left(\mathbf{S}\,\mathbf{\Theta}\right) - \mathrm{J}_\lambda(\mathbf{\Theta}).$$

Graphical SLOPE problem

$$\widehat{\mathbf{\Theta}} \in \underset{\Theta \in S_+^p}{\arg\max} \left\{\log\det \mathbf{\Theta} - \mathrm{tr}\left(\mathbf{S}\,\mathbf{\Theta}\right) - \mathrm{J}_\lambda(\mathbf{\Theta})\right\},$$

Graphical SLOPE parameter choice (1/2)

Holm lambda for Graphical SLOPE

$$m = \frac{p(p-1)}{2},$$

$$\lambda_k^{\text{Holm}} = \frac{\mathrm{qt}_{n-2}(1 - \frac{\alpha k}{m})}{\sqrt{n - 2 + \mathrm{qt}^2_{n-2}(1 - \frac{\alpha k}{m})}},$$

$$\lambda^{\text{Holm}} = \{\lambda_1^{\text{Holm}}, \lambda_2^{\text{Holm}}, ..., \lambda_m^{\text{Holm}}\}.$$

It is based on Holm method for multiple testing.

Graphical SLOPE parameter choice (2/2)

BH lambda for Graphical SLOPE

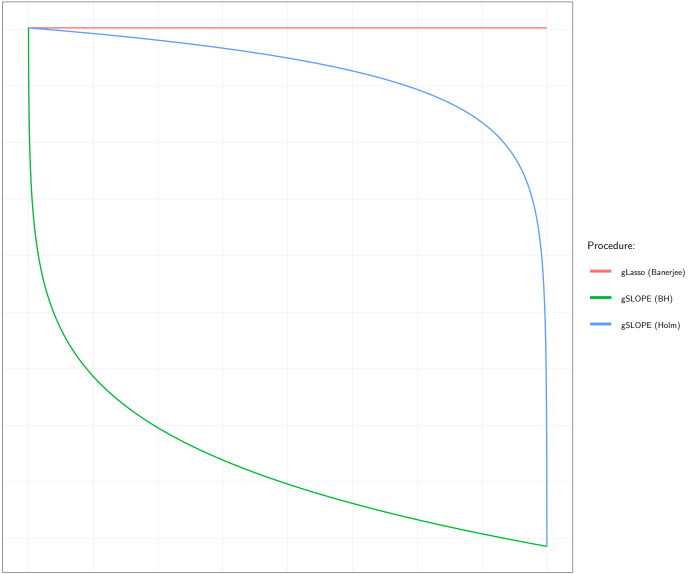
$$m = \frac{p(p-1)}{2},$$

$$\lambda_k^{\mathrm{BH}} = \frac{\mathrm{qt}_{n-2}(1 - \frac{\alpha}{m+1-k})}{\sqrt{n - 2 + \mathrm{qt}^2_{n-2}(1 - \frac{\alpha}{m+1-k})}},$$

$$\lambda^{\mathrm{BH}} = \{\lambda_1^{\mathrm{BH}}, \lambda_2^{\mathrm{BH}}, ..., \lambda_m^{\mathrm{BH}}\}.$$

It is based on Benjamini-Hochberg procedure for multiple testing.

Lambda comparison



Algorithms

For solving the Graphical SLOPE problem we used the *Alternating direction method of multipliers*, it can solve convex problems of the form

$$\begin{aligned} \text{minimize} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax + By = c. \end{aligned}$$

Algorithms

For solving the Graphical SLOPE problem we used the *Alternating direction method of multipliers*, it can solve convex problems of the form

$$\begin{aligned} \text{minimize} \quad & f(x) + g(y) \\ \text{subject to} \quad & Ax + By = c. \end{aligned}$$

For solving the Graphical Lasso problem we used an algorithm proposed by Friedman et al. in theirs first work about this method. Although we derived an ADMM-based algorithm, it was orders of magnitude slower than original one.

Overview

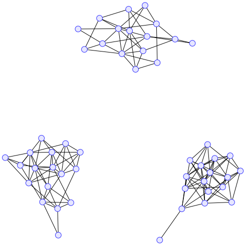
- Implementation with R, **huge** package for simulation.

Overview

- Implementation with R, **huge** package for simulation.
- Various types of graphs structure:

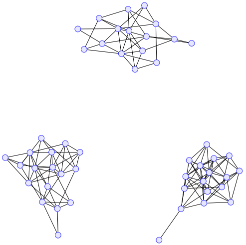
Overview

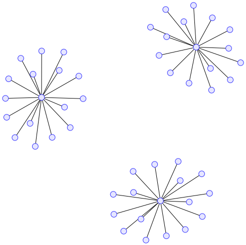
- Implementation with R, **huge** package for simulation.
- Various types of graphs structure: cluster



Overview

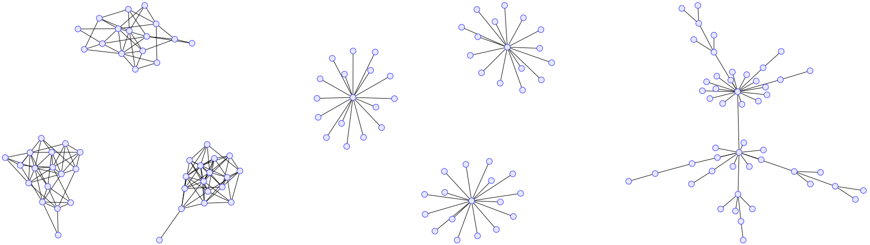
- Implementation with R, **huge** package for simulation.
- Various types of graphs structure: cluster, hub





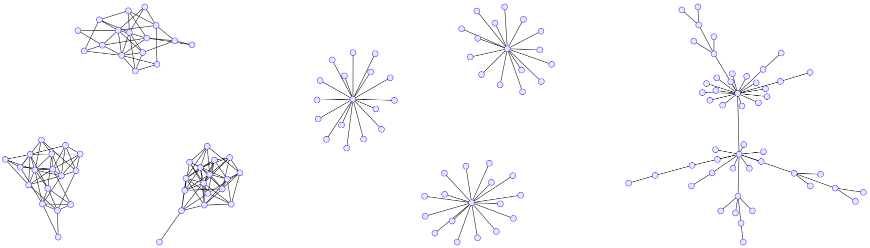
Overview

- Implementation with R, **huge** package for simulation.
- Various types of graphs structure: cluster, hub, and scale-free.



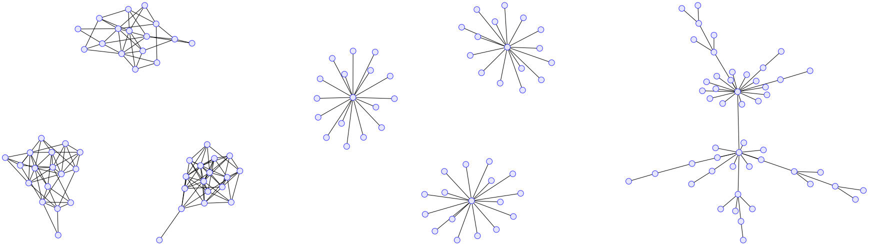
Overview

- Implementation with R, **huge** package for simulation.
- Various types of graphs structure: cluster, hub, and scale-free.
- Data: $p = 100$, $n \in \{50, 100, 200, 400\}$; different magnitude ratio; different sparsity and size of component.



Overview

- Implementation with R, **huge** package for simulation.
- Various types of graphs structure: cluster, hub, and scale-free.
- Data: $p = 100$, $n \in \{50, 100, 200, 400\}$; different magnitude ratio; different sparsity and size of component.
- Two levels of desirable FDR control: 0.05 and 0.2 .



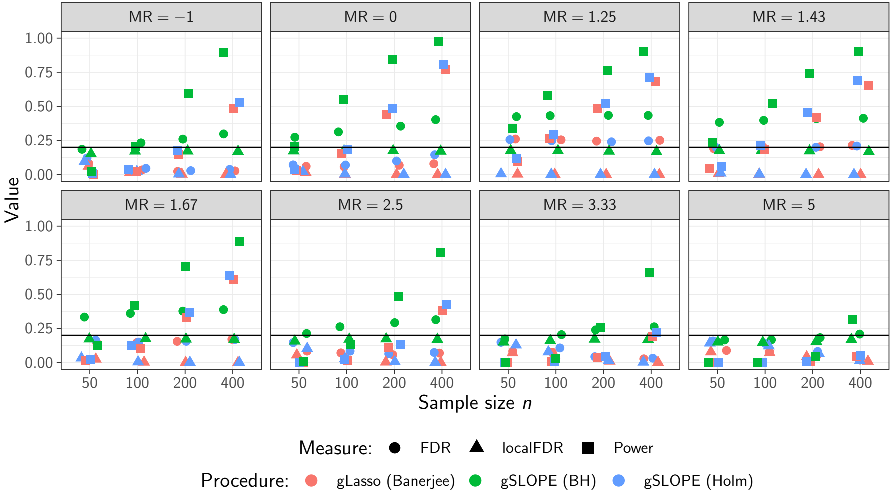
Measures

$$\mathrm{FDR} = \mathbb{E}\left[\frac{\#[\text{False positive}]}{\#[\text{False positive}] + \#[\text{True positive}]}\right]$$

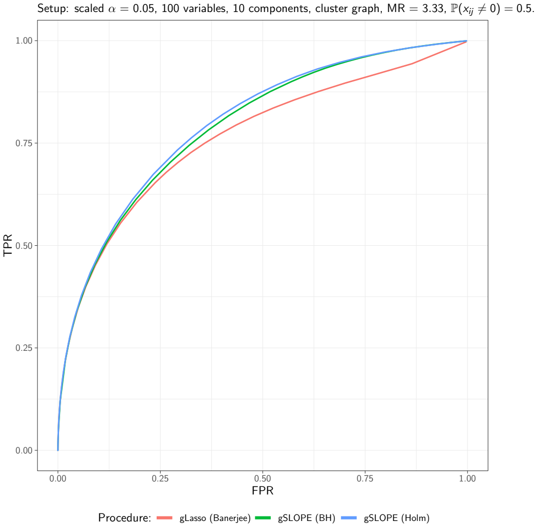
$$\mathrm{localFDR} = \mathbb{E}\left[\frac{\#[\text{False positive outside the component}]}{\#[\text{False positive}] + \#[\text{True positive}]}\right]$$

Cluster results

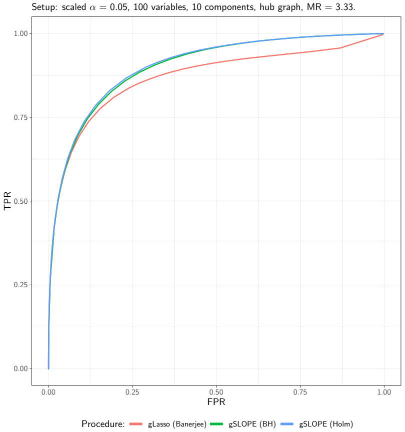
Setup: $\alpha = 0.2$, 100 variables, 10 components, cluster graph, $\mathbb{P}(x_{ij} \neq 0) = 0.5$.



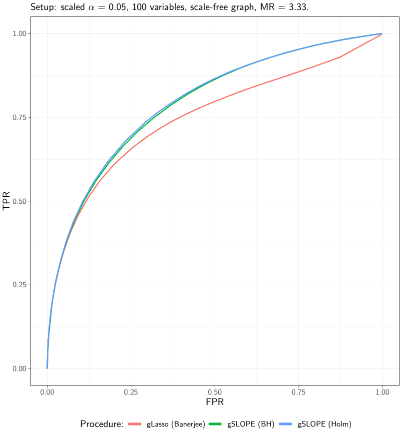
Cluster ROC


Setup: scaled $\alpha = 0.05$, 100 variables, 10 components, cluster graph, MR = 3.33, $\mathbb{P}(x_{ij} \neq 0) = 0.5$.
TPR
1.00
0.75
0.50
0.25
0.00
FPR
0.00
0.25
0.50
0.75
1.00
Procedure:
gLasso (Banerjee)
gSLOPE (BH)
gSLOPE (Holm)

Hub ROC



Setup: scaled $\alpha = 0.05$, 100 variables, 10 components, hub graph, MR = 3.33.
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Procedure: gLasso (Banerjee) gSLOPE (BH) gSLOPE (Holm)



Setup: scaled $\alpha = 0.05$, 100 variables, scale-free graph, MR = 3.33.
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Procedure: gLasso (Banerjee) gSLOPE (BH) gSLOPE (Holm)

Bibliography

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data". In: *The Journal of Machine Learning Research* 9 (2008), pp. 485–516.

Małgorzata Bogdan et al. "SLOPE - Adaptive variable selection via convex optimization". In: *The annals of applied statistics* 9.3 (2015), pp. 1103–1140. DOI: 10.1214/15-AOAS842.

Emmanuel Candes. *Advanced Topics in Convex Optimization*. 2015.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics (Oxford, England)* 9.3 (July 2008), pp. 432–41. DOI: 10.1093/biostatistics/kxm045.

Trevor Hastie et al. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015, p. 367. ISBN: 9781498712163.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Second. Springer-Verlag New York, 2009, p. 745. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7.

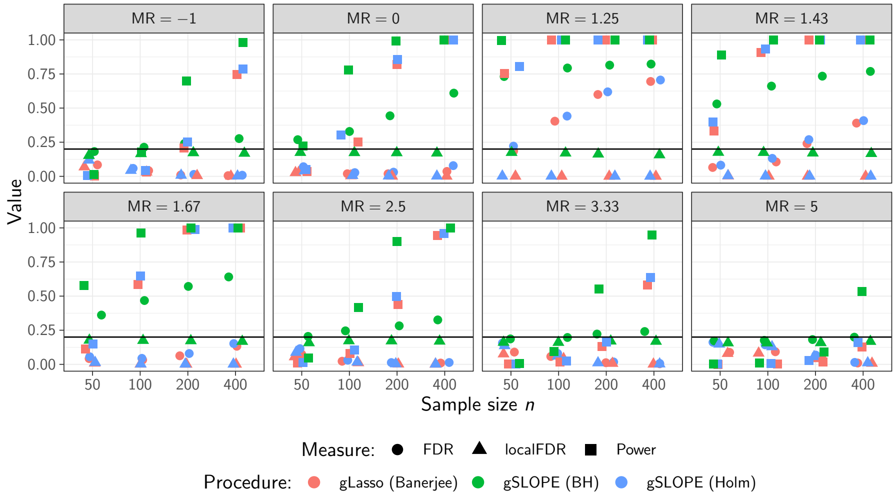
Kathryn Roeder, John Lafferty, and Larry Wasserman. *The huge Package for High-dimensional Undirected Graph Estimation in R*. 2012.

Piotr Sobczyk. "Identifying low-dimensional structures through model selection in high-dimensional data". PhD thesis. Wrocław University of Science and Technology, 2018.

Thank you!

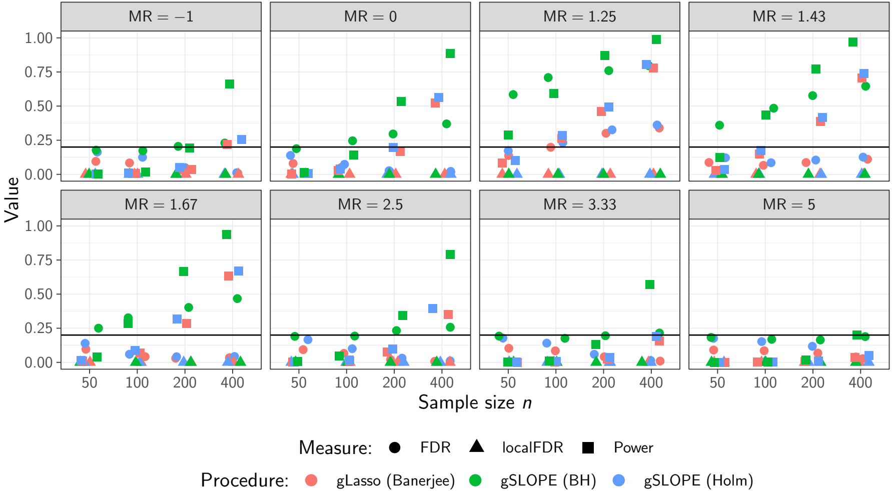
Hub results

Setup: $\alpha = 0.2$, 100 variables, 10 components, hub graph.



Scale-free results

Setup: $\alpha = 0.2$, 100 variables, scale-free graph.



Factorization theorem

Compatibility function

Let $G = (V, E)$ be a graph with a vertex set $V = 1, 2, \ldots, p$ and $\mathfrak{C}$ be its clique set. Let $\mathbb{X} = (X_1, \ldots, X_p)$ be a random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, indexed by the graph nodes.

Definition (Compatibility function)

Let $C \in \mathfrak{C}$ be a clique of the graph G and let $\mathbb{X}_C$ be a subvector of the vector $\mathbb{X}$ indexed by the elements of the clique C, that is $\mathbb{X}_C = (X_s, s \in C)$. A real-valued function ψ_C of the vector $\mathbb{X}_C$ taking positive real values is called a *compatibility function*.

Factorization property

Definition (Factorization)

Let $C \in \mathfrak{C}$ be a clique of the graph G and let $\mathbb{X}_C$ be a subvector of the vector $\mathbb{X}$ indexed by the elements of the clique C, that is $\mathbb{X}_C = (X_s, s \in C)$. A real-valued function ψ_C of the vector $\mathbb{X}_C$ taking positive real values is called a *compatibility function*.

Given a collection of compatibility functions, we say that probability distribution $\mathbb{P}$ *factorizes over G* if it has decomposition

$$\mathbb{P}(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{C \in \mathfrak{C}} \psi_C(x_C), \tag{2}$$

where Z is the normalizing constant, known as the *partition function*. It is given by

$$Z = \sum_x \prod_{C \in \mathfrak{C}} \psi_C(x_C), \tag{3}$$

where the sum goes over all possible realizations of $\mathbb{X}$.

Markov property

Consider a cut set S of the given graph and let introduce a symbol $\perp\!\!\!\perp$ to denote the relation *is conditionally independent of*. With this notation, we say that the random vector $\mathbb{X}$ is Markov with respect to G if

$$\mathbb{X}_A \perp\!\!\!\perp \mathbb{X}_B \mid \mathbb{X}_S \qquad \text{for all cut sets } S \subset V, \tag{4}$$

where $\mathbb{X}_A$ denotes the subvector indexed by the subgraph A.

Canonical formulation

Canonical formulation

Any nondegenerated multivariate normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ can reparametrized into canonical parameters of the form

$$\gamma = \mathbf{\Sigma}^{-1}\mu \quad \text{and} \quad \mathbf{\Theta} = \mathbf{\Sigma}^{-1}.$$

Then density function is given by

$$\mathbb{P}_{\gamma,\mathbf{\Theta}}(x) = \exp\left\{\sum_{s=1}^{p}\gamma_s x_s - \frac{1}{2}\sum_{s,t=1}^{p}\theta_{st}x_s x_t - A(\gamma, \mathbf{\Theta})\right\},$$

where $A(\gamma, \mathbf{\Theta}) = -\frac{1}{2}\left(\det[(2\pi)^{-1}\,\mathbf{\Theta}] + \gamma^T\,\mathbf{\Theta}^{-1}\,\gamma\right)$.

Canonical formula derivation

$$\mathbb{P}_{\mu,\mathbf{\Sigma}}(x) = \left(\sqrt{\det[2\pi\mathbf{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right\}$$

Canonical formula derivation

$$\mathbb{P}_{\mu,\mathbf{\Sigma}}(x) = \left(\sqrt{\det[2\pi\mathbf{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right\}$$
$$= \left(\sqrt{\det[(2\pi\mathbf{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\mathbf{\Sigma}^{-1}x + x^T\mathbf{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\mathbf{\Sigma}^{-1}\mu\right\}$$

Canonical formula derivation

$$
\begin{aligned}
\mathbb{P}_{\mu,\mathbf{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\mathbf{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\mathbf{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\mathbf{\Sigma}^{-1}x + x^T\mathbf{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\mathbf{\Sigma}^{-1}\mu\right\} \\
&= \left(\sqrt{\det[(2\pi)^{-1}\,\mathbf{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\mathbf{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\mathbf{\Theta}^{-1}\,\gamma\right\}
\end{aligned}
$$

Canonical formula derivation

$$\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) = \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\}$$

$$= \left(\sqrt{\det[(2\pi\boldsymbol{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\boldsymbol{\Sigma}^{-1}x + x^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu\right\}$$

$$= \left(\sqrt{\det[(2\pi)^{-1}\,\boldsymbol{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right\}$$

$$= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\left(\det[(2\pi)^{-1}\,\boldsymbol{\Theta}] + \gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right)\right\}$$

Canonical formula derivation

$$
\begin{aligned}
\mathbb{P}_{\mu,\mathbf{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\mathbf{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\mathbf{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\mathbf{\Sigma}^{-1}x + x^T\mathbf{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\mathbf{\Sigma}^{-1}\mu\right\} \\
&= \left(\sqrt{\det[(2\pi)^{-1}\,\mathbf{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\mathbf{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\mathbf{\Theta}^{-1}\,\gamma\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\mathbf{\Theta}\,x + x^T\gamma - \frac{1}{2}\left(\det[(2\pi)^{-1}\,\mathbf{\Theta}] + \gamma^T\,\mathbf{\Theta}^{-1}\,\gamma\right)\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\mathbf{\Theta}\,x + x^T\gamma - A(\gamma,\mathbf{\Theta})\right\}
\end{aligned}
$$

Canonical formula derivation

$$
\begin{aligned}
\mathbb{P}_{\mu,\boldsymbol{\Sigma}}(x) &= \left(\sqrt{\det[2\pi\boldsymbol{\Sigma}]}\right)^{-1} \exp\left\{(-\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(x-\mu)\right\} \\
&= \left(\sqrt{\det[(2\pi\boldsymbol{\Sigma})^{-1}]}\right) \exp\left\{-\frac{1}{2}x^T\boldsymbol{\Sigma}^{-1}x + x^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu\right\} \\
&= \left(\sqrt{\det[(2\pi)^{-1}\,\boldsymbol{\Theta}]}\right)^{-1} \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - \frac{1}{2}\left(\det[(2\pi)^{-1}\,\boldsymbol{\Theta}] + \gamma^T\,\boldsymbol{\Theta}^{-1}\,\gamma\right)\right\} \\
&= \exp\left\{-\frac{1}{2}x^T\,\boldsymbol{\Theta}\,x + x^T\gamma - A(\gamma,\boldsymbol{\Theta})\right\} \\
&= \mathbb{P}_{\gamma,\boldsymbol{\Theta}}(x)
\end{aligned}
$$

Log-likelihood derivation

Log-likelihood derivation (1/2)

$$\mathbb{L}(\mathbf{\Theta}, \mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\mathbf{\Theta}}(x_i)$$

Log-likelihood derivation (1/2)

$$
\begin{aligned}
\mathbb{L}(\boldsymbol{\Theta}, \mathbf{X}) &= \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\boldsymbol{\Theta}}(x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i - A(\boldsymbol{\Theta})
\end{aligned}
$$

Log-likelihood derivation (1/2)

$$
\begin{aligned}
\mathbb{L}(\mathbf{\Theta}, \mathbf{X}) &= \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\mathbf{\Theta}}(x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \mathbf{\Theta} \, x_i - A(\mathbf{\Theta}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \log \det[(2\pi)^{-1} \, \mathbf{\Theta}] - \frac{1}{2} x_i^T \, \mathbf{\Theta} \, x_i
\end{aligned}
$$

Log-likelihood derivation (1/2)

$$
\begin{aligned}
\mathbb{L}(\boldsymbol{\Theta}, \mathbf{X}) &= \frac{1}{N}\sum_{i=1}^{N} \log \mathbb{P}_{\boldsymbol{\Theta}}(x_i) \\
&= \frac{1}{N}\sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i - A(\boldsymbol{\Theta}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2} \log \det[(2\pi)^{-1} \, \boldsymbol{\Theta}] - \frac{1}{2} x_i^T \, \boldsymbol{\Theta} \, x_i \\
&= \frac{1}{2N}\sum_{i=1}^{N} \log \Big((2\pi)^{-N} \det[\boldsymbol{\Theta}] \Big) - x_i^T \, \boldsymbol{\Theta} \, x_i
\end{aligned}
$$

Log-likelihood derivation (1/2)

$$
\begin{aligned}
\mathbb{L}(\mathbf{\Theta}, \mathbf{X}) &= \frac{1}{N} \sum_{i=1}^{N} \log \mathbb{P}_{\mathbf{\Theta}}(x_i) \\
&= \frac{1}{N} \sum_{i=1}^{N} -\frac{1}{2} x_i^T \, \mathbf{\Theta} \, x_i - A(\mathbf{\Theta}) \\
&= \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \log \det[(2\pi)^{-1} \, \mathbf{\Theta}] - \frac{1}{2} x_i^T \, \mathbf{\Theta} \, x_i \\
&= \frac{1}{2N} \sum_{i=1}^{N} \log \left((2\pi)^{-N} \det[\mathbf{\Theta}] \right) - x_i^T \, \mathbf{\Theta} \, x_i \\
&= \frac{1}{2N} \sum_{i=1}^{N} \log \det \mathbf{\Theta} - N \log 2\pi - x_i^T \, \mathbf{\Theta} \, x_i = \ldots
\end{aligned}
$$

Log-likelihood derivation (2/2)

$$\ldots = \frac{1}{2N} \sum_{i=1}^{N} \log \det \boldsymbol{\Theta} - N \log 2\pi - x_i^T \, \boldsymbol{\Theta} \, x_i$$

Log-likelihood derivation (2/2)

$$
\begin{aligned}
\ldots &= \frac{1}{2N}\sum_{i=1}^{N} \log\det \boldsymbol{\Theta} - N\log 2\pi - x_i^T\, \boldsymbol{\Theta}\, x_i \\
&= \frac{1}{2N}\sum_{i=1}^{N} \log\det \boldsymbol{\Theta} - N\log 2\pi - \operatorname{tr}\left(x_i^T\, \boldsymbol{\Theta}\, x_i\right)
\end{aligned}
$$

Log-likelihood derivation (2/2)

$$
\begin{aligned}
\dots &= \frac{1}{2N}\sum_{i=1}^{N} \log\det \mathbf{\Theta} - N\log 2\pi - x_i^T\,\mathbf{\Theta}\,x_i \\
&= \frac{1}{2N}\sum_{i=1}^{N} \log\det \mathbf{\Theta} - N\log 2\pi - \operatorname{tr}\left(x_i^T\,\mathbf{\Theta}\,x_i\right) \\
&= \frac{1}{2}\log\det \mathbf{\Theta} - \frac{N}{2}\log 2\pi - \frac{1}{2N}\sum_{i=1}^{N} \operatorname{tr}\left(x_i x_i^T\,\mathbf{\Theta}\right)
\end{aligned}
$$

Log-likelihood derivation (2/2)

$$
\begin{aligned}
\ldots &= \frac{1}{2N}\sum_{i=1}^{N} \log\det \boldsymbol{\Theta} - N\log 2\pi - x_i^T\, \boldsymbol{\Theta}\, x_i \\
&= \frac{1}{2N}\sum_{i=1}^{N} \log\det \boldsymbol{\Theta} - N\log 2\pi - \operatorname{tr}\left(x_i^T\, \boldsymbol{\Theta}\, x_i\right) \\
&= \frac{1}{2}\log\det \boldsymbol{\Theta} - \frac{N}{2}\log 2\pi - \frac{1}{2N}\sum_{i=1}^{N} \operatorname{tr}\left(x_i x_i^T\, \boldsymbol{\Theta}\right) \\
&= \frac{1}{2}\log\det \boldsymbol{\Theta} - \frac{N}{2}\log 2\pi - \frac{1}{2}\operatorname{tr}\left(\mathbf{S}\, \boldsymbol{\Theta}\right),
\end{aligned}
$$

where $\mathbf{S}$ is an empirical covariance matrix given by $\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$.

ADMM for Graphical SLOPE

ADMM for Graphical SLOPE

Graphical SLOPE problem - ADMM formulation

$$\begin{aligned} \text{minimize} \quad & -\log\det X + \operatorname{tr}(XS) + \mathbb{I}[X \succeq 0] + \mathrm{J}_\lambda(Y) \\ \text{subject to} \quad & X = Y. \end{aligned}$$

ADMM for Graphical SLOPE

Graphical SLOPE problem - ADMM formulation

$$\begin{aligned} \text{minimize} \quad & -\log\det X + \operatorname{tr}(XS) + \mathbb{I}[X \succeq 0] + \mathrm{J}_\lambda(Y) \\ \text{subject to} \quad & X = Y. \end{aligned}$$

Graphical SLOPE problem - Augmented Lagrangian

$$\begin{aligned} \mathcal{L}_\rho(X, Y, N) = & -\log\det X + \operatorname{tr}(XS) + \mathbb{I}[X \succeq 0] \\ & + \lambda \|Y\|_1 + \rho\langle N, X - Y\rangle_F + \frac{\rho}{2}\|X - Y\|_F^2 \end{aligned}$$

X-update (1/3)

We have

$$X_k = \arg\min_{X} \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = \arg\min_{X \succeq 0} \left\{ -\log\det X + \frac{\rho}{2} \left\| X - \tilde{S}_{k-1} \right\|_F^2 \right\},$$

where

$$\tilde{S}_{k-1} = -N_{k-1} + Y_{k-1} - \frac{1}{\rho} S,$$

X-update (1/3)

We have

$$X_k = \arg\min_{X} \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = \arg\min_{X \succeq 0} \left\{ -\log\det X + \frac{\rho}{2} \left\| X - \tilde{S}_{k-1} \right\|_F^2 \right\},$$

where

$$\tilde{S}_{k-1} = -N_{k-1} + Y_{k-1} - \frac{1}{\rho} S,$$

The X-gradient of the augmented Lagrangian is given by

$$\nabla_X \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = -X^{-1} + \rho X - \rho \tilde{S}_{k-1}.$$

X-update (1/3)

We have

$$X_k = \arg\min_{X} \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = \arg\min_{X \succeq 0} \left\{ -\log\det X + \frac{\rho}{2} \left\| X - \tilde{S}_{k-1} \right\|_F^2 \right\},$$

where

$$\tilde{S}_{k-1} = -N_{k-1} + Y_{k-1} - \frac{1}{\rho} S,$$

The X-gradient of the augmented Lagrangian is given by

$$\nabla_X \mathcal{L}_\rho(X, Y_{k-1}, N_{k-1}) = -X^{-1} + \rho X - \rho \tilde{S}_{k-1}.$$

As the augmented Lagrangian is convex, it is clear that for some $X^* \succeq 0$

$$\nabla_X \mathcal{L}_\rho(X^*, Y_{k-1}, N_{k-1}) = -(X^*)^{-1} + \rho X^* - \rho \tilde{S}_{k-1} = 0.$$

X-update (2/3)

Rewriting equation as

$$-(X^*)^{-1} + \rho X^* = \rho \tilde{S}_{k-1},$$

we can find a matrix that meets this condition.

X-update (2/3)

Rewriting equation as

$$-(X^*)^{-1} + \rho X^* = \rho \tilde{S}_{k-1},$$

we can find a matrix that meets this condition.

At first, lets take the eigenvalue decomposition of right side

$$\rho \tilde{S}_{k-1} = \rho Q \Lambda Q^T.$$

X-update (2/3)

Rewriting equation as

$$-(X^*)^{-1} + \rho X^* = \rho \tilde{S}_{k-1},$$

we can find a matrix that meets this condition.

At first, lets take the eigenvalue decomposition of right side

$$\rho \tilde{S}_{k-1} = \rho Q \Lambda Q^T.$$

Then by multiplying right and left side by Q and Q^T respectively, we obtain

$$-(\tilde{X}^*)^{-1} + \rho \tilde{X}^* = \rho \Lambda,$$

where $\tilde{X}^* = Q^T X^* Q$.

X-update (3/3)

We have to find positive numbers $\tilde{x}_{ii}^*$ that satisfy

$$(\tilde{x}_{ii}^*)^2 - l_{ii}\tilde{x}_{ii}^* - \frac{1}{\rho} = 0.$$

It is obvious that

$$\tilde{x}_{ii} = \frac{l_i + \sqrt{l_i^2 + 4/\rho}}{2}.$$

Thus X^* is given by $X^* = Q^T \tilde{X}^* Q$. All diagonals are positive since $\rho > 0$.
Define $\mathcal{F}_\rho(\Lambda)$ as

$$\mathcal{F}_\rho(\Lambda) = \frac{1}{2}\,\mathrm{diag}\left\{ l_i + \sqrt{l_i^2 + 4/\rho} \right\}.$$

Since that

$$X^* = Q^T \tilde{X}^* Q = Q^T \mathcal{F}_\rho(\Lambda) Q = \mathcal{F}_\rho(\tilde{S}_{k-1}) = \mathcal{F}_\rho\left(-N_{k-1} + Y_{k-1} - \frac{1}{\rho}S\right),$$

we obtain a formula for updating X_k in each step.

Y-update

A formula for Y_k is different. We have

$$
\begin{aligned}
Y_k &= \arg\min_{Y} \mathcal{L}_\rho(X_k, Y, N_{k-1}) \\
&= \arg\min_{Y} \left\{ \mathrm{J}_\lambda(Y) + \frac{\rho}{2} \| Y - (X_k + N_{k-1}) \|_F^2 \right\}
\end{aligned}
$$

Y-update

A formula for Y_k is different. We have

$$\begin{aligned} Y_k &= \arg\min_Y \mathcal{L}_\rho(X_k, Y, N_{k-1}) \\ &= \arg\min_Y \left\{ \mathrm{J}_\lambda(Y) + \frac{\rho}{2} \| Y - (X_k + N_{k-1}) \|_F^2 \right\} \end{aligned}$$

The last line of Y-update can be represented as a **proximity operator** which has closed form formula for SLOPE

$$\arg\min_Y \left\{ \mathrm{J}_\lambda(Y) + \frac{\rho}{2} \| Y - (X_k + N_{k-1}) \|_F^2 \right\} = \mathbf{prox}_{\mathrm{J}_\lambda, \rho}\left(X_k + N_{k-1}\right). \quad (5)$$

ADMM for Graphical SLOPE

Algorithm 4 Alternative direction method of multipliers for gSLOPE

$Y_0 \leftarrow \tilde{Y}$, $N_0 \leftarrow \tilde{N}$, $k \leftarrow 1$ ▷ initialize (loosely)
$\mu \leftarrow \tilde{\mu} > 0$ ▷ initialize
while convergence criterion is not meet **do**
  $X_k \leftarrow \mathcal{F}_\rho(N_{k-1} + Y_{k-1} - \frac{1}{\rho}S)$ ▷ x-minimization
  $Y_k \leftarrow \mathbf{prox}_{J_\lambda,\rho}(X_k + N_{k-1})$ ▷ y-minimization
  $N_k \leftarrow N_{k-1} + \rho(X_k - Y_k)$ ▷ dual update
  $k \leftarrow k + 1$
end while

FWER

FWER definition

Definition (Familywise error rate)

A *family-wise error rate* (FWER) is the probability of making one or more false discoveries, that is,

$$\mathrm{FWER} = \mathbb{P}(\text{type I error}).$$