Precision matrix estimation in Gaussian graphical models

Michał Makowski

Wydział Matematyki i Informatyki
Uniwersytet Wrocławski

m.makowski@aol.com

Presentation plan

1. Overview
   - Intuition and examples
   - Background
   - Implementation
   - Results
2. Theory
   - Factorization
   - Multiple testing
   - Alternating direction method of multipliers
3. Even more theory

Graphical models

A graphical models theory generalizes and is able to describe a broad range of statistical models

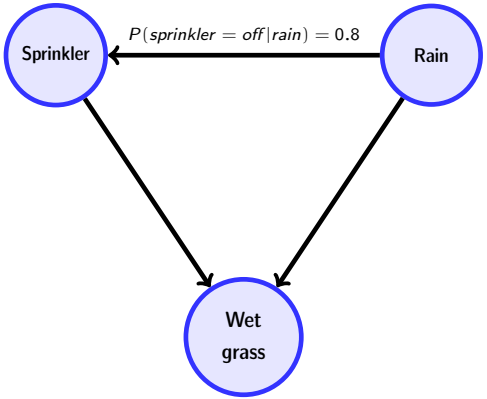
- Markov models/hidden Markov models
- Bayesian networks
- Kalman filters
- neural networks

Intuitions

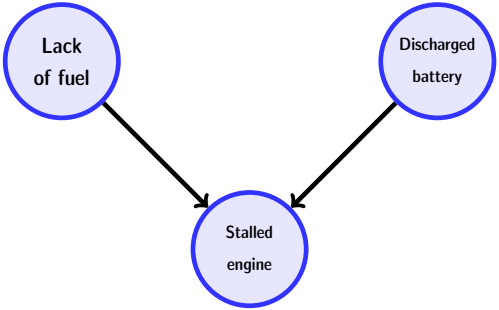
Graphical models bound probability and graph theory

- probability - uncertainty/randomness
- graph theory - dependence/correlation

Directed graph

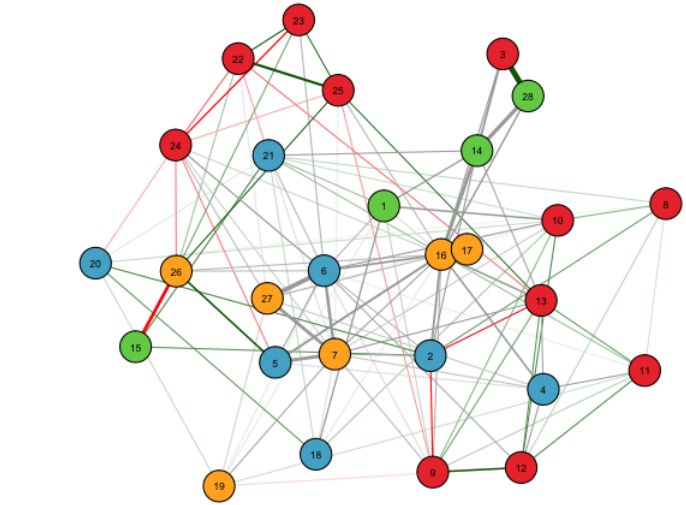


(a) 'Garden' model



(b) 'Car' model

Undirected graph [1/3]



- **Demographics**
- 1: Gender
- 14: Type of Housing
- 15: No of unfinished Educations
- 28: Age
- **Psychological**
- 2: IQ
- 4: Openness about Diagnosis
- 5: Success selfrating
- 6: Well being
- 18: No of Interests
- 20: Good Characteristics due to Autism
- 21: No of Transition Problems
- **Social environment**
- 7: Integration in Society
- 16: Type of work
- 17: Workinghours
- 19: No of Social Contacts
- 26: Satisfaction: Work
- 27: Satisfaction: Social Contacts
- **Medical**
- 3: Age diagnosis
- 8: No of family members with autism
- 9: No of Comorbidities
- 10: No of Physical Problems
- 11: No of Treatments
- 12: No of Medications
- 13: No of Care Units
- 22: Satisfaction: Treatment
- 23: Satisfaction: Medication
- 24: Satisfaction: Care
- 25: Satisfaction: Education

Undirected graph [2/3]

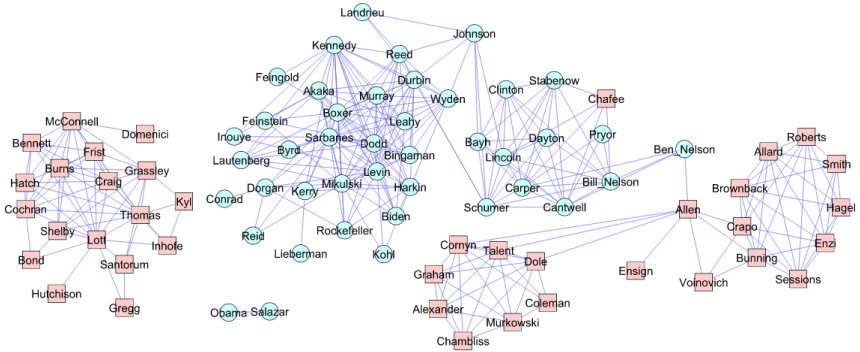


Figure 16: US Senate, 109th Congress (2004-2006). The graph displays the solution to (12) obtained using the log determinant relaxation to the log partition function of Wainwright and Jordan (2006). Democratic senators are represented by round nodes and Republican senators are represented by square nodes.

Undirected graph [3/3]

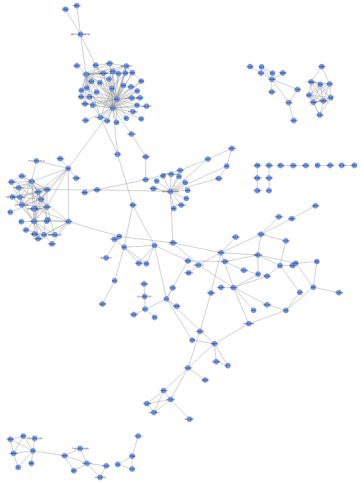


Figure 13: Application to Hughes compendium. The above graph results from solving (1) for this data set with a penalty parameter of $\lambda = 0.0313$.

Gaussian distribution factorization

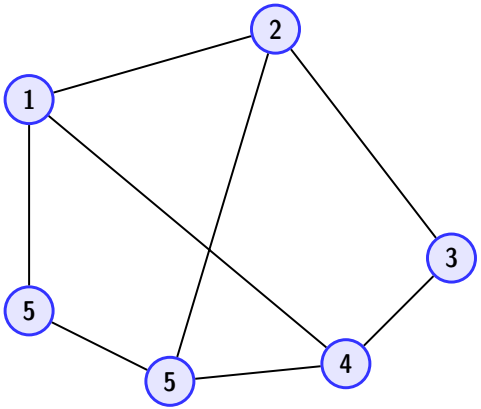
Any multivariate normal distribution $\mathcal{N}(\mu, \mathbf{\Sigma})$ can be parameterized by canonical parameters in the form

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

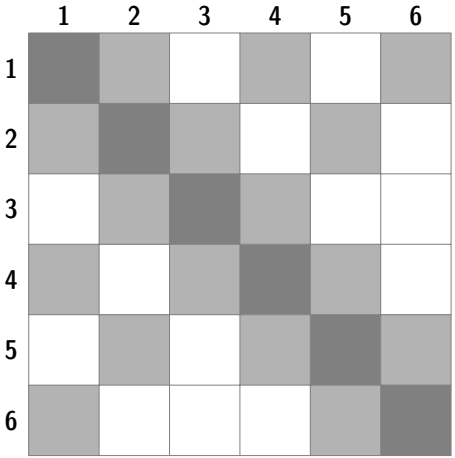
If $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ factorizes according to some graph G, then $\theta_{st} = 0$ for any pair $(s, t) \notin E$.

This 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



(c) The undirected graph G on six vertices.



(d) The associated sparsity pattern of the precision matrix Θ. White squares correspond to zero entries.

MLE...

Let $\mathbf{S}$ be a sample covariance matrix.

MLE

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

MLE exists *iff* the matrix $\mathbf{S}$ is nonsingular. Then the solution to problem above is simply given by

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

...and its problems

If a number of variables p is comparable or greater than a number of observations N, then the sample covariance matrix $\mathbf{S}$ is singular, thus the MLE does not exist.

Moreover, it is obvious that $\mathbb{P}(\exists_{i,j}\ \widehat{\sigma}_{ij} = 0) = 0$, but in many applications a *sparse* solution is demanded.

Regularization

The number of edges can be controlled by the 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

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

Unfortunately, the ℓ_0-based constrained defines a highly nonconvex constraint set, what makes the problem hard to solve.

Convex regularization

Convex relaxation of ℓ_0-based constrain leads to

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

where $F(\cdot)$ denotes any convex function, which might be applied in the considered problem.

Main motivation - FDR control

A performance of every binary classifier could be summarized in a confusion matrix

Test outcome \ Real value	$(+)$	$(-)$
$(+)$	True positive	False positive
$(-)$	False negative	True negative

(Local) False discovery rate (FDR)

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

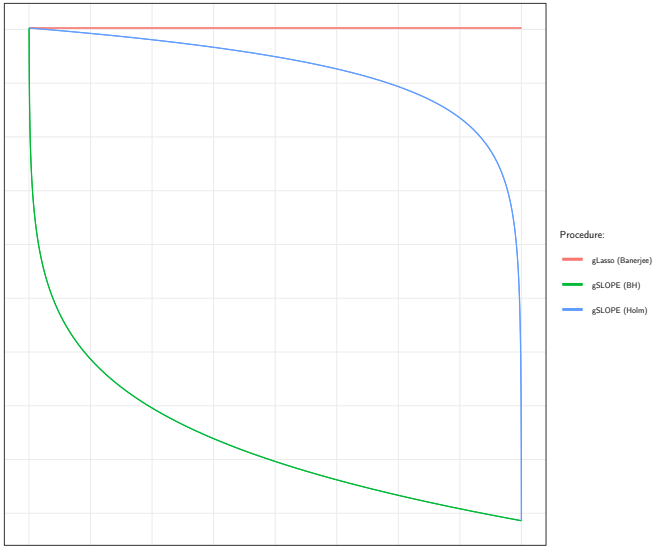
$$\text{localFDR} = \mathbb{E}\left[\frac{\#[\text{False positive outside the component}]}{\#[\text{False positive}] + \#[\text{True positive}]}\right]$$

Penalty

Lambda based on multiple testing theory

- gLasso - Bonferonni correction (Banerjee et. al.)
- **gSLOPE - Holm method**
- **gSLOPE - Banjamini-Hochberg procedure**

Lambda comparision



Algorithms

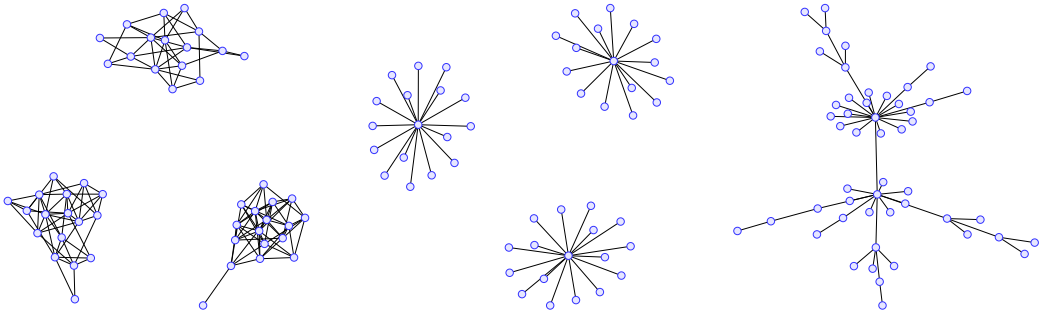
For solving the Graphical SLOPE problem we used the *Alternating direction method of multipliers*, it can solve convex problems in the form

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

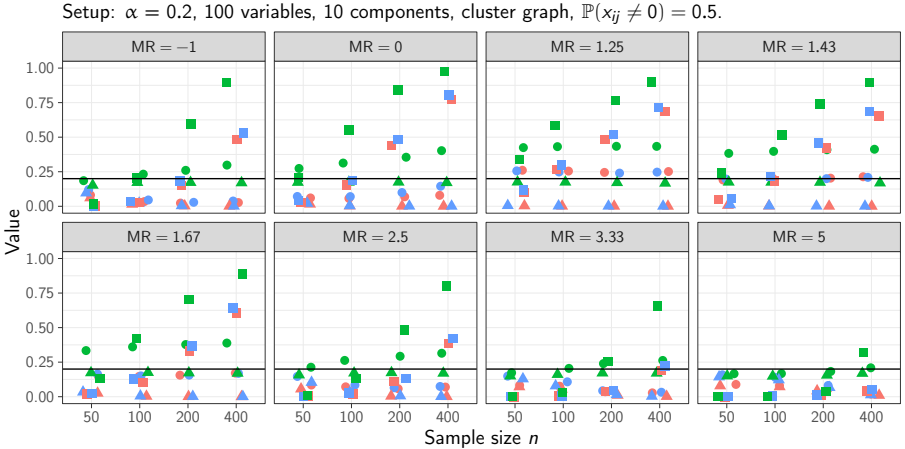
For solving the Graphical Lasso problem we used an algorithm proposed by Friedman et al. in theirs first work about this method. Although we derived an ADMM-based algorithm, it was orders of magnitude slower than original one.

Implementation overview

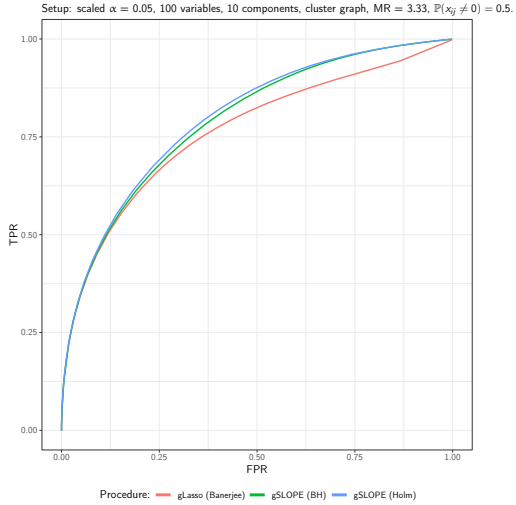
- Implementation with R, the **huge** package for simulations.
- Various types of graphs structure: cluster, hub, and scale-free.
- Data: $p = 100$, $n \in \{50, 100, 200, 400\}$; different magnitude ratios; different sparsity and size of components.
- Two levels of a desirable FDR control: 0.05 and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



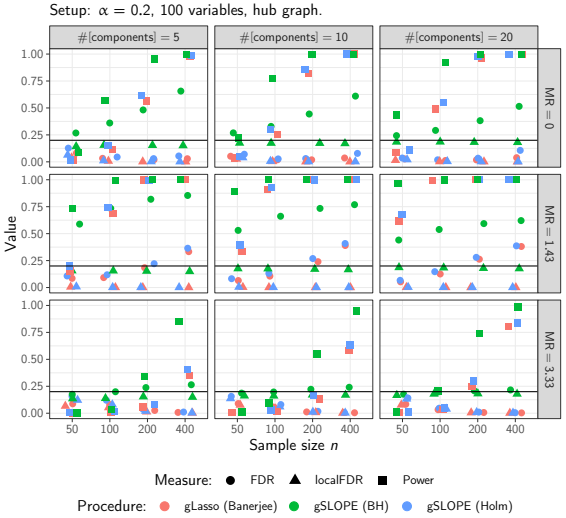
Cluster results



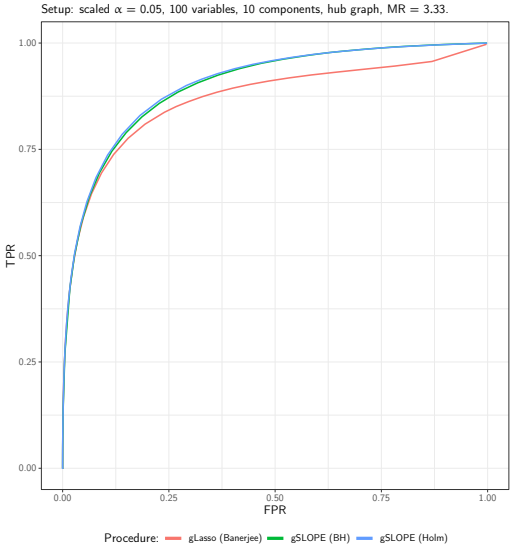
Cluster ROC

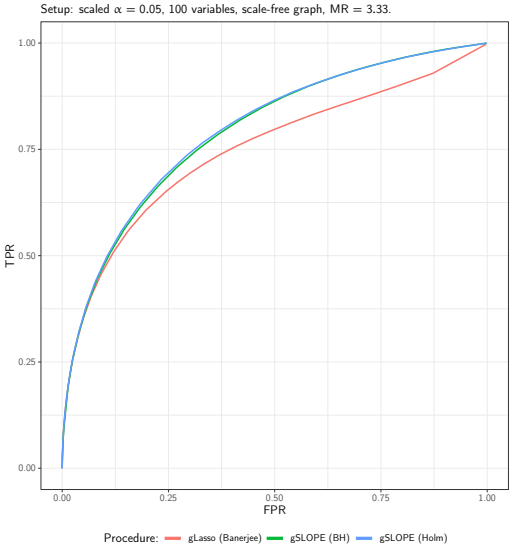

Setup: scaled $\alpha = 0.05$, 100 variables, 10 components, cluster graph, MR = 3.33, $\mathbb{P}(x_{ij} \neq 0) = 0.5$.
TPR
FPR
0.00
0.25
0.50
0.75
1.00
Procedure:
gLasso (Banerjee)
gSLOPE (BH)
gSLOPE (Holm)

Hub results


Setup: α = 0.2, 100 variables, hub graph.
#[components] = 5
#[components] = 10
#[components] = 20
MR = 0
MR = 1.43
MR = 3.33
Value
1.00
0.75
0.50
0.25
0.00
50
100
200
400
Sample size n
Measure: FDR localFDR Power
Procedure: gLasso (Banerjee) gSLOPE (BH) gSLOPE (Holm)

Non-cluster ROC





Let's go deeper into theory...

Conditional independence

Two random variables X, Y are conditionally independent *wrt* random vector $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ *iff* their distributions are independent *wrt* $\mathbf{Z}$.
This relationship is denoted by symbol $\perp\!\!\!\perp$.

Formally, for all triples x, y, z

$$(X \perp\!\!\!\perp Y) \mid \mathbf{Z} \iff F_{X,Y|\mathbf{Z}=z}(x, y) = F_{X|\mathbf{Z}=z}(x) \cdot F_{Y|\mathbf{Z}=z}(y),$$

where $F_{X,Y \mid \mathbf{Z}=z}(x, y)$ is the conditional CDF of X and Y for a given $\mathbf{Z}$.

Conditional independence and graph structure

For a family of multivariate probability distributions there could be constructed a graph which constraints a PDF factorization and the conditional independence property

$$\text{node } A \text{ is not connected with node } B \quad \Longleftrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_{-AB},$$

where X_{-AB} denotes all random variables except X_A and X_B.

Conditional independence and precision matrix

Canonical parameterization of Multivariate Gaussian Distribution constraint a precision matrix and a conditional independence property.

Let $(X_1, \ldots, X_n) \sim \mathcal{N}(\gamma, \mathbf{\Theta})$, then

$$X_s \perp\!\!\!\perp X_t \mid X_{-st} \iff \theta_{st} = 0,$$

where X_{-st} denotes all random variables except X_s and X_t.

Graph structure and precision matrix

There is connection between an underlying graph and a precision matrix structure

$$\text{node } S \text{ is not connected with node } S \quad \Longleftrightarrow \quad \theta_{st} = 0,$$

where s and t are variable indexes of Multivariate Gaussian Distribution.

Multiple testing

Bonferonni correction

The Bonferroni correction rejects the null hypothesis for each $p_i \leqslant \frac{\alpha}{m}$, thereby controls the FWER at level α, that is FWER $\leqslant \alpha$. No other assumptions are required.

Holm method

Sort p-values ascending and reject the first hypothesis for which $p_{(k)} > \frac{\alpha}{m+1-k}$ is true, then reject every hypothesis before. Holm method controls FWER at level α.

Benjamini-Hochberg method

Sort p-values ascending and find first the hypothesis for which $p_{(k)} \leqslant \alpha \frac{k}{m}$ is true, then reject every hypothesis before. BH method controls FDR at level α.

Graphical Lasso

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

Graphical Lasso problem

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

Graphical Lasso parameter choice

Banerjee lambda for Graphical Lasso

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

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*

$$\mathbb{P}(\text{False Discovery}) \leqslant \alpha.$$

SLOPE

In the [Bog+15] Bogdan et al. proposed novel approach for regularization in regression analysis. SLOPE uses the *OL1* norm instead of the *L1* norm for a coefficient choice.

OL1 norm

Penalty based on sorted-ℓ_1 (known as *OL1*, *OWL* or *OSCAR*) for $\beta \in \mathbb{R}^p$ and $\lambda \in \mathbb{R}^p, \lambda_1 \geqslant \ldots \geqslant \lambda_p$ is defined as

$$J_\lambda(\beta) = \sum_{i=1}^{p} \lambda_i |\beta|_{(i)}.$$

It was shown that under some assumptions and a construction of λ parameters based on BH procedure, proposed method controls FDR in multivariate regression settings.

Graphical SLOPE

In the graphical SLOPE the *L1* norm from a graphical Lasso algorithm is changed for the off-diagonal *OL1* norm

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

Graphical SLOPE problem

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

In the [Sob19] P. Sobczyk showed that *OL1* brings promising results in terms of a FDR control.

Graphical SLOPE parameter choice (1/2)

Holm lambda for Graphical SLOPE

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

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

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

It is based on the Holm method for the multiple testing.

Graphical SLOPE parameter choice (2/2)

BH lambda for Graphical SLOPE

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

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

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

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

Precursors

The alternating direction method of multipliers (ADMM) is an algorithm that solves convex optimization problems by breaking them into smaller pieces, each of which are then easier to handle.

The precursors of ADMM

- Dual Ascent and Dual Decomposition algorithms (decomposability properties)
- Method of multpliers (convergence properties)

ADMM

The ADMM algorithm can solve problems in the form

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

where convexity of functions f and g is assumed.
Augmented Lagrangian with parameter $\rho > 0$ is defined as

$$\mathcal{L}_\rho(x, y, \nu) = f(x) + g(y) + \nu^T(Ax + By - c) + \frac{\rho}{2}\|Ax + By - b\|^2.$$

Algorithm 1 Alternative direction method of multipliers

$y_0 \leftarrow \tilde{y}$, $\nu_0 \leftarrow \tilde{\nu}$, $k \leftarrow 1$
$\mu \leftarrow \tilde{\rho} > 0$ ▷ initialize
while convergence criterion is not met **do**
  $x_k \leftarrow \arg\min_x L_\rho(x, y_{k-1}, \nu_{k-1})$ ▷ x-minimization
  $y_k \leftarrow \arg\min_y L_\rho(x_k, y, \nu_{k-1})$ ▷ y-minimization
  $\nu_k \leftarrow \nu_{k-1} + \rho(Ax_k + By_k - b)$ ▷ dual update
  $k \leftarrow k + 1$
end while

ADMM for gSLOPE

Optimisation problem for graphical SLOPE is in the form

$$\begin{aligned} \text{minimize} \quad & -\log\det\boldsymbol{\Theta} + \operatorname{tr}(\boldsymbol{S}\,\boldsymbol{\Theta}) + \mathbb{I}[\boldsymbol{\Theta} \succeq 0] + J_\lambda(Y) \\ \text{subject to} \quad & Y = \boldsymbol{\Theta}\,. \end{aligned}$$

Augmented Lagrangian $\mathcal{L}_\rho : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \to \mathbb{R}$ with parameter $\rho > 0$ is given by

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

Bibliography

Onurena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. *Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data*. 2008.

Małgorzata Bogdan et al. *SLOPE - Adaptive variable selection via convex optimization*. 2015.

Stephen Boyd et al. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. 2010.

Emmanuel Candes. *Advanced Topics in Convex Optimization*.

Trevor Hastie et al. *Statistical Learning with Sparsity: The Lasso and Generalizations*. 2015.

Piotr Sobczyk. *Identifying low-dimensional structures through model selection in high-dimensional data*. 2019.

Let's dive even deeper...

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 be reparameterized into canonical parameters in 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

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

Log-likelihood derivation

Log-likelihood derivation (1/2)

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

Log-likelihood derivation (2/2)

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

where $\boldsymbol{S}$ 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

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

Graphical SLOPE problem - Augmented Lagrangian

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

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.

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

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

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

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

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

Thank you for your attention