# Proximal Approaches for Matrix Optimization Problems. Application to Robust Graphical LASSO.

Emilie CHOUZENOUX
(joint work with J.C. Pesquet and A. Benfenati)

CENTER FOR VISUAL COMPUTING
École Centrale Paris / INRIA

PROJET FINANCÉ PAR L'ANR - PROJECT FUNDED BY THE ANR
ANR

SPARS, 4th July 2019

# Outline

# 1) Introduction

Several problems lead to find the minimum of a matrix functional:

- shape classification (Duchi et al, 2008)
- gene expression (Ma et al, 2013)
- model selection (Chandrasekaran et al, 2012)
- matrix completion (McRae and Davenport, 2019)
- computer vision (Guo et al, 2011)
- phase retrieval (Candes et al, 2015)
- inverse covariance estimation (d'Aspremont et al, 2008)
- graph estimation (Meinshausen et al, 2006),
- brain network analysis (Yang et al, 2015)

**Challenge:** How to deal with versatile functionals, involving non necessarily convex terms, acting both on the matrix entries and its eigenvalues ?

Several problems lead to find the minimum of a matrix functional:

- shape classification (Duchi et al, 2008)
- gene expression (Ma et al, 2013)
- model selection (Chandrasekaran et al, 2012)
- matrix completion (McRae and Davenport, 2019)
- computer vision (Guo et al, 2011)
- phase retrieval (Candes et al, 2015)
- inverse covariance estimation (d'Aspremont et al, 2008)
- graph estimation (Meinshausen et al, 2006),
- brain network analysis (Yang et al, 2015)

**Challenge:** How to deal with versatile functionals, involving non necessarily convex terms, acting both on the matrix entries and its eigenvalues ?

**This talk:**

✓ new optimization tools for dealing with minimization problems in a **symmetric matrix space**;

Several problems lead to find the minimum of a matrix functional:

- shape classification (Duchi et al, 2008)
- gene expression (Ma et al, 2013)
- model selection (Chandrasekaran et al, 2012)
- matrix completion (McRae and Davenport, 2019)
- computer vision (Guo et al, 2011)
- phase retrieval (Candes et al, 2015)
- inverse covariance estimation (d'Aspremont et al, 2008)
- graph estimation (Meinshausen et al, 2006),
- brain network analysis (Yang et al, 2015)

**Challenge:** How to deal with versatile functionals, involving non necessarily convex terms, acting both on the matrix entries and its eigenvalues ?

**This talk:**

- ✓ new optimization tools for dealing with minimization problems in a **symmetric matrix space**;
- ✓ new **proximal algorithm** for minimizing convex penalized cost with regularization split in two parts, one being a **spectral function** while the other is arbitrary;

# Introduction

Several problems lead to find the minimum of a matrix functional:

- shape classification (Duchi et al, 2008)
- gene expression (Ma et al, 2013)
- model selection (Chandrasekaran et al, 2012)
- matrix completion (McRae and Davenport, 2019)
- computer vision (Guo et al, 2011)
- phase retrieval (Candes et al, 2015)
- inverse covariance estimation (d'Aspremont et al, 2008)
- graph estimation (Meinshausen et al, 2006),
- brain network analysis (Yang et al, 2015)

**Challenge:** How to deal with versatile functionals, involving non necessarily convex terms, acting both on the matrix entries and its eigenvalues ?

**This talk:**

- ✓ new optimization tools for dealing with minimization problems in a **symmetric matrix space**;
- ✓ new **proximal algorithm** for minimizing convex penalized cost with regularization split in two parts, one being a **spectral function** while the other is arbitrary;
- ✓ new minimization approach for **non-convex** problem arising in **covariance matrix estimation**, combining **majorization-minimization** framework and Douglas-Rachford proximal scheme.

∗ Aim: **Inferring Gaussian graphical model parameters** $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from $N$ i.i.d realizations: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ symmetric definite positive.

- Sample mean and empirical covariance matrix:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)}, \quad \mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})^\top.$$

- Negative Gaussian log-likelihood:

$$-\frac{1}{N} \ell(\boldsymbol{\Sigma}^{-1} | \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}) = -\log\det \boldsymbol{\Sigma}^{-1} + \text{trace}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + \text{constant}.$$

∗ Aim: **Inferring Gaussian graphical model parameters** $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from $N$ i.i.d realizations: $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}$ of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ symmetric definite positive.

- Sample mean and empirical covariance matrix:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)}, \quad \mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})^\top.$$

- Negative Gaussian log-likelihood:

$$-\frac{1}{N} \ell(\boldsymbol{\Sigma}^{-1} | \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}) = -\log\det \boldsymbol{\Sigma}^{-1} + \text{trace}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + \text{constant}.$$

∗ GLASSO: Estimator of $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$ based on the use of $\ell_1$ penalty (Meinshausen and Buhlmann, 2006)

$$\widehat{\boldsymbol{C}} = \text{argmin}_{\boldsymbol{C} \succ 0} - \log\det \boldsymbol{C} + \text{trace}(\mathbf{S}\boldsymbol{C}) + \lambda \|\boldsymbol{C}\|_1$$

with $\|\boldsymbol{C}\|_1 = \sum_{j,k} |C_{jk}|$, and $\lambda > 0$ regularization parameter.

## Important example: Graphical Lasso

∗ Aim: **Inferring Gaussian graphical model parameters** $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from $N$ i.i.d realizations: $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$ of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ symmetric definite positive.

- Sample mean and empirical covariance matrix:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)}, \quad \mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \widehat{\boldsymbol{\mu}})^\top.$$

- Negative Gaussian log-likelihood:

$$-\frac{1}{N}\ell(\boldsymbol{\Sigma}^{-1} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = -\log\det \boldsymbol{\Sigma}^{-1} + \text{trace}(\mathbf{S}\boldsymbol{\Sigma}^{-1}) + \text{constant}.$$

∗ GLASSO: Estimator of $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$ based on the use of $\ell_1$ penalty (Meinshausen and Buhlmann, 2006)

$$\widehat{\boldsymbol{C}} = \text{argmin}_{\boldsymbol{C} \succ 0} - \log\det \boldsymbol{C} + \text{trace}(\mathbf{S}\boldsymbol{C}) + \lambda \|\boldsymbol{C}\|_1$$

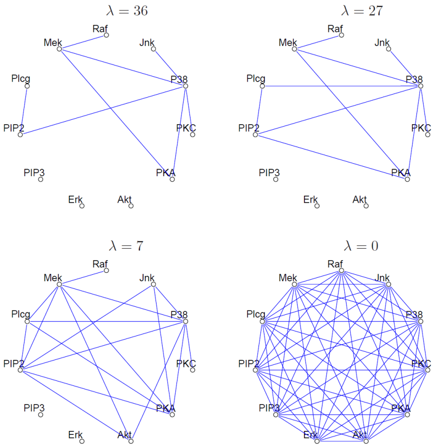
with $\|\boldsymbol{C}\|_1 = \sum_{j,k} |C_{jk}|$, and $\lambda > 0$ regularization parameter.

- **Convex optimization** problem in **symmetric matrix space**.

Several solvers available. (Banerjee et al, 2007)(Friedman et al, 2007)(Boyd et al, 2011)(Duchi et al, 2008).

**Challenges:** Which optimization method for **more sophisticated penalties** ? How to account for the **noise** possibly degrading the input data ?

# Example




Four different GLASSO solutions for the flow-cytometry data with $n = 11$ proteins measured on $N = 7466$ cells (Friedman et al, 2007).

# 2) Douglas-Rachford algorithm for matrix optimization problems

# Definition: Spectral function

Let

$$f : \mathcal{S}_n \to ]-\infty, +\infty], \quad \mathcal{S}_n = \{\mathbf{C} \in \mathbb{R}^{n\times n} | \mathbf{C}^\top = \mathbf{C}\}.$$

f is a **spectral function** if, for every permutation matrix $\mathbf{P} \in \mathbb{R}^{n\times n}$,

$$f(\mathbf{C}) = \varphi(\mathbf{Pd}),$$

with $\mathbf{d} \in \mathbb{R}^n$ a vector of eigenvalues of $\mathbf{C}$, and $\varphi : \mathbb{R}^n \to ]-\infty, +\infty]$ is proper, lower semicontinuous (lsc).

| | $f(\mathbf{C})$ | $\varphi(\mathbf{Pd})$ |
|---|---|---|
| Logdet function | $\begin{cases} -\log\det(\mathbf{C}) & \text{if } \mathbf{C} \in \mathcal{S}_n^{++} \\ +\infty & \text{else} \end{cases}$ | $\begin{cases} -\sum_{i=1}^n \log(d_i) & \text{if } \mathbf{d} \in ]0, +\infty[^n \\ +\infty & \text{else} \end{cases}$ |
| Froebenius norm | $\frac{1}{2}\|\mathbf{C}\|_{\mathrm{F}}^2$ | $\frac{1}{2}\sum_{i=1}^n d_i^2$ |
| Nuclear norm | $\mathcal{R}_1(\mathbf{C})$ | $\sum_{i=1}^n \|d_i\|$ |
| Rank | $\mathrm{rank}(\mathbf{C})$ | $\mathrm{Card}\,\{i \in \{1, \ldots, n\} \text{s.t. } d_i \neq 0\}$ |

Let us consider the following minimization problem:

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\text{minimize}}\; f(\mathbf{C}) - \operatorname{tr}(\mathbf{TC}) + g_0(\mathbf{C}) \tag{1}$$

with

- $f$ a **spectral function** associated to $\varphi$, lsc function;
- $g_0$ a **spectral function** associated to $\psi$, lsc function;
- $\mathbf{T} \in \mathcal{S}_n$ and $\operatorname{tr}(\cdot)$ the trace operator.

**Theorem**

Let $\mathbf{t} \in \mathbb{R}^n$ be a vector of eigenvalues of $\mathbf{T}$ and let $\mathbf{U_T} \in \mathcal{O}_n$ be such that $\mathbf{T} = \mathbf{U}_T \mathrm{Diag}(\mathbf{t}) \mathbf{U}_T^\top$. Assume that dom$\varphi \cap$ dom$\psi \neq \varnothing$ and that the function $\mathbf{d} \mapsto \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d})$ is coercive. Then a solution to Problem (1) exists, and is given by

$$\widehat{\mathbf{C}} = \mathbf{U_T} \mathrm{Diag}(\widehat{\mathbf{d}}) \mathbf{U_T}^\top$$

where $\widehat{\mathbf{d}}$ is any solution to the following problem:

$$\underset{\mathbf{d}\in\mathbb{R}^n}{\text{minimize}}\; \varphi(\mathbf{d}) - \mathbf{d}^\top \mathbf{t} + \psi(\mathbf{d}).$$

Let $f$ **convex**, differentiable on $\text{int}(\text{dom} f) \neq \varnothing$.
The $f$-Bregman divergence between $\mathbf{C} \in \mathcal{S}_n$ and $\mathbf{Y} \in \text{int}(\text{dom} f)$ is

$$D^f(\mathbf{C}, \mathbf{Y}) = f(\mathbf{C}) - f(\mathbf{Y}) - \text{tr}(\mathbf{T}(\mathbf{C} - \mathbf{Y})) \quad \text{with} \quad \mathbf{T} = \nabla f(\mathbf{Y}).$$

Computing the $D^f$-proximity operator of $g_0$ with $g_0$ proper, lsc, at $\overline{\mathbf{C}} \in \text{int}(\text{dom} f)$ amounts to solve

$$\underset{\mathbf{C} \in \mathcal{S}_n}{\text{minimize}}\ g_0(\mathbf{C}) + D^f(\mathbf{C}, \mathbf{Y}) \tag{2}$$

$*$ For particular choices of $f$ and $\mathbf{T}$, Problem (2) is equivalent to Problem (1).

Let $f$ **convex**, differentiable on $\mathrm{int}(\mathrm{dom} f) \neq \varnothing$.
The $f$-Bregman divergence between $\mathbf{C} \in \mathcal{S}_n$ and $\mathbf{Y} \in \mathrm{int}(\mathrm{dom} f)$ is

$$D^f(\mathbf{C}, \mathbf{Y}) = f(\mathbf{C}) - f(\mathbf{Y}) - \mathrm{tr}(\mathbf{T}(\mathbf{C} - \mathbf{Y})) \quad \text{with} \quad \mathbf{T} = \nabla f(\mathbf{Y}).$$

Computing the $D^f$-proximity operator of $g_0$ with $g_0$ proper, lsc, at $\overline{\mathbf{C}} \in \mathrm{int}(\mathrm{dom} f)$ amounts to solve

$$\underset{\mathbf{C} \in \mathcal{S}_n}{\mathrm{minimize}}\; g_0(\mathbf{C}) + D^f(\mathbf{C}, \mathbf{Y}) \tag{2}$$

∗ For particular choices of $f$ and $\mathbf{T}$, Problem (2) is equivalent to Problem (1).

## Corollary

Let $f$ and $g_0$ spectral functions associated, respectively, to $\varphi \in \Gamma_0(\mathbb{R}^n)$ Legendre function, and $\psi \in \Gamma_0(\mathbb{R}^n)$ with $\mathrm{int}(\mathrm{dom}\varphi) \cap \mathrm{int}(\mathrm{dom}\psi) \neq \varnothing$ and either $\psi$ is bounded from below or $\varphi + \psi$ is supercoercive. Then, the solution to (2) exists, and is unique, for every $\mathbf{Y} \in \mathcal{S}_n$ such that $\mathbf{Y} = \mathbf{U}_\mathbf{Y} \mathrm{Diag}(\mathbf{y}) \mathbf{U}_\mathbf{Y}^\top$ with $\mathbf{U}_\mathbf{Y} \in \mathcal{O}_n$ and $\mathbf{y} \in \mathrm{int}(\mathrm{dom}\varphi)$, and it is expressed as

$$\mathrm{prox}^f_{g_0}(\mathbf{Y}) = \mathbf{U}_\mathbf{Y} \mathrm{Diag}(\mathrm{prox}^\varphi_\psi(\mathbf{y})) \mathbf{U}_\mathbf{Y}^\top,$$

with $\mathrm{prox}^\varphi_\psi \colon \mathbf{y} \mapsto \underset{\mathbf{x} \in \mathbb{R}^n}{\mathrm{argmin}}\; \psi(\mathbf{x}) + \varphi(\mathbf{x}) - \varphi(\mathbf{y}) - \langle \nabla\varphi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$

✓ Extend (Bauschke and Combettes, 2017) to **Bregman divergence** setting.

Let us now consider the following minimization problem:

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\text{minimize}}\;\; f(\mathbf{C}) - \operatorname{tr}(\mathbf{TC}) + g_0(\mathbf{C}) + \frac{1}{2\gamma}\|\mathbf{C} - \overline{\mathbf{C}}\|_F^2 \tag{3}$$

with $\gamma > 0$, $\overline{\mathbf{C}} \in \mathcal{S}_n$, $\mathbf{T} \in \mathcal{S}_n$ and

- $f$ a **spectral function** associated to $\varphi$, lsc function;
- $g_0$ a **spectral function** associated to $\psi$, lsc function.

The (possibly empty) set of solutions is denoted $\operatorname{Prox}_{\gamma(f - \operatorname{tr}(\mathbf{T}\,\cdot) + g_0)}$

# Proximity operator

Let us now consider the following minimization problem:

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\text{minimize}} \;\; f(\mathbf{C}) - \operatorname{tr}(\mathbf{T}\mathbf{C}) + g_0(\mathbf{C}) + \frac{1}{2\gamma}\|\mathbf{C}-\overline{\mathbf{C}}\|_F^2 \tag{3}$$

with $\gamma > 0$, $\overline{\mathbf{C}} \in \mathcal{S}_n$, $\mathbf{T} \in \mathcal{S}_n$ and

- $f$ a **spectral function** associated to $\varphi$, lsc function;
- $g_0$ a **spectral function** associated to $\psi$, lsc function.

The (possibly empty) set of solutions is denoted $\operatorname{Prox}_{\gamma(f-\operatorname{tr}(\mathbf{T}\,\cdot)+g_0)}$

## Proposition

Assume that dom$\varphi$ $\cap$ dom$\psi \neq \varnothing$. Let $\boldsymbol{\lambda} \in \mathbb{R}^n$ and $\mathbf{U} \in \mathcal{O}_n$ be such that $\overline{\mathbf{C}} + \gamma\mathbf{T} = \mathbf{U}\operatorname{Diag}(\boldsymbol{\lambda})\mathbf{U}^\top$.

(i) If $\psi$ is lower bounded by an affine function then $\operatorname{Prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda}) \neq \varnothing$ and, for every $\widehat{\boldsymbol{\lambda}} \in \operatorname{Prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$,

$$\mathbf{U}\operatorname{Diag}(\widehat{\boldsymbol{\lambda}})\mathbf{U}^\top \in \operatorname{Prox}_{\gamma(f-\operatorname{tr}(\mathbf{T}\,\cdot)+g_0)}(\overline{\mathbf{C}}).$$

(ii) If $\psi$ is convex, then

$$\operatorname{prox}_{\gamma(f-\operatorname{tr}(\mathbf{T}\,\cdot)+g_0)}(\overline{\mathbf{C}}) = \mathbf{U}\operatorname{Diag}\Big(\operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})\Big)\mathbf{U}^\top.$$

**Frobenius norm:**

$f(\cdot) = \|\cdot\|_{\mathrm{F}}^2/2$, spectral function associated with $\varphi = \|\cdot\|^2/2$.

**Log-determinant:**

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad f(\mathbf{C}) = \begin{cases} -\log\det(\mathbf{C}) & \text{if } \mathbf{C} \in \mathcal{S}_n^{++} \\ +\infty & \text{otherwise.} \end{cases}$$

Spectral function associated with

$$\left(\forall \boldsymbol{\lambda} = (\lambda_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n\right) \qquad \varphi(\boldsymbol{\lambda}) = \begin{cases} -\sum_{i=1}^{n} \log(\lambda_i) & \text{if } \boldsymbol{\lambda} \in ]0, +\infty[^n \\ +\infty & \text{otherwise.} \end{cases}$$

**Van Neumann entropy:**

$$(\forall \mathbf{C} \in \mathcal{S}_n) \quad f(\mathbf{C}) = \begin{cases} \operatorname{tr}\left(\mathbf{C}\log(\mathbf{C})\right) & \text{if } \mathbf{C} \in \mathcal{S}_n^{+} \\ +\infty & \text{otherwise.} \end{cases}$$

Spectral function associated with

$$\left(\forall \boldsymbol{\lambda} = (\lambda_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n\right) \qquad \varphi(\boldsymbol{\lambda}) = \begin{cases} \sum_{i=1}^{n} \lambda_i \log(\lambda_i) & \text{if } \boldsymbol{\lambda} \in [0, +\infty[^n \\ +\infty & \text{otherwise.} \end{cases}$$

## Proximity operators for different choices for $g_0$ and $f$ Frobenius norm

| $g_0(\mathbf{C}),\ \mu > 0$ | $\operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$ |
|---|---|
| Nuclear norm<br>$\mu\mathcal{R}_1(\mathbf{C})$ | $\left(\operatorname{soft}_{\frac{\mu\gamma}{\gamma+1}}\left(\frac{\lambda_i}{\gamma+1}\right)\right)_{1\leqslant i\leqslant n}$ |
| Squared Frobenius norm<br>$\mu\\|\mathbf{C}\\|_{\mathrm{F}}^2$ | $\dfrac{\boldsymbol{\lambda}}{1+\gamma\,(1+2\mu)}$ |
| Schatten $p$–penalty<br>$\mu\mathcal{R}_p^p(\mathbf{C}),\ p \geqslant 1$ | $\big(\operatorname{sign}(\lambda_i)d_i\big)_{1\leqslant i\leqslant n}$<br>with $(\forall i\in\{1,\ldots,n\})\ d_i \geqslant 0$<br>and $\mu\gamma p d_i^{p-1} + (\gamma+1)d_i = \lambda_i$ |
| Inverse Schatten $p$–penalty<br>$\mu\mathcal{R}_p^p(\mathbf{C}^{-1}),\ p > 0$ | $\big(d_i\big)_{1\leqslant i\leqslant n}$<br>with $(\forall i\in\{1,\ldots,n\})\ d_i > 0$<br>and $(\gamma+1)d_i^{p+2} - \lambda_i d_i^{p+1} = \mu\gamma p$ |
| Bounds on eigenvalues<br>$\iota_{\mathbf{E}}(\mathbf{C})$ | $(\min(\max(\lambda_i/(\gamma+1),\alpha),\beta))_{1\leqslant i\leqslant n}$<br>$[\alpha,\beta]\subset[0,+\infty]$ |
| Rank<br>$\mu\,\operatorname{rank}(\mathbf{C})$ | $\left(\operatorname{hard}_{\sqrt{\frac{2\mu\gamma}{1+\gamma}}}\left(\dfrac{\lambda_i}{1+\gamma}\right)\right)_{1\leqslant i\leqslant n}$ |
| Cauchy<br>$\mu\log\det(\mathbf{C}^2+\varepsilon I),\ \varepsilon > 0$ | $\in\big\{(\operatorname{sign}(\lambda_i)d_i)_{1\leqslant i\leqslant n} \mid (\forall i\in\{1,\ldots,n\})\ d_i \geqslant 0$ and<br>$(\gamma+1)d_i^3 - \lvert\lambda_i\rvert d_i^2 + \big(2\gamma\mu+\varepsilon(\gamma+1)\big)d_i = \lvert\lambda_i\rvert\varepsilon\big\}$ |

$\mathbf{E}$ denotes the set of matrices in $\mathcal{S}_n$ with eigenvalues between $\alpha$ and $\beta$.

**Proximity operators for different choices for $g_0$ and $f$ log determinant**

| $g_0(\mathbf{C})$, $\mu > 0$ | $\text{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$ |
|---|---|
| Nuclear norm<br>$\mu \mathcal{R}_1(\mathbf{C})$ | $\frac{1}{2}\left(\lambda_i - \gamma\mu + \sqrt{(\lambda_i - \gamma\mu)^2 + 4\gamma}\right)_{1\leqslant i\leqslant n}$ |
| Squared Frobenius norm<br>$\mu\\|\mathbf{C}\\|_{\mathrm{F}}^2$ | $\frac{1}{2(2\gamma\mu+1)}\left(\lambda_i + \sqrt{\lambda_i^2 + 4\gamma(2\gamma\mu+1)}\right)_{1\leqslant i\leqslant n}$ |
| Schatten $p$–penalty<br>$\mu \mathcal{R}_p^p(\mathbf{C})$, $p \geqslant 1$ | $(d_i)_{1\leqslant i\leqslant n}$<br>$\mu\gamma p d_i^p + d_i^2 - \lambda_i d_i = \gamma$ |
| Inverse Schatten $p$–penalty<br>$\mu \mathcal{R}_p^p(\mathbf{C}^{-1})$, $p > 0$ | $(d_i)_{1\leqslant i\leqslant n}$<br>$d_i^{p+2} - \lambda_i d_i^{p+1} - \gamma d_i^p = \mu\gamma p$ |
| Bounds on eigenvalues<br>$\iota_{\mathbf{E}}(\mathbf{C})$ | $\left(\min\left(\max\left(\frac{1}{2}\left(\lambda_i + \sqrt{\lambda_i^2 + 4\gamma}\right), \alpha\right), \beta\right)\right)_{1\leqslant i\leqslant n}$<br>$[\alpha, \beta] \subset [0, +\infty]$ |
| Cauchy<br>$\mu \log\det(\mathbf{C}^2 + \varepsilon I)$, $\varepsilon > 0$ | $\in \{(d_i)_{1\leqslant i\leqslant n} \mid (\forall i \in \{1,\ldots,n\})\ d_i > 0$ and<br>$d_i^4 - \lambda d_i^3 + \big(\varepsilon + \gamma(2\mu - 1)\big)d_i^2 - \varepsilon\lambda_i d_i = \gamma\varepsilon\}$ |

## Proximity operators for different choices for $g_0$ and $f$ VN entropy

| $g_0(\mathbf{C})$, $\mu > 0$ | $\operatorname{prox}_{\gamma(\varphi+\psi)}(\boldsymbol{\lambda})$ |
|---|---|
| Nuclear norm<br>$\mu\mathcal{R}_1(\mathbf{C})$ | $\gamma\left(\mathrm{W}\left(\frac{1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma}-\mu-1\right)\right)\right)_{1\leqslant i\leqslant n}$ |
| Squared Frobenius norm<br>$\mu\\|\mathbf{C}\\|_{\mathrm{F}}^2$ | $\frac{\gamma}{2\mu\gamma+1}\left(\mathrm{W}\left(\frac{2\mu\gamma+1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma}-1\right)\right)\right)_{1\leqslant i\leqslant n}$ |
| Schatten $p$–penalty<br>$\mu\mathcal{R}_p^p(\mathbf{C})$, $p\geqslant 1$ | $(d_i)_{1\leqslant i\leqslant n}$<br>$d_i > 0$ s.t. $p\mu\gamma d_i^{p-1} + d_i + \gamma\log d_i + \gamma = \lambda_i$ |
| Bounds on eigenvalues<br>$\iota_{\mathbf{E}}(\mathbf{C})$ with $[\alpha,\beta]\subset[0,+\infty]$ | $\left(\min\left(\max\left(\gamma\mathrm{W}\left(\frac{1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma}-1\right)\right),\alpha\right),\beta\right)\right)_{1\leqslant i\leqslant n}$ |
| Rank<br>$\mu\,\mathrm{rank}(\mathbf{C})$ | $(d_i)_{1\leqslant i\leqslant n}$ with<br>$d_i = \begin{cases}\rho_i & \text{if } \rho_i > \chi\\ 0 \text{ or } \rho_i & \text{if } \rho_i = \chi\\ 0 & \text{otherwise}\end{cases}$ and $\begin{cases}\chi = \sqrt{\gamma(\gamma+2\mu)}-\gamma,\\ \rho_i = \gamma\mathrm{W}\left(\frac{1}{\gamma}\exp\left(\frac{\lambda_i}{\gamma}-1\right)\right)\end{cases}$ |

$\mathrm{W}(\cdot)$ denotes the W-Lambert function.

## Minimization problem

Now, let us consider:

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\text{minimize}}\ f(\mathbf{C}) - \operatorname{tr}(\mathbf{TC}) + g(\mathbf{C}) \tag{4}$$

with

$$g(\mathbf{C}) = \mu_0 g_0(\mathbf{C}) + \mu_1 g_1(\mathbf{C}), \quad \mu_0, \mu_1 > 0$$

and

- $f$ a **spectral function** associated to $\varphi \in \Gamma_0(\mathbb{R}^n)$;
- $g_0$ a **spectral function** associated to $\psi \in \Gamma_0(\mathbb{R}^n)$;
- $g_1 \in \Gamma_0(\mathbb{R}^{n\times n})$ acting on the whole matrix $\mathbf{C}$ (e.g., the $\ell_1$ norm)

$\rightsquigarrow$ The "spectral terms" of the functional can be gathered together:

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\operatorname{argmin}}\ \underbrace{f(\mathbf{C}) - \operatorname{tr}(\mathbf{TC}) + \mu_0 g_0(\mathbf{C})}_{h_0(\mathbf{C})} + \underbrace{\mu_1 g_1(\mathbf{C})}_{h_1\mathbf{C}}$$

$\Rightarrow$ Douglas-Rachford algorithm (Combettes and Pesquet, 2007).

## Douglas-Rachford Algorithm

Let $\mathbf{T}$ be a given matrix in $\mathcal{S}_n$, set $\gamma > 0$ and $\mathbf{C}^{(0)} \in \mathcal{S}_n$.
For $k = 0, 1, \ldots$
Diagonalize $\mathbf{C}^{(k)} + \gamma\mathbf{T}$, i.e. find $\mathbf{U}^{(k)} \in \mathcal{O}_n$ and $\boldsymbol{\lambda}^{(k)} \in \mathbb{R}^n$ such that

$$\mathbf{C}^{(k)} + \gamma\mathbf{T} = \mathbf{U}^{(k)}\mathrm{Diag}(\boldsymbol{\lambda}^{(k)})(\mathbf{U}^{(k)})^\top$$

$\mathbf{d}^{(k+\frac{1}{2})} \in \mathrm{Prox}_{\gamma(\varphi+\psi)}\left(\boldsymbol{\lambda}^{(k)}\right)$
$\mathbf{C}^{(k+\frac{1}{2})} = \mathbf{U}^{(k)}\mathrm{Diag}(\mathbf{d}^{(k+\frac{1}{2})})(\mathbf{U}^{(k)})^\top$
Choose $\alpha^{(k)} \in [0, 2]$
$\mathbf{C}^{(k+1)} \in \mathbf{C}^{(k)} + \alpha^{(k)}\Big(\mathrm{Prox}_{\gamma g_1}(2\mathbf{C}^{(k+\frac{1}{2})} - \mathbf{C}^{(k)}) - \mathbf{C}^{(k+\frac{1}{2})}\Big).$

## Douglas-Rachford Algorithm

Let $\mathbf{T}$ be a given matrix in $\mathcal{S}_n$, set $\gamma > 0$ and $\mathbf{C}^{(0)} \in \mathcal{S}_n$.
For $k = 0, 1, \ldots$

Diagonalize $\mathbf{C}^{(k)} + \gamma \mathbf{T}$, i.e. find $\mathbf{U}^{(k)} \in \mathcal{O}_n$ and $\boldsymbol{\lambda}^{(k)} \in \mathbb{R}^n$ such that

$$\mathbf{C}^{(k)} + \gamma \mathbf{T} = \mathbf{U}^{(k)} \mathrm{Diag}(\boldsymbol{\lambda}^{(k)}) (\mathbf{U}^{(k)})^\top$$

$\mathbf{d}^{(k+\frac{1}{2})} \in \mathrm{Prox}_{\gamma(\varphi+\psi)} \left(\boldsymbol{\lambda}^{(k)}\right)$

$\mathbf{C}^{(k+\frac{1}{2})} = \mathbf{U}^{(k)} \mathrm{Diag}(\mathbf{d}^{(k+\frac{1}{2})}) (\mathbf{U}^{(k)})^\top$

Choose $\alpha^{(k)} \in [0, 2]$

$\mathbf{C}^{(k+1)} \in \mathbf{C}^{(k)} + \alpha^{(k)} \Big( \mathrm{Prox}_{\gamma g_1} (2\mathbf{C}^{(k+\frac{1}{2})} - \mathbf{C}^{(k)}) - \mathbf{C}^{(k+\frac{1}{2})} \Big).$

## Theorem

Let $f$ and $g_0$ be spectral functions associated to $\varphi \in \Gamma_0(\mathbb{R}^n)$ and $\psi \in \Gamma_0(\mathbb{R}^n)$. Let $g_1 \in \Gamma_0(\mathcal{S}_n)$ be such that $f - \mathrm{tr}\,(\mathbf{T}\cdot) + g_0 + g_1$ is coercive. Assume that the intersection of the relative interiors of the domains of $f + g_0$ and $g_1$ is non empty. Let $(\alpha^{(k)})_{k \geqslant 0}$ be a sequence in $[0, 2]$ such that $\sum_{k=0}^{+\infty} \alpha^{(k)} (2 - \alpha^{(k)}) = +\infty$. Then, the sequences $(\mathbf{C}^{(k+\frac{1}{2})})_{k \geqslant 0}$ and $\left(\mathrm{prox}_{\gamma g_1} (2\mathbf{C}^{(k+\frac{1}{2})} - \mathbf{C}^{(k)})\right)_{k \geqslant 0}$ generated by the DR Algorithm converge to a solution to Problem (4).

3) Majorization-Minimization algorithm for robust graphical lasso

# Graphical lasso with noisy data

Let us consider the following signal model (Sun et al, 2017):

$$(\forall i \in \{1, \dots, N\}) \quad \mathbf{x}^{(i)} = \mathbf{A}\mathbf{s}^{(i)} + \mathbf{e}^{(i)}$$

where

- $\mathbf{A} \in \mathbb{R}^{n \times m}$ with $m \leqslant n$
- $\mathbf{s}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{E})$, $\mathbf{s}^{(i)} \in \mathbb{R}^m$
- $\mathbf{e}^{(i)} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I_d}\right)$, $\mathbf{e}^{(i)} \in \mathbb{R}^n$
- $\mathbf{s}^{(i)}$ and $\mathbf{e}^{(i)}$ are iid

Such observation model is encountered in several practical applications, e.g. in the context of "Relevant Vector Machine" (Tipping et al, 2001), (Wipf et al, 2004)

Covariance matrix of observed signal:

$$\begin{aligned} \mathbf{\Sigma} &= \mathbf{A}^\top \mathbf{E} \mathbf{A} + \sigma^2 \mathbf{I_d} \\ &= \mathbf{Y} + \sigma^2 \mathbf{I_d} \end{aligned}$$

**Goal:** Penalized maximum likelihood approach for finding an estimate $\mathbf{C}$ of $\mathbf{Y}^{-1}$, given the knowledge of $\sigma^2$ and the **empirical** covariance matrix

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^\top$$

**Prior:** Sparsity and low-rank structure of $\mathbf{C}$.

## Minimization problem

$$\underset{\mathbf{C}\in\mathcal{S}_n^{++}}{\text{minimize}} \left(\mathcal{F}(\mathbf{C}) = f(\mathbf{C}) + \mathcal{T}_{\mathbf{S}}(\mathbf{C}) + g_0(\mathbf{C}) + g_1(\mathbf{C})\right) \tag{5}$$

where

- $(\forall \mathbf{C} \in \mathcal{S}_n) \qquad f(\mathbf{C}) = \begin{cases} \log\det\left(\mathbf{C}^{-1} + \sigma^2 \mathbf{I_d}\right) & \text{if } \mathbf{C} \in \mathcal{S}_n^+ \\ +\infty & \text{otherwise,} \end{cases}$
- $(\forall \mathbf{C} \in \mathcal{S}_n) \qquad \mathcal{T}_{\mathbf{S}}(\mathbf{C}) = \begin{cases} \operatorname{tr}\left(\left(\mathbf{I_d} + \sigma^2\mathbf{C}\right)^{-1}\mathbf{CS}\right) & \text{if } \mathbf{C} \in \mathcal{S}_n^+ \\ +\infty & \text{otherwise,} \end{cases}$
- $g_0 \in \Gamma_0(\mathcal{S}_n)$ is a spectral function associated with $\psi \in \Gamma_0(\mathbb{R}^n)$, and $g_1 \in \Gamma_0(\mathcal{S}_n)$.

$\rightsquigarrow$ $f + g_0 + g_1$ is a **convex** function on $\mathcal{S}_n$.

$\rightsquigarrow$ The trace term $\mathcal{T}_{\mathbf{S}}$ is **concave** on $\mathcal{S}_n^+$

The whole functional $\mathcal{F}$ is nonconvex.

**Definition**

Let $\mathbf{C}' \in \mathcal{S}_n$. $\mathcal{G}(\cdot|\mathbf{C}')$ is a tangent majorant function for $\mathcal{F}$ at $\mathbf{C}'$ if, for every $\mathbf{C} \in \mathcal{S}_n$,

$$\mathcal{F}(\mathbf{C}) \leqslant \mathcal{G}(\mathbf{C}|\mathbf{C}') \quad \text{and} \quad \mathcal{F}(\mathbf{C}') = \mathcal{G}(\mathbf{C}'|\mathbf{C}')$$

**Majorization–Minimization algorithm:**

$$(\forall \ell \in \mathbb{N}) \quad \mathbf{C}^{(\ell+1)} = \underset{\mathbf{C}\in\mathcal{S}_n}{\operatorname{argmin}}\ \mathcal{G}(\mathbf{C}|\mathbf{C}^{(\ell)})$$

$\rightsquigarrow$ Ensures monotone decrease of $\left(\mathcal{F}(\mathbf{C}^{(\ell)})\right)_{\ell\in\mathbb{N}}$.

**Proposed strategy:**

- $\mathcal{F}$ reads as the sum of **convex** and **concave** terms
- Majoration of the concave term $\mathcal{T}_{\mathbf{S}}$ by a linear function
- Convex majorant function minimized by our Douglas-Rachford scheme.

- Construction of a majorizing approximation of $\mathcal{T}_{\mathbf{S}}$ at $\mathbf{C}' \in \mathcal{S}_n^+$:

$$(\forall \mathbf{C} \in \mathcal{S}_n^+) \quad \mathcal{T}_{\mathbf{S}}(\mathbf{C}) \leqslant \mathcal{T}_{\mathbf{S}}(\mathbf{C}') + \operatorname{tr}\left(\nabla \mathcal{T}_{\mathbf{S}}(\mathbf{C}')\left(\mathbf{C} - \mathbf{C}'\right)\right).$$

- As $f$ is finite only on $\mathcal{S}_n^+$, a tangent majorant of the cost function $\mathcal{F}$ at $\mathbf{C}'$ reads:

$$(\forall \mathbf{C} \in \mathcal{S}_n^+) \quad \mathcal{G}(\mathbf{C} \mid \mathbf{C}') = \\ f(\mathbf{C}) + \mathcal{T}_{\mathbf{S}}(\mathbf{C}') + \operatorname{tr}\left(\nabla \mathcal{T}_{\mathbf{S}}(\mathbf{C}')\left(\mathbf{C} - \mathbf{C}'\right)\right) + g_0(\mathbf{C}) + g_1(\mathbf{C}).$$

$\rightsquigarrow$ $\mathcal{F}(\mathbf{C}) \leqslant \mathcal{G}(\mathbf{C} \mid \mathbf{C}')$ for all $\mathbf{C} \in \mathcal{S}_n^+$ and $\mathcal{G}(\mathbf{C}' \mid \mathbf{C}') = \mathcal{F}(\mathbf{C}')$ at $\mathbf{C}' \in \mathcal{S}_n^+$.

- This leads to the general MM scheme:

$$(\forall \ell \in \mathbb{N}) \quad \mathbf{C}^{(\ell+1)} \in \operatorname{Argmin}_{\mathbf{C} \in \mathcal{S}_n} f(\mathbf{C}) + \operatorname{tr}(\nabla \mathcal{T}_{\mathbf{S}}(\mathbf{C}^{(\ell)})\mathbf{C}) + g_0(\mathbf{C}) + g_1(\mathbf{C})$$

with $\mathbf{C}^{(0)} \in \mathcal{S}_n^+$.

✓ At each iteration of the MM algorithm: Convex optimization problem of the form (4) ⇒ Douglas–Rachford approach.

✓ Convergence guarantee to a critical point of $\mathcal{F}$.

## 4) Numerical experiments

The dataset is generated by a slight modification of Boyd's code[1]:

- a **sparse precision matrix** $\mathbf{C}_0$ of dimension $n \times n$ is generated (n=100)
- its inverse $\mathbf{\Sigma_0}$ is employed to generate $N = 10000$ realizations of a Gaussian mvrv $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_0})$
- Gaussian noise of variance $\sigma^2$ is added to the realizations, in order to satisfy $\mathbf{x}^{(i)} = \mathbf{A}\mathbf{s}^{(i)} + \mathbf{e}^{(i)}$ $(\mathbf{A} = \mathbf{I_d})$ and hence the true covariance matrix is

$$\mathbf{\Sigma} = \mathbf{\Sigma_0} + \sigma^2 \mathbf{I_d}$$

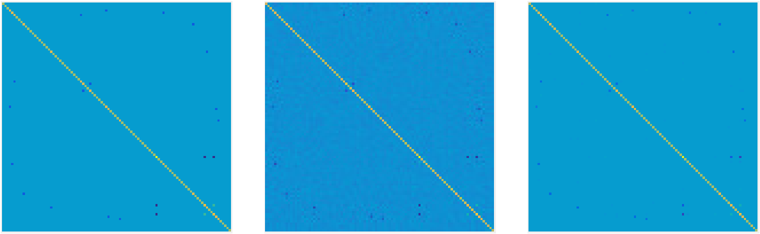
- the empirical covariance matrix $\mathbf{S}$ is obtained by

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^\top$$

Three type of error measurements:

| False Positive Rate on Precision Matrix (fpr) | True Positive Rate on Precision Matrix (tpr) | Relative Mean Square Error on $\Sigma$ (RMSE) |
| --- | --- | --- |

[1] http://stanford.edu/~boyd/papers/admm/covsel/covsel_example.html

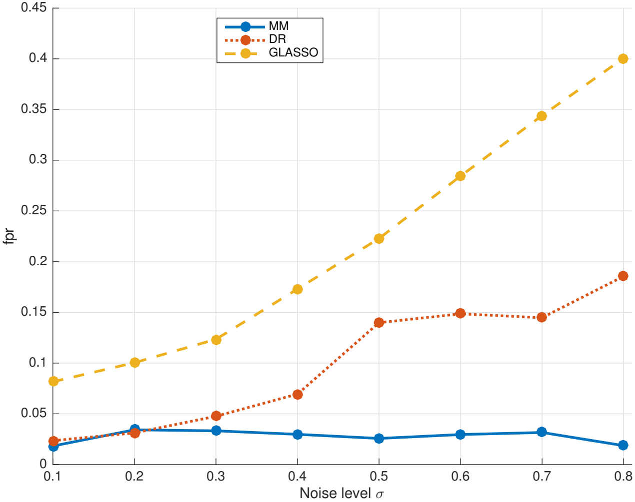


$\Sigma_0$ $\mathbf{S}$ $\Sigma_{\text{rec}}$

Settings:

- $g_0(\mathbf{C}) = \mu_0 \mathcal{R}_1(\mathbf{C}^{-1})$ (Schatten 1–norm, nuclear norm)
- $g_1(\mathbf{C}) = \mu_1 \|\mathbf{C}\|_1$ (component–wise $\ell_1$ norm)
- $\mu_0 = 0.0716$, $\mu_1 = 0.0278$, $\alpha = 1.5$
- Noise level: $\sigma = 0.5$
- RMSE: 0.1180
- FPR (on precision matrix): 0.0257
- TPR (on precision matrix): 100%

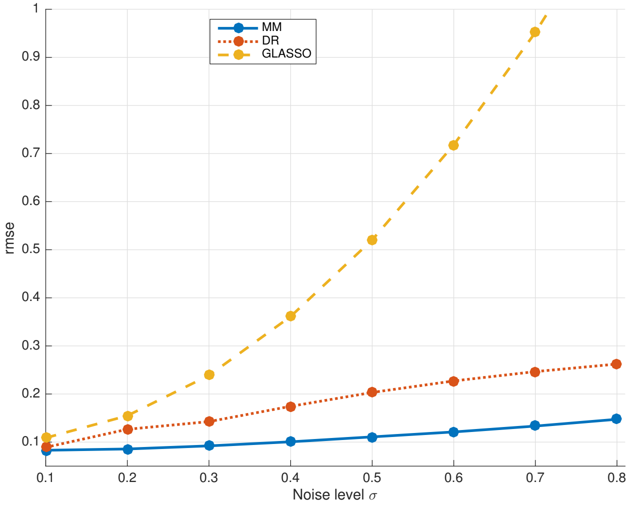
# Numerical Experiments: Stability Under Increasing Noise




## Comparisons:

- GLASSO: $\sigma = 0,\ g_0 = 0$
- DR: $\sigma = 0$

# Numerical Experiments: Stability Under Increasing Noise




## Comparisons:

- GLASSO: $\sigma = 0,\ g_0 = 0$
- DR: $\sigma = 0$

# Conclusions

Three main contributions:

- ✓ proximity operators for different coupling of spectral fidelity and regularization functions
- ✓ a **nonconvex formulation** of matrix estimation problem arising in the context of noisy Graphical LASSO
- ✓ a Majorization–Minimization approach proposed to solve the nonconvex model.

The comparison with state–of–the–art algorithms has shown that the proposed model is stable w.r.t. increasing noise perturbing the data.

Future work:

- Extension to complex Hermitian matrices.
- Extension to non–squared matrices via SVD.

All the presented results are collected in:

A.Benfenati, E. Chouzenoux, J.–C. Pesquet, *A Proximal Approach for a Class of Matrix Optimization Problems*, submitted. [hal-01673027]