A Nonconvex Variational Approach For Robust Graphical LASSO

A. Benfenati, E. Chouzenoux, J.–C. Pesquet

ESIEE PARIS

CENTER FOR VISUAL COMPUTING
École Centrale Paris / INRIA

IEEE ICASSP 2018, 19 April 2018

Outline

Several problems such as shape classification (Duchi et al, 2008) gene expression (Ma et al, 2013) model selection (Chandrasekaran et al, 2012) computer vision (Guo et al, 2011) inverse covariance estimation (Aspremont et al, 2008) graph estimation (Meinshausen et al, 2006), brain network analysis (Yang et al, 2015) lead to find the minimum of a matrix functional $\mathcal{G}$.

Classical Example: Graphical LASSO Model

$$(\forall\, \mathbf{C} \in \mathcal{S}_n) \quad \mathcal{G}(\mathbf{C}) = f\left(\mathbf{C}\right) + \operatorname{tr}\left(\mathbf{CS}\right) + \mu\|\mathbf{C}\|_1, \quad \mu \in \mathbb{R}^+$$

$$\operatorname{tr}\left(\mathbf{C}\right) = \sum_{i=1}^{n} \mathbf{C}_{ii}, \quad \|\mathbf{C}\|_1 = \sum_{i,j=1}^{n} |\mathbf{C}_{ij}|, \quad f\left(\mathbf{C}\right) = \begin{cases} -\log\det\left(\mathbf{C}\right) & \text{if } \mathbf{C} \in \mathcal{S}_n^{++} \\ +\infty & \text{otherwise} \end{cases}$$

and $\mathbf{S}$ is a given matrix.

Algorithms

Several algorithms were proposed to solve the above problem:

- popular GLASSO algorithm (Friedman et al, 2008),
- ADMM (Boyd et al, 2011)
- Gradient Projection method (Duchi et al, 2008)
- ...

Main Problem

The GLASSO model **does not take into account** the noise on the data $\mathbf{S}$.

New variational approach:

(i) a more versatile regularization function g, consisting in a sum of two terms:

$$g = g_0 + g_1$$

(ii) a novel fidelity function which includes the information about the noise

Main Problem

The GLASSO model **does not take into account** the noise on the data $\mathbf{S}$.

New variational approach:

(i) a more versatile regularization function g, consisting in a sum of two terms:

$$g = g_0 + g_1$$

(ii) a novel fidelity function which includes the information about the noise

(i) $\rightarrow$ Douglas–Rachford proximal minimization approach

Main Problem

The GLASSO model **does not take into account** the noise on the data $\mathbf{S}$.

New variational approach:

(i) a more versatile regularization function g, consisting in a sum of two terms:

$$g = g_0 + g_1$$

(ii) a novel fidelity function which includes the information about the noise

(i) $\rightarrow$ Douglas–Rachford proximal minimization approach

(ii) $\rightarrow$ the new functional is **nonconvex**, then a **Majorization–Minimization approach** is adopted

Let

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

f is a **spectral function** iff

$$f(\mathbf{C}) = \varphi(\mathbf{Pd}), \quad \mathbf{C} = \mathbf{U}^\top \mathrm{Diag}(\mathbf{d})\mathbf{V}, \quad \varphi \in \Gamma_0(\mathbb{R}^n)$$

and $\mathbf{P}$ is a permutation matrix.

	$f(\mathbf{C})$	$\varphi(\mathbf{Pd})$
GLASSO function	$-\log\det(\mathbf{C})$	$-\sum_{i=1}^{n}\log(\mathbf{d}_i)$
Froebenius norm	$\frac{1}{2}\|\mathbf{C}\|_{\mathrm{F}}^2$	$\sum_{i=1}^{n}\mathbf{d}_i^2$
Von Neumann Entropy	$\mathrm{tr}(\mathbf{C}\log(\mathbf{C}))$	$\sum_{i=1}^{n}\mathbf{d}_i\log(\mathbf{d}_i)$

Let consider

$$\underset{\mathbf{C}\in\mathcal{S}_n}{\operatorname{argmin}}\ f(\mathbf{C}) + \operatorname{tr}(\mathbf{TC}) + g(\mathbf{C})$$

with f being a spectral function and g s.t.

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

where

- $g_0(\mathbf{C}) = \psi(\mathbf{Pd}), \mathbf{C} = \mathbf{U}^\top \operatorname{Diag}(\mathbf{d})\mathbf{V}, \psi \in \Gamma_0(\mathbb{R}^n)$, i.e. g_0 is a spectral function
- $g_1 \in \Gamma_0(\mathbb{R}^{n\times n})$ acts on the whole matrix $\mathbf{C}$ (e.g., the ℓ_1 norm)

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

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

The problem

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

can be solved by the **Douglas–Rachford** algorithm ($\gamma > 0$):

$$\begin{aligned}
\mathbf{C}^{(k+1/2)} &= \operatorname{prox}_{\gamma h_0}(\mathbf{C}^{(k)}) \\
\mathbf{C}^{(k+1)} &= \mathbf{C}^{(k)} + \alpha_k \left(\operatorname{prox}_{\gamma h_1}\left(2\mathbf{C}^{(k+1/2)} - \mathbf{C}^{(k)}\right) - \mathbf{C}^{(k+1/2)} \right),\ \alpha_k \in [0,2)
\end{aligned}$$

where

$$\operatorname{prox}_{\gamma f}(\widehat{\mathbf{C}}) = \underset{\mathbf{C}}{\operatorname{argmin}}\ \gamma f(\mathbf{C}) + \frac{1}{2}\|\mathbf{C} - \widehat{\mathbf{C}}\|_{\mathrm{F}}^2$$

The burden of the computation of the $\operatorname{prox}_{\gamma h_0}$ can be loaded on the sole eigenvalues:

The problem

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

can be solved by the **Douglas–Rachford** algorithm ($\gamma > 0$):

$$\begin{aligned}
\mathbf{C}^{(k+1/2)} &= \operatorname{prox}_{\gamma h_0}(\mathbf{C}^{(k)}) \\
\mathbf{C}^{(k+1)} &= \mathbf{C}^{(k)} + \alpha_k \left(\operatorname{prox}_{\gamma h_1}\left(2\mathbf{C}^{(k+1/2)} - \mathbf{C}^{(k)}\right) - \mathbf{C}^{(k+1/2)} \right),\ \alpha_k \in [0,2)
\end{aligned}$$

where

$$\operatorname{prox}_{\gamma f}(\widehat{\mathbf{C}}) = \underset{\mathbf{C}}{\operatorname{argmin}}\ \gamma f(\mathbf{C}) + \frac{1}{2}\|\mathbf{C} - \widehat{\mathbf{C}}\|_{\mathrm{F}}^2$$

The burden of the computation of the $\operatorname{prox}_{\gamma h_0}$ can be loaded on the sole eigenvalues:

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

The problem

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

can be solved by the **Douglas–Rachford** algorithm ($\gamma > 0$):

$$\begin{aligned}
\mathbf{C}^{(k+1/2)} &= \operatorname{prox}_{\gamma h_0}(\mathbf{C}^{(k)}) \\
\mathbf{C}^{(k+1)} &= \mathbf{C}^{(k)} + \alpha_k \left(\operatorname{prox}_{\gamma h_1}\left(2\mathbf{C}^{(k+1/2)} - \mathbf{C}^{(k)}\right) - \mathbf{C}^{(k+1/2)} \right),\ \alpha_k \in [0,2)
\end{aligned}$$

where

$$\operatorname{prox}_{\gamma f}(\widehat{\mathbf{C}}) = \underset{\mathbf{C}}{\operatorname{argmin}}\ \gamma f(\mathbf{C}) + \frac{1}{2}\|\mathbf{C} - \widehat{\mathbf{C}}\|_{\mathrm{F}}^2$$

The burden of the computation of the $\operatorname{prox}_{\gamma h_0}$ can be loaded on the sole eigenvalues:

$$\begin{aligned}
\mathbf{C}^{(k)} - \gamma\mathbf{T} &= \mathbf{U}^{(k)}\operatorname{Diag}(\boldsymbol{\lambda}^{(k)})(\mathbf{U}^{(k)})^\top \\
\mathbf{d}^{(k+\frac{1}{2})} &\in \operatorname{prox}_{\gamma(\varphi+\psi)}\left(\boldsymbol{\lambda}^{(k)}\right)
\end{aligned}$$

The problem

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

can be solved by the **Douglas–Rachford** algorithm ($\gamma > 0$):

$$\begin{aligned}
\mathbf{C}^{(k+1/2)} &= \operatorname{prox}_{\gamma h_0}(\mathbf{C}^{(k)}) \\
\mathbf{C}^{(k+1)} &= \mathbf{C}^{(k)} + \alpha_k \left(\operatorname{prox}_{\gamma h_1}\left(2\mathbf{C}^{(k+1/2)} - \mathbf{C}^{(k)}\right) - \mathbf{C}^{(k+1/2)} \right),\ \alpha_k \in [0,2)
\end{aligned}$$

where

$$\operatorname{prox}_{\gamma f}(\widehat{\mathbf{C}}) = \underset{\mathbf{C}}{\operatorname{argmin}}\ \gamma f(\mathbf{C}) + \frac{1}{2}\|\mathbf{C} - \widehat{\mathbf{C}}\|_{\mathrm{F}}^2$$

The burden of the computation of the $\operatorname{prox}_{\gamma h_0}$ can be loaded on the sole eigenvalues:

$$\begin{aligned}
\mathbf{C}^{(k)} - \gamma\mathbf{T} &= \mathbf{U}^{(k)}\operatorname{Diag}(\boldsymbol{\lambda}^{(k)})(\mathbf{U}^{(k)})^\top \\
\mathbf{d}^{(k+\frac{1}{2})} &\in \operatorname{prox}_{\gamma(\varphi+\psi)}\left(\boldsymbol{\lambda}^{(k)}\right) \\
\mathbf{C}^{(k+1/2)} &= \mathbf{U}^{(k)}\operatorname{Diag}(\mathbf{d}^{(k+\frac{1}{2})})(\mathbf{U}^{(k)})^\top
\end{aligned}$$

Example of Prox Operators

Proximity operators for different choices for g_0 and

$$f = -\log\det$$

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

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

The **true** covariance matrix $\mathbf{\Sigma}$ of the observed signal is

$$\mathbf{\Sigma} = \mathbf{A}^\top \mathbf{E} \mathbf{A} + \sigma^2 \mathbf{I_d}.$$

The **empirical** covariance matrix of the $\mathbf{x}^{(i)}$s

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

can be seen as a rough approximation of Σ.

$$\underset{\mathbf{C}\in\mathcal{S}_n^{++}}{\operatorname{argmin}}\ f(\mathbf{C}) + \mathcal{T}_{\mathbf{S}}(\mathbf{C}) + \mu_0 g_0(\mathbf{C}) + \mu_1 g_1(\mathbf{C}) \qquad (*)$$

where

$$\begin{aligned} (\forall \mathbf{C} \in \mathcal{S}_n^{++}) \quad f(\mathbf{C}) &\triangleq \log\det\left(\mathbf{C}^{-1} + \sigma^2 \mathbf{I_d}\right) \\ (\forall \mathbf{C} \in \mathcal{S}_n^{+}) \quad \mathcal{T}_{\mathbf{S}}(\mathbf{C}) &\triangleq \operatorname{tr}\left(\left(\mathbf{I_d} + \sigma^2 \mathbf{C}\right)^{-1} \mathbf{CS}\right), \end{aligned}$$

and $g_0, g_1 \in \Gamma_0(\mathbb{R}^{n\times n})$, $g_0(\mathbf{C}) = \psi(\mathbf{Pd})$.

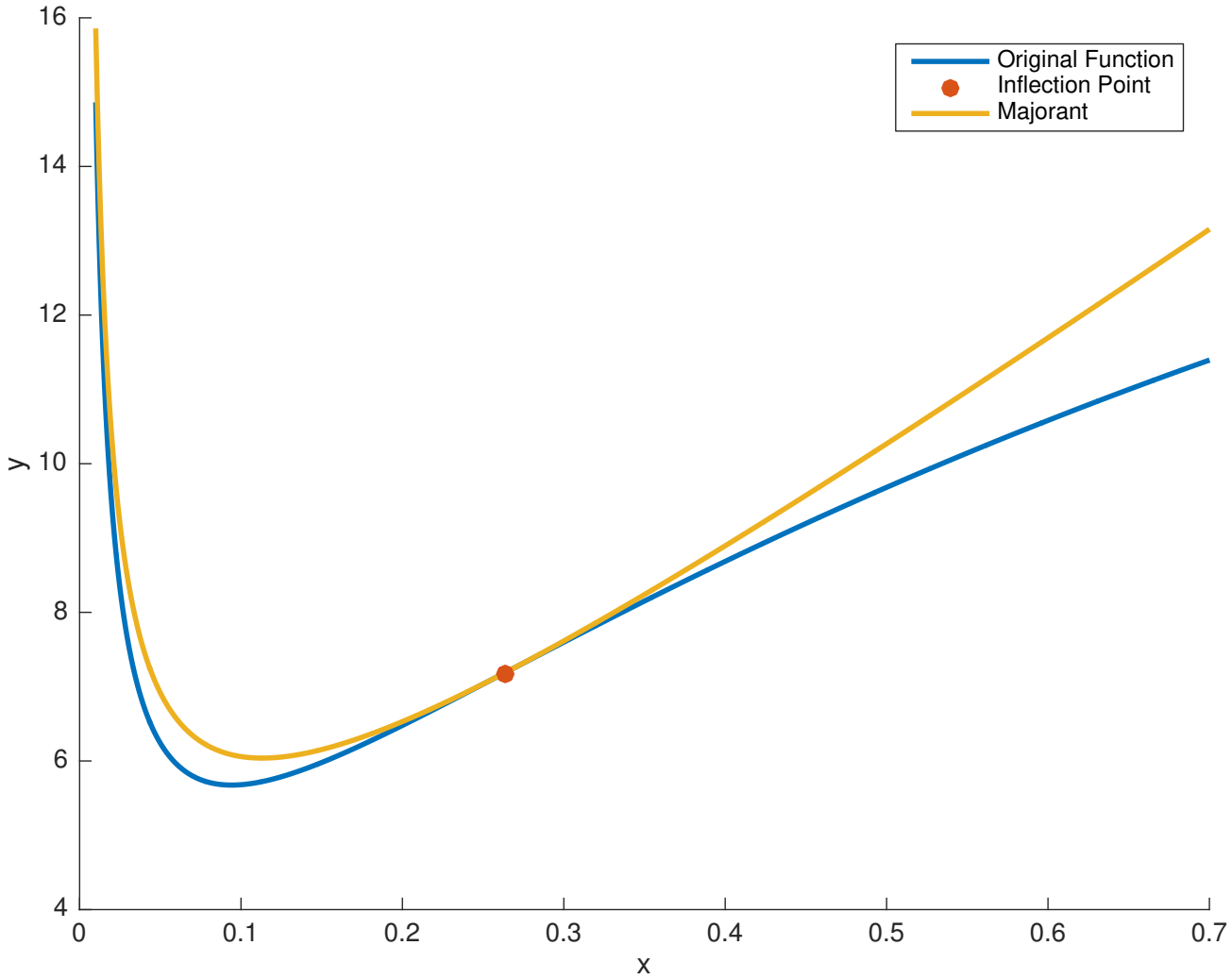
Lemma

Consider (*) *with* g_0 *and* g_1 *in* $\Gamma_0(\mathbb{R}^{n\times n})$.

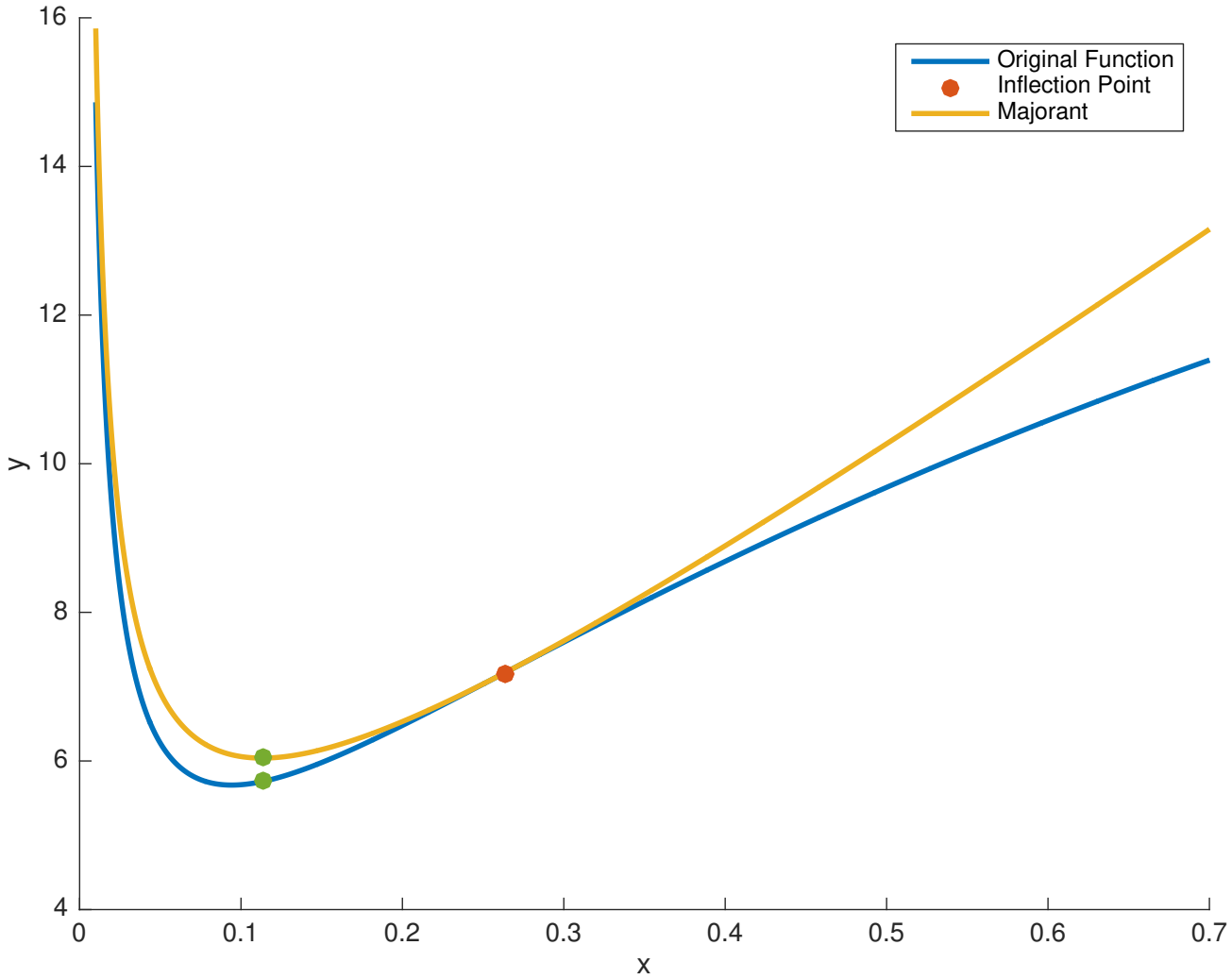
1. $f + g_0 + g_1$ *is a convex function.*
2. *The trace term* $\mathcal{T}_{\mathbf{S}}$ *is concave on* $\mathcal{S}_n^+$.

The whole functional is nonconvex.

- Majorize the original function via a convex approximation.

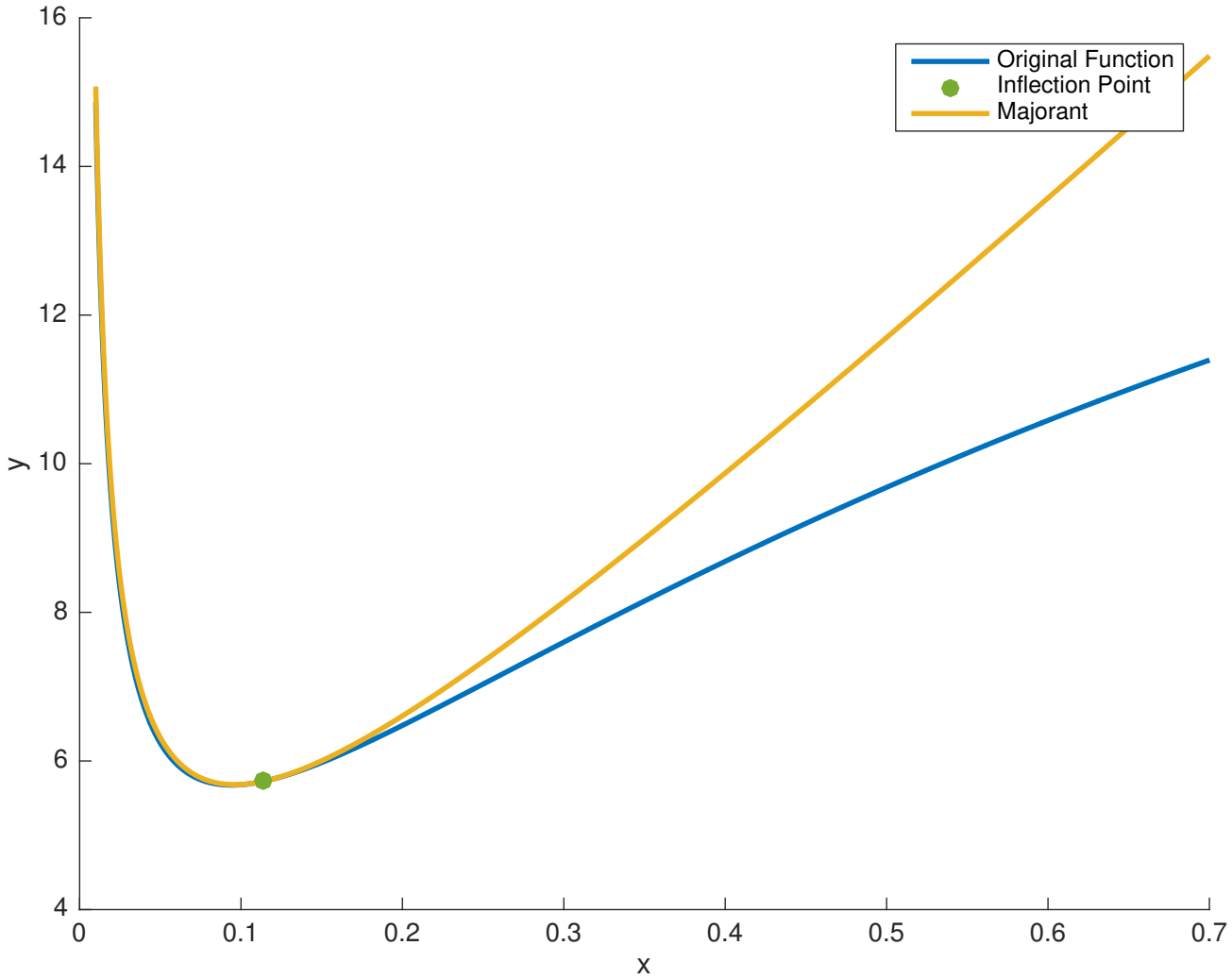


- Majorize the original function via a convex approximation.
- Minimize the approximated functional.



Majorization–Minimization Approach

- Majorize the original function via a convex approximation.
- Minimize the approximated functional.
- Repeat until convergence.



The problem to be solved hence is

$$\mathbf{C}^{(\ell+1)} = \underset{\mathbf{C}\in\mathcal{S}_n^{++}}{\operatorname{argmin}} \left[f(\mathbf{C}) + \operatorname{tr}\left(\nabla\mathcal{T}_{\mathbf{S}}\left(\mathbf{C}^{(\ell)}\right)\mathbf{C}\right) + \mu_0 g_0(\mathbf{C}) + \mu_1 g_1(\mathbf{C}) \right]. \quad (\dagger)$$

with

$$\nabla\mathcal{T}_{\mathbf{S}}(\mathbf{C}^{(\ell)}) = \left(\mathbf{I_d} + \sigma^2\mathbf{C}^{(\ell)}\right)^{-1} \mathbf{S} \left(\mathbf{I_d} + \sigma^2\mathbf{C}^{(\ell)}\right)^{-1}.$$

i.e. $\mathcal{T}_{\mathbf{S}}$ has been substituted by its linear approximation.

Observe that

- f is convex and spectral
- $f + \mu_0 g_0$ is convex
- $g_1 \in \Gamma_0(\mathbb{R}^{n\times n})$

then

the solution to $(\dagger)$ can be found via the Douglas–Rachford approach.

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 $N = 10000$ realization of a Gaussian mvrv $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma_0})$
- Gaussian noise of variance σ^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

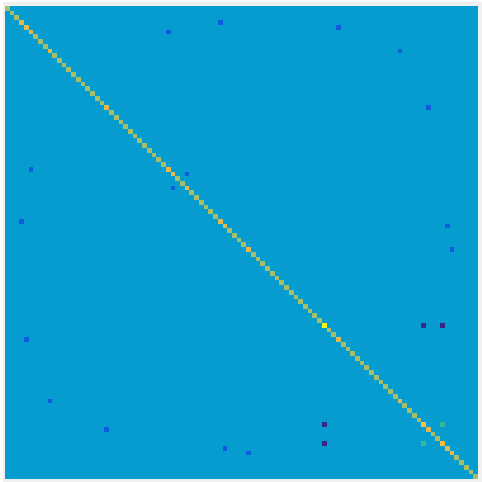
$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} (\mathbf{x}^{(i)})^\top$$

Three type of error measurements:

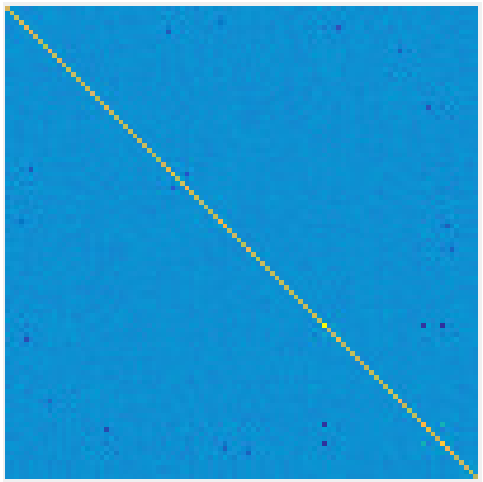
False Positive Rate on Precision Matrix (fpr)	True Positive Rate on Precision Matrix (tpr)	Relative Mean Square Error on Σ (RMSE)

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

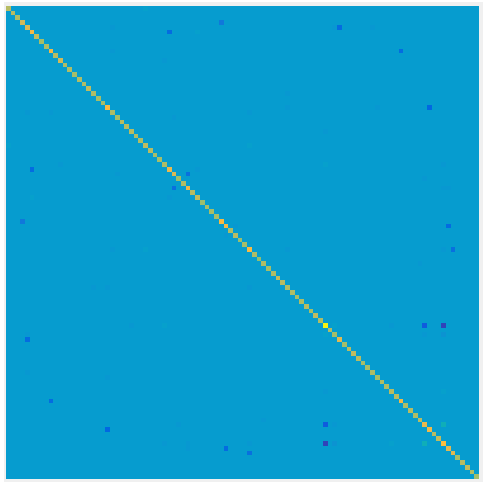
Numerical Experiments: Quality of Reconstruction



Σ_0



$\mathbf{S}$

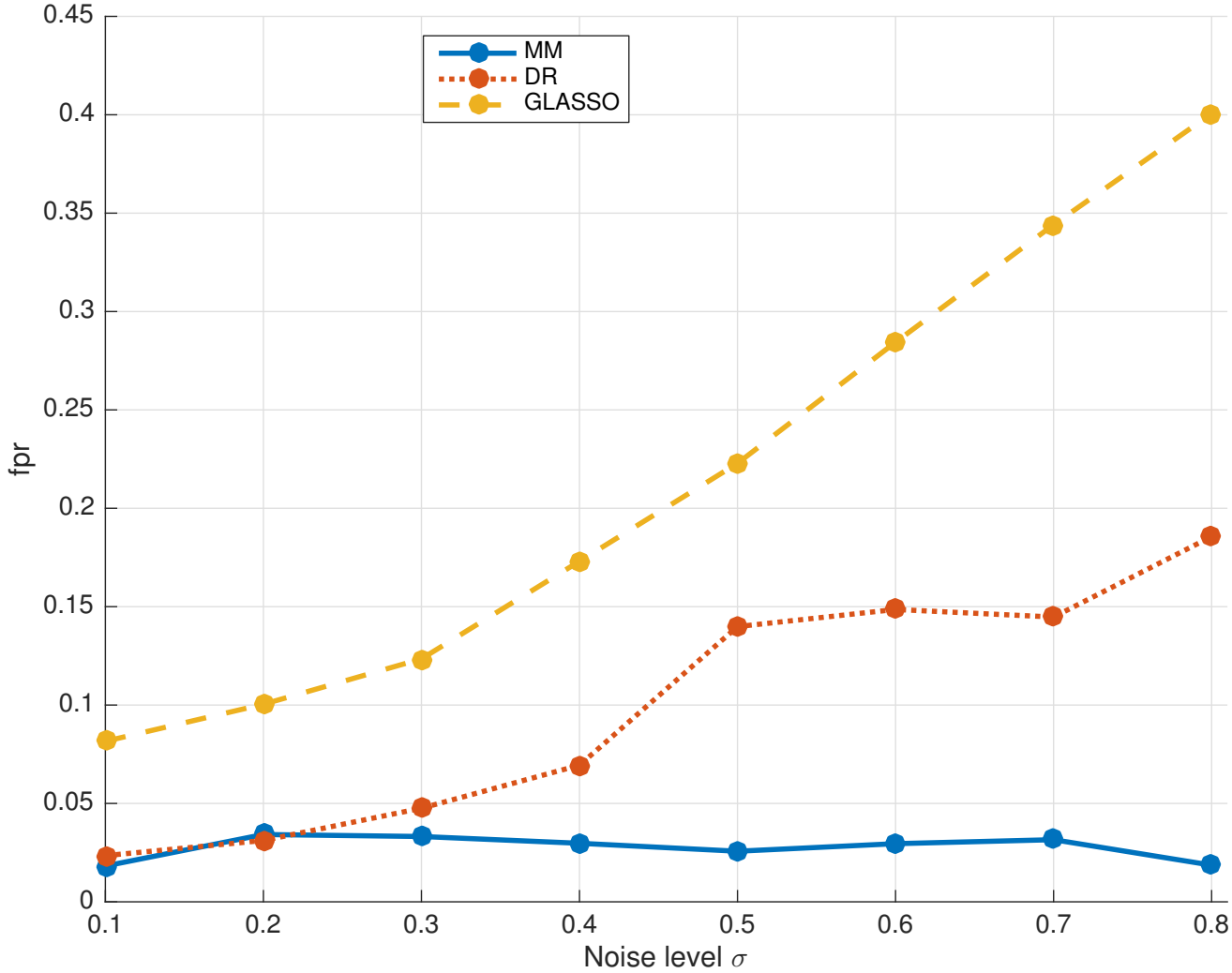


Σ_{rec}

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 ℓ_1 norm)
- $\mu_0 = 0.0716$, $\mu_1 = 0.0278$
- Noise level: $\sigma = 0.5$
- RMSE: 0.1180
- FPR (on precision matrix): 0.0257
- TPR (on precision matrix): 100%

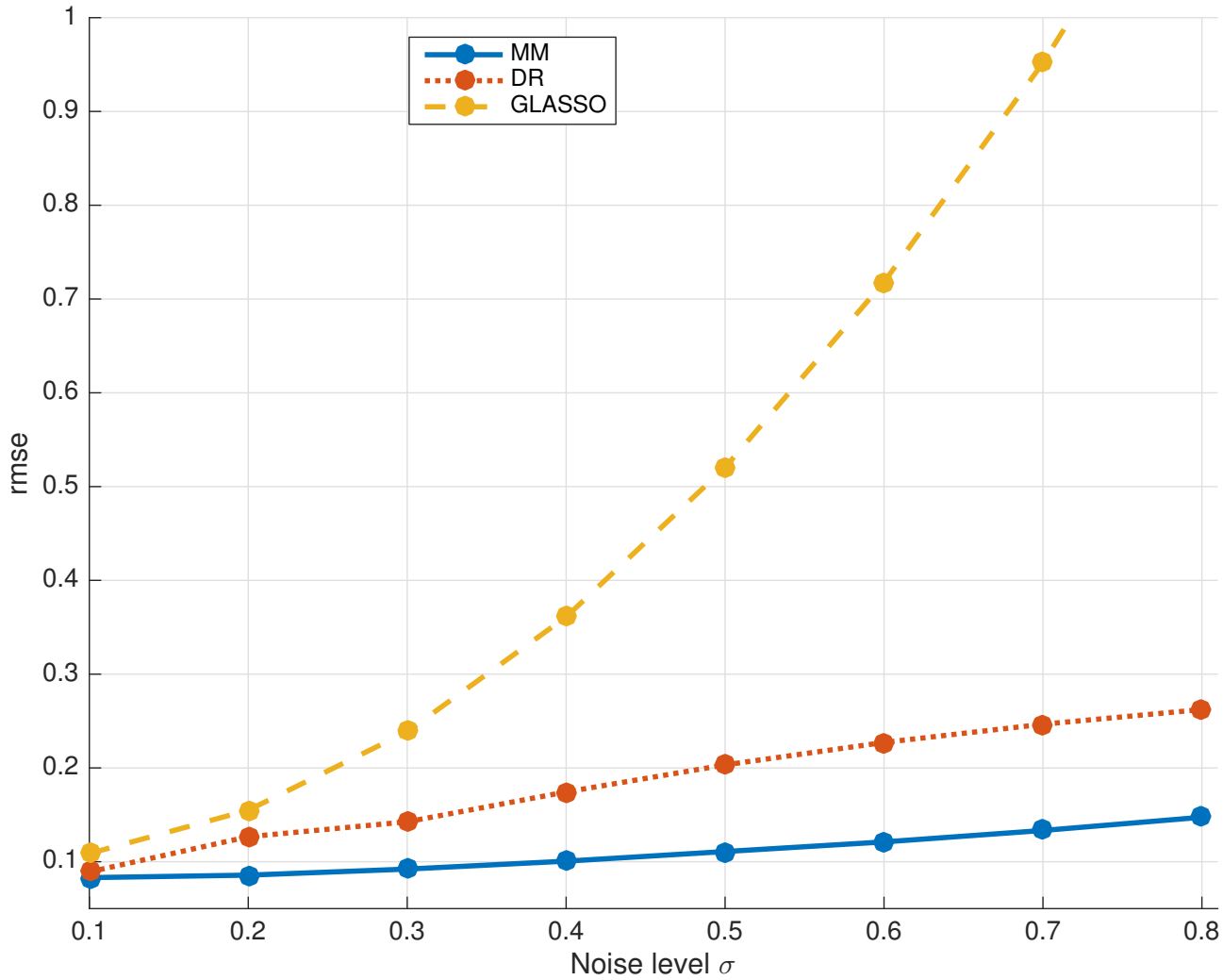
Numerical Experiments: Stability Under Increasing Noise



Reminder: GLASSO Approach

$$-\log\det(\mathbf{C}) + \mathrm{tr}(\mathbf{CS}) + \mu\|\mathbf{C}\|_1, \quad \mu \in \mathbb{R}^+$$

Numerical Experiments: Stability Under Increasing Noise



Reminder: GLASSO Approach

$$-\log\det(\mathbf{C}) + \mathrm{tr}(\mathbf{CS}) + \mu\|\mathbf{C}\|_1, \quad \mu \in \mathbb{R}^+$$

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