Self-Concordant Function Minimization: Sparse Graph Selection

Anastasios Kyrillidis   Volkan Cevher

Laboratory for Information and Inference Systems (LIONS), EPFL

{anastasios.kyrillidis, volkan.cevher}@epfl.ch

Motivation: Sparse Gaussian Graph Selection

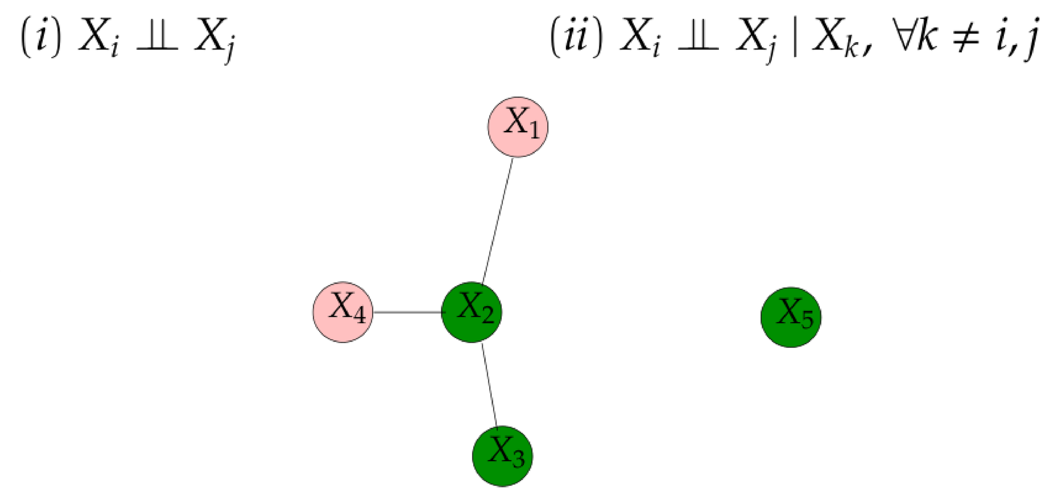
- **Setting:** $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ with joint pdf $f(X_1, \dots, X_n) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- **Goal:** infer (conditional/unconditional) independence among $\mathbf{X}$, given a set of samples.
- **Sampling:** $\{\mathbf{x}_j\}_{j=1}^p$ is a collection of p i.i.d. n-variate vectors where $\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\forall j$.

  Empirical mean: $\widehat{\boldsymbol{\mu}} = \frac{1}{p}\sum_{j=1}^p \mathbf{x}_j$ **Sample covariance:** $\widehat{\boldsymbol{\Sigma}} = \frac{1}{p}\sum_{j=1}^p (\mathbf{x}_j - \widehat{\boldsymbol{\mu}})(\mathbf{x}_j - \widehat{\boldsymbol{\mu}})^T$.

- **Connection to graph learning:**

$G = (V, E)$ is a Gaussian Markov network where

1. $V = \{1, \dots, n\}$: set of variables.
2. E: contains edge (i, j) iff X_i is *conditionally* independent to X_j.



Why Gaussian? $\boldsymbol{\Sigma}_{ij}^{-1} = 0 \iff X_i \perp\!\!\!\perp X_j \mid X_k$, $\forall k \neq i, j$

- **"Old" estimation techniques:** using $\widehat{\boldsymbol{\Sigma}}$ only or through unregularized ML techniques.

  CAVEAT: (i) usually, results in a non-robust estimator,

  (ii) no easy interpretation since the estimator is usually fully dense!

- **Parsimony principle:** select the *simplest* graphical model that adequately explains the data.

Let $\boldsymbol{\Theta} = \widehat{\boldsymbol{\Sigma}}^{-1}$. Given $\widehat{\boldsymbol{\Sigma}}$ and $\rho > 0$:

$$\boldsymbol{\Theta}^* = \operatorname*{argmin}_{\boldsymbol{\Theta} \succ 0} \Big\{ \underbrace{-\log\det(\boldsymbol{\Theta}) + \operatorname{trace}(\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}})}_{=f(\boldsymbol{\Theta})} + \underbrace{\rho\|\boldsymbol{\Theta}\|_1}_{=g(\boldsymbol{\Theta})} \Big\} \quad (1)$$

- **Contributions**:
  1. **New** first-order gradient scheme: fast convergence (# iter.) as compared to state-of-the-art.
  2. **New ingredient: Adaptive** step size selection using **self-concordance** of the objective.

Challenges and Related Work

- Challenge #1: High-dimensional statistical problems have become the norm.
- Challenge #2: Neither $f(\boldsymbol{\Theta})$ nor $g(\boldsymbol{\Theta})$ is Lipschitz-continuous gradient functions; $g(\boldsymbol{\Theta})$ is a nonsmooth regularizer.
- Challenge #3: (1) is defined over the positive-definite cone $\mathbb{S}_{++}^n$.
- Challenge #4: The selection of regularization parameter ρ is crucial.

- Why not use off-the-self Interior-point methods (IPM)?

(BANERJEE, EL GHAOUI, AND D'ASPREMONT (2007)): "... the resulting complexity for existing IPMs is $\mathcal{O}(n^6)$ where n is the number of variables..."

- Plenty of efficient approaches:
  - Block coordinate descent/ascent schemes (Graphical Lasso, CovSel, SINCO, etc.)
  - Lagrangian schemes (ALM, ADMM, etc.)
  - Second-order schemes (QUIC, Newton-based methods, etc.)
  - ...

Gradient descent scheme

- $\boldsymbol{\Theta}^* = \operatorname*{argmin}_{\boldsymbol{\Theta} \succ 0} f(\boldsymbol{\Theta}) + g(\boldsymbol{\Theta})$.
- Use $\boldsymbol{\Delta} := -\nabla f(\boldsymbol{\Theta}_i)$ (ignore $g(\boldsymbol{\Theta}_i)$).
- Quadratic surrogate for $f(\boldsymbol{\Theta})$ at $\boldsymbol{\Theta}_i \in \mathbb{R}^{n \times n}$:

$$f(\boldsymbol{\Theta}) \leq U(\boldsymbol{\Theta}, \boldsymbol{\Theta}_i)$$
$$:= f(\boldsymbol{\Theta}_i) + \operatorname{trace}\left(\nabla f(\boldsymbol{\Theta}_i)(\boldsymbol{\Theta} - \boldsymbol{\Theta}_i)\right) + \frac{1}{2\tau_i}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}_i\|_F^2$$

for some $\tau_i > 0$.

- Iteratively, solve:

$$\boldsymbol{\Theta}_{i+1} = \arg\min_{\boldsymbol{\Theta} \succ 0}\{U(\boldsymbol{\Theta}, \boldsymbol{\Theta}_i) + g(\boldsymbol{\Theta})\}$$

or in "proximity operator" form:

$$\boldsymbol{\Theta}_{i+1} = \underbrace{\arg\min_{\boldsymbol{\Theta} \succ 0}\Big\{\frac{1}{2\tau_i}\|\boldsymbol{\Theta} - \overbrace{(\boldsymbol{\Theta}_i - \tau_i\nabla f(\boldsymbol{\Theta}_i))}^{:\text{gradient descent}}\|_F^2 + g(\boldsymbol{\Theta})\Big\}}_{=\text{Soft}(\cdot, \tau_i\rho)}$$

Step size selection τ_i - Part 1

- Gradient descent: $\boldsymbol{X}_i = \boldsymbol{\Theta}_i - \tau_i\nabla f(\boldsymbol{\Theta}_i)$.
- Bregman divergence between $\boldsymbol{X}_i$ and $\boldsymbol{\Theta}_i$:

$$\mathcal{D}_f(\boldsymbol{X}_i \parallel \boldsymbol{\Theta}_i) = -\sum_{j=1}^n \log(1 - \tau_i\lambda_j) - \tau_i \cdot \operatorname{trace}\left(\boldsymbol{\Theta}_i^{-1}\nabla f(\boldsymbol{\Theta}_i)\right) =: \phi(\tau_i), \text{ where } \lambda_j : \text{ eigs. of } \boldsymbol{\Theta}_i^{-1/2}\nabla f(\boldsymbol{\Theta}_i)\boldsymbol{\Theta}_i^{-1/2}$$

- Condition on τ_i to be satisfied: $\tau_i \leq 1/\lambda_j$, $\forall j$.
- **KEY INGREDIENTS:**

A convex function $h : \mathbb{R} \to \mathbb{R}$ is self-concordant if $|h'''(x)| \leq 2h''(x)^{3/2}$ for all $x \in \mathbb{R}$. Furthermore, a function $h : \mathbb{R}^{n \times n} \to \mathbb{R}$ is self-concordant if, for any $t \in \mathbb{R}$, the function $\phi(t) := h(\mathbf{X} + t\mathbf{V})$ is self-concordant for all $\mathbf{X}, \mathbf{V} \in \mathbb{R}^{n \times n}$. Given h_1, h_2 are self-concordant functions, then $h_1 + h_2$ is self-concordant.

Let $h : \mathbb{R} \to \mathbb{R}$ be a *strictly convex*, self-concordant function. Then: $\frac{h''(0)}{\left(1+t\sqrt{h''(0)}\right)^2} \leq h''(t) \leq \frac{h''(0)}{\left(1-t\sqrt{h''(0)}\right)^2}$, where the lower bound holds for $t \geq 0$ and the upper bound is valid for $0 \leq t \leq 1/\sqrt{h''(0)}$.

LEMMA: The function $\phi(\tau_i)$ is strictly convex and self-concordant.

Step size selection τ_i - Part 2

- By the second order expansion of $\phi(\tau_i)$:

LEMMA: The function $\phi(\tau_i)$ satisfies: $\phi(\tau_i) = \frac{1}{2} \cdot \tau_i^2 \cdot \phi''(\widehat{\tau}_i)$, for $\widehat{\tau}_i \in (0, \tau_i]$ and $\phi''(\widehat{\tau}_i) = \sum_{j=1}^n \frac{\lambda_j^2}{(1-\widehat{\tau}_i\lambda_j)^2}$

- Since $\phi(\tau_i) := \mathcal{D}_f(\boldsymbol{X}_i \parallel \boldsymbol{\Theta}_i)$ and using $\frac{\delta}{\left(1+\tau_i\sqrt{\delta}\right)^2} \leq \phi''(\widehat{\tau}_i) \leq \frac{\delta}{\left(1-\tau_i\sqrt{\delta}\right)^2}$, we obtain:

$$\frac{\widetilde{\mu}}{2} \leq \frac{\mathcal{D}_f(\boldsymbol{X}_i \parallel \boldsymbol{\Theta}_i)}{\|\boldsymbol{X}_i - \boldsymbol{\Theta}_i\|_F^2} \leq \frac{\widetilde{L}}{2} \longleftarrow \text{Local Lipschitz constants and strong convexity parameter}$$

where $\frac{\widetilde{L}}{2} = \frac{\delta}{2\left(1-\tau_i\sqrt{\delta}\right)^2\|\nabla f(\boldsymbol{\Theta}_i)\|_F^2}$ and $\frac{\widetilde{\mu}}{2} = \frac{\delta}{2\left(1+\tau_i\sqrt{\delta}\right)^2\|\nabla f(\boldsymbol{\Theta}_i)\|_F^2}$.

- Two Nesterov-based step size selection schemes:

LEMMA: For convex and strongly convex (unconstrained) minimization, the step size τ_i^* is uniquely determined as the *minimum and maximum* (resp.) root of the quadratic forms:

$$\tau_i = 1/\widetilde{L} \iff \tau_i^2 - 2\left(\frac{1}{\sqrt{\delta}} + \frac{1}{2\epsilon}\right)\tau_i + \frac{1}{\delta} = 0 \text{ and } \tau_i = \frac{2}{\widetilde{\mu} + \widetilde{L}} \iff \tau_i^2 + \frac{1}{\sqrt{\epsilon}}\tau_i - \frac{1}{\delta} = 0$$

respectively, where $\delta := \phi''(0)$ and $\epsilon := \|\boldsymbol{X}_i - \boldsymbol{\Theta}_i\|_F^2$. Moreover, τ_i^* satisfies $0 \leq \tau_i^* < 1/\sqrt{\phi''(0)}$.

Experiments

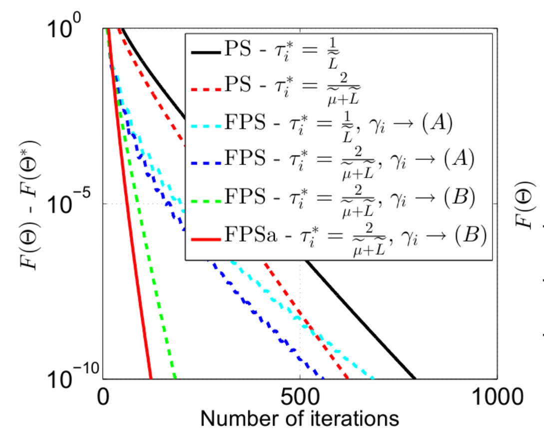


Fig. 1: *Convergence rates*

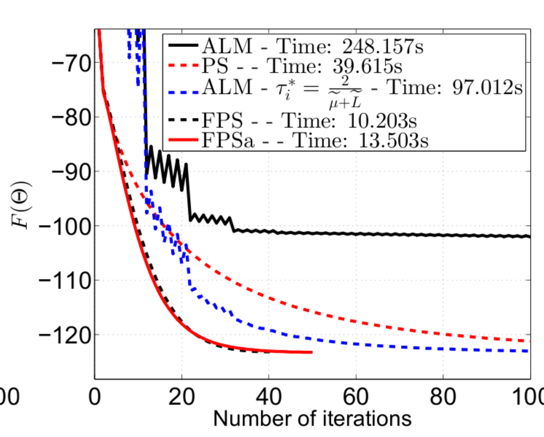


Fig. 2: *Comparison plot*

Setting (i)	ALM	PS	FPS	FPSa
$\frac{\|\boldsymbol{\Theta}^* - \boldsymbol{\Sigma}^{-1}\|_F}{\|\boldsymbol{\Sigma}^{-1}\|_F}$	0.44	0.414	**0.413**	**0.413**
Correct	1705	**1893**	**1893**	**1893**
Missed	291	**103**	**103**	**103**
Extra	365	232	**228**	**228**
Iterations	400	379	129	**114**
#Inversions	400	379	129	**114**
Setting (ii)	**ALM**	**PS**	**FPS**	**FPSa**
$\frac{\|\boldsymbol{\Theta}^* - \boldsymbol{\Sigma}^{-1}\|_F}{\|\boldsymbol{\Sigma}^{-1}\|_F}$	-	0.444	**0.43**	**0.43**
Correct	-	8710	**8725**	8724
Missed	-	290	**275**	276
Extra	-	**4**	**4**	**4**
Iterations	-	300	100	**92**
#Inversions	-	300	100	**92**