Speeding Up Latent Variable Gaussian Graphical Model Estimation via Nonconvex Optimization

Pan Xu* and **Jian Ma†** and **Quanquan Gu***

* University of Virginia † Carnegie Mellon University

Introduction

Latent Variable Gaussian Graphical Model (LVGGM):

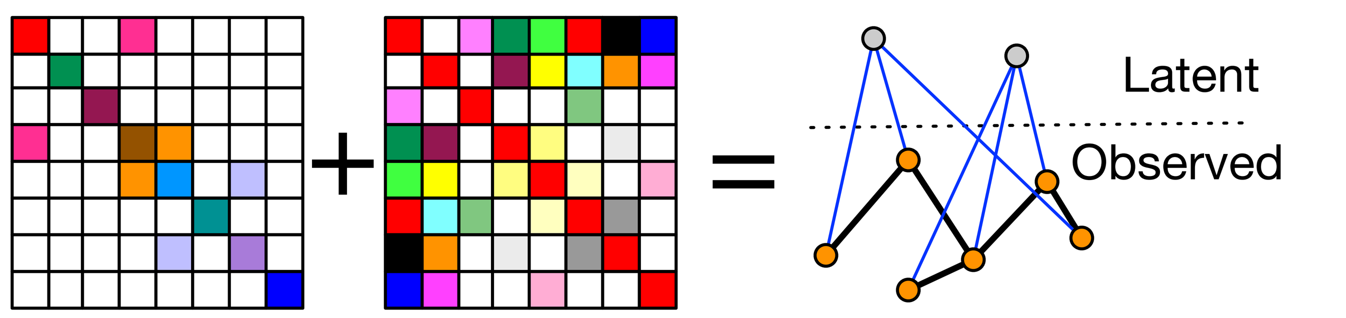
$\boldsymbol{X}_O \in \mathbb{R}^d$ is the observed variables and $\boldsymbol{X}_L \in \mathbb{R}^r$ the latent variables. $\boldsymbol{X} = (\boldsymbol{X}_O^\top, \boldsymbol{X}_L^\top)^\top \sim N(\boldsymbol{\mu}, \widetilde{\boldsymbol{\Sigma}})$ and sparse precision matrix $\widetilde{\boldsymbol{\Omega}} = \widetilde{\boldsymbol{\Sigma}}^{-1}$. Then $\boldsymbol{X}_O$ follows a normal distribution with marginal covariance matrix $\boldsymbol{\Sigma}^* = \widetilde{\boldsymbol{\Sigma}}_{OO}$ being the top-left block matrix in $\widetilde{\boldsymbol{\Sigma}}$. By Schur complement

$$\boldsymbol{\Omega}^* = (\widetilde{\boldsymbol{\Sigma}}_{OO})^{-1} = \widetilde{\boldsymbol{\Omega}}_{OO} - \widetilde{\boldsymbol{\Omega}}_{OL}\widetilde{\boldsymbol{\Omega}}_{LL}^{-1}\widetilde{\boldsymbol{\Omega}}_{LO}.$$

Let $\mathbf{S}^* := \widetilde{\boldsymbol{\Omega}}_{OO}$ and $\mathbf{L}^* := -\widetilde{\boldsymbol{\Omega}}_{OL}\widetilde{\boldsymbol{\Omega}}_{LL}^{-1}\widetilde{\boldsymbol{\Omega}}_{LO}$. Thus, the precision matrix of LVGGM can be written as

$$\boldsymbol{\Omega}^* = \mathbf{S}^* + \mathbf{L}^*,$$

where $\|\mathbf{S}^*\|_{0,0} = s^*$ and $\mathrm{rank}(\mathbf{L}^*) = r$.



Sparse plus Low-rank Matrix Gene Regulatory Network

The Proposed Estimator

Suppose that we observe i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $N(\mathbf{0}, \boldsymbol{\Sigma}^*)$.

- the negative log-likelihood function

$$p_n(\mathbf{S}, \mathbf{L}) = \mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}(\mathbf{S} + \mathbf{L})\right] - \log|\mathbf{S} + \mathbf{L}|,$$

  where $\widehat{\boldsymbol{\Sigma}} = 1/n\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$ is the sample covariance matrix, and $|\mathbf{S} + \mathbf{L}|$ is the determinant of $\boldsymbol{\Omega} = \mathbf{S} + \mathbf{L}$.
- Due to the symmetry and low-rankness of $\mathbf{L}$, we reparameterize it as $\mathbf{L} = \mathbf{Z}\mathbf{Z}^\top$, where $\mathbf{Z} \in \mathbb{R}^{d \times r}$ and $r > 0$ is the number of latent variables.

Estimator: we propose a nonconvex estimator using sparsity constrained maximum likelihood:

$$\min_{\mathbf{S},\mathbf{Z}} \quad q_n(\mathbf{S}, \mathbf{Z}) = \mathrm{tr}\left[\widehat{\boldsymbol{\Sigma}}(\mathbf{S} + \mathbf{Z}\mathbf{Z}^\top)\right] - \log|\mathbf{S} + \mathbf{Z}\mathbf{Z}^\top|, \quad \text{s.t. } \|\mathbf{S}\|_{0,0} \le s,$$

where $s > 0$ is a tuning parameter that controls the sparsity of $\mathbf{S}$.

The Proposed Algorithm

We present the proposed algorithm here, which consists of two stages: initialization and alternating gradient descent.

Algorithm 1 Alternating Thresholded Gradient Descent (AltGD) for LVGGM

Input: i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, max number of iterations T, and parameters η, η', r, s.

Stage I: Initialization

$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top$.

$\widehat{\mathbf{S}}^{(0)} = \mathcal{HT}_s(\widehat{\boldsymbol{\Sigma}}^{-1})$, which preserves the s largest magnitudes of $\widehat{\boldsymbol{\Sigma}}^{-1}$.

Compute SVD: $\widehat{\boldsymbol{\Sigma}}^{-1} - \widehat{\mathbf{S}}^{(0)} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, where $\mathbf{D}$ is a diagonal matrix. Let $\widehat{\mathbf{Z}}^{(0)} = \mathbf{U}\mathbf{D}_r^{1/2}$, where $\mathbf{D}_r$ is the first r columns of $\mathbf{D}$.

Stage II: Alternating Gradient Descent

for $t = 0, \ldots, T-1$ **do**

$\quad\widehat{\mathbf{S}}^{(t+0.5)} = \widehat{\mathbf{S}}^{(t)} - \eta\nabla_{\mathbf{S}}q_n\big(\widehat{\mathbf{S}}^{(t)}, \widehat{\mathbf{Z}}^{(t)}\big)$;

$\quad\widehat{\mathbf{S}}^{(t+1)} = \mathcal{HT}_s\Big(\widehat{\mathbf{S}}^{(t+0.5)}\Big)$, which preserves the s largest magnitudes of $\widehat{\mathbf{S}}^{(t+0.5)}$;

$\quad\widehat{\mathbf{Z}}^{(t+1)} = \widehat{\mathbf{Z}}^{(t)} - \eta'\nabla_{\mathbf{Z}}q_n\big(\widehat{\mathbf{S}}^{(t)}, \widehat{\mathbf{Z}}^{(t)}\big)$;

end for

output: $\widehat{\mathbf{S}}^{(T)}, \widehat{\mathbf{Z}}^{(T)}$.

Theoretical Analysis

- Assumptions
  - **A1 (Bounded Eigenvalues):** $\exists\, \nu > 0$ such that the eigenvalues of $\boldsymbol{\Sigma}^*$ are bounded, i.e., $0 < 1/\nu \le \lambda_{\min}(\boldsymbol{\Sigma}^*) \le \lambda_{\max}(\boldsymbol{\Sigma}^*) \le \nu < \infty$.
  - **A2 (Spikiness Condition):** the *spikiness ratio* is defined as $\alpha_{sp}(\mathbf{L}) := d\|\mathbf{L}\|_{\infty,\infty}/\|\mathbf{L}\|_F$. We assume $\exists\, \alpha^* > 0$ such that

$$\|\mathbf{L}^*\|_{\infty,\infty} = \frac{\alpha_{sp}(\mathbf{L}^*) \cdot \|\mathbf{L}^*\|_F}{d} \le \frac{\alpha^*}{d}.$$

  - **FOS (First-Order Stability):** If $\max\{\|\mathbf{S} - \mathbf{S}^*\|_F, d(\mathbf{Z}, \mathbf{Z}^*)\} \le R$ for some $R > 0$ and $\mathbf{L} = \mathbf{Z}\mathbf{Z}^\top$ and $\mathbf{L}^* = \mathbf{Z}^*\mathbf{Z}^{*\top}$. It holds that

$$\big\|\nabla_{\mathbf{S}}p(\mathbf{S}, \mathbf{L}) - \nabla_{\mathbf{S}}p(\mathbf{S}, \mathbf{L}^*)\big\|_F \le \gamma_2 \cdot \|\mathbf{L} - \mathbf{L}^*\|_F,$$
$$\big\|\nabla_{\mathbf{L}}p(\mathbf{S}, \mathbf{L}) - \nabla_{\mathbf{L}}p(\mathbf{S}^*, \mathbf{L})\big\|_F \le \gamma_1 \cdot \|\mathbf{S} - \mathbf{S}^*\|_F,$$

  where γ_1, γ_2 are constants and $d(\mathbf{Z}, \mathbf{Z}^*) = \min_{\mathbf{U} \in \mathbb{O}(r^*)}\|\mathbf{Z} - \mathbf{Z}^*\mathbf{U}\|_F$.
- **Main Theory**

Validation of Initialization: Suppose **A1** and **A2** hold. Assume $n \ge c\nu^2 r s^* \log d/R^2$ and $s^* \le c'd^2R^2/(r\alpha^{*2})$, where R is a constant depending on ν. Then with probability at least $1 - C/d$, we have

$$\boxed{\big\|\widehat{\mathbf{S}}^{(0)} - \mathbf{S}^*\big\|_F \le R, \quad \text{and} \quad d(\widehat{\mathbf{Z}}^{(0)}, \mathbf{Z}^*) \le R,}$$

where $C > 0$ is an absolute constant.

Convergence Rate: Furthermore, suppose **FOS** holds. Let the step sizes $\eta \le C_0/\nu^2$ and $\eta' \le C_0/\nu^4$, and the sparsity parameter satisfy $s \ge \big(4(1/(2\sqrt{\rho}) - 1)^2 + 1\big)s^*$. Let ρ and τ be

$$\rho = \max\left\{1 - \frac{\eta}{\nu^2}, 1 - \frac{\eta'}{\nu^2}\right\}, \qquad \tau = \max\left\{\frac{cs^*\log d}{\nu^4 n}, \frac{crd}{\nu^6 n}\right\}.$$

Then for any $t \ge 1$, with probability at least $1 - C_1/d$, we have

$$\boxed{\max\left\{\big\|\widehat{\mathbf{S}}^{(t+1)} - \mathbf{S}^*\big\|_F^2, d^2(\widehat{\mathbf{Z}}^{(t+1)}, \mathbf{Z}^*)\right\} \le \underbrace{\frac{\tau}{1 - \sqrt{\rho}}}_{\text{statistical error}} + \underbrace{\sqrt{\rho^{t+1}} \cdot R}_{\text{optimization error}},}$$

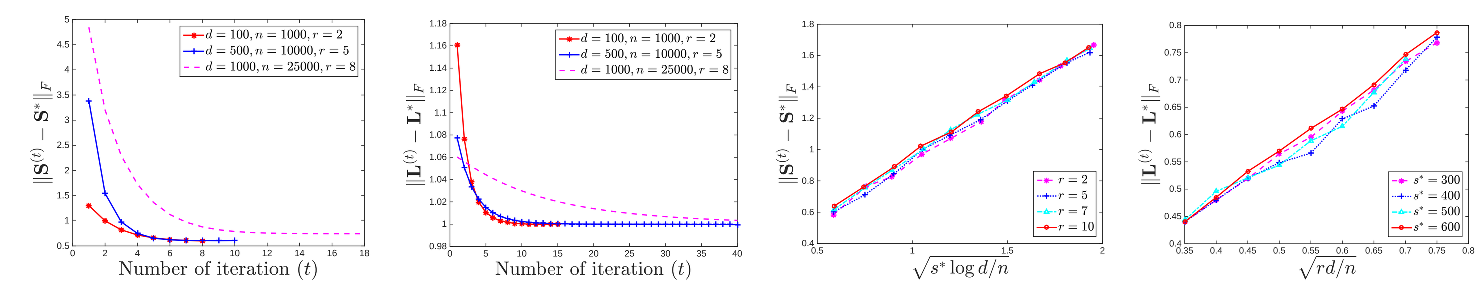
where $C_1 > 0$ is an absolute constant.

Main Remarks

- The initial points returned by the initialization stage of AltGD fall in small neighborhoods of $\mathbf{S}^*$ and $\mathbf{Z}^*$ if $n = O(s^*\log d)$, which essentially attains the optimal sample complexity for LVGGM estimation. In addition, we require $s^* \lesssim d^2/(r\alpha^{*2})$, which means the unknown sparse matrix cannot be too dense.
- The statistical error scales as $\max\left\{O_p\big(\sqrt{s^*\log d/n}\big), O_p\big(\sqrt{rd/n}\big)\right\}$, where $O_p(\sqrt{s^*\log d/n})$ corresponds to the statistical error of $\mathbf{S}^*$, and $O_p(\sqrt{rd/n})$ corresponds to that of $\mathbf{L}^*$ (or equivalently $\mathbf{Z}^*$). This matches the minimax optimal rate of estimation errors in Frobenius norm for LVGGM estimation.
- AltGD enjoys linear convergence rate for optimization error. After $T \ge \max\{O(\log(\nu^4 n/(s^*\log d))), O(\log(\nu^6 n/(rd)))\}$ iterations, the total estimation error achieves the same order as the statistical error.

Numerical Simulations

- **Data Generation:** We randomly generated a sparse positive definite matrix $\widetilde{\boldsymbol{\Omega}} \in \mathbb{R}^{(d+r)\times(d+r)}$, with sparsity $s^* = 0.02d^2$. Set $\mathbf{S}^* := \widetilde{\boldsymbol{\Omega}}_{1:d;1:d}$ and $\mathbf{L}^* := -\widetilde{\boldsymbol{\Omega}}_{1:d;(d+1):(d+r)}[\widetilde{\boldsymbol{\Omega}}_{(d+1):(d+r);(d+1):(d+r)}]^{-1}\widetilde{\boldsymbol{\Omega}}_{(d+1):(d+r);1:d}$. Then we sampled $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim N(\mathbf{0}, (\boldsymbol{\Omega}^*)^{-1})$, where $\boldsymbol{\Omega}^* = \mathbf{S}^* + \mathbf{L}^*$.
- Validation of Convergence Rate:



(a) Estimation error for $\mathbf{S}^*$ (b) Estimation error for $\mathbf{L}^*$ (c) r fixed and varying n, d and s^* (d) s^* fixed and varying n, d and r

Figure: (a)-(b): Evolution of estimation errors with number of iterations t going up with $s^* = 0.02d^2$ and varying d, n and r. (c)-(d): Estimation errors $\|\widehat{\mathbf{S}}^{(T)} - \mathbf{S}^*\|_F$ and $\|\widehat{\mathbf{L}}^{(T)} - \mathbf{L}^*\|_F$ versus scaled statistical errors $\sqrt{s^*\log d/n}$ and $\sqrt{rd/n}$.

- Comparisons with Convex Methods: **AltGD** is nearly 50 times faster than the other two methods based on convex algorithms.

Table: Estimation errors in terms of Frobenius norm on different synthetic datasets. Results were reported on 10 replicates in each setting.

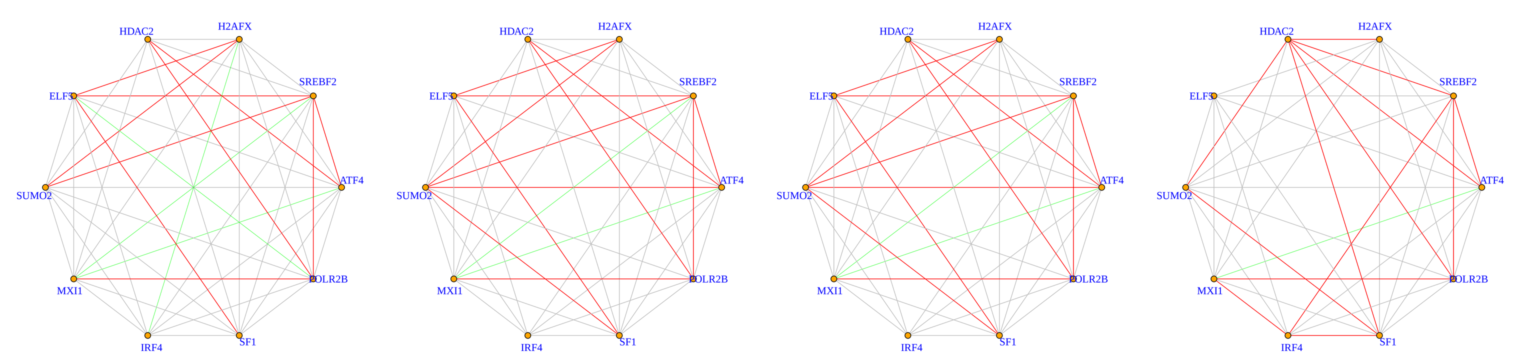
Setting	Method	$\|\widehat{\mathbf{S}}^{(T)} - \mathbf{S}^*\|_F$	$\|\widehat{\mathbf{L}}^{(T)} - \mathbf{L}^*\|_F$	$\|\widehat{\boldsymbol{\Omega}}^{(T)} - \boldsymbol{\Omega}^*\|_F$	Time (s)
$d = 100, r = 2, n = 2000$	PPA	0.7335±0.0352	0.0170±0.0125	0.7350±0.0359	1.1610
	ADMM	0.7521±0.0288	0.0224±0.0115	0.7563±0.0298	1.1120
	AltGD	0.6241±0.0668	0.0113±0.0014	0.6236±0.0669	0.0250
$d = 500, r = 5, n = 10000$	PPA	0.9803±0.0192	0.0195±0.0046	0.9813±0.0192	35.7220
	ADMM	1.0571±0.0135	0.0294±0.0041	1.0610±0.0134	25.8010
	AltGD	0.8212±0.0143	0.0125±0.0000	0.8210±0.0143	0.4800
$d = 1000, r = 8, n = 2.5 \times 10^4$	PPA	1.1620±0.0177	0.0224±0.0034	1.1639±0.0179	356.7360
	ADMM	1.1867±0.0253	0.0356±0.0033	1.1869±0.0254	156.5550
	AltGD	0.9016±0.0245	0.0167±0.0030	0.9021±0.0244	7.4740
$d = 5000, r = 10, n = 2 \times 10^5$	PPA	1.4822±0.0302	0.0371±0.0052	1.4824±0.0120	33522.0200
	ADMM	1.5010±0.0240	0.0442±0.0068	1.5012±0.0240	21090.7900
	AltGD	1.3449±0.0073	0.0208±0.0014	1.3449±0.0084	445.6730

Experiments on Genomic Datasets

Experiments on TCGA breast cancer gene expression data ($n = 601$ samples and $d = 299$ TFs) to infer the regulatory network. Methods based on LVGGMs are able to recover more edges accurately than graphical Lasso because of the intervention of latent variables. AltGD runs much faster than the convex methods.

Table: Summary of CPU time on luminal subtype breast cancer dataset.

Method	GLasso	PPA	ADMM	AltGD
Time (s)	38.63	85.01	7.67	0.15



(a) GLasso (b) PPA (c) ADMM (d) AltGD

Figure: An example of subnetwork in the transcriptional regulatory network of luminal breast cancer. Gray edges are the interactions from the Cistrome Cancer Database; red edges are the ones inferred by the respective methods; green edges are incorrectly inferred interactions.