HDS28 - Graphical models in exponential form and with corrupted or hidden variables

Yangjianchen Xu

Department of Biostatistics
University of North Carolina at Chapel Hill

10/01/2021

Overview

- Graphical models in exponential form
  - A general form of neighbourhood regression
  - Graph selection for Ising model
- Graphs with corrupted or hidden variables
  - Gaussian graph estimation with corrupted data
  - Gaussian graph selection with hidden variables

Graphical models in exponential form

- Consider the graph estimation problem for a more general class of graphical model in the exponential form:

$$p_{\Theta^*}\left(x_1, \ldots, x_d\right) \propto \exp\left\{\sum_{j \in V} \phi_j\left(x_j; \Theta_j^*\right) + \sum_{(j,k) \in E} \phi_{jk}\left(x_j, x_k; \Theta_{jk}^*\right)\right\}.$$

- For instance, the Gaussian graphical model is a special case where $\Theta_j^* = \theta_j^*$ and $\Theta_{jk}^* = \theta_{jk}^*$ with potential functions

$$\phi_j\left(x_j; \theta_j^*\right) = \theta_j^* x_j, \quad \phi_{jk}\left(x_j, x_k; \theta_{jk}^*\right) = \theta_{jk}^* x_j x_k.$$

- Ising model: take values in the binary hypercube $\{0, 1\}^d$.

Example: Potts model

- Each variable X_s takes values in the discrete set $\{0, \ldots, M-1\}$.
- Factorization form: $\Theta_j^* = \{\Theta_{j;a}, a = 1, \ldots, M-1\}$ is an $(M-1)$-vector, $\Theta_{jk}^* = \left\{\Theta_{jk;ab}^*, a, b = 1, \ldots, M-1\right\}$ is an $(M-1) \times (M-1)$ matrix.
- The potential functions are

$$\phi_j\left(x_j; \Theta_j^*\right) = \sum_{a=1}^{M-1} \Theta_{j;a}^* I\left[x_j = a\right]$$

and

$$\phi_{jk}\left(x_j, x_k; \Theta_{jk}^*\right) = \sum_{a=1}^{M-1} \sum_{b=1}^{M-1} \Theta_{jk,ab}^* I\left[x_j = a, x_k = b\right]$$

- Generalization of the Ising model.

Example: Poisson graphical model

- Model $(X_1, \ldots, X_d)$ with count values $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$.
- To build a graphical model, specify the conditional distribution of each variable given its neighbors.
- Suppose variable X_j, conditioned on its neighbors, is a Poisson random variable with mean

$$\mu_j = \exp\left(\theta_j^* + \sum_{k \in \mathcal{N}(j)} \theta_{jk}^* x_k\right)$$

- Lead to a Markov radom field of the exponential form with

$$\begin{aligned} \phi_j\left(x_j; \theta_j^*\right) &= \theta_j^* x_j - \log(x!) && \text{for all } j \in V \\ \phi_{jk}\left(x_j, x_k; \theta_{jk}^*\right) &= \theta_{jk}^* x_j x_k && \text{for all } (j,k) \in E \end{aligned}$$

- In order for the density to be normalizable, require $\theta_{jk}^* \le 0$ for all $(j,k) \in E$. The model can only capture competitive interactions between variables.

A general form of neighborhood regression

- Consider the conditional likelihood of $X_j \in \mathbb{R}^n$ given $\mathrm{X}_{\setminus\{j\}} \in \mathbb{R}^{n\times(d-1)}$, which only depends on

$$\Theta_{j+} := \{\Theta_j, \Theta_{jk}, k \in V\setminus\{j\}\}$$

- Observation: in the true model Θ^*, we have $\Theta^*_{jk} = 0$ whenever $(j,k) \notin E$.
- Impose some type of block-based sparsity penalty on Θ_{j+}.
- General form of neighborhood regression:

$$\widehat{\Theta}_{j+} = \arg\min_{\Theta_{j+}} \left\{ \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \log p_{\Theta_{j+}}\left(x_{ij} \mid x_{i\setminus\{j\}}\right)}_{\mathcal{L}_n\left(\Theta_{j+}; x_j, x_{\setminus\{j\}}\right)} + \lambda_n \sum_{k\in V\setminus\langle j\}} |||\Theta_{jk}||| \right\}$$

- Frobenius norm, a general form of the group Lasso.

Graph selection for Ising models

- Recall the Ising distribution is over binary variables

$$p_{\theta^*}(x_1, \ldots, x_d) \propto \exp\left\{\sum_{j\in V} \theta_j^* x_j + \sum_{(j,k)\in E} \theta_{jk}^* x_j x_k\right\}$$

- For any node $j \in V$, define

$$\theta_{j+} := \{\theta_j, \theta_{jk}, k \in V\backslash\{j\}\}$$

- The neighborhood regression reduced to a form of logistic regression

$$\widehat{\theta}_{j+} = \arg\min_{\theta_{j+}\in\mathbb{R}^d} \left\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} f\left(\theta_j x_{ij} + \sum_{k\in V\backslash\langle j\rangle} \theta_{jk} x_{ij} x_{ik}\right)}_{\mathcal{L}_n(\theta_{j+}; x_j, x_{\backslash\{j\}})} + \lambda_n \sum_{k\in V\backslash\langle j\rangle} |\theta_{jk}| \right\},$$

where $f(t) = \log(1 + e^t)$ is the logistic function.

Conditions for the consistency under Ising models

- Under what conditions does the estimate recover the correct neighborhood set $\mathcal{N}(j)$? Limit the influence of irrelevant variables–those outside $\mathcal{N}(j)$–on variables inside the set.
- Let θ^*_{j+} be the minimizer of the population objective function $\overline{\mathcal{L}}\left(\theta_{j+}\right) = \mathbb{E}\left[\mathcal{L}_n\left(\theta_{j+}; X_j, X_{\backslash\{j\}}\right)\right]$.
- Hessian of the cost function $\mathrm{J} := \nabla^2 \overline{\mathcal{L}}(\theta^*_{j+})$.
- J satisfies an α-incoherence condition at $j \in V$ if

$$\max_{k \notin S} \left\| J_{kS}\left(\mathrm{J}_{SS}\right)^{-1} \right\|_1 \leqslant 1 - \alpha$$

- The submatrix J_{SS} has its smallest eigenvalue lower bounded by some $c_{\min} > 0$.
- A graph G with d vertices and maximum degree at most m.

Consistency for Ising models

Theorem (11.15)

Given n i.i.d. samples with $n > c_0 m^2 \log d$, consider the estimator in the above neighborhood regression with $\lambda_n = \frac{32}{\alpha}\sqrt{\frac{\log d}{n}} + \delta$ for some $\delta \in [0, 1]$. Then with probability at least $1 - c_1 e^{-c_2\left(n\delta^2 + \log d\right)}$, the estimate $\widehat{\theta}_{j+}$ has the following properties:

(a) *It has a support $\widehat{S} = \text{supp}(\widehat{\theta})$ that is contained within the neighborhood set $\mathcal{N}(j)$.*

(b) *It satisfies the ℓ_∞-bound $\|\widehat{\theta}_{j+} - \theta^*_{j+}\|_\infty \leq \frac{c_3}{c_{\min}}\sqrt{m}\lambda_n$.*

- Part (a) guarantees that the method leads to no false inclusions.
- The ℓ_∞-bound in part (b) ensures that the method picks up all significant variables.
- The proof is based on the same type of primal–dual witness construction used in the proof of Theorem 11.12.

Graphs with corrupted or hidden variables

- Thus far, we have assumed that the samples $\{x_i\}_{i=1}^n$ are observed perfectly.
- This idealized setting can be violated in a number of ways:
  - The samples may be corrupted by some type of measurement noise, or certain entries may be missing.
  - In the most extreme case, some subset of the variables are never observed, and so are known as hidden or latent variables.
- We focus primarily on the Gaussian case for simplicity.

Gaussian graph estimation with corrupted data

- Suppose we observe $Z = X + V$, where the matrix V represents some type of measurement error.
- The naive approach (graphical Lasso) would be to solve the convex program:

$$\widehat{\Theta}_{\mathrm{NAI}} = \arg \min_{\Theta \in \mathcal{S}^{d \times d}} \left\{ \langle\langle \Theta, \widehat{\Sigma}_z \rangle\rangle - \log \det \Theta + \lambda_n |||\Theta|||_{1,\mathrm{off}} \right\},$$

  where $\widehat{\Sigma}_z = \frac{1}{n} Z^{\mathrm{T}} Z = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^{\mathrm{T}}$ is now the sample covariance based on the observed data matrix Z.
- Exercise 11.8: the addition of noise does not preserve Markov properties, so that the estimate $\widehat{\Theta}_{\mathrm{NAI}}$ will not lead to consistent estimates of either the edge set, or the underlying precision matrix Θ^*.
- We need to replace $\widehat{\Sigma}_z$ with an unbiased estimator of $\mathrm{cov}(x)$ based on the observed data matrix Z.

Unbiased covariance estimate for additive corruptions

- Suppose that each row v_i of the noise matrix V is drawn i.i.d. from a zero-mean distribution with covariance Σ_v.
- In this case, a natural estimate of $\Sigma_x := \mathrm{cov}(x)$ is given by

$$\widehat{\Gamma} := \frac{1}{n} Z^{\mathrm{T}} Z - \Sigma_v$$

- $\widehat{\Gamma}$ is an unbiased estimate of Σ_x as long as the noise matrix V is independent of X.
- When both X and V have sub-Gaussian rows, a deviation condition of the form $\|\widehat{\Gamma} - \Sigma_x\|_{\max} \precsim \sqrt{\frac{\log d}{n}}$ holds with high probability. (Exercise 11.12)

Missing data

- Some entries of the data matrix X might be missing.
- In the simplest model–missing completely at random (MCAR)–entry (i, j) of the data matrix is missing with some probability $v \in [0, 1)$.
- We can construct a new matrix $\widetilde{Z} \in \mathbb{R}^{n \times d}$ with entries

$$\widetilde{Z}_{ij} = \begin{cases} \frac{Z_{ij}}{1-v} & \text{if entry } (i,j) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

- With this choice, it can be verified that

$$\widehat{\Gamma} = \frac{1}{n}\widetilde{Z}^{\mathrm{T}}\widetilde{Z} - v\,\mathrm{diag}(\widetilde{Z}^{\mathrm{T}}\widetilde{Z}/n)$$

is an unbiased estimate of the covariance matrix $\Sigma_x = \mathrm{cov}(x)$.

- Under suitable tail conditions, it also satisfies the deviation condition $\|\widehat{\Gamma} - \Sigma_x\|_{\max} \lesssim \sqrt{\frac{\log d}{n}}$ with high probability. (Exercise 11.13)

Correcting the Gaussian graphical Lasso

- Any unbiased estimate $\widehat{\Gamma}$ of Σ_x defines a form of the corrected graphical Lasso estimator

$$\widetilde{\Theta} = \arg \min_{\Theta \in \mathcal{S}_+^{d \times d}} \left\{ \langle\langle \Theta, \widehat{\Gamma} \rangle\rangle - \log\det \Theta + \lambda_n |||\Theta|||_{1,\text{off}} \right\}$$

- Depending on the nature of the covariance estimate $\widehat{\Gamma}$, the program may not have any solution.
- Exercise 11.9: as long as $\lambda_n > \|\widehat{\Gamma} - \Sigma_x\|_{\max}$, this optimization problem has a unique optimum that is achieved.
- Moreover, by inspecting the proofs of the claims in Section 11.2.1, it can be seen that the estimator $\widetilde{\Theta}$ obeys similar Frobenius norm and edge selection bounds as the usual graphical Lasso.

Correcting neighborhood regression

- Use X to denote the $n \times (d-1)$ matrix with $\{X_k, k \in V \backslash \{j\}\}$ as its columns, and use $y = X_j$ to denote the response vector.
- With this notation, we have an instance of a corrupted linear regression model:

$$y = \mathrm{X}\theta^* + w \quad \text{and} \quad \mathrm{Z} \sim \mathbb{Q}(\cdot \mid \mathrm{X}),$$

  where the conditional probability distribution $\mathbb{Q}$ varies according to the nature of the corruption.
- The response vector y might also be further corrupted, but this case can often be reduced to an instance of the previous model.

Correcting neighborhood regression

- The naive approach would be simply to solve a least-squares problem involving the cost function $\frac{1}{2n}\|y - Z\theta\|_2^2$.
- Exercise 11.10: doing so will lead to an inconsistent estimate of the neighborhood regression vector θ^*.
- Question: what types of quantities need to be "corrected" in order to obtain a consistent form of linear regression?
- Consider the following population-level objective function

$$\overline{\mathcal{L}}(\theta) = \frac{1}{2}\theta^{\mathrm{T}}\Gamma\theta - \langle \theta, \gamma \rangle,$$

where $\Gamma := \mathrm{cov}(x)$ and $\gamma := \mathrm{cov}(x, y)$. By construction, the true regression vector is the unique global minimizer of $\overline{\mathcal{L}}$.

Correcting neighborhood regression

- Thus, a natural strategy is to solve a penalized regression problem in which the pair (γ, Γ) are replaced by data-dependent estimates $(\widehat{\gamma}, \widehat{\Gamma})$.
- Doing so leads to the empirical objective function

$$\mathcal{L}_n(\theta) = \frac{1}{2}\theta^{\mathrm{T}}\widehat{\Gamma}\theta - \langle \theta, \widehat{\gamma} \rangle,$$

  where the estimates $(\widehat{\gamma}, \widehat{\Gamma})$ must be based on the observed data (y, Z).
- We are led to study the following corrected Lasso estimator

$$\min_{\|\theta\|_1 \leq \sqrt{\frac{n}{\log d}}} \left\{ \frac{1}{2}\theta^{\mathrm{T}}\widehat{\Gamma}\theta - \langle \widehat{\gamma}, \theta \rangle + \lambda_n \|\theta\|_1 \right\}$$

- In the high-dimensional regime $(n < d)$, the previously described choices of $\widehat{\Gamma}$ given have negative eigenvalues. The constrain $\|\theta\|_1 \leq \sqrt{\frac{n}{\log d}}$ is actually needed when the objective function $\mathcal{L}_n(\theta)$ is non-convex. (Exercise 11.11)

Non-convex problem: local optima

- A local optimum for the previous program is any vector $\widetilde{\theta} \in \mathbb{R}^d$ such that

$$\langle \nabla \mathcal{L}_n(\widetilde{\theta}), \theta - \widetilde{\theta} \rangle \geq 0 \quad \text{for all } \theta \text{ such that } \|\theta\|_1 \leq \sqrt{\frac{n}{\log d}}.$$

- Under suitable conditions, any local optimum is relatively close to the true regression vector.
- Restricted eigenvalue (RE) condition: assume that there exists a constant $\kappa > 0$ such that

$$\langle \Delta, \widehat{\Gamma} \Delta \rangle \geq \kappa \|\Delta\|_2^2 - c_0 \frac{\log d}{n} \|\Delta\|_1^2 \quad \text{for all } \Delta \in \mathbb{R}^d$$

- Assume that the minimizer θ^* of $\bar{\mathcal{L}}(\theta)$ has sparsity s and ℓ_2-norm at most 1, and that $n \geq s \log d$. These assumptions ensure that $\|\theta^*\|_1 \leq \sqrt{s} \leq \sqrt{\frac{n}{\log d}}$, so that θ^* is feasible for the non-convex Lasso.

Theoretical results of local optima

Proposition (11.18)

Under the RE condition, suppose that the pair $(\widehat{\gamma}, \widehat{\Gamma})$ satisfy the deviation condition

$$\|\widehat{\Gamma}\theta^* - \widehat{\gamma}\|_{\max} \leq \varphi(\mathbb{Q}, \sigma_w)\sqrt{\frac{\log d}{n}} \tag{1}$$

for a pre-factor $\varphi(\mathbb{Q}, \sigma_w)$ depending on the conditional distribution $\mathbb{Q}$ and noise standard deviation σ_w. Then for any regularization parameter $\lambda_n \geq 2\left(2c_0 + \varphi(\mathbb{Q}, \sigma_w)\right)\sqrt{\frac{\log d}{n}}$, any local optimum $\widetilde{\theta}$ to the corrected Lasso program satisfies the bound

$$\|\widetilde{\theta} - \theta^*\|_2 \leq \frac{2}{\kappa}\sqrt{s}\lambda_n. \tag{2}$$

Observe that $\nabla\overline{\mathcal{L}}(\theta^*) = \Gamma\theta^* - \gamma = 0$. Condition (1) is the sample-based and approximate equivalent of this optimality condition.

Proof of Proposition 11.18

Proof. We prove this result in the special case when the optimum occurs in the interior of the set $\|\theta\|_1 \leq \sqrt{\frac{n}{\log d}}$. In this case, any local optimum $\widetilde{\theta}$ must satisfy the condition $\nabla \mathcal{L}_n(\widetilde{\theta}) + \lambda_n \widehat{z} = 0$, where $\widehat{z}$ belongs to the subdifferential of the ℓ_1-norm at $\widetilde{\theta}$. Define the error vector $\widehat{\Delta} := \widetilde{\theta} - \theta^*$. Adding and subtracting terms and then taking inner products with $\widehat{\Delta}$ yields the inequality

$$\begin{aligned}
\langle \widehat{\Delta}, \widehat{\Gamma} \widehat{\Delta} \rangle &= \langle \widehat{\Delta}, \nabla \mathcal{L}_n(\theta^* + \widehat{\Delta}) - \nabla \mathcal{L}_n(\theta^*) \rangle \\
&\leq |\langle \widehat{\Delta}, \nabla \mathcal{L}_n(\theta^*) \rangle| - \lambda_n \langle \widehat{z}, \widehat{\Delta} \rangle \\
&\leq \|\widehat{\Delta}\|_1 \|\nabla \mathcal{L}_n(\theta^*)\|_\infty + \lambda_n \{\|\theta^*\|_1 - \|\widetilde{\theta}\|_1\},
\end{aligned}$$

where we have used the facts that $\langle \widehat{z}, \widetilde{\theta} \rangle = \|\widetilde{\theta}\|_1$ and $\langle \widehat{z}, \theta^* \rangle \leq \|\theta^*\|_1$. From the proof of Theorem 7.8, since the vector θ^* is S-sparse, we have

$$\|\theta^*\|_1 - \|\widetilde{\theta}\|_1 \leq \|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1.$$

Proof of Proposition 11.18 (cont.)

Since $\nabla \mathcal{L}_n(\theta) = \widehat{\Gamma}\theta - \widehat{\gamma}$, the deviation condition (1) is equivalent to the bound

$$\|\nabla \mathcal{L}_n(\theta^*)\|_\infty \leq \varphi(\mathbb{Q}, \sigma_w)\sqrt{\frac{\log d}{n}},$$

which is less than $\lambda_n/2$ by our choice of regularization parameter. Consequently, we have

$$\langle \widehat{\Delta}, \widehat{\Gamma}\widehat{\Delta} \rangle \leq \frac{\lambda_n}{2}\|\widehat{\Delta}\|_1 + \lambda_n\{\|\widehat{\Delta}_S\|_1 - \|\widehat{\Delta}_{S^c}\|_1\} = \frac{3}{2}\lambda_n\|\widehat{\Delta}_S\|_1 - \frac{1}{2}\lambda_n\|\widehat{\Delta}_{S^c}\|_1 \tag{3}$$

Since θ^* is s-sparse, we have $\|\theta^*\|_1 \leq \sqrt{s}\,\|\theta^*\|_2 \leq \sqrt{\frac{n}{\log d}}$, where the final inequality follows from the assumption that $n \geq s \log d$. Consequently, we have

$$\|\widehat{\Delta}\|_1 \leq \|\widehat{\theta}\|_1 + \|\theta^*\|_1 \leq 2\sqrt{\frac{n}{\log d}}$$

Combined with the RE condition, we have

$$\langle \widehat{\Delta}, \widehat{\Gamma}\widehat{\Delta} \rangle \geq \kappa \|\widehat{\Delta}\|_2^2 - c_0 \frac{\log d}{n} \|\widehat{\Delta}\|_1^2 \geq \kappa \|\widehat{\Delta}\|_2^2 - 2c_0 \sqrt{\frac{\log d}{n}} \|\widehat{\Delta}\|_1$$

Recombining with our earlier bound (3), we have

$$\begin{aligned}
\kappa \|\widehat{\Delta}\|_2^2 &\leq 2c_0 \sqrt{\frac{\log d}{n}} \|\widehat{\Delta}\|_1 + \frac{3}{2}\lambda_n \|\widehat{\Delta}_S\|_1 - \frac{1}{2}\lambda_n \|\widehat{\Delta}_S\|_1 \\
&\leq \frac{1}{2}\lambda_n \|\widehat{\Delta}\|_1 + \frac{3}{2}\lambda_n \|\widehat{\Delta}_S\|_1 - \frac{1}{2}\lambda_n \|\widehat{\Delta}_{S^c}\|_1 \\
&= 2\lambda_n \|\widehat{\Delta}_S\|_1
\end{aligned}$$

Since $\|\widehat{\Delta}_S\|_1 \leq \sqrt{s}\|\widehat{\Delta}\|_2$, the claim follows.

Gaussian graph selection with hidden variables

- In certain settings, a given set of random variables might not be accurately described using a sparse graphical model on their own, but can be when augmented with an additional set of hidden variables.
- For instance, the random variables $X_1 =$ Shoe size and $X_2 =$ Gray hair are likely to be dependent: few children have gray hair.
- However, it might be reasonable to model them as being conditionally independent given a third variable-namely $X_3 =$ Age.
- Consider a family of $d + r$ random variables $X := (X_1, \ldots, X_d, X_{d+1}, \ldots, X_{d+r})$ and suppose that this full vector can be modeled by a sparse graphical model with $d + r$ vertices.
    - Observed variables: the subvector $X_{\mathrm{O}} := (X_1, \ldots, X_d)$
    - Hidden variables: $X_{\mathrm{H}} := (X_{d+1}, \ldots, X_{d+r})$
- Given this partial information, our goal is to recover useful information about the underlying graph.

Matrix-theoretic formulation for the Gaussian case

- Let Σ^*_{OO} denote the covariance matrix of X_{o}. Θ° is the inverse covariance matrix of the full vector $X = (X_{\mathrm{O}}, X_{\mathrm{H}})$, which can be written in the block-partitioned form

$$\Theta^\circ = \begin{bmatrix} \Theta^\circ_{\mathrm{OO}} & \Theta^\circ_{\mathrm{OH}} \\ \Theta^\circ_{\mathrm{HO}} & \Theta^\circ_{\mathrm{HH}} \end{bmatrix}$$

- By the block-matrix inversion formula,

$$(\Sigma^*_{\mathrm{OO}})^{-1} = \underbrace{\Theta^\diamond_{\mathrm{OO}}}_{\Gamma^*} - \underbrace{\Theta^\diamond_{\mathrm{OH}} (\Theta^\diamond_{\mathrm{HH}})^{-1} \Theta^\diamond_{\mathrm{HO}}}_{\Lambda^*}.$$

- By our modeling assumptions, the matrix $\Gamma^* := \Theta^\circ_{\mathrm{OO}}$ is sparse and $\Lambda^* := \Theta^\circ_{\mathrm{OH}} (\Theta^\circ_{\mathrm{HH}})^{-1} \Theta^\circ_{\mathrm{HO}}$ has rank at most $\min\{r, d\}$.
- If r is substantially less than d, the inverse covariance matrix can be decomposed as the sum of a sparse and a low-rank matrix.

Matrix-theoretic formulation for the Gaussian case

- Suppose $x_i \in \mathbb{R}^d$ $(i = 1, \ldots, n)$ are i.i.d. samples from a zero-mean Gaussian with covariance Σ^*_{OO}. We require $n > d$ due to the absence of any sparsity in the low-rank component.
- When $n > d$, the sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ will be invertible with high probability, and hence setting $\mathrm{Y} := (\widehat{\Sigma})^{-1}$, we can consider an observation model of the form

$$\mathrm{Y} = \Gamma^* - \Lambda^* + \mathrm{W}$$

Here $\mathrm{W} \in \mathbb{R}^{d \times d}$ is a stochastic noise matrix.

- A very simple two-step estimator:

$$\widehat{\Gamma} := T_{v_n}((\widehat{\Sigma})^{-1}) \quad \text{and} \quad \widehat{\Lambda} := \widehat{\Gamma} - (\widehat{\Sigma})^{-1},$$

where the hard-thresholding operator is given by $T_{v_n}(v) = vI[|v| > v_n]$ and $v_n > 0$ to be chosen.

Assumptions and choice of v_n

- As with our earlier study of matrix decompositions in Section 10.7, we assume here that the low-rank component satisfies a "spikiness" constraint: $\|\Lambda^*\|_{\max} \leq \frac{\alpha}{d}$.
- In addition, we assume that the matrix square root of the true precision matrix $\Theta^* = \Gamma^* - \Lambda^*$ has a bounded ℓ_∞-operator norm:

$$|||\sqrt{\Theta^*}|||_\infty = \max_{j=1,\ldots,d} \sum_{k=1}^{d} |\sqrt{\Theta^*}|_{jk} \leq \sqrt{M}$$

- In terms of the parameters (α, M), we then choose the threshold parameter v_n in our estimates as

$$v_n := M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{\alpha}{d} \quad \text{for some } \delta \in [0, 1]$$

Theoretical results of the two-step estimator

Proposition (11.19)

Consider a precision matrix Θ^* *that can be decomposed as the difference* $\Gamma^* - \Lambda^*$, *where* Γ^* *has most* s *non-zero entries per row, and* Λ^* *is* α*-spiky. Given* $n > d$ *i.i.d. samples from the* $\mathcal{N}(0, (\Theta^*)^{-1})$ *distribution and any* $\delta \in (0, 1]$, *the estimates* $(\widehat{\Gamma}, \widehat{\Lambda})$ *satisfy the bounds*

$$\|\widehat{\Gamma} - \Gamma^*\|_{\max} \leq 2M \left(4\sqrt{\frac{\log d}{n}} + \delta \right) + \frac{2\alpha}{d} \tag{4}$$

and

$$|||\widehat{\Lambda} - \Lambda^*|||_2 \leq M \left(2\sqrt{\frac{d}{n}} + \delta \right) + s\|\widehat{\Gamma} - \Gamma^*\|_{\max} \tag{5}$$

with probability at least $1 - c_1 e^{-c_2 n \delta^2}$.

Proof. We first prove that the inverse sample covariance matrix $\mathrm{Y} := (\widehat{\Sigma})^{-1}$ is itself a good estimate of Θ^*, in the sense that, for all $\delta \in (0, 1]$,

$$|||\mathrm{Y} - \Theta^*|||_2 \leq M \left(2\sqrt{\frac{d}{n}} + \delta \right) \tag{6}$$

and

$$\|\mathrm{Y} - \Theta^*\|_{\max} \leq M \left(4\sqrt{\frac{\log d}{n}} + \delta \right) \tag{7}$$

with probability at least $1 - c_1 e^{-c_2 n \delta^2}$.

Proof of Proposition 11.19 (cont.)

To prove the first bound (6), we note that

$$(\widehat{\Sigma})^{-1} - \Theta^* = \sqrt{\Theta^*}\left\{n^{-1}\mathrm{V}^{\mathrm{T}}\mathrm{V} - \mathrm{I}_d\right\}\sqrt{\Theta^*} \tag{8}$$

where $\mathrm{V} \in \mathbb{R}^{n \times d}$ is a standard Gaussian random matrix. Consequently, by sub-multiplicativity of the operator norm, we have

$$\begin{aligned}
|||(\widehat{\Sigma})^{-1} - \Theta^*|||_2 &\leq |||\sqrt{\Theta^*}|||_2 |||n^{-1}\mathrm{V}^{\mathrm{T}}\mathrm{V} - I_d|||_2 |||\sqrt{\Theta^*}|||_2 \\
&= |||\Theta^*|||_2 |||n^{-1}\mathrm{V}^{\mathrm{T}}\mathrm{V} - I_d|||_2 \\
&\leq |||\Theta^*|||_2 \left(2\sqrt{\frac{d}{n}} + \delta\right),
\end{aligned}$$

where the final inequality holds with probability $1 - c_1 e^{-n\delta^2}$, via an application of Theorem 6.1. To complete the proof, we note that

$$|||\Theta^*|||_2 \leq |||\Theta^*|||_\infty \leq (|||\sqrt{\Theta^*}|||_\infty)^2 \leq M$$

from which the bound (6) follows.

Turning to the bound (7), using the decomposition (8) and introducing the shorthand $\widetilde{\Sigma} = \frac{V^{\mathrm{T}}V}{n} - I_d$, we have

$$\begin{aligned}\|(\widehat{\Sigma})^{-1} - \Theta^*\|_{\max} &= \max_{j,k=1,\dots,d} \left| e_j^{\mathrm{T}} \sqrt{\Theta^*} \widetilde{\Sigma} \sqrt{\Theta^*} e_k \right| \\ &\leq \max_{j,k=1,\dots,d} \|\sqrt{\Theta^*} e_j\|_1 \|\tilde{\Sigma} \sqrt{\Theta^*} e_k\|_\infty \\ &\leq \|\widetilde{\Sigma}\|_{\max} \max_{j=1,\dots,d} \|\sqrt{\Theta^*} e_j\|_1^2.\end{aligned}$$

Now observe that

$$\max_{j=1,\dots,d} \|\sqrt{\Theta^*} e_j\|_1 \leq \max_{\|u\|_1=1} \|\sqrt{\Theta^*} u\|_1 = \max_{l=1,\dots,d} \sum_{k=1}^{d} |[\sqrt{\Theta^*}]|_{kl} = \|\!|\sqrt{\Theta^*}|\!\|_\infty.$$

This yields that $\|(\widehat{\Sigma})^{-1} - \Theta^*\|_{\max} \leq M \|\tilde{\Sigma}\|_{\max}$. We have $\|\widetilde{\Sigma}\|_{\max} \leq 4\sqrt{\frac{\log d}{n}} + \delta$ with probability at least $1 - c_1 e^{-c_2 n \delta^2}$ for all $\delta \in [0, 1]$. This completes the proof of the bound (7).

Proof of Proposition 11.19 (cont.)

Next we establish bounds on the estimates $(\widehat{\Gamma}, \widehat{\Lambda})$ defined in

$$\widehat{\Gamma} := T_{v_n}((\widehat{\Sigma})^{-1}) \quad \text{and} \quad \widehat{\Lambda} := \widehat{\Gamma} - (\widehat{\Sigma})^{-1}.$$

Recalling our shorthand $\mathrm{Y} = (\widehat{\Sigma})^{-1}$, by the definition of $\widehat{\Gamma}$ and the triangle inequality, we have

$$\begin{aligned}
\|\widehat{\Gamma} - \Gamma^*\|_{\max} &\leq \|\mathrm{Y} - \Theta^*\|_{\max} + \|\mathrm{Y} - T_{v_n}(\mathrm{Y})\|_{\max} + \|\Lambda^*\|_{\max} \\
&\leq M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + v_n + \frac{\alpha}{d} \\
&\leq 2M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{2\alpha}{d}
\end{aligned}$$

thereby establishing inequality (4).

Proof of Proposition 11.19 (cont.)

Turning to the operator norm bound, the triangle inequality implies that

$$|||\widehat{\Lambda} - \Lambda^*|||_2 \leq |||\mathrm{Y} - \Theta^*|||_2 + |||\widehat{\Gamma} - \Gamma^*|||_2 \leq M\left(2\sqrt{\frac{d}{n}} + \delta\right) + |||\widehat{\Gamma} - \Gamma^*|||_2.$$

Recall that Γ^* has at most s-non-zero entries per row. For any index (j,k) such that $\Gamma^*_{jk} = 0$, we have $\Theta^*_{jk} = \Lambda^*_{jk}$, and hence

$$|Y_{jk}| \leq \left|Y_{jk} - \Theta^*_{jk}\right| + \left|\Lambda^*_{jk}\right| \leq M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{\alpha}{d} \leq v_n$$

Consequently $\widehat{\Gamma}_{jk} = T_{v_n}(Y_{jk}) = 0$ by construction. Therefore, the error matrix $\widehat{\Gamma} - \Gamma^*$ has at most s non-zero entries per row, whence

$$|||\widehat{\Gamma} - \Gamma^*|||_2 \leq |||\widehat{\Gamma} - \Gamma^*|||_\infty = \max_{j=1,\ldots,d} \sum_{k=1}^{d} |\widehat{\Gamma}_{jk} - \Gamma^*_{jk}| \leq s\|\widehat{\Gamma} - \Gamma^*\|_{\max}.$$

Putting together the pieces yields the claimed bound (5).