Graphical models for high-dimensional data

Tianhao Wang, Liangchen He

Department of Statistics and Finance, USTC

November 20, 2022

Undirected graphical models

- An undirected graph $G = (V, E)$ consists of a set of vertices $V = 1, 2, \ldots, d$ joined together by a collection of edges E. An edge (j, k) is an unordered pair of distinct vertices $j, k \in V$.
- We associate to each vertex $j \in V$ a random variable X_j, taking values in some space $\mathcal{X}_j$. We then consider the distribution $\mathbb{P}$ of the d-dimensional random vector $X = (X_1, \ldots, X_d)$.
- A clique C means that $(j, k) \in E$ for all distinct vertices $j, k \in C$. A maximal clique is a clique that is not a subset of any other clique. We use $\mathfrak{C}$ to denote the set of all cliques in G, and for each clique $C \in \mathfrak{C}$, we use ψ_C to denote a function of the subvector $x_C := (x_j, j \in C)$. This clique compatibility function takes inputs from the Cartesian product space $\mathcal{X}^C := \bigotimes_{j \in C} \mathcal{X}_j$, and returns non-negative real numbers.

Factorization

Definition (11.1)

The random vector $(X_1, \ldots, X_d)$ factorizes according to the graph G if its density function p can be represented as

$$p(x_1, \ldots, x_d) \propto \prod_{C \in \mathfrak{C}} \psi_C(x_C) \tag{11.1}$$

for some collection of clique compatibility functions $\psi_C : \mathcal{X}^C \to [0, \infty)$.

Example 11.3

Any non-degenerate Gaussian distribution with zero mean can be parameterized in terms of its inverse covariance matrix $\mathbf{\Theta}^* = \mathbf{\Sigma}^{-1}$, also known as the precision matrix. In particular, its density can be written as

$$p\left(x_1, \ldots, x_d; \mathbf{\Theta}^*\right) = \frac{\sqrt{\det\left(\mathbf{\Theta}^*\right)}}{(2\pi)^{d/2}} e^{-\frac{1}{2}x^{\mathrm{T}}\mathbf{\Theta}^* x} \tag{11.2}$$

By expanding the quadratic form, we see that

$$e^{-\frac{1}{2}x^{\mathrm{T}}\mathbf{\Theta}^* x} = \exp\left(-\frac{1}{2}\sum_{(j,k)\in E}\Theta^*_{jk}x_j x_k\right) = \prod_{(j,k)\in E}\underbrace{-\frac{1}{2}\Theta^*_{jk}x_j x_k}_{\psi_{jk}\left(x_j, x_k\right)},$$

showing that any zero-mean Gaussian distribution can be factorized in terms of functions on edges, or cliques of size two.

Conditional independence

- A vertex cutset S is a subset of vertices whose removal from the graph breaks it into two or more disjoint pieces.
- Removing S from the vertex set V leads to the vertex-induced subgraph $G(V\backslash S)$, consisting of the vertex set $V\backslash S$, and the residual edge set

$$E(V\backslash S) := \{(j,k) \in E \mid j,k \in V\backslash S\}. \tag{11.4}$$

- The set S is a vertex cutset if the residual graph $G(V\backslash S)$ consists of two or more disconnected non-empty components.

For any subset $A \subseteq V$, let $X_A := (X_j, j \in A)$ represent the subvector of random variables indexed by vertices in A. For any three disjoint subsets, say A, B and S, of the vertex set V, we use $X_A \perp\!\!\!\perp X_B \mid X_S$ to mean that the subvector X_A is conditionally independent of X_B given X_S.

Definition (11.5)

A random vector $X = (X_1, \ldots, X_d)$ is Markov with respect to a graph G if, for all vertex cutsets S breaking the graph into disjoint pieces A and B, the conditional independence statement $X_A \perp\!\!\!\perp X_B \mid X_S$ holds.

Example 11.6

The Markov chain graph on vertex set $V = \{1, 2, \ldots, d\}$ contains the edges $(j, j+1)$ for $j = 1, 2, \ldots, d-1$. For such a chain graph, each vertex $j \in \{2, 3, \ldots, d-1\}$ is a non-trivial cutset, breaking the graph into the "past" $P = \{1, 2, \ldots, j-1\}$ and "future" $F = \{j+1, \ldots, d\}$. These singleton cutsets define the essential Markov property of a Markov time-series model-namely, that the past X_P and future X_F are conditionally independent given the present X_j.

Theorem (11.8 Hammersley-Clifford)

For a given undirected graph and any random vector $X = (X_1, \ldots, X_d)$ with strictly positive density p, the following two properties are equivalent:
(a) The random vector X factorizes according to the structure of the graph G, as in Definition 11.1.
(b) The random vector X is Markov with respect to the graph G, as in Definition 11.5.

Proof

(a) $\Rightarrow$ (b): Let S be an arbitrary vertex cutset of the graph such that subsets A and B are separated by S . We may assume without loss of generality that both A and B are non-empty, and we need to show that $X_A \perp\!\!\!\perp X_B \mid X_S$.

- $\mathfrak{C}_A := \{C \in \mathfrak{C} \mid C \cap A \neq \emptyset\}, \mathfrak{C}_B := \{C \in \mathfrak{C} \mid C \cap B \neq \emptyset\}$ and $\mathfrak{C}_S := \{C \in \mathfrak{C} \mid C \subseteq S\}$.
- $\mathfrak{C} = \mathfrak{C}_A \cup \mathfrak{C}_S \cup \mathfrak{C}_B$.

$$p\left(x_A, x_S, x_B\right) = \frac{1}{Z} \underbrace{\left[\prod_{C \in \mathfrak{C}_A} \psi_C\left(x_C\right)\right]}_{\Psi_A\left(x_A, x_s\right)} \underbrace{\left[\prod_{C \in \mathfrak{C}_S} \psi_C\left(x_C\right)\right]}_{\Psi_S\left(x_s\right)} \underbrace{\left[\prod_{C \in \mathfrak{C}_B} \psi_C\left(x_C\right)\right]}_{\Psi_B\left(x_B, x_S\right)}.$$

Defining the quantities

$$Z_A(x_S) := \sum_{x_A} \Psi_A(x_A, x_S) \text{ and } Z_B(x_S) := \sum_{x_B} \Psi_B(x_B, x_S),$$

we then obtain the following expressions for the marginal distributions of interest:

$$\begin{aligned} p(x_S) &= \frac{Z_A(x_S)\,Z_B(x_S)}{Z}\Psi_S(x_S), \\ p(x_A, x_S) &= \frac{Z_B(x_S)}{Z}\Psi_A(x_A, x_S)\,\Psi_S(x_S), \\ p(x_B, x_S) &= \frac{Z_A(x_S)}{Z}\Psi_B(x_B, x_S)\,\Psi_S(x_S). \end{aligned}$$

for any x_S for which $p(x_S) > 0$

$$\frac{p(x_A, x_S, x_B)}{p(x_S)} = \frac{\Psi_A(x_A, x_S)\,\Psi_B(x_B, x_S)}{Z_A(x_S)\,Z_B(x_S)},$$

$$\frac{p(x_A, x_S)}{p(x_S)} = \frac{\Psi_A(x_A, x_S)}{Z_A(x_S)},$$

$$\frac{p(x_B, x_S)}{p(x_S)} = \frac{\Psi_B(x_B, x_S)}{Z_B(x_S)}.$$

$$\Rightarrow p(x_A, x_B \mid x_S) = \frac{p(x_A, x_B, x_S)}{p(x_S)} = \frac{p(x_A, x_S)}{p(x_S)}\frac{p(x_B, x_S)}{p(x_S)}$$

$$= p(x_A \mid x_S)\,p(x_B \mid x_S).$$

The graphical Lasso estimator

$$\widehat{\mathbf{\Theta}} \in \arg\min_{\mathbf{\Theta} \in \mathcal{S}^{d\times d}} \{ \underbrace{\langle\langle \mathbf{\Theta}, \widehat{\mathbf{\Sigma}} \rangle\rangle - \log\det \mathbf{\Theta}}_{\mathcal{L}_n(\mathbf{\Theta})} + \lambda_n \|\mathbf{\Theta}\|_{1,\text{off}} \},$$

where $\|\mathbf{\Theta}\|_{1,\text{ off}} := \sum_{j\neq k} \left|\Theta_{jk}\right|$ corresponds to the ℓ_1 -norm applied to the off-diagonal entries of $\mathbf{\Theta}$.

The following result is based on a sample covariance matrix $\widehat{\Sigma}$ formed from n i.i.d. samples $\{x_i\}_{i=1}^n$ of a zero-mean random vector in which each coordinate has σ-sub-Gaussian tails.

Theorem (11.9 Frobenius norm bounds for graphical Lasso)

Suppose that the inverse covariance matrix Θ^* has at most m non-zero entries per row, and we solve the graphical Lasso (11.10) with regularization parameter $\lambda_n = 8\sigma^2 \left(\sqrt{\frac{\log d}{n}} + \delta \right)$ for some $\delta \in (0,1]$. Then as long as $\left(\left\| \mathbf{\Theta}^* \right\|_2 + 1 \right)^2 \lambda_n \sqrt{md} < 1$, the graphical Lasso estimate $\widehat{\mathbf{\Theta}}$ satisfies

$$\left\| \widehat{\mathbf{\Theta}} - \mathbf{\Theta}^* \right\|_{\mathrm{F}}^2 \leq \frac{9}{\left(\left\| \mathbf{\Theta}^* \right\|_2 + 1 \right)^4} md\lambda_n^2 \tag{11.11}$$

with probability at least $1 - 8e^{-\frac{1}{16}n\delta^2}$.

Proof

- $\mathbb{B}_F(1) = \left\{\Delta \in \mathcal{S}^{d\times d} \mid \|\Delta\|_{\mathrm{F}} \leq 1\right\}$
- $\nabla \mathcal{L}_n(\mathbf{\Theta}) = \widehat{\Sigma} - \mathbf{\Theta}^{-1}$ and $\nabla^2 \mathcal{L}_n(\mathbf{\Theta}) = \mathbf{\Theta}^{-1} \otimes \mathbf{\Theta}^{-1}$
- $\underbrace{\mathcal{L}_n\left(\mathbf{\Theta}^* + \Delta\right) - \mathcal{L}_n\left(\mathbf{\Theta}^*\right) - \langle\langle \nabla \mathcal{L}_n\left(\mathbf{\Theta}^*\right), \Delta\rangle\rangle}_{\mathcal{E}_n(\Delta)} =$
  $\frac{1}{2} \operatorname{vec}(\Delta)^{\mathrm{T}} \nabla^2 \mathcal{L}_n\left(\mathbf{\Theta}^* + t\Delta\right) \operatorname{vec}(\Delta)$ for some $t \in [0, 1]$
- $\left\|\mathbf{A}^{-1} \otimes \mathbf{A}^{-1}\right\|_2 = \frac{1}{\|\mathbf{A}\|_2^2}$ for any symmetric invertible matrix

Verifying restricted strong convexity

For any $\Delta \in \mathbb{B}_F(1)$,

$$\mathcal{E}_n(\Delta) \geq \frac{1}{2}\gamma_{\min}\left(\nabla^2\mathcal{L}_n\left(\mathbf{\Theta}^* + t\Delta\right)\right)\|\operatorname{vec}(\Delta)\|_2^2 = \frac{1}{2}\frac{\|\Delta\|_{\mathrm{F}}^2}{\left\|\mathbf{\Theta}^* + t\Delta\right\|_2^2}.$$

$t\|\Delta\|_2 \leq t\|\Delta\|_{\mathrm{F}} \leq 1$ implies that $\left\|\mathbf{\Theta}^* + t\Delta\right\|_2^2 \leq (\|\Theta^*\|_2 + 1)^2$. Then

$$\mathcal{E}_n(\Delta) \geq \frac{\kappa}{2}\|\Delta\|_{\mathrm{F}}^2 \quad \text{where } \kappa := \left(\left\|\mathbf{\Theta}^*\right\|_2 + 1\right)^{-2}$$

showing that the RSC condition from Definition 9.15 holds over $\mathbb{B}_F(1)$ with tolerance $\tau_n^2 = 0$.

Computing the subspace Lipschitz constant

Letting S denote the support set of $\boldsymbol{\Theta}^*$, we define the subspace $\mathbb{M}(S) = \left\{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} \mid \Theta_{jk} = 0 \text{ for all } (j,k) \notin S\right\}$. Then we have

$$\Psi^2(\mathbb{M}(S)) = \sup_{\boldsymbol{\Theta}\in M(S)} \frac{\left(\sum_{j\neq k} \left|\Theta_{jk}\right|\right)^2}{\|\boldsymbol{\Theta}\|_{\mathrm{F}}^2} \leq |S| \overset{\text{(i)}}{\leq} md.$$

where inequality (i) follows since $\boldsymbol{\Theta}^*$ has at most m non-zero entries per row.

Verifying event $\mathbb{G}(\lambda_n)$

Using Lemma 6.26, we have

$$\mathbb{P}\left[\|\widehat{\Sigma} - \Sigma\|_{\text{max, off}} \geq \sigma^2 t\right] \leq 8e^{-\frac{\pi}{16}\min\{t, t^2\} + 2\log d} \quad \text{for all } t > 0.$$

Setting $t = \lambda_n/\sigma^2$ shows that the event $\mathbb{G}(\lambda_n)$ from Corollary 9.20 holds with the claimed probability. Consequently, Proposition 9.13 implies that the error matrix $\widehat{\Delta}$ satisfies the bound $\left\|\widehat{\Delta}_{S^c}\right\|_1 \leq 3\left\|\widehat{\Delta}_S\right\|_1$, and hence $\|\widehat{\Delta}\|_1 \leq 4\left\|\widehat{\Delta}_S\right\|_1 \leq 4\sqrt{md}\|\widehat{\Delta}\|_{\mathrm{F}}$, where the final inequality again uses the fact that $|S| \leq md$. In order to apply Corollary 9.20 , the only remaining detail to verify is that $\widehat{\Delta}$ belongs to the Frobenius ball $\mathbb{B}_F(1)$.

Localizing the error matrix

The result of Exercise 9.10 then implies that

$$\langle\langle \nabla \mathcal{L}_n(\mathbf{\Theta}^* + \Delta) - \nabla \mathcal{L}_n(\mathbf{\Theta}^*), \Delta \rangle\rangle \geq \kappa \|\Delta\|_{\mathrm{F}} \quad \text{for all } \Delta \in \mathcal{S}^{d\times d} \backslash \mathbb{B}_F(1).$$

By the optimality of $\widehat{\mathbf{\Theta}}$, we have $0 = \left\langle\left\langle \nabla \mathcal{L}_n\left(\mathbf{\Theta}^* + \widehat{\Delta}\right) + \lambda_n \widehat{\mathbf{Z}}, \widehat{\mathbf{\Delta}}\right\rangle\right\rangle$, where $\widehat{\mathbf{Z}} \in \partial \|\widehat{\mathbf{\Theta}}\|_{1,\text{ off}}$. By adding and subtracting terms, we find that

$$\begin{aligned}\left\langle\left\langle \nabla \mathcal{L}_n\left(\mathbf{\Theta}^* + \widehat{\Delta}\right) - \nabla \mathcal{L}_n(\mathbf{\Theta}^*), \widehat{\Delta}\right\rangle\right\rangle &\leq \lambda_n |\langle \widehat{\mathbf{Z}}, \widehat{\Delta}\rangle| + \left\langle\left\langle \nabla \mathcal{L}_n(\mathbf{\Theta}^*), \widehat{\Delta}\right\rangle\right\rangle \\ &\leq \left\{\lambda_n + \left\|\nabla \mathcal{L}_n(\mathbf{\Theta}^*)\right\|_{\max}\right\} \|\widehat{\Delta}\|_1.\end{aligned}$$

the right-hand side is at most $\frac{3\lambda_n}{2}\|\widehat{\mathbf{\Delta}}\|_1 \leq 6\lambda_n\sqrt{md}\|\widehat{\mathbf{\Delta}}\|_{\mathrm{F}}$. If $\|\widehat{\Delta}\|_{\mathrm{F}} > 1$, then we obtain $\kappa\|\widehat{\Delta}\|_{\mathrm{F}} \leq \frac{3\lambda_n}{2}\|\widehat{\mathbf{\Delta}}\|_1 \leq 6\lambda_n\sqrt{md}\|\widehat{\Delta}\|_{\mathrm{F}}$ This inequality leads to a contradiction whenever $\frac{6\lambda_n\sqrt{md}}{\kappa} < 1$, which completes the proof.

- $S := E \cup \{(j,j) \mid j \in V\}, S^c = (V \times V) \backslash S.$
- the matrix $\Gamma^* := \nabla^2 \mathcal{L}_n(\boldsymbol{\Theta}^*)$ is α-incoherent if
  $\max_{e \in S^c} \left\| \Gamma^*_{eS} \left(\boldsymbol{\Gamma}^*_{SS} \right)^{-1} \right\|_1 \leq 1 - \alpha \quad$ for some $\alpha \in (0,1]$.
- $\widehat{E} := \left\{ (j,k) \in [d] \times [d] \mid j < k \text{ and } \widehat{\boldsymbol{\Theta}}_{jk} \neq 0 \right\}.$
- $\|\mathbf{A}\|_2 \leq m + 1$ for any graph of degree at most m.

Theorem (11.10)

Consider a zero-mean d-dimensional Gaussian distribution based on an α-incoherent inverse covariance matrix $\mathbf{\Theta}^*$. Given a sample size lower bounded as $n > c_0 \left(1 + 8\alpha^{-1}\right)^2 m^2 \log d$, suppose that we solve the graphical Lasso (11.10) with a regularization parameter $\lambda_n = \frac{c_1}{\alpha} \sqrt{\frac{\log d}{n}} + \delta$ for some $\delta \in (0, 1]$. Then with probability at least $1 - c_2 e^{-c_3 n\delta^2}$, we have the following:
(a) The graphical Lasso solution leads to no false inclusions-that is, $\widehat{\Theta}_{jk} = 0$ for all $(j, k) \notin E$.
(b) It satisfies the sup-norm bound

$$\left\|\widehat{\mathbf{\Theta}} - \mathbf{\Theta}^*\right\|_{\max} \leq c_4\{\underbrace{\left(1 + 8\alpha^{-1}\right) \sqrt{\frac{\log d}{n}}}_{\tau(n,d,\alpha)} + \lambda_n\}. \tag{11.16}$$

Corollary (11.11)

Under the conditions of Proposition 11.10, consider the graphical Lasso estimate $\widehat{\boldsymbol{\Theta}}$ with regularization parameter $\lambda_n = \frac{c_1}{\alpha}\sqrt{\frac{\log d}{n}} + \delta$ for some $\delta \in (0, 1]$. Then with probability at least $1 - c_2 e^{-c_3 n\delta^2}$, we have

$$\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_2 \leq c_4 \|\mathbf{A}\|_2 \left\{\left(1 + 8\alpha^{-1}\right)\sqrt{\frac{\log d}{n}} + \lambda_n\right\}, \qquad (11.17a)$$

where $\mathbf{A}$ denotes the adjacency matrix of the graph G (including ones on the diagonal). In particular, if the graph has maximum degree m , then

$$\left\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\right\|_2 \leq c_4 (m + 1) \left\{\left(1 + 8\alpha^{-1}\right)\sqrt{\frac{\log d}{n}} + \lambda_n\right\}. \qquad (11.17b)$$

Neighborhood-based methods

- $\mathcal{N}^+(j) := \{j\} \cup \mathcal{N}(j), X_j \perp\!\!\!\perp X_{V \setminus \mathcal{N}^+(j)} \mid X_{\mathcal{N}(j)}$.

Lasso-based neighborhood regression:

1 For each node $j \in V$:

(a) Extract the column vector $X_j \in \mathbb{R}^n$ and the submatrix $\mathbf{X}_{\setminus\{j\}} \in \mathbb{R}^{n\times(d-1)}$.

(b) Solve the Lasso problem:

$$\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^{d-1}} \left\{ \frac{1}{2n} \left\| X_j - \mathbf{X}_{\setminus\{j\}}\theta \right\|_2^2 + \lambda_n \|\theta\|_1 \right\}.$$

(c) Return the neighborhood estimate $\widehat{\mathcal{N}}(j) = \left\{ k \in V \setminus \{j\} \mid \widehat{\theta}_k \neq 0 \right\}$.

2 Combine the neighborhood estimates to form an edge estimate $\widehat{E}$, using either the OR rule or the AND rule.

Assume $\mathrm{diag}(\boldsymbol{\Sigma}^*) \leq 1$ and $n \gtrsim m \log d$.

Theorem (11.12)

Consider a zero-mean Gaussian random vector with covariance $\boldsymbol{\Sigma}^*$ such that for each $j \in V$, the submatrix $\boldsymbol{\Sigma}^*_{\backslash\{j\}} := \mathrm{cov}(\mathbf{X}_{\backslash\{j\}})$ is α-incoherent with respect to $\mathcal{N}(j)$, and $\left|\!\left|\!\left|\left(\Sigma^*_{\mathcal{N}(j),\mathcal{N}(j)}\right)^{-1}\right|\!\right|\!\right|_\infty \leq b$ for some $b \geq 1$. Suppose that the neighborhood Lasso selection method is implemented with $\lambda_n = c_0\left\{\frac{1}{\alpha}\sqrt{\frac{\log d}{n}} + \delta\right\}$ for some $\delta \in (0,1]$. Then with probability greater than $1 - c_2 e^{-c_3 n \min\{\delta^2, \frac{1}{m}\}}$, the estimated edge set $\widehat{E}$, based on either the AND or OR rules, has the following properties: (a) No false inclusions: it includes no false edges, so that $\widehat{E} \subseteq E$. (b) All significant edges are captured: it includes all edges (j,k) for which $\left|\Theta^*_{jk}\right| \geq 7b\lambda_n$.

- $\boldsymbol{\Gamma}^* = \text{cov}\left(X_{\backslash\{j\}}\right), \widehat{\boldsymbol{\Gamma}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}_{\backslash\{j\}}\mathbf{X}_{\backslash\{j\}}, S = \mathcal{N}(j), S^c = V \backslash \mathcal{N}^+(j).$
- $\widehat{\boldsymbol{\Gamma}}_{SS} :=$ the submatrix indexed by the subset S.

Proof of part (a)

We follow the proof of Theorem 7.21 until equation (7.53), namely

$$\widehat{z}_{S^c} = \underbrace{\widehat{\boldsymbol{\Gamma}}_{S^cS}\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1}\widehat{z}_S}_{\mu\in\mathbb{R}^{d-s}} + \underbrace{\mathbf{X}_{S^c}^{\mathrm{T}}\left[\mathbf{I}_n - \mathbf{X}_S\left(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\right)^{-1}\mathbf{X}_S^{\mathrm{T}}\right]\left(\frac{W_j}{\lambda_n n}\right)}_{V_{S^c}\in\mathbb{R}^{d-s}}$$

As argued in Chapter 7, in order to establish that the Lasso support is included within S , it suffices to establish the strict dual feasibility condition $\|\widehat{z}_S\|_\infty < 1$. We do so by establishing that

$$\mathbb{P}\left[\|\mu\|_\infty \geq 1 - \frac{3}{4}\alpha\right] \leq c_1 e^{-c_2 n\alpha^2 - \log d}, \tag{11.25a}$$

$$\mathbb{P}\left[\|V_{S^c}\|_\infty \geq \frac{\alpha}{4}\right] \leq c_1 e^{-c_2 n\delta^2\alpha^2 - \log d}. \tag{11.25b}$$

Taken together, these bounds ensure that $\|\bar{z}_{S^c}\|_\infty \leq 1 - \frac{\alpha}{2} < 1$, and hence that the Lasso support is contained within $S = \mathcal{N}(j)$, with probability at least $1 - c_1 e^{-c_2 n\delta^2\alpha^2 - \log d}$, where the values of the universal constants may change from line to line. Taking the union bound over all d vertices, we conclude that $\widehat{E} \subseteq E$ with probability at least $1 - c_1 e^{-c_2 n\delta^2\alpha^2}$.

Proof of (11.25a): $\mathbf{X}_{S^c}^{\mathrm{T}} = \mathbf{\Gamma}_{S^cS}^* \left(\mathbf{\Gamma}_{SS}^*\right)^{-1} \mathbf{X}_S^{\mathrm{T}} + \tilde{\mathbf{W}}_{S^c}^{\mathrm{T}}$, where $\tilde{\mathbf{W}}_{S^c} \in \mathbb{R}^{n\times|S^c|}$ is a zero-mean Gaussian random matrix that is independent of $\mathbf{X}_S$. Since $\operatorname{cov}\left(\tilde{\mathbf{W}}_{S^c}\right) = \mathbf{\Gamma}_{S^cS^c}^* - \Gamma_{S^cS}^* \left(\mathbf{\Gamma}_{SS}^*\right)^{-1} \mathbf{\Gamma}_{SS^c}^* \leq \mathbf{\Gamma}^*$ and $\operatorname{diag}\left(\mathbf{\Gamma}^*\right) \leq 1$, we see that the elements of $\tilde{\mathbf{W}}_{S^c}$ have variance at most 1.

$$
\begin{aligned}
\|\mu\|_\infty &= \left\| \boldsymbol{\Gamma}^*_{S^cS} \left(\boldsymbol{\Gamma}^*_{SS}\right)^{-1} \widehat{z}_S + \frac{\tilde{\mathbf{W}}^{\mathrm{T}}_{S^c}}{\sqrt{n}} \frac{\mathbf{X}_S}{\sqrt{n}} \left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \widehat{z}_S \right\|_\infty \\
&\overset{\text{(i)}}{\leq} (1-\alpha) + \| \underbrace{\frac{\tilde{\mathbf{W}}^{\mathrm{T}}_{S^c}}{\sqrt{n}} \frac{\mathbf{X}_S}{\sqrt{n}} \left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \widehat{z}_S}_{\widetilde{V} \in \mathbb{R}^{|S^c|}} \|_\infty
\end{aligned}
\tag{11.27}
$$

$$
\begin{aligned}
\frac{1}{\sqrt{n}} \left\| \frac{\mathbf{X}_S}{\sqrt{n}} \left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \widehat{z}_S \right\|_2 &\leq \frac{1}{\sqrt{n}} \left\| \frac{\mathbf{X}_S}{\sqrt{n}} \left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \right\|_2 \left\|\widehat{z}_S\right\|_2 \\
&\leq \frac{1}{\sqrt{n}} \sqrt{\left\| \left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \right\|_2} \sqrt{m} \\
&\leq 2\sqrt{\frac{bm}{n}},
\end{aligned}
$$

where inequality (i) follows with probability at least $1 - 4e^{-c_1 n}$, using standard bounds on Gaussian random matrices (see Theorem 6.1). Using this upper bound to control the conditional variance of $\widetilde{V}$, standard Gaussian tail bounds and the union bound then ensure that

$$\mathbb{P}\left[\|\widetilde{V}\|_\infty \geq t\right] \leq 2\left|S^c\right| e^{-\frac{nn^2}{8bm}} \leq 2e^{-\frac{nn^2}{8bm}+\log d}.$$

We now set $t = \left[\frac{64bm\log d}{n} + \frac{1}{64}\alpha^2\right]^{1/2}$, a quantity which is less than $\frac{\alpha}{4}$ as long as $n \geq c\frac{bm\log d}{\alpha}$ for a sufficiently large universal constant. Thus, we have established that $\|\widetilde{V}\|_\infty \leq \frac{\alpha}{4}$ with probability at least $1 - c_1 e^{-c_2 n\alpha^2 - \log d}$. Combined with the earlier bound (11.27), the claim (11.25a) follows.

Proof of (11.25b): note that the matrix $\mathbf{\Pi} := \mathbf{I}_n - \mathbf{X}_S\left(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\right)^{-1}\mathbf{X}_S^{\mathrm{T}}$ has the range of $\mathbf{X}_S$ as its nullspace. Thus, using the decomposition (11.26), we have

$$V_{S^c} = \tilde{\mathbf{W}}_{S^c}^{\mathrm{T}}\mathbf{\Pi}\left(\frac{W_j}{\lambda_n n}\right),$$

where $\tilde{\mathbf{W}}_{S^c} \in \mathbb{R}^{|S^c|}$ is independent of $\mathbf{\Pi}$ and W_j . Since $\mathbf{\Pi}$ is a projection matrix, we have $\left\|\mathbf{\Pi}W_j\right\|_2 \leq \left\|W_j\right\|_2$. The vector $W_j \in \mathbb{R}^n$ has i.i.d. Gaussian entries with variance at most 1 , and hence the event $\mathcal{E} = \left\{\frac{\|W_j\|_2}{\sqrt{n}} \leq 2\right\}$ holds with probability at least $1 - 2e^{-n}$. Conditioning on this event and its complement, we find that

$$\mathbb{P}\left[\|V_{S^c}\|_\infty \geq t\right] \leq \mathbb{P}\left[\|V_{S^c}\|_\infty \geq t \mid \mathcal{E}\right] + 2e^{-c_3 n}.$$

Conditioned on $\mathcal{E}$, each element of V_{S^c} has variance at most $\frac{4}{\lambda_n^2 n}$, and hence

$$\mathbb{P}\left[\|V_{S^c}\|_\infty \geq \frac{\alpha}{4}\right] \leq 2e^{-\frac{\lambda_n^2 n^2}{256}+\log|S^c|} + 2e^{-n},$$

where we have combined the union bound with standard Gaussian tail bounds. Since $\lambda_n = c_0 \left\{ \frac{1}{\alpha} \sqrt{\frac{\log d}{n}} + \delta \right\}$ for a universal constant c_0 that may be chosen, we can ensure that $\frac{\lambda_n^2 n \alpha^2}{256} \geq c_2 n \alpha^2 \delta^2 + 2 \log d$ for some constant c_2 , for which it follows that

$$\mathbb{P}\left[\|V_{S^c}\|_\infty \geq \frac{\alpha}{4}\right] \leq c_1 e^{-c_2 n \delta^2 \alpha^2 - \log d} + 2e^{-n}.$$

Proof of part (b)

It suffices to establish ℓ_∞-bounds on the error in the Lasso solution. Here we provide a proof in the case $m \le \log d$. Again returning to the proof of Theorem 7.21 , equation (7.54) guarantees that

$$\left\|\widehat{\theta}_S - \theta_S^*\right\|_\infty \le \left\|\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \mathbf{X}_S^{\mathrm{T}} \frac{W_j}{n}\right\|_\infty + \lambda_n \left\|\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1}\right\|_\infty$$
$$\le \left\|\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \mathbf{X}_S^{\mathrm{T}} \frac{W_j}{n}\right\|_\infty + \lambda_n \left\{\left\|\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} - \left(\boldsymbol{\Gamma}_{SS}^*\right)^{-1}\right\|_\infty + \left\|\left(\boldsymbol{\Gamma}_{SS}^*\right)^{-1}\right\|_\infty\right. \tag{11.28}$$

Now for any symmetric $m \times m$ matrix, we have

$$\|\mathbf{A}\|_\infty = \max_{i=1,\ldots,m} \sum_{\ell=1}^m |A_{i\ell}| \le \sqrt{m} \max_{i=1,\ldots,m} \sqrt{\sum_{\ell=1}^m |A_{i\ell}|^2} \le \sqrt{m}\|\mathbf{A}\|_2$$

Applying this bound to the matrix $\mathbf{A} = \left(\widehat{\mathbf{\Gamma}}_{SS}\right)^{-1} - \left(\mathbf{\Gamma}^*_{SS}\right)^{-1}$, we find

$$\left\| \left(\widehat{\mathbf{\Gamma}}_{SS}\right)^{-1} - \left(\mathbf{\Gamma}^*_{SS}\right)^{-1} \right\|_\infty \leq \sqrt{m} \left\| \left(\widehat{\mathbf{\Gamma}}_{SS}\right)^{-1} - \left(\mathbf{\Gamma}^*_{SS}\right)^{-1} \right\|_2 . \tag{11.29}$$

Since $\left\|\Gamma^*_{SS}\right\|_2 \leq \left\|\Gamma^*_{SS}\right\|_\infty \leq b$, applying the random matrix bound from Theorem 6.1 allows us to conclude that

$$\left\| \left(\widehat{\mathbf{\Gamma}}_{SS}\right)^{-1} - \left(\mathbf{\Gamma}^*_{SS}\right)^{-1} \right\|_2 \leq 2b \left(\sqrt{\frac{m}{n}} + \frac{1}{\sqrt{m}} + 10 \sqrt{\frac{\log d}{n}} \right),$$

with probability at least $1 - c_1 e^{-c_2 \frac{n}{m} - \log d}$. Combined with the earlier bound (11.29), we find that

$$\left\| \left(\widehat{\mathbf{\Gamma}}_{SS}\right)^{-1} - \left(\mathbf{\Gamma}^*_{SS}\right)^{-1} \right\|_\infty \leq 2b \left(\sqrt{\frac{m^2}{n}} + 1 + 10 \sqrt{\frac{m \log d}{n}} \right) \overset{\text{(i)}}{\leq} 6b, \tag{11.30}$$

where inequality (i) uses the assumed lower bound $n \gtrsim m \log d \geq m^2$. Putting together the pieces in the bound (11.28) leads to

$$\left\|\widehat{\theta}_S - \theta_S^*\right\|_\infty \leq \| \underbrace{\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} \mathbf{X}_S^{\mathrm{T}} \frac{W_j}{n}}_{U_S} \|_\infty + 7b\lambda_n. \tag{11.31}$$

Now the vector $W_j \in \mathbb{R}^n$ has i.i.d. Gaussian entries, each zero-mean with variance at most $\operatorname{var}\left(X_j\right) \leq 1$, and is independent of $\mathbf{X}_S$. Consequently, conditioned on $\mathbf{X}_S$, the quantity U_S is a zero-mean Gaussian m -vector, with maximal variance

$$\frac{1}{n}\left\|\operatorname{diag}\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1}\right\|_\infty \leq \frac{1}{n}\left\{\left\|\left(\widehat{\boldsymbol{\Gamma}}_{SS}\right)^{-1} - \left(\boldsymbol{\Gamma}_{SS}^*\right)^{-1}\right\|_\infty + \left\|\left(\boldsymbol{\Gamma}_{SS}^*\right)^{-1}\right\|_\infty\right\} \leq \frac{7b}{n},$$

where we have combined the assumed bound $\left\|\left(\boldsymbol{\Gamma}_{SS}^*\right)^{-1}\right\|_\infty \leq b$ with the inequality (11.30).

Therefore, the union bound combined with Gaussian tail bounds implies that

$$\mathbb{P}\left[\|U_S\|_\infty \geq b\lambda_n\right] \leq 2|S|e^{-\frac{n\lambda_n^2}{14}} \overset{(i)}{\leq} c_1 e^{-c_2 nb\delta^2 - \log d},$$

where, as in our earlier argument, inequality (i) can be guaranteed by a sufficiently large choice of the pre-factor c_0 in the definition of λ_n . Substituting back into the earlier bound (11.31), we find that $\left\|\widehat{\theta}_S - \theta_S^*\right\|_\infty \leq 7b\lambda_n$ with probability at least $1 - c_1 e^{-c_2 n\{\delta^2 \wedge \frac{1}{m}\} - \log d}$. Finally, taking the union bound over all vertices $j \in V$ causes a loss of at most a factor log d in the exponent.

Let us now move beyond the Gaussian case, and consider the graph estimation problem for a more general class of graphical models that can be written in an exponential form. In particular, for a given graph $G = (V, E)$, consider probability densities that have a pairwise factorization of the form

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

Example 11.4(Ising model)

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

A general form of neighborhood regression

- We can form a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with x_i^{T} as the i th row. For $j = 1, \ldots, d$, we let $X_j \in \mathbb{R}^n$ denote the j th column of $\mathbf{X}$. Neighborhood regression is based on predicting the column $X_j \in \mathbb{R}^n$ using the columns of the submatrix $\mathbf{X}_{\backslash(j\}} \in \mathbb{R}^{n \times (d-1)}$.
- Consider the conditional likelihood of $X_j \in \mathbb{R}^n$ given $\mathbf{X}_{\backslash(j)} \in \mathbb{R}^{n \times (d-1)}$,this conditional likelihood depends only on the vector of parameters

$$\mathbf{\Theta}_{j+} := \left\{\Theta_j, \mathbf{\Theta}_{jk}, k \in V \backslash \{j\}\right\}$$

Moreover, in the true model $\mathbf{\Theta}^*$, we are guaranteed that $\mathbf{\Theta}^*_{jk} = 0$ whenever $(j, k) \notin E$, so that it is natural to impose some type of block-based sparsity penalty on $\mathbf{\Theta}_{j+}$.

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

Especially,or the Ising model, the neighborhood regression estimate:

$$\widehat{\theta}_{j+} = \arg\min_{\theta_{j+}\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^{n} f\left(\theta_j x_{ij} + \sum_{k\in V\setminus\{j\}} \theta_{jk} x_{ij} x_{ik}\right) + \lambda_n \sum_{k\in V\setminus\{j\}} \left|\theta_{jk}\right| \right\}$$

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

Let θ_{j+}^* denote the minimizer of the population objective function $\overline{\mathcal{L}}(\theta_{j+}) = \mathbb{E}\left[\mathcal{L}_n\left(\theta_{j+}; X_j, \mathbf{X}_{\setminus\{j\}}\right)\right]$. We then consider the Hessian of the cost function $\overline{\mathcal{L}}$ evaluated at the "true parameter" θ_{j+}^*-namely, the d-dimensional matrix $\mathbf{J} := \nabla^2 \overline{\mathcal{L}}\left(\theta_{j+}^*\right)$.

- For a given $\alpha \in (0, 1]$, we say that $\mathbf{J}$ satisfies an α-incoherence condition at node $j \in V$ if

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

- We assume the submatrix $\mathbf{J}_{SS}$ has its smallest eigenvalue lower bounded by some $c_{\min} > 0$.
- The following result applies to an Ising model defined on a graph G with d vertices and maximum degree at most m.

Theorem 11.15 Given n i.i.d. samples with $n > c_0 m^2 \log d$, consider the estimator 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 $\widehat{S} = \mathrm{supp}(\widehat{\theta})$ that is contained within the neighborhood set $\mathcal{N}(j)$.
- (b) It satisfies the ℓ_∞-bound $\left\|\widehat{\theta}_{j+} - \theta^*_{j+}\right\|_\infty \leq \frac{c_3}{c_{\min}}\sqrt{m}\lambda_n$.

Part (a) guarantees that the method leads to no false inclusions. On the other hand, the ℓ_∞-bound in part (b) ensures that the method picks up all significant variables.

Let us begin our exploration with the case of corrupted data. Letting $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote the data matrix corresponding to the original samples, suppose that we instead observe a corrupted version $\mathbf{Z}$. In the simplest case, we might observe $\mathbf{Z} = \mathbf{X} + \mathbf{V}$, where the matrix $\mathbf{V}$ represents some type of measurement error.

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

where $\widehat{\mathbf{\Sigma}}_z = \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n} z_i z_i^{\mathrm{T}}$.
More generally, any unbiased estimate $\widehat{\mathbf{\Gamma}}$ of $\mathbf{\Sigma}_x$:

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

Unbiased covariance estimate

In order to obtain a consistent estimator, we need to replace $\widehat{\mathbf{\Sigma}}_z$ with an unbiased estimator of $\mathrm{cov}(x)$ based on the observed data matrix $\mathbf{Z}$. In order to develop intuition, let us explore an example. Suppose that each row v_i of the noise matrix $\mathbf{V}$ is drawn i.i.d. from a zero-mean distribution, say with covariance $\mathbf{\Sigma}_v$.

$$\widehat{\mathbf{\Gamma}} := \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} - \mathbf{\Sigma}_v.$$

As long as the noise matrix $\mathbf{V}$ is independent of $\mathbf{X}$, then $\widehat{\mathbf{\Gamma}}$ is an unbiased estimate of $\mathbf{\Sigma}_x$.

Missing data

Example 11.17 (Missing data) In other settings, some entries of the data matrix $\mathbf{X}$ might be missing, with the remaining entries observed. In the simplest model of missing data known as missing completely at random-entry (i, j) of the data matrix is missing with some probability $v \in [0, 1)$.

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

With this choice, it can be verified that

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

linear regression

In order to obtain a consistent form of linear regression, consider the following population-level objective function

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

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

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

As in our analysis of the ordinary Lasso from Chapter 7 , we impose a restricted eigenvalue (RE) condition on the covariance estimate $\widehat{\mathbf{\Gamma}}$: more precisely, we assume that there exists a constant $\kappa > 0$ such that

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

Proposition 11.18 Under the RE condition, suppose that the pair $(\widehat{\gamma}, \widehat{\Gamma})$ satisfy the deviation condition

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

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 program (11.48) satisfies the bound

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

Proof:

- $$\left\langle \widehat{\Delta}, \nabla \mathcal{L}_n \left(\theta^* + \widehat{\Delta}\right) - \nabla \mathcal{L}_n \left(\theta^*\right) \right\rangle \leq \left| \left\langle \widehat{\Delta}, \nabla \mathcal{L}_n \left(\theta^*\right) \right\rangle \right| - \lambda_n \langle \widehat{z}, \widehat{\Delta} \rangle$$
$$\leq \|\widehat{\Delta}\|_1 \left\| \nabla \mathcal{L}_n \left(\theta^*\right) \right\|_\infty + \lambda_n \left\{ \left\| \theta^* \right\|_1 - \|\widetilde{\theta}\|_1 \right.$$
- $\|\theta^*\|_1 - \|\widetilde{\theta}\|_1 \leq \left\| \widehat{\Delta}_S \right\|_1 - \left\| \widehat{\Delta}_S \right\|_1$.(Theorem 7.8)
- 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

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

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

Gaussian graph selection with hidden variables

Consider a family of $d+r$ random variables-say written as $X := (X_1, \ldots, X_d, X_{d+1}, \ldots, X_{d+r})$ and suppose that this full vector can be modeled by a sparse graphical model with $d+r$ vertices. Now suppose that we observe only the subvector $X_{\mathrm{O}} := (X_1, \ldots, X_d)$, with the other components $X_{\mathrm{H}} := (X_{d+1}, \ldots, X_{d+r})$ staying hidden.

$$\mathbf{\Theta}^{\diamond} = \begin{bmatrix} \mathbf{\Theta}^{\diamond}_{\mathrm{OO}} & \mathbf{\Theta}^{\diamond}_{\mathrm{OH}} \\ \mathbf{\Theta}^{\circ}_{\mathrm{HO}} & \mathbf{\Theta}^{\diamond}_{\mathrm{HH}} \end{bmatrix}.$$

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

The Hammersley-Clifford theorem implies that the inverse covariance matrix $\mathbf{\Theta}^{\diamond}$ of the full vector $X = (X_{\mathrm{O}}, X_{\mathrm{H}})$ is sparse. By our modeling assumptions, the matrix $\Gamma^{*} := \mathbf{\Theta}^{\circ}_{\mathrm{OO}}$ is sparse.

When $n > d$, then the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ will be invertible with high probability, and hence setting $\mathbf{Y} := (\widehat{\mathbf{\Sigma}})^{-1}$, we can consider an observation model of the form

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

For a threshold $v_n > 0$ to be chosen, we define the estimates

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

Here the hard-thresholding operator is given by $T_{v_n}(v) = vI\left[|v| > v_n\right].$

- (A1) $\|\Lambda^*\|_{\max} \leq \frac{\alpha}{d}$
- (A2) $\left\| \sqrt{\mathbf{\Theta}^*} \right\|_{\infty} = \max_{j=1,\dots,d} \sum_{k=1}^{d} \left| \sqrt{\Theta^*} \right|_{jk} \leq \sqrt{M}$
- (A3) $v_n := M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{\alpha}{d}$ for some $\delta \in [0,1]$

Proposition 11.19 Consider a precision matrix $\mathbf{\Theta}^*$ that can be decomposed as the difference $\mathbf{\Gamma}^* - \Lambda^*$. Given $n > d$ i.i.d. samples from the $\mathcal{N}\left(0, (\mathbf{\Theta}^*)^{-1}\right)$ distribution and any $\delta \in (0,1]$

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

and

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

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

Proof

- $(\widehat{\mathbf{\Sigma}})^{-1} - \mathbf{\Theta}^* = \sqrt{\mathbf{\Theta}^*}\left\{n^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathbf{I}_d\right\}\sqrt{\mathbf{\Theta}^*}$
- $$\begin{aligned}\left\|(\widehat{\mathbf{\Sigma}})^{-1} - \mathbf{\Theta}^*\right\|_{\max} &= \max_{j,k=1,\ldots,d}\left|e_j^{\mathrm{T}}\sqrt{\mathbf{\Theta}^*}\widetilde{\mathbf{\Sigma}}\sqrt{\mathbf{\Theta}^*}e_k\right| \\ &\le \max_{j,k=1,\ldots,d}\left\|\sqrt{\mathbf{\Theta}^*}e_j\right\|_1\left\|\widetilde{\mathbf{\Sigma}}\sqrt{\mathbf{\Theta}^*}e_k\right\|_\infty \\ &\le \|\widetilde{\mathbf{\Sigma}}\|_{\max}\max_{j=1,\ldots,d}\left\|\sqrt{\mathbf{\Theta}^*}e_j\right\|_1^2.\end{aligned}$$
- $$\begin{aligned}\left\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*\right\|_{\max} &\le \left\|\mathbf{Y} - \mathbf{\Theta}^*\right\|_{\max} + \left\|\mathbf{Y} - T_{v_n}(\mathbf{Y})\right\|_{\max} + \left\|\mathbf{\Lambda}^*\right\|_{\max} \\ &\le M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + v_n + \frac{\alpha}{d} \\ &\le 2M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{2\alpha}{d}\end{aligned}$$