11

Graphical models for high-dimensional data

Graphical models are based on a combination of ideas from both probability theory and graph theory, and are useful in modeling high-dimensional probability distributions. They have been developed and studied in a variety of fields, including statistical physics, spatial statistics, information and coding theory, speech processing, statistical image processing, computer vision, natural language processing, computational biology and social network analysis among others. In this chapter, we discuss various problems in high-dimensional statistics that arise in the context of graphical models.

11.1 Some basics

We begin with a brief introduction to some basic properties of graphical models, referring the reader to the bibliographic section for additional references. There are various types of graphical models, distinguished by the type of underlying graph used—directed, undirected, or a hybrid of the two. Here we focus exclusively on the case of *undirected graphical models*, also known as Markov random fields. These models are based on an undirected graph $G = (V, E)$, which consists of a set of vertices $V = \{1, 2, \ldots, d\}$ joined together by a collection of edges E. In the undirected case, an edge (j, k) is an unordered pair of distinct vertices $j, k \in V$.

In order to introduce a probabilistic aspect to our models, 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 $X = (X_1, \ldots, X_d)$. Of primary interest to us are connections between the structure of $\mathbb{P}$, and the structure of the underlying graph G. There are two ways in which to connect the probabilistic and graphical structures: one based on factorization, and the second based on conditional independence properties. A classical result in the field, known as the Hammersley–Clifford theorem, asserts that these two characterizations are essentially equivalent.

11.1.1 Factorization

One way to connect the undirected graph G to the random variables is by enforcing a certain factorization of the probability distribution. A *clique* C is a subset of vertices that are all joined by edges, meaning 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. See Figure 11.1(b) for an illustration of these concepts. 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*

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

takes inputs from the Cartesian product space $\mathcal{X}^C := \bigotimes_{j \in C} \mathcal{X}_j$, and returns non-negative real numbers. With this notation, we have the following:

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 $\psi_C : \mathcal{X}^C \to [0, \infty)$.

Here the density function is taken with respect either to the counting measure for discrete-valued random variables, or to some (possibly weighted) version of the Lebesgue measure for continuous random variables. As an illustration of Definition 11.1, any density that factorizes according to the graph shown in Figure 11.1(a) must have the form

$$p(x_1, \ldots, x_7) \propto \psi_{123}(x_1, x_2, x_3)\, \psi_{345}(x_3, x_4, x_5)\, \psi_{46}(x_4, x_6)\, \psi_{57}(x_5, x_7).$$

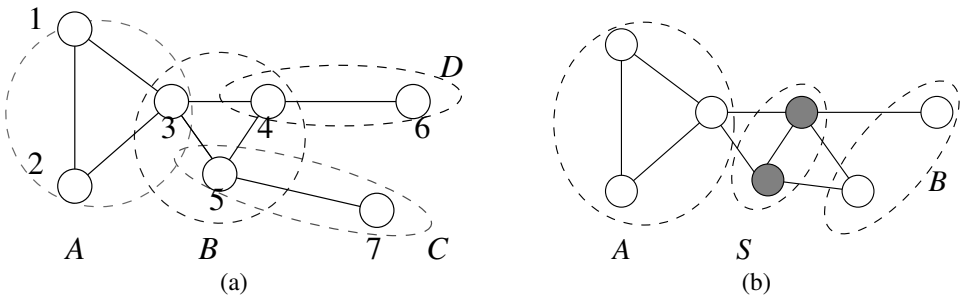

Figure 11.1 Illustration of basic graph-theoretic properties. (a) Subsets A and B are 3-cliques, whereas subsets C and D are 2-cliques. All of these cliques are maximal. Each vertex is a clique as well, but none of these singleton cliques are maximal for this graph. (b) Subset S is a vertex cutset, breaking the graph into two disconnected subgraphs with vertex sets A and B, respectively.

Without loss of generality—redefining the clique compatibility functions as necessary—the product over cliques can always be restricted to the set of all maximal cliques. However, in practice, it can be convenient to allow for terms associated with non-maximal cliques as well, as illustrated by the following.

Example 11.2 (Markov chain factorization) The standard way of factoring the distribution of a Markov chain on variables $(X_1, \ldots, X_d)$ is as

$$p(x_1, \ldots, x_d) = p_1(x_1)\ p_{2|1}(x_2 \mid x_1) \cdots p_{d|(d-1)}(x_d \mid x_{d-1}),$$

where p_1 denotes the marginal distribution of X_1, and for $j \in \{1, 2, \ldots, d-1\}$, the term $p_{j+1|j}$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

denotes the conditional distribution of X_{j+1} given X_j. This representation can be understood as a special case of the factorization (11.1), using the vertex-based functions

$$\psi_1(x_1) = p_1(x_1) \quad \text{at vertex } 1 \quad \text{and} \quad \psi_j(x_j) = 1 \quad \text{for all } j = 2, \ldots, d,$$

combined with the edge-based functions

$$\psi_{j,j+1}(x_j, x_{j+1}) = p_{j+1|j}(x_{j+1} \mid x_j) \qquad \text{for } j = 1, \ldots, d-1.$$

But this factorization is by no means unique. We could just as easily adopt the symmetrized factorization $\widetilde{\psi}_j(x_j) = p_j(x_j)$ for all $j = 1, \ldots, d$, and

$$\widetilde{\psi}_{jk}(x_j, x_k) = \frac{p_{jk}(x_j, x_k)}{p_j(x_j) p_k(x_k)} \qquad \text{for all } (j,k) \in E,$$

where p_{jk} denotes the joint distribution over the pair (X_j, X_k). ♣

Example 11.3 (Multivariate Gaussian factorization) Any non-degenerate Gaussian distribution with zero mean can be parameterized in terms of its inverse covariance matrix $\boldsymbol{\Theta}^* = \boldsymbol{\Sigma}^{-1}$, also known as the *precision matrix*. In particular, its density can be written as

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

By expanding the quadratic form, we see that

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

showing that any zero-mean Gaussian distribution can be factorized in terms of functions on edges, or cliques of size two. The Gaussian case is thus special: the factorization can always be restricted to cliques of size two, even if the underlying graph has higher-order cliques. ♣

We now turn to a non-Gaussian graphical model that shares a similar factorization:

Example 11.4 (Ising model) Consider a vector $X = (X_1, \ldots, X_d)$ of binary random variables, with each $X_j \in \{0, 1\}$. The *Ising model* is one of the earliest graphical models, first introduced in the context of statistical physics for modeling interactions in a magnetic field. Given an undirected graph $G = (V, E)$, it posits a factorization of the form

$$p(x_1, \ldots, x_d; \theta^*) = \frac{1}{Z(\theta^*)} \exp\left\{ \sum_{j \in V} \theta^*_j x_j + \sum_{(j,k) \in E} \theta^*_{jk} x_j x_k \right\}, \tag{11.3}$$

where the parameter θ^*_j is associated with vertex $j \in V$, and the parameter θ^*_{jk} is associated with edge $(j,k) \in E$. The quantity $Z(\theta^*)$ is a constant that serves to enforce that the probability mass function p normalizes properly to one; more precisely, we have

$$Z(\theta^*) = \sum_{x \in \{0,1\}^d} \exp\left\{ \sum_{j \in V} \theta^*_j x_j + \sum_{(j,k) \in E} \theta^*_{jk} x_j x_k \right\}.$$

See the bibliographic section for further discussion of the history and uses of this model. ♣

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

11.1.2 Conditional independence

We now turn to an alternative way in which to connect the probabilistic and graphical structures, involving certain conditional independence statements defined by the graph. These statements are based on the notion of a *vertex cutset* S, which (loosely stated) is a subset of vertices whose removal from the graph breaks it into two or more disjoint pieces. More formally, removing S from the vertex set V leads to the vertex-induced subgraph $G(V \setminus S)$, consisting of the vertex set $V \setminus S$, and the residual edge set

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

The set S is a vertex cutset if the residual graph $G(V \setminus S)$ consists of two or more disconnected non-empty components. See Figure 11.1(b) for an illustration.

We now define a conditional independence relationship associated with each vertex cutset of the graph. 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.

Let us consider some examples to illustrate.

Example 11.6 (Markov chain conditional independence) The Markov chain provides the simplest (and most classical) illustration of this definition. A chain graph on vertex set $V = \{1, 2, \ldots, d\}$ contains the edges $(j, j+1)$ for $j = 1, 2, \ldots, d-1$; the case $d = 5$ is illustrated in Figure 11.2(a). 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. ♣

Example 11.7 (Neighborhood-based cutsets) Another canonical type of vertex cutset is provided by the neighborhood structure of the graph. For any vertex $j \in V$, its *neighborhood set* is the subset of vertices

$$\mathcal{N}(j) := \{k \in V \mid (j,k) \in E\} \tag{11.5}$$

that are joined to j by an edge. It is easy to see that $\mathcal{N}(j)$ is always a vertex cutset, a non-trivial one as long as j is not connected to every other vertex; it separates the graph into the two disjoint components $A = \{j\}$ and $B = V \setminus (\mathcal{N}(j) \cup \{j\})$. This particular choice of vertex cutset plays an important role in our discussion of neighborhood-based methods for graphical model selection later in the chapter. ♣

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

11.1.3 Hammersley–Clifford equivalence

Thus far, we have introduced two (ostensibly distinct) ways of relating the random vector X to the underlying graph structure, namely the Markov property and the factorization property. We now turn to a fundamental theorem that establishes that these two properties are equivalent for any strictly positive distribution:

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 Here we show that the factorization property (Definition 11.1) implies the Markov property (Definition 11.5). See the bibliographic section for references to proofs of the converse. Suppose that the factorization (11.1) holds, and 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 $X_A \perp\!\!\!\perp X_B \mid X_S$. Let us define subsets of cliques by $\mathfrak{C}_A := \{C \in \mathfrak{C} \mid C \cap A \neq \emptyset\}$, $\mathfrak{C}_B := \{C \in \mathfrak{C} \mid C \cap B \neq \emptyset\}$ and $\mathfrak{C}_S := \{C \in \mathfrak{C} \mid C \subseteq S\}$. We claim that these three subsets form a disjoint partition of the full clique set—namely, $\mathfrak{C} = \mathfrak{C}_A \cup \mathfrak{C}_S \cup \mathfrak{C}_B$. Given any clique C, it is either contained entirely within S, or must have non-trivial intersection with either A or B, which proves the union property. To establish disjointedness, it is immediate that $\mathfrak{C}_S$ is disjoint from $\mathfrak{C}_A$ and $\mathfrak{C}_B$. On the other hand, if there were some clique $C \in \mathfrak{C}_A \cap \mathfrak{C}_B$, then there would exist nodes $a \in A$ and $b \in B$ with $\{a, b\} \in C$, which contradicts the fact that A and B are separated by the cutset S.

Given this disjoint partition, we may write

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

Defining the quantities

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

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

$$p(x_S) = \frac{Z_A(x_S)\, Z_B(x_S)}{Z} \Psi_S(x_S) \quad \text{and} \quad p(x_A, x_S) = \frac{Z_B(x_S)}{Z} \Psi_A(x_A, x_S)\, \Psi_S(x_S),$$

with a similar expression for $p(x_B, x_S)$. Consequently, for any x_S for which $p(x_S) > 0$, we

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

may write

$$\frac{p(x_A, x_S, x_B)}{p(x_S)} = \frac{\frac{1}{Z}\Psi_A(x_A, x_S)\Psi_S(x_S)\Psi_B(x_B, x_S)}{\frac{Z_A(x_S)\,Z_B(x_S)}{Z}\Psi_S(x_S)} = \frac{\Psi_A(x_A, x_S)\Psi_B(x_B, x_S)}{Z_A(x_S)\,Z_B(x_S)}. \tag{11.6}$$

Similar calculations yield the relations

$$\frac{p(x_A, x_S)}{p(x_S)} = \frac{\frac{Z_B(x_S)}{Z}\Psi_A(x_A, x_S)\,\Psi_S(x_S)}{\frac{Z_A(x_S)Z_B(x_S)}{Z}\Psi_S(x_S)} = \frac{\Psi_A(x_A, x_S)}{Z_A(x_S)} \tag{11.7a}$$

and

$$\frac{p(x_B, x_S)}{p(x_S)} = \frac{\frac{Z_A(x_S)}{Z}\Psi_B(x_B, x_S)\,\Psi_S(x_S)}{\frac{Z_A(x_S)Z_B(x_S)}{Z}\Psi_S(x_S)} = \frac{\Psi_B(x_B, x_S)}{Z_B(x_S)}. \tag{11.7b}$$

Combining equation (11.6) with equations (11.7a) and (11.7b) yields

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

thereby showing that $X_A \perp\!\!\!\perp X_B \mid X_S$, as claimed. □

11.1.4 Estimation of graphical models

Typical applications of graphical models require solving some sort of inverse problem of the following type. Consider a collection of samples $\{x_i\}_{i=1}^n$, where each $x_i = (x_{i1}, \ldots, x_{id})$ is a d-dimensional vector, hypothesized to have been drawn from some graph-structured probability distribution. The goal is to estimate certain aspects of the underlying graphical model. In the problem of *graphical parameter estimation*, the graph structure itself is assumed to be known, and we want to estimate the compatibility functions $\{\psi_C, C \in \mathfrak{C}\}$ on the graph cliques. In the more challenging problem of *graphical model selection*, the graph structure itself is unknown, so that we need to estimate both it *and* the clique compatibility functions. In the following sections, we consider various methods for solving these problems for both Gaussian and non-Gaussian models.

11.2 Estimation of Gaussian graphical models

We begin our exploration of graph estimation for the case of Gaussian Markov random fields. As previously discussed in Example 11.3, for a Gaussian model, the factorization property is specified by the inverse covariance or precision matrix $\mathbf{\Theta}^*$. Consequently, the Hammersley–Clifford theorem is especially easy to interpret in this case: it ensures that $\Theta^*_{jk} = 0$ for any $(j, k) \notin E$. See Figure 11.2 for some illustrations of this correspondence between graph structure and the sparsity of the inverse covariance matrix.

Now let us consider some estimation problems that arise for Gaussian Markov random fields. Since the mean is easily estimated, we take it to be zero for the remainder of our development. Thus, the only remaining parameter is the precision matrix $\mathbf{\Theta}^*$. Given an estimate $\widehat{\mathbf{\Theta}}$ of $\mathbf{\Theta}^*$, its quality can be assessed in different ways. In the problem of graphical model selection, also known as *(inverse) covariance selection*, the goal is to recover the

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

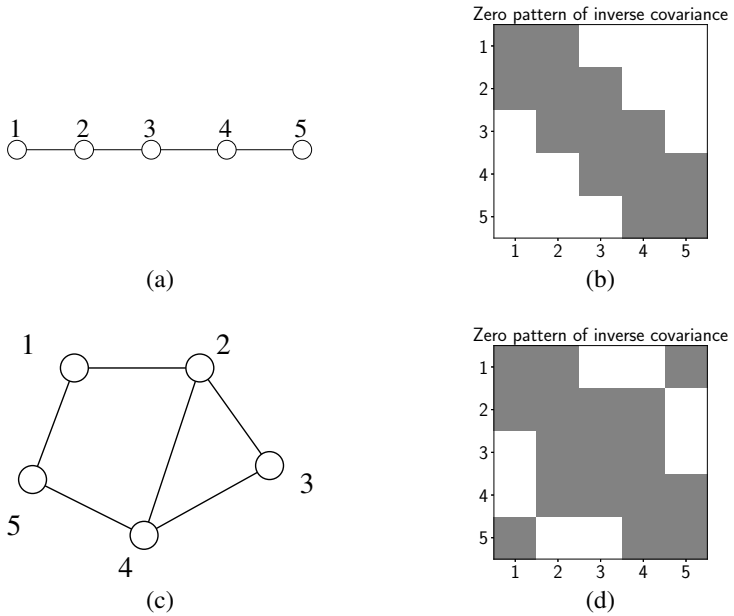

Figure 11.2 For Gaussian graphical models, the Hammersley–Clifford theorem guarantees a correspondence between the graph structure and the sparsity pattern of the inverse covariance matrix or precision matrix $\boldsymbol{\Theta}^*$. (a) Chain graph on five vertices. (b) Inverse covariance for a Gauss–Markov chain must have a tri-diagonal structure. (c), (d) More general Gauss–Markov random field and the associated inverse covariance matrix.

edge set E of the underlying graph G. More concretely, letting $\widehat{E}$ denote an estimate of the edge set based on $\widehat{\boldsymbol{\Theta}}$, one figure of merit is the error probability $\mathbb{P}[\widehat{E} \neq E]$, which assesses whether or not we have recovered the true underlying edge set. A related but more relaxed criterion would focus on the probability of recovering a fraction $1 - \delta$ of the edge set, where $\delta \in (0, 1)$ is a user-specified tolerance parameter. In other settings, we might be interested in estimating the inverse covariance matrix itself, and so consider various types of matrix norms, such as the operator norm $|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_2$ or the Frobenius norm $|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_{\mathrm{F}}$. In the following sections, we consider these different choices of metrics in more detail.

11.2.1 Graphical Lasso: ℓ_1-regularized maximum likelihood

We begin with a natural and direct method for estimating a Gaussian graphical model, namely one based on the global likelihood. In order to do so, let us first derive a convenient form of the rescaled negative log-likelihood, one that involves the log-determinant function. For any two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, recall that we use $\langle\langle \mathbf{A}, \mathbf{B} \rangle\rangle := \mathrm{trace}(\mathbf{AB})$ to denote the trace inner product. The negative log-determinant function is defined on the space $\mathcal{S}^{d \times d}$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

of symmetric matrices as

$$-\log\det(\boldsymbol{\Theta}) := \begin{cases} -\sum_{j=1}^{d} \log \gamma_j(\boldsymbol{\Theta}) & \text{if } \boldsymbol{\Theta} \succ 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{11.8}$$

where $\gamma_1(\boldsymbol{\Theta}) \geq \gamma_2(\boldsymbol{\Theta}) \geq \cdots \geq \gamma_d(\boldsymbol{\Theta})$ denote the ordered eigenvalues of the symmetric matrix Θ. In Exercise 11.1, we explore some basic properties of the log-determinant function, including its strict convexity and differentiability.

Using the parameterization (11.2) of the Gaussian distribution in terms of the precision matrix, the rescaled negative log-likelihood of the multivariate Gaussian, based on samples $\{x_i\}_{i=1}^n$, takes the form

$$\mathcal{L}_n(\boldsymbol{\Theta}) = \langle\langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}} \rangle\rangle - \log\det(\boldsymbol{\Theta}), \tag{11.9}$$

where $\widehat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n x_i x_i^{\mathrm{T}}$ is the sample covariance matrix. Here we have dropped some constant factors in the log-likelihood that have no effect on the maximum likelihood solution, and also rescaled the log-likelihood by $-\frac{2}{n}$ for later theoretical convenience.

The unrestricted maximum likelihood solution $\widehat{\boldsymbol{\Theta}}_{\mathrm{MLE}}$ takes a very simple form for the Gaussian model. If the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ is invertible, we have $\widehat{\boldsymbol{\Theta}}_{\mathrm{MLE}} = \widehat{\boldsymbol{\Sigma}}^{-1}$; otherwise, the maximum likelihood solution is undefined (see Exercise 11.2 for more details). Whenever $n < d$, the sample covariance matrix is always rank-deficient, so that the maximum likelihood estimate does not exist. In this setting, some form of regularization is essential. When the graph G is expected to have relatively few edges, a natural form of regularization is to impose an ℓ_1-constraint on the entries of $\boldsymbol{\Theta}$. (If computational considerations were not a concern, it would be natural to impose ℓ_0-constraint, but as in Chapter 7, we use the ℓ_1-norm as a convex surrogate.)

Combining ℓ_1-regularization with the negative log-likelihood yields the *graphical Lasso estimator*

$$\widehat{\boldsymbol{\Theta}} \in \arg\min_{\boldsymbol{\Theta} \in \mathcal{S}^{d\times d}} \left\{ \underbrace{\langle\langle \boldsymbol{\Theta}, \widehat{\boldsymbol{\Sigma}} \rangle\rangle - \log\det \boldsymbol{\Theta}}_{\mathcal{L}_n(\boldsymbol{\Theta})} + \lambda_n |||\boldsymbol{\Theta}|||_{1,\mathrm{off}} \right\}, \tag{11.10}$$

where $|||\boldsymbol{\Theta}|||_{1,\mathrm{off}} := \sum_{j\neq k} |\Theta_{jk}|$ corresponds to the ℓ_1-norm applied to the off-diagonal entries of $\boldsymbol{\Theta}$. One could also imagine penalizing the diagonal entries of $\boldsymbol{\Theta}$, but since they must be positive for any non-degenerate inverse covariance, doing so only introduces additional bias. The convex program (11.10) is a particular instance of a log-determinant program, and can be solved in polynomial time with various generic algorithms. Moreover, there is also a line of research on efficient methods specifically tailored to the graphical Lasso problem; see the bibliographic section for further discussion.

Frobenius norm bounds

We begin our investigation of the graphical Lasso (11.10) by deriving bounds on the Frobenius norm error $|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_{\mathrm{F}}$. The following result is based on a sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$ formed from n i.i.d. samples $\{x_i\}_{i=1}^n$ of a zero-mean random vector in which each coordinate

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

has σ-sub-Gaussian tails (recall Definition 2.2 from Chapter 2).

Proposition 11.9 (Frobenius norm bounds for graphical Lasso) *Suppose that the inverse covariance matrix $\boldsymbol{\Theta}^*$ has at most m non-zero entries per row, and we solve the graphical Lasso* (11.10) *with regularization parameter $\lambda_n = 8\sigma^2(\sqrt{\frac{\log d}{n}} + \delta)$ for some $\delta \in (0, 1]$. Then as long as $6(|||\boldsymbol{\Theta}^*|||_2 + 1)^2 \lambda_n \sqrt{md} < 1$, the graphical Lasso estimate $\widehat{\boldsymbol{\Theta}}$ satisfies*

$$|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_{\mathrm{F}}^2 \leq \frac{9}{(|||\boldsymbol{\Theta}^*|||_2 + 1)^4} m d \lambda_n^2 \tag{11.11}$$

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

Proof We prove this result by applying Corollary 9.20 from Chapter 9. In order to do so, we need to verify the restricted strong convexity of the loss function (see Definition 9.15), as well as other technical conditions given in the corollary.

Let $\mathbb{B}_F(1) = \{\boldsymbol{\Delta} \in \mathcal{S}^{d\times d} \mid |||\boldsymbol{\Delta}|||_{\mathrm{F}} \leq 1\}$ denote the set of symmetric matrices with Frobenius norm at most one. Using standard properties of the log-determinant function (see Exercise 11.1), the loss function underlying the graphical Lasso is twice differentiable, with

$$\nabla \mathcal{L}_n(\boldsymbol{\Theta}) = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Theta}^{-1} \quad \text{and} \quad \nabla^2 \mathcal{L}_n(\boldsymbol{\Theta}) = \boldsymbol{\Theta}^{-1} \otimes \boldsymbol{\Theta}^{-1},$$

where $\otimes$ denotes the Kronecker product between matrices.

Verifying restricted strong convexity: Our first step is to establish that restricted strong convexity holds over the Frobenius norm ball $\mathbb{B}_F(1)$. Let $\mathrm{vec}(\cdot)$ denote the vectorized form of a matrix. For any $\boldsymbol{\Delta} \in \mathbb{B}_F(1)$, a Taylor-series expansion yields

$$\underbrace{\mathcal{L}_n(\boldsymbol{\Theta}^* + \boldsymbol{\Delta}) - \mathcal{L}_n(\boldsymbol{\Theta}^*) - \langle\langle \nabla \mathcal{L}_n(\boldsymbol{\Theta}^*),\ \boldsymbol{\Delta} \rangle\rangle}_{\mathcal{E}_n(\boldsymbol{\Delta})} = \frac{1}{2} \mathrm{vec}(\boldsymbol{\Delta})^{\mathrm{T}} \nabla^2 \mathcal{L}_n(\boldsymbol{\Theta}^* + t\boldsymbol{\Delta}) \mathrm{vec}(\boldsymbol{\Delta})$$

for some $t \in [0, 1]$. Thus, we have

$$\mathcal{E}_n(\boldsymbol{\Delta}) \geq \frac{1}{2} \gamma_{\min}(\nabla^2 \mathcal{L}_n(\boldsymbol{\Theta}^* + t\boldsymbol{\Delta}))\, \|\mathrm{vec}(\boldsymbol{\Delta})\|_2^2 = \tfrac{1}{2} \frac{|||\boldsymbol{\Delta}|||_{\mathrm{F}}^2}{|||\boldsymbol{\Theta}^* + t\boldsymbol{\Delta}|||_2^2},$$

using the fact that $|||\mathbf{A}^{-1} \otimes \mathbf{A}^{-1}|||_2 = \frac{1}{|||\mathbf{A}|||_2^2}$ for any symmetric invertible matrix. The triangle inequality, in conjunction with the bound $t|||\boldsymbol{\Delta}|||_2 \leq t|||\boldsymbol{\Delta}|||_{\mathrm{F}} \leq 1$, implies that $|||\boldsymbol{\Theta}^* + t\boldsymbol{\Delta}|||_2^2 \leq (|||\boldsymbol{\Theta}^*|||_2 + 1)^2$. Combining the pieces yields the lower bound

$$\mathcal{E}_n(\boldsymbol{\Delta}) \geq \frac{\kappa}{2} |||\boldsymbol{\Delta}|||_{\mathrm{F}}^2 \qquad \text{where } \kappa := (|||\boldsymbol{\Theta}^*|||_2 + 1)^{-2}, \tag{11.12}$$

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: Next we introduce a subspace suitable for

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

application of Corollary 9.20 to the graphical Lasso. Letting S denote the support set of $\mathbf{\Theta}^*$, we define the subspace

$$\mathbb{M}(S) = \{\mathbf{\Theta} \in \mathbb{R}^{d \times d} \mid \Theta_{jk} = 0 \text{ for all } (j,k) \notin S\}.$$

With this choice, we have

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

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

Verifying event $\mathbb{G}(\lambda_n)$: Next we verify that the stated choice of regularization parameter λ_n satisfies the conditions of Corollary 9.20 with high probability: in order to do so, we need to compute the score function and obtain a bound on its dual norm. Since $(\mathbf{\Theta}^*)^{-1} = \mathbf{\Sigma}$, the score function is given by $\nabla \mathcal{L}_n(\mathbf{\Theta}^*) = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}$, corresponding to the deviations between the sample covariance and population covariance matrices. The dual norm defined by $|||\cdot|||_{1,\mathrm{off}}$ is given by the ℓ_∞-norm applied to the off-diagonal matrix entries, which we denote by $|||\cdot|||_{\max,\mathrm{off}}$. Using Lemma 6.26, we have

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

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 $\widehat{\mathbf{\Delta}}$ satisfies the bound $\|\widehat{\mathbf{\Delta}}_{S^c}\|_1 \leq 3\|\widehat{\mathbf{\Delta}}_S\|_1$, and hence

$$\|\widehat{\mathbf{\Delta}}\|_1 \leq 4\|\widehat{\mathbf{\Delta}}_S\|_1 \leq 4\sqrt{md}\,|||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}}, \tag{11.13}$$

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 $\widehat{\mathbf{\Delta}}$ belongs to the Frobenius ball $\mathbb{B}_F(1)$.

Localizing the error matrix: By an argument parallel to the earlier proof of RSC, we have

$$\mathcal{L}_n(\mathbf{\Theta}^*) - \mathcal{L}_n(\mathbf{\Theta}^* + \mathbf{\Delta}) + \langle\!\langle \nabla \mathcal{L}_n(\mathbf{\Theta}^* + \mathbf{\Delta}),\ -\mathbf{\Delta} \rangle\!\rangle \geq \frac{\kappa}{2}|||\mathbf{\Delta}|||_{\mathrm{F}}^2.$$

Adding this lower bound to the inequality (11.12), we find that

$$\langle\!\langle \nabla \mathcal{L}_n(\mathbf{\Theta}^* + \mathbf{\Delta}) - \nabla \mathcal{L}_n(\mathbf{\Theta}^*),\ \mathbf{\Delta} \rangle\!\rangle \geq \kappa\,|||\mathbf{\Delta}|||_{\mathrm{F}}^2.$$

The result of Exercise 9.10 then implies that

$$\langle\!\langle \nabla \mathcal{L}_n(\mathbf{\Theta}^* + \mathbf{\Delta}) - \nabla \mathcal{L}_n(\mathbf{\Theta}^*),\ \mathbf{\Delta} \rangle\!\rangle \geq \kappa\,|||\mathbf{\Delta}|||_{\mathrm{F}} \qquad \text{for all } \mathbf{\Delta} \in \mathcal{S}^{d \times d} \setminus \mathbb{B}_F(1). \tag{11.14}$$

By the optimality of $\widehat{\mathbf{\Theta}}$, we have $0 = \langle\!\langle \nabla \mathcal{L}_n(\mathbf{\Theta}^* + \widehat{\mathbf{\Delta}}) + \lambda_n \widehat{\mathbf{Z}},\ \widehat{\mathbf{\Delta}} \rangle\!\rangle$, where $\widehat{\mathbf{Z}} \in \partial |||\widehat{\mathbf{\Theta}}|||_{1,\mathrm{off}}$ is a subgradient matrix for the elementwise ℓ_1-norm. By adding and subtracting terms, we find that

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

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

Since $\|\nabla \mathcal{L}_n(\mathbf{\Theta}^*)\|_{\max} \leq \frac{\lambda_n}{2}$ under the previously established event $\mathbb{G}(\lambda_n)$, the right-hand side is at most

$$\frac{3\lambda_n}{2}\|\widehat{\mathbf{\Delta}}\|_1 \leq 6\lambda_n \sqrt{md}\, |||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}},$$

where we have applied our earlier inequality (11.13). If $|||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}} > 1$, then our earlier lower bound (11.14) may be applied, from which we obtain

$$\kappa |||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}} \leq \frac{3\lambda_n}{2}\|\widehat{\mathbf{\Delta}}\|_1 \leq 6\lambda_n \sqrt{md}\, |||\widehat{\mathbf{\Delta}}|||_{\mathrm{F}}.$$

This inequality leads to a contradiction whenever $\frac{6\lambda_n\sqrt{md}}{\kappa} < 1$, which completes the proof. □

Edge selection and operator norm bounds

Proposition 11.9 is a relatively crude result, in that it only guarantees that the graphical Lasso estimate $\widehat{\mathbf{\Theta}}$ is close in Frobenius norm, but not that the edge structure of the underlying graph is preserved. Moreover, the result actually precludes the setting $n < d$: indeed, the conditions of Proposition 11.9 imply that the sample size n must be lower bounded by a constant multiple of $md \log d$, which is larger than d.

Accordingly, we now turn to a more refined type of result, namely one that allows for high-dimensional scaling ($d \gg n$), and moreover guarantees that the graphical Lasso estimate $\widehat{\mathbf{\Theta}}$ correctly selects all the edges of the graph. Such an edge selection result can be guaranteed by first proving that $\widehat{\mathbf{\Theta}}$ is close to the true precision matrix $\mathbf{\Theta}^*$ in the elementwise ℓ_∞-norm on the matrix elements (denoted by $\|\cdot\|_{\max}$). In turn, such max-norm control can also be converted to bounds on the ℓ_2-matrix operator norm, also known as the spectral norm.

The problem of edge selection in a Gaussian graphical model is closely related to the problem of variable selection in a sparse linear model. As previously discussed in Chapter 7, variable selection with an ℓ_1-norm penalty requires a certain type of incoherence condition, which limits the influence of irrelevant variables on relevant ones. In the case of least-squares regression, these incoherence conditions were imposed on the design matrix, or equivalently on the Hessian of the least-squares objective function. Accordingly, in a parallel manner, here we impose incoherence conditions on the Hessian of the objective function $\mathcal{L}_n$ in the graphical Lasso (11.10). As previously noted, this Hessian takes the form $\nabla^2 \mathcal{L}_n(\mathbf{\Theta}) = \mathbf{\Theta}^{-1} \otimes \mathbf{\Theta}^{-1}$, a $d^2 \times d^2$ matrix that is indexed by ordered pairs of vertices (j, k).

More specifically, the incoherence condition must be satisfied by the d^2-dimensional matrix $\mathbf{\Gamma}^* := \nabla^2 \mathcal{L}_n(\mathbf{\Theta}^*)$, corresponding to the Hessian evaluated at the true precision matrix. We use $S := E \cup \{(j, j) \mid j \in V\}$ to denote the set of row/column indices associated with edges in the graph (including both (j, k) and (k, j)), along with all the self-edges (j, j). Letting $S^c = (V \times V) \setminus S$, we say that the matrix $\mathbf{\Gamma}^*$ is *α-incoherent* if

$$\max_{e \in S^c} \|\Gamma^*_{eS}(\mathbf{\Gamma}^*_{SS})^{-1}\|_1 \leq 1 - \alpha \qquad \text{for some } \alpha \in (0, 1]. \tag{11.15}$$

With this definition, we have the following result:

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

Proposition 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 $n > c_0(1 + 8\alpha^{-1})^2 m^2 \log d$, 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*

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

Note that part (a) guarantees that the edge set estimate

$$\widehat{E} := \{(j, k) \in [d] \times [d] \mid j < k \text{ and } \widehat{\mathbf{\Theta}}_{jk} \neq 0\}$$

is always a subset of the true edge set E. Part (b) guarantees that $\widehat{\mathbf{\Theta}}$ is uniformly close to $\mathbf{\Theta}^*$ in an elementwise sense. Consequently, if we have a lower bound on the minimum non-zero entry of $|\mathbf{\Theta}^*|$—namely the quantity $\tau^*(\mathbf{\Theta}^*) = \min_{(j,k)\in E} |\Theta^*_{jk}|$—then we can guarantee that the graphical Lasso recovers the full edge set correctly. In particular, using the notation of part (b), as long as this minimum is lower bounded as $\tau^*(\mathbf{\Theta}^*) > c_4(\tau(n, d, \alpha) + \lambda_n)$, then the graphical Lasso recovers the correct edge set with high probability.

The proof of Proposition 11.10 is based on an extension of the primal–dual witness technique used to prove Theorem 7.21 in Chapter 7. In particular, it involves constructing a pair of matrices $(\widetilde{\mathbf{\Theta}}, \widetilde{\mathbf{Z}})$, where $\widetilde{\mathbf{\Theta}} \succ 0$ is a primal optimal solution and $\widetilde{\mathbf{Z}}$ a corresponding dual optimum. This pair of matrices is required to satisfy the zero subgradient conditions that define the optimum of the graphical Lasso (11.10)—namely

$$\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Theta}}^{-1} + \lambda_n \widetilde{\mathbf{Z}} = 0 \quad \text{or equivalently} \quad \widetilde{\mathbf{\Theta}}^{-1} = \widehat{\mathbf{\Sigma}} + \lambda_n \widetilde{\mathbf{Z}}.$$

The matrix $\widetilde{\mathbf{Z}}$ must belong to the subgradient of the $|||\cdot|||_{1,\text{off}}$ function, evaluated at $\widetilde{\mathbf{\Theta}}$, meaning that $|||\widetilde{\mathbf{Z}}|||_{\max,\text{off}} \leq 1$, and that $\widetilde{Z}_{jk} = \text{sign}(\widetilde{\Theta}_{jk})$ whenever $\widetilde{\Theta}_{jk} \neq 0$. We refer the reader to the bibliographic section for further details and references for the proof.

Proposition 11.10 also implies bounds on the operator norm error in the estimate $\widehat{\mathbf{\Theta}}$.

Corollary 11.11 (Operator norm bounds) *Under the conditions of Proposition 11.10, consider the graphical Lasso estimate $\widehat{\mathbf{\Theta}}$ with regularization parameter $\lambda_n = \frac{c_1}{\alpha}\sqrt{\frac{\log d}{n}} + \delta$*

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

for some $\delta \in (0, 1]$. Then with probability at least $1 - c_2 e^{-c_3 n \delta^2}$, we have

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

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

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

Proof These claims follow in a straightforward way from Proposition 11.10 and certain properties of the operator norm exploited previously in Chapter 6. In particular, Proposition 11.10 guarantees that for any pair $(j, k) \notin E$, we have $|\widehat{\Theta}_{jk} - \Theta^*_{jk}| = 0$, whereas the bound (11.16) ensures that for any pair $(j, k) \in E$, we have $|\widehat{\Theta}_{jk} - \Theta^*_{jk}| \leq c_4 \{\tau(n, d, \alpha) + \lambda_n\}$. Note that the same bound holds whenever $j = k$. Putting together the pieces, we conclude that

$$|\widehat{\Theta}_{jk} - \Theta^*_{jk}| \leq c_4 \{\tau(n, d, \alpha) + \lambda_n\} \, A_{jk}, \tag{11.18}$$

where $\mathbf{A}$ is the adjacency matrix, including ones on the diagonal. Using the matrix-theoretic properties from Exercise 6.3(c), we conclude that

$$|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_2 \leq |||\,|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|\,|||_2 \leq c_4 \{\tau(n, d, \alpha) + \lambda_n\} \, |||\mathbf{A}|||_2,$$

thus establishing the bound (11.17a). The second inequality (11.17b) follows by noting that $|||\mathbf{A}|||_2 \leq m+1$ for any graph of degree at most m. (See the discussion following Corollary 6.24 for further details.) □

As we noted in Chapter 6, the bound (11.17b) is not tight for a general graph with maximum degree m. In particular, a star graph with one hub connected to m other nodes (see Figure 6.1(b)) has maximum degree m, but satisfies $|||\mathbf{A}|||_2 = 1 + \sqrt{m-1}$, so that the bound (11.17a) implies the operator norm bound $|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_2 \precsim \sqrt{\frac{m \log d}{n}}$. This guarantee is tighter by a factor of $\sqrt{m}$ than the conservative bound (11.17b).

It should also be noted that Proposition 11.10 also implies bounds on the Frobenius norm error. In particular, the elementwise bound (11.18) implies that

$$|||\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*|||_{\mathrm{F}} \leq c_3 \sqrt{2s + d} \left\{ (1 + 8\alpha^{-1}) \sqrt{\frac{\log d}{n}} + \lambda_n \right\}, \tag{11.19}$$

where s is the total number of edges in the graph. We leave the verification of this claim as an exercise for the reader.

11.2.2 Neighborhood-based methods

The Gaussian graphical Lasso is a global method, one that estimates the full graph simultaneously. An alternative class of procedures, known as neighborhood-based methods, are instead local. They are based on the observation that recovering the full graph is equivalent

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

to recovering the neighborhood set (11.5) of each vertex $j \in V$, and that these neighborhoods are revealed via the Markov properties of the graph.

Neighborhood-based regression

Recall our earlier Definition 11.5 of the Markov properties associated with a graph. In our discussion following this definition, we also noted that for any given vertex $j \in V$, the neighborhood $\mathcal{N}(j)$ is a vertex cutset that breaks the graph into the disjoint pieces $\{j\}$ and $V \backslash \mathcal{N}^+(j)$, where we have introduced the convenient shorthand $\mathcal{N}^+(j) := \{j\} \cup \mathcal{N}(j)$. Consequently, by applying the definition (11.5), we conclude that

$$X_j \perp\!\!\!\perp X_{V \backslash \mathcal{N}^+(j)} \mid X_{\mathcal{N}(j)}. \tag{11.20}$$

Thus, the neighborhood structure of each node is encoded in the structure of the conditional distribution. What is a good way to detect these conditional independence relationships and hence the neighborhood? A particularly simple method is based on the idea of neighborhood regression: for a given vertex $j \in V$, we use the random variables $X_{\backslash\{j\}} := \{X_k \mid k \in V \setminus \{j\}\}$ to predict X_j, and keep only those variables that turn out to be useful.

Let us now formalize this idea in the Gaussian case. In this case, by standard properties of multivariate Gaussian distributions, the conditional distribution of X_j given $X_{\backslash\{j\}}$ is also Gaussian. Therefore, the random variable X_j has a decomposition as the sum of the best linear prediction based on $X_{\backslash\{j\}}$ plus an error term—namely

$$X_j = \left\langle X_{\backslash\{j\}}, \theta_j^* \right\rangle + W_j, \tag{11.21}$$

where $\theta_j^* \in \mathbb{R}^{d-1}$ is a vector of regression coefficients, and W_j is a zero-mean Gaussian variable, independent of $X_{\backslash\{j\}}$. (See Exercise 11.3 for the derivation of these and related properties.) Moreover, the conditional independence relation (11.20) ensures that $\theta_{jk}^* = 0$ for all $k \notin \mathcal{N}(j)$. In this way, we have reduced the problem of Gaussian graph selection to that of detecting the support in a sparse linear regression problem. As discussed in Chapter 7, the Lasso provides a computationally efficient approach to such support recovery tasks.

In summary, the neighborhood-based approach to Gaussian graphical selection proceeds as follows. Given n samples $\{x_1, \ldots, x_n\}$, we use $\mathbf{X} \in \mathbb{R}^{n \times d}$ to denote the design matrix with $x_i \in \mathbb{R}^d$ as its ith row, and then perform the following steps.

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 $\mathbf{X}_{\backslash\{j\}} \in \mathbb{R}^{n \times (d-1)}$.

(b) Solve the Lasso problem:

$$\widehat{\theta} = \arg\min_{\theta \in \mathbb{R}^{d-1}} \left\{ \frac{1}{2n} \|X_j - \mathbf{X}_{\backslash\{j\}}\theta\|_2^2 + \lambda_n \|\theta\|_1 \right\}. \tag{11.22}$$

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

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

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

Note that the first step returns a neighborhood estimate $\widehat{N}(j)$ for each vertex $j \in V$. These neighborhood estimates may be inconsistent, meaning that for a given pair of distinct vertices (j, k), it may be the case that $k \in \widehat{N}(j)$ whereas $j \notin \widehat{N}(k)$. Some rules to resolve this issue include:

- the OR rule that declares that $(j, k) \in \widehat{E}_{\text{OR}}$ if either $k \in \widehat{N}(j)$ or $j \in \widehat{N}(k)$;
- the AND rule that declares that $(j, k) \in \widehat{E}_{\text{AND}}$ if $k \in \widehat{N}(j)$ and $j \in \widehat{N}(k)$.

By construction, the AND rule is more conservative than the OR rule, meaning that $\widehat{E}_{\text{AND}} \subseteq \widehat{E}_{\text{OR}}$. The theoretical guarantees that we provide end up holding for either rule, since we control the behavior of each neighborhood regression problem.

Graph selection consistency

We now state a result that guarantees selection consistency of neighborhood regression. As with our previous analysis of the Lasso in Chapter 7 and the graphical Lasso in Section 11.2.1, we require an incoherence condition. Given a positive definite matrix $\mathbf{\Gamma}$ and a subset S of its columns, we say $\mathbf{\Gamma}$ is α-incoherent with respect to S if

$$\max_{k \notin S} \|\Gamma_{kS}(\mathbf{\Gamma}_{SS})^{-1}\|_1 \leq 1 - \alpha. \tag{11.23}$$

Here the scalar $\alpha \in (0, 1]$ is the incoherence parameter. As discussed in Chapter 7, if we view $\mathbf{\Gamma}$ as the covariance matrix of a random vector $Z \in \mathbb{R}^d$, then the row vector $\Gamma_{kS}(\mathbf{\Gamma}_{SS})^{-1}$ specifies the coefficients of the optimal linear predictor of Z_k given the variables $Z_S := \{Z_j, j \in S\}$. Thus, the incoherence condition (11.23) imposes a limit on the degree of dependence between the variables in the correct subset S and any variable outside of S.

The following result guarantees graph selection consistency of the Lasso-based neighborhood procedure, using either the AND or the OR rules, for a Gauss–Markov random field in which the covariance matrix $\mathbf{\Sigma}^* = (\mathbf{\Theta}^*)^{-1}$ has maximum degree m, and diagonals scaled such that $\text{diag}(\mathbf{\Sigma}^*) \leq 1$. This latter inequality entails no loss of generality, since it can always be guaranteed by rescaling the variables. Our statement involves the ℓ_∞-matrix-operator norm $|||A|||_2 := \max_{i=1,\ldots,d} \sum_{j=1}^d |A_{ij}|$.

Finally, in stating the result, we assume that the sample size is lower bounded as $n \gtrsim m \log d$. This assumption entails no loss of generality, because a sample size of this order is actually necessary for any method. See the bibliographic section for further details on such information-theoretic lower bounds for graphical model selection.

Theorem 11.12 (Graph selection consistency) *Consider a zero-mean Gaussian random vector with covariance $\mathbf{\Sigma}^*$ such that for each $j \in V$, the submatrix $\mathbf{\Sigma}^*_{\backslash\{j\}} := \text{cov}(\mathbf{X}_{\backslash\{j\}})$ is α-incoherent with respect to $N(j)$, and $|||(\mathbf{\Sigma}^*_{N(j),N(j)})^{-1}|||_\infty \leq b$ for some $b \geq 1$. Suppose that the neighborhood Lasso selection method is implemented with $\lambda_n = c_0\{\frac{1}{\alpha}\sqrt{\frac{\log d}{n}} + \delta\}$*

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

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 $|\Theta^*_{jk}| \geq 7\, b\, \lambda_n$.*

Of course, if the non-zero entries of the precision matrix are bounded below in absolute value as $\min_{(j,k)\in E} |\Theta^*_{jk}| > 7\, b\, \lambda_n$, then in fact Theorem 11.12 guarantees that $\widehat{E} = E$ with high probability.

Proof It suffices to show that for each $j \in V$, the neighborhood $\mathcal{N}(j)$ is recovered with high probability; we can then apply the union bound over all the vertices. The proof requires an extension of the primal–dual witness technique used to prove Theorem 7.21. The main difference is that Theorem 11.12 applies to random covariates, as opposed to the case of deterministic design covered by Theorem 7.21. In order to reduce notational overhead, we adopt the shorthand $\mathbf{\Gamma}^* = \operatorname{cov}(X_{\backslash\{j\}})$ along with the two subsets $S = \mathcal{N}(j)$ and $S^c = V \setminus \mathcal{N}^+(j)$. In this notation, we can write our observation model as $X_j = \mathbf{X}_{\backslash\{j\}}\theta^* + W_j$, where $\mathbf{X}_{\backslash\{j\}} \in \mathbb{R}^{n\times(d-1)}$ while X_j and W_j are both n-vectors. In addition, we let $\widehat{\mathbf{\Gamma}} = \frac{1}{n}\mathbf{X}^{\mathrm{T}}_{\backslash\{j\}}\mathbf{X}_{\backslash\{j\}}$ denote the sample covariance defined by the design matrix, and we use $\widehat{\mathbf{\Gamma}}_{SS}$ to denote the submatrix indexed by the subset S, with the submatrix $\widehat{\mathbf{\Gamma}}_{S^cS}$ defined similarly.

Proof of part (a): We follow the proof of Theorem 7.21 until equation (7.53), namely

$$\widehat{z}_{S^c} = \underbrace{\widehat{\mathbf{\Gamma}}_{S^cS}(\widehat{\mathbf{\Gamma}}_{SS})^{-1}\widehat{z}_S}_{\mu\in\mathbb{R}^{d-s}} + \underbrace{\mathbf{X}^{\mathrm{T}}_{S^c}\left[\mathbf{I}_n - \mathbf{X}_S(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S)^{-1}\mathbf{X}_S^{\mathrm{T}}\right]\left(\frac{W_j}{\lambda_n\, n}\right)}_{V_{S^c}\in\mathbb{R}^{d-s}}. \tag{11.24}$$

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 $\|\widehat{z}_{S^c}\|_\infty < 1$. We do so by establishing that

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

and

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

Taken together, these bounds ensure that $\|\widehat{z}_{S^c}\|_\infty \leq 1 - \frac{\alpha}{2} < 1$, 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 $1 - c_1 e^{-c_2 n\delta^2\alpha^2}$.

Let us begin by establishing the bound (11.25a). By standard properties of multivariate Gaussian vectors, we can write

$$\mathbf{X}^{\mathrm{T}}_{S^c} = \mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{X}_S + \tilde{\mathbf{W}}^{\mathrm{T}}_{S^c}, \tag{11.26}$$

where $\tilde{\mathbf{W}}_{S^c} \in \mathbb{R}^{n\times|S^c|}$ is a zero-mean Gaussian random matrix that is independent of $\mathbf{X}_S$.

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

Observe moreover that

$$\operatorname{cov}(\tilde{\mathbf{W}}_{S^c}) = \mathbf{\Gamma}^*_{S^cS^c} - \mathbf{\Gamma}^*_{S^cS}(\mathbf{\Gamma}^*_{SS})^{-1}\mathbf{\Gamma}^*_{SS^c} \preceq \mathbf{\Gamma}^*.$$

Recalling our assumption that $\operatorname{diag}(\mathbf{\Gamma}^*) \leq 1$, we see that the elements of $\tilde{\mathbf{W}}_{S^c}$ have variance at most 1.

Using the decomposition (11.26) and the triangle inequality, we have

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

where step (i) uses the population-level α-incoherence condition. Turning to the remaining stochastic term, conditioned on the design matrix, the vector $\widetilde{V}$ is a zero-mean Gaussian random vector, each entry of which has standard deviation at most

$$\begin{aligned} \frac{1}{\sqrt{n}} \|\frac{\mathbf{X}_S}{\sqrt{n}} (\widehat{\mathbf{\Gamma}}_{SS})^{-1}\widehat{z}_S\|_2 &\leq \frac{1}{\sqrt{n}} |||\frac{\mathbf{X}_S}{\sqrt{n}} (\widehat{\mathbf{\Gamma}}_{SS})^{-1}|||_2 \, \|\widehat{z}_S\|_2 \\ &\leq \frac{1}{\sqrt{n}} \sqrt{|||(\widehat{\mathbf{\Gamma}}_{SS})^{-1}|||_2} \; \sqrt{m} \\ &\overset{\text{(i)}}{\leq} 2\sqrt{\frac{b\,m}{n}}, \end{aligned}$$

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

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

We now set $t = \big[\frac{64bm\log d}{n} + \frac{1}{64}\alpha^2\big]^{1/2}$, 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.

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

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

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 $\|\mathbf{\Pi}W_j\|_2 \leq \|W_j\|_2$. The vector $W_j \in \mathbb{R}^n$ has i.i.d. Gaussian entries with variance at most 1, and hence the event $\mathcal{E} = \{\frac{\|W_j\|_2}{\sqrt{n}} \leq 2\}$ holds with probability at least $1 - 2e^{-n}$. Conditioning on this event and its complement, we find that

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

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

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

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

where we have combined the union bound with standard Gaussian tail bounds. Since $\lambda_n = c_0\{\frac{1}{\alpha}\sqrt{\frac{\log d}{n}} + \delta\}$ 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

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

Proof of part (b): In order to prove part (b) of the theorem, it suffices to establish ℓ_∞-bounds on the error in the Lasso solution. Here we provide a proof in the case $m \leq \log d$, referring the reader to the bibliographic section for discussion of the general case. Again returning to the proof of Theorem 7.21, equation (7.54) guarantees that

$$\begin{aligned}\|\widehat{\theta}_S - \theta^*_S\|_\infty &\leq \left\|(\widehat{\mathbf{\Gamma}}_{SS})^{-1}\mathbf{X}_S^{\mathrm{T}}\frac{W_j}{n}\right\|_\infty + \lambda_n |||(\widehat{\mathbf{\Gamma}}_{SS})^{-1}|||_\infty \\ &\leq \left\|(\widehat{\mathbf{\Gamma}}_{SS})^{-1}\mathbf{X}_S^{\mathrm{T}}\frac{W_j}{n}\right\|_\infty + \lambda_n \left\{|||(\widehat{\mathbf{\Gamma}}_{SS})^{-1} - (\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty + |||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty\right\}. \qquad (11.28)\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

Now the vector $W_j \in \mathbb{R}^n$ has i.i.d. Gaussian entries, each zero-mean with variance at most

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

$\mathrm{var}(X_j) \leq 1$, 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

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

where we have combined the assumed bound $|||(\mathbf{\Gamma}^*_{SS})^{-1}|||_\infty \leq b$ with the inequality (11.30). Therefore, the union bound combined with Gaussian tail bounds implies that

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

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 $\|\widehat{\theta}_S - \theta^*_S\|_\infty \leq 7b\lambda_n$ 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. □

11.3 Graphical models in exponential form

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

$$p_{\mathbf{\Theta}^*}(x_1, \ldots, x_d) \propto \exp\left\{\sum_{j\in V} \phi_j(x_j; \Theta^*_j) + \sum_{(j,k)\in E} \phi_{jk}(x_j, x_k; \mathbf{\Theta}^*_{jk})\right\}, \tag{11.32}$$

where Θ^*_j is a vector of parameters for node $j \in V$, and $\mathbf{\Theta}^*_{jk}$ is a matrix of parameters for edge (j, k). For instance, the Gaussian graphical model is a special case in which $\Theta^*_j = \theta^*_j$ and $\mathbf{\Theta}^*_{jk} = \theta^*_{jk}$ are both scalars, the potential functions take the form

$$\phi_j(x_j; \theta^*_j) = \theta^*_j x_j, \qquad \phi_{jk}(x_j, x_k; \theta^*_{jk}) = \theta^*_{jk} x_j x_k, \tag{11.33}$$

and the density (11.32) is taken with respect to Lebesgue measure over $\mathbb{R}^d$. The Ising model (11.3) is another special case, using the same choice of potential functions (11.33), but taking the density with respect to the counting measure on the binary hypercube $\{0, 1\}^d$.

Let us consider a few more examples of this factorization:

Example 11.13 (Potts model) The *Potts model*, in which each variable X_s takes values in the discrete set $\{0, \ldots, M-1\}$ is another special case of the factorization (11.32). In this case, the parameter $\Theta^*_j = \{\Theta_{j;a}, a = 1, \ldots, M-1\}$ is an $(M-1)$-vector, whereas the parameter $\mathbf{\Theta}^*_{jk} = \{\Theta^*_{jk;ab}, a, b = 1, \ldots, M-1\}$ is an $(M-1) \times (M-1)$ matrix. The potential functions take the form

$$\phi_j(x_j; \Theta^*_j) = \sum_{a=1}^{M-1} \Theta^*_{j;a}\, \mathbb{I}[x_j = a] \tag{11.34a}$$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

and

$$\phi_{jk}(x_j, x_k; \mathbf{\Theta}^*_{jk}) = \sum_{a=1}^{M-1} \sum_{b=1}^{M-1} \Theta^*_{jk;ab} \, \mathbb{I}[x_j = a, x_k = b]. \tag{11.34b}$$

Here $\mathbb{I}[x_j = a]$ is a zero–one indicator function for the event that $\{x_j = a\}$, with the indicator function $\mathbb{I}[x_j = a, x_k = b]$ defined analogously. Note that the Potts model is a generalization of the Ising model (11.3), to which it reduces for variables taking $M = 2$ states. ♣

Example 11.14 (Poisson graphical model) Suppose that we are interested in modeling a collection of random variables $(X_1, \ldots, X_d)$, each of which represents some type of count data taking values in the set of positive integers $\mathbb{Z}_+ = \{0, 1, 2, \ldots\}$. One way of building a graphical model for such variables is by specifying the conditional distribution of each variable given its neighbors. In particular, suppose that variable X_j, when conditioned on its neighbors, is a Poisson random variable with mean

$$\mu_j = \exp\left(\theta^*_j + \sum_{k \in N(j)} \theta^*_{jk} x_k\right).$$

This form of conditional distribution leads to a Markov random field of the form (11.32) with

$$\phi_j(x_j; \theta^*_j) = \theta^*_j x_j - \log(x!) \qquad \text{for all } j \in V, \tag{11.35a}$$

$$\phi_{jk}(x_j, x_k; \theta^*_{jk}) = \theta^*_{jk} x_j x_k \qquad \text{for all } (j,k) \in E. \tag{11.35b}$$

Here the density is taken with respect to the counting measure on $\mathbb{Z}_+$ for all variables. A potential deficiency of this model is that, in order for the density to be normalizable, we must necessarily have $\theta^*_{jk} \leq 0$ for all $(j,k) \in E$. Consequently, this model can only capture competitive interactions between variables. ♣

One can also consider various types of mixed graphical models, for instance in which some of the nodes take discrete values, whereas others are continuous-valued. Gaussian mixture models are one important class of such models.

11.3.1 A general form of neighborhood regression

We now consider a general form of neighborhood regression, applicable to any graphical model of the form (11.32). Let $\{x_i\}_{i=1}^n$ be a collection of n samples drawn i.i.d. from such a graphical model; here each x_i is a d-vector. Based on these samples, we can form a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ with x_i^{T} as the ith row. For $j = 1, \ldots, d$, we let $X_j \in \mathbb{R}^n$ denote the jth 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 $\mathbf{X}_{\backslash\{j\}} \in \mathbb{R}^{n \times (d-1)}$. As we show in Exercise 11.6, for any distribution of the form (11.32), this conditional likelihood depends only on the vector of parameters

$$\mathbf{\Theta}_{j+} := \{\Theta_j, \mathbf{\Theta}_{jk}, \, k \in V \setminus \{j\}\} \tag{11.36}$$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

that involve node j. 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 $\mathbf{\Theta}_{j+}$. Letting $|||\cdot|||$ denote some matrix norm, we arrive at a general form of neighborhood regression:

$$\widehat{\mathbf{\Theta}}_{j+} = \arg\min_{\mathbf{\Theta}_{j+}} \Big\{ \underbrace{-\frac{1}{n}\sum_{i=1}^{n} \log p_{\mathbf{\Theta}_{j+}}(x_{ij} \mid x_{i\setminus\{j\}})}_{\mathcal{L}_n(\mathbf{\Theta}_{j+};\, x_j, x_{\setminus\{j\}})} + \lambda_n \sum_{k \in V\setminus\{j\}} |||\mathbf{\Theta}_{jk}||| \Big\}. \tag{11.37}$$

This formulation actually describes a family of estimators, depending on which norm $|||\cdot|||$ that we impose on each matrix component $\mathbf{\Theta}_{jk}$. Perhaps the simplest is the Frobenius norm, in which case the estimator (11.37) is a general form of the group Lasso; for details, see equation (9.66) and the associated discussion in Chapter 9. Also, as we verify in Exercise 11.5, this formula reduces to ℓ_1-regularized linear regression (11.22) in the Gaussian case.

11.3.2 Graph selection for Ising models

In this section, we consider the graph selection problem for a particular type of non-Gaussian distribution, namely the Ising model. Recall that the Ising distribution is over binary variables, and takes the form

$$p_{\theta^*}(x_1, \ldots, x_d) \propto \exp\left\{ \sum_{j \in V} \theta^*_j x_j + \sum_{(j,k) \in E} \theta^*_{jk} x_j x_k \right\}. \tag{11.38}$$

Since there is only a single parameter per edge, imposing an ℓ_1-penalty suffices to encourage sparsity in the neighborhood regression. For any given node $j \in V$, we define the subset of coefficients associated with it—namely, the set

$$\theta_{j+} := \big\{ \theta_j, \theta_{jk}, k \in V \setminus \{j\} \big\}.$$

For the Ising model, the neighborhood regression estimate reduces to a form of logistic regression—specifically

$$\widehat{\theta}_{j+} = \arg\min_{\theta_{j+} \in \mathbb{R}^d} \Big\{ \underbrace{\frac{1}{n}\sum_{i=1}^{n} f\Big(\theta_j x_{ij} + \sum_{k \in V\setminus\{j\}} \theta_{jk} x_{ij} x_{ik}\Big)}_{\mathcal{L}_n(\theta_{j+};\, x_j, x_{\setminus\{j\}})} + \lambda_n \sum_{k \in V\setminus\{j\}} |\theta_{jk}| \Big\}, \tag{11.39}$$

where $f(t) = \log(1 + e^t)$ is the logistic function. See Exercise 11.7 for details.

Under what conditions does the estimate (11.39) recover the correct neighborhood set $\mathcal{N}(j)$? As in our earlier analysis of neighborhood linear regression and the graphical Lasso, such a guarantee requires some form of incoherence condition, limiting the influence of irrelevant variables—those outside $\mathcal{N}(j)$—on variables inside the set. Recalling the cost function $\mathcal{L}_n$ in the optimization problem (11.39), let θ^*_{j+} denote the minimizer of the population objective function $\bar{\mathcal{L}}(\theta_{j+}) = \mathbb{E}[\mathcal{L}_n(\theta_{j+};\ X_j, \mathbf{X}_{\setminus\{j\}})]$. We then consider the Hessian of the cost function $\bar{\mathcal{L}}$ evaluated at the "true parameter" θ^*_{j+}—namely, the d-dimensional matrix $\mathbf{J} := \nabla^2 \bar{\mathcal{L}}(\theta^*_{j+})$. For a given $\alpha \in (0, 1]$, we say that $\mathbf{J}$ satisfies an α-incoherence condition at

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

node $j \in V$ if

$$\max_{k \notin S} \|J_{kS}(\mathbf{J}_{SS})^{-1}\|_1 \leq 1 - \alpha, \tag{11.40}$$

where we have introduced the shorthand $S = \mathcal{N}(j)$ for the neighborhood set of node j. In addition, we assume the submatrix $\mathbf{J}_{SS}$ has its smallest eigenvalue lower bounded by some $c_{\min} > 0$. With this set-up, the following result applies to an Ising model (11.38) defined on a graph G with d vertices and maximum degree at most m, with Fisher information $\mathbf{J}$ at node j satisfying the $c_{\min}$-eigenvalue bound, and the α-incoherence condition (11.40).

Theorem 11.15 *Given n i.i.d. samples with $n > c_0 m^2 \log d$, consider the estimator* (11.39) *with $\lambda_n = \frac{32}{\alpha}\sqrt{\frac{\log d}{n}} + \delta$ for some $\delta \in [0, 1]$. Then with probability at least $1 - c_1 e^{-c_2(n\delta^2 + \log d)}$, 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 $\|\widehat{\theta}_{j+} - \theta^*_{j+}\|_\infty \leq \frac{c_3}{c_{\min}}\sqrt{m}\lambda_n$.*

As with our earlier results on the neighborhood and graphical Lasso, 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. The proof of Theorem 11.15 is based on the same type of primal–dual witness construction used in the proof of Theorem 11.12. See the bibliographic section for further details.

11.4 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. In this section, we discuss some methods for addressing these types of problems, focusing primarily on the Gaussian case for simplicity.

11.4.1 Gaussian graph estimation with corrupted data

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. A naive approach would be simply to apply a standard Gaussian graph estimator to the observed data, but, as we will see, doing so typically leads to inconsistent estimates.

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

Correcting the Gaussian graphical Lasso

Consider the graphical Lasso (11.10), which is usually based on the sample covariance matrix $\widehat{\mathbf{\Sigma}}_x = \frac{1}{n}\mathbf{X}^{\mathrm{T}}\mathbf{X} = \frac{1}{n}\sum_{i=1} x_i x_i^{\mathrm{T}}$ of the raw samples. The naive approach would be instead to solve the convex program

$$\widehat{\mathbf{\Theta}}_{\mathrm{NAI}} = \arg\min_{\mathbf{\Theta}\in\mathcal{S}^{d\times d}} \left\{ \langle\!\langle \mathbf{\Theta}, \widehat{\mathbf{\Sigma}}_z \rangle\!\rangle - \log\det\mathbf{\Theta} + \lambda_n |||\mathbf{\Theta}|||_{1,\mathrm{off}} \right\}, \tag{11.41}$$

where $\widehat{\mathbf{\Sigma}}_z = \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \frac{1}{n}\sum_{i=1}^n z_i z_i^{\mathrm{T}}$ is now the sample covariance based on the observed data matrix $\mathbf{Z}$. However, as we explore in Exercise 11.8, the addition of noise does not preserve Markov properties, so that—at least in general—the estimate $\widehat{\mathbf{\Theta}}_{\mathrm{NAI}}$ will not lead to consistent estimates of either the edge set, or the underlying precision matrix $\mathbf{\Theta}^*$. 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 a few examples.

Example 11.16 (Unbiased covariance estimate for additive corruptions) In the additive noise setting ($\mathbf{Z} = \mathbf{X} + \mathbf{V}$), 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 $\mathbf{\Sigma}_v$. In this case, a natural estimate of $\mathbf{\Sigma}_x := \mathrm{cov}(x)$ is given by

$$\widehat{\mathbf{\Gamma}} := \frac{1}{n}\mathbf{Z}^{\mathrm{T}}\mathbf{Z} - \mathbf{\Sigma}_v. \tag{11.42}$$

As long as the noise matrix $\mathbf{V}$ is independent of $\mathbf{X}$, then $\widehat{\mathbf{\Gamma}}$ is an unbiased estimate of $\mathbf{\Sigma}_x$. Moreover, as we explore in Exercise 11.12, when both $\mathbf{X}$ and $\mathbf{V}$ have sub-Gaussian rows, then a deviation condition of the form $|||\widehat{\mathbf{\Gamma}} - \mathbf{\Sigma}_x|||_{\max} \precsim \sqrt{\frac{\log d}{n}}$ holds with high probability. ♣

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 $\nu \in [0, 1)$. Based on the observed matrix $\mathbf{Z} \in \mathbb{R}^{n\times d}$, we can construct a new matrix $\widetilde{\mathbf{Z}} \in \mathbb{R}^{n\times d}$ with entries

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

With this choice, it can be verified that

$$\widehat{\mathbf{\Gamma}} = \frac{1}{n}\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}} - \nu\,\mathrm{diag}\left(\frac{\widetilde{\mathbf{Z}}^{\mathrm{T}}\widetilde{\mathbf{Z}}}{n}\right) \tag{11.43}$$

is an unbiased estimate of the covariance matrix $\mathbf{\Sigma}_x = \mathrm{cov}(x)$, and moreover, under suitable tail conditions, it also satisfies the deviation condition $|||\widehat{\mathbf{\Gamma}} - \mathbf{\Sigma}_x|||_{\max} \precsim \sqrt{\frac{\log d}{n}}$ with high probability. See Exercise 11.13 for more details. ♣

More generally, any unbiased estimate $\widehat{\mathbf{\Gamma}}$ of $\mathbf{\Sigma}_x$ defines a form of the *corrected graphical Lasso* estimator

$$\widetilde{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}\in\mathcal{S}_+^{d\times d}} \left\{ \langle\!\langle \mathbf{\Theta}, \widehat{\mathbf{\Gamma}} \rangle\!\rangle - \log\det\mathbf{\Theta} + \lambda_n |||\mathbf{\Theta}|||_{1,\mathrm{off}} \right\}. \tag{11.44}$$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

As with the usual graphical Lasso, this is a strictly convex program, so that the solution (when it exists) must be unique. However, depending on the nature of the covariance estimate $\widehat{\mathbf{\Gamma}}$, it need not be the case that the program (11.44) has any solution at all! In this case, equation (11.44) is nonsensical, since it presumes the existence of an optimal solution. However, in Exercise 11.9, we show that as long as $\lambda_n > \|\widehat{\mathbf{\Gamma}} - \mathbf{\Sigma}_x\|_{\max}$, then this optimization problem has a unique optimum that is achieved, so that the estimator is meaningfully defined. Moreover, by inspecting the proofs of the claims in Section 11.2.1, it can be seen that the estimator $\widehat{\mathbf{\Theta}}$ obeys similar Frobenius norm and edge selection bounds as the usual graphical Lasso. Essentially, the only differences lie in the techniques used to bound the deviation $\|\widehat{\mathbf{\Gamma}} - \mathbf{\Sigma}_x\|_{\max}$.

Correcting neighborhood regression

We now describe how the method of neighborhood regression can be corrected to deal with corrupted or missing data. Here the underlying optimization problem is typically nonconvex, so that the analysis of the estimator becomes more interesting than the corrected graphical Lasso.

As previously described in Section 11.2.2, the neighborhood regression approach involves solving a linear regression problem, in which the observation vector $X_j \in \mathbb{R}^n$ at a given node j plays the role of the response variable, and the remaining $(d-1)$ variables play the role of the predictors. Throughout this section, we use $\mathbf{X}$ to denote the $n \times (d-1)$ matrix with $\{X_k, k \in V \setminus \{j\}\}$ as its columns, and we use $y = X_j$ to denote the response vector. With this notation, we have an instance of a corrupted linear regression model, namely

$$y = \mathbf{X}\theta^* + w \quad \text{and} \quad \mathbf{Z} \sim \mathbb{Q}(\cdot \mid \mathbf{X}), \tag{11.45}$$

where the conditional probability distribution $\mathbb{Q}$ varies according to the nature of the corruption. In application to graphical models, the response vector y might also be further corrupted, but this case can often be reduced to an instance of the previous model. For instance, if some entries of $y = X_j$ are missing, then we can simply discard those data points in performing the neighborhood regression at node j, or if y is subject to further noise, it can be incorporated into the model.

As before, the naive approach would be simply to solve a least-squares problem involving the cost function $\frac{1}{2n}\|y - \mathbf{Z}\theta\|_2^2$. As we explore in Exercise 11.10, doing so will lead to an inconsistent estimate of the neighborhood regression vector θ^*. However, as with the graphical Lasso, the least-squares estimator can also be corrected. 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

$$\bar{\mathcal{L}}(\theta) = \tfrac{1}{2}\theta^T \mathbf{\Gamma}\theta - \langle \theta, \gamma \rangle, \tag{11.46}$$

where $\mathbf{\Gamma} := \operatorname{cov}(x)$ and $\gamma := \operatorname{cov}(x, y)$. By construction, the true regression vector is the unique global minimizer of $\bar{\mathcal{L}}$. Thus, a natural strategy is to solve a penalized regression problem in which the pair $(\gamma, \mathbf{\Gamma})$ are replaced by data-dependent estimates $(\widehat{\gamma}, \widehat{\mathbf{\Gamma}})$. Doing so leads to the empirical objective function

$$\mathcal{L}_n(\theta) = \tfrac{1}{2}\theta^T \widehat{\mathbf{\Gamma}}\theta - \langle \theta, \widehat{\gamma} \rangle. \tag{11.47}$$

To be clear, the estimates $(\widehat{\gamma}, \widehat{\mathbf{\Gamma}})$ must be based on the observed data $(y, \mathbf{Z})$. In Examples 11.16

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

and 11.17, we described suitable unbiased estimators $\widehat{\mathbf{\Gamma}}$ for the cases of additive corruptions and missing entries, respectively. Exercises 11.12 and 11.13 discuss some unbiased estimators $\widehat{\gamma}$ of the cross-covariance vector γ.

Combining the ingredients, we are led to study the following *corrected Lasso* estimator

$$\min_{\|\theta\|_1 \leq \sqrt{\frac{n}{\log d}}} \left\{ \tfrac{1}{2}\theta^{\mathrm{T}}\widehat{\mathbf{\Gamma}}\theta - \langle \widehat{\gamma}, \theta \rangle + \lambda_n \|\theta\|_1 \right\}. \tag{11.48}$$

Note that it combines the objective function (11.47) with an ℓ_1-penalty, as well as an ℓ_1-constraint. At first sight, including both the penalty and constraint might seem redundant, but as shown in Exercise 11.11, this combination is actually needed when the objective function (11.47) is non-convex. Many of the standard choices of $\widehat{\mathbf{\Gamma}}$ lead to non-convex programs: for instance, in the high-dimensional regime ($n < d$), the previously described choices of $\widehat{\mathbf{\Gamma}}$ given in equations (11.42) and (11.43) both have negative eigenvalues, so that the associated optimization problem is non-convex.

When the optimization problem (11.48) is non-convex, it may have local optima in addition to global optima. Since standard algorithms such as gradient descent are only guaranteed to converge to local optima, it is desirable to have theory that applies them. More precisely, a *local optimum* for the program (11.48) is any vector $\widetilde{\theta} \in \mathbb{R}^d$ such that

$$\left\langle \nabla \mathcal{L}_n(\widetilde{\theta}), \theta - \widetilde{\theta} \right\rangle \geq 0 \qquad \text{for all } \theta \text{ such that } \|\theta\|_1 \leq \sqrt{\tfrac{n}{\log d}}. \tag{11.49}$$

When $\widetilde{\theta}$ belongs to the interior of the constraint set—that is, when it satisfies the inequality $\|\widetilde{\theta}\|_1 < \sqrt{\frac{n}{\log d}}$ strictly—then this condition reduces to the usual zero-gradient condition $\nabla \mathcal{L}_n(\widetilde{\theta}) = 0$. Thus, our specification includes both local minima, local maxima and saddle points.

We now establish an interesting property of the corrected Lasso (11.48): under suitable conditions—ones that still permit non-convexity—*any* local optimum is relatively close to the true regression vector. 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

$$\left\langle \Delta, \widehat{\mathbf{\Gamma}}\Delta \right\rangle \geq \kappa \|\Delta\|_2^2 - c_0 \frac{\log d}{n} \|\Delta\|_1^2 \qquad \text{for all } \Delta \in \mathbb{R}^d. \tag{11.50}$$

Interestingly, such an RE condition can hold for matrices $\widehat{\mathbf{\Gamma}}$ that are indefinite (with both positive and negative eigenvalues), including our estimators for additive corruptions and missing data from Examples 11.16 and 11.17. See Exercises 11.12 and 11.13, respectively, for further details on these two cases.

Moreover, we assume that the minimizer θ^* of the population objective (11.46) has sparsity s and ℓ_2-norm at most one, and that the sample size n is lower bounded as $n \geq s \log d$. 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 (11.48).

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

Proposition 11.18 *Under the RE condition* (11.50)*, suppose that the pair* $(\widehat{\gamma}, \widehat{\mathbf{\Gamma}})$ *satisfy the deviation condition*

$$\|\widehat{\mathbf{\Gamma}}\theta^* - \widehat{\gamma}\|_{\max} \leq \varphi(\mathbb{Q}, \sigma_w) \sqrt{\frac{\log d}{n}}, \tag{11.51}$$

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* $\lambda_n \geq 2(2c_0 + \varphi(\mathbb{Q}, \sigma_w)) \sqrt{\frac{\log d}{n}}$*, any local optimum* $\widetilde{\theta}$ *to the program* (11.48) *satisfies the bound*

$$\|\widetilde{\theta} - \theta^*\|_2 \leq \frac{2}{\kappa} \sqrt{s}\, \lambda_n. \tag{11.52}$$

In order to gain intuition for the constraint (11.51), observe that the optimality of θ^* for the population-level objective (11.46) implies that $\nabla \overline{\mathcal{L}}(\theta^*) = \mathbf{\Gamma}\theta^* - \gamma = 0$. Consequently, condition (11.51) is the sample-based and approximate equivalent of this optimality condition. Moreover, under suitable tail conditions, it is satisfied with high probability by our previous choices of $(\widehat{\gamma}, \widehat{\mathbf{\Gamma}})$ for additively corrupted or missing data. Again, see Exercises 11.12 and 11.13 for further details.

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

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

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

$$\|\theta^*\|_1 - \|\widetilde{\theta}\|_1 \leq \|\widetilde{\Delta}_S\|_1 - \|\widetilde{\Delta}_{S^c}\|_1. \tag{11.53}$$

Since $\nabla \mathcal{L}_n(\theta) = \widehat{\mathbf{\Gamma}}\theta - \widehat{\gamma}$, the deviation condition (11.51) 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

$$\left\langle \widetilde{\Delta}, \widehat{\mathbf{\Gamma}}\widetilde{\Delta} \right\rangle \leq \frac{\lambda_n}{2} \|\widetilde{\Delta}\|_1 + \lambda_n \{\|\widetilde{\Delta}_S\|_1 - \|\widetilde{\Delta}_{S^c}\|_1\} = \frac{3}{2}\lambda_n \|\widetilde{\Delta}_S\|_1 - \frac{1}{2}\lambda_n \|\widetilde{\Delta}_{S^c}\|_1. \tag{11.54}$$

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

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

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

Combined with the RE condition (11.50), we have

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

Recombining with our earlier bound (11.54), we have

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

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

11.4.2 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. The extreme case of this phenomenon is the distinction between independence and conditional independence: for instance, the random variables $X_1 = \texttt{Shoe size}$ and $X_2 = \texttt{Gray hair}$ are likely to be dependent, since few children have gray hair. However, it might be reasonable to model them as being conditionally independent given a third variable—namely $X_3 = \texttt{Age}$.

How to estimate a sparse graphical model when only a subset of the variables are observed? More precisely, 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. Given this partial information, our goal is to recover useful information about the underlying graph.

In the Gaussian case, this problem has an attractive matrix-theoretic formulation. In particular, the observed samples of X_{O} give us information about the covariance matrix $\mathbf{\Sigma}^*_{\mathrm{OO}}$. On the other hand, since we have assumed that the full vector is Markov with respect to a sparse graph, 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. This $(d + r)$-dimensional matrix can be written in the block-partitioned form

$$\mathbf{\Theta}^\diamond = \begin{bmatrix} \mathbf{\Theta}^\diamond_{\mathrm{OO}} & \mathbf{\Theta}^\diamond_{\mathrm{OH}} \\ \mathbf{\Theta}^\diamond_{\mathrm{HO}} & \mathbf{\Theta}^\diamond_{\mathrm{HH}} \end{bmatrix}. \tag{11.55}$$

The block-matrix inversion formula (see Exercise 11.3) ensures that the inverse of the d-dimensional covariance matrix $\mathbf{\Sigma}^*_{\mathrm{OO}}$ has the decomposition

$$(\mathbf{\Sigma}^*_{\mathrm{OO}})^{-1} = \underbrace{\mathbf{\Theta}^\diamond_{\mathrm{OO}}}_{\mathbf{\Gamma}^*} - \underbrace{\mathbf{\Theta}^\diamond_{\mathrm{OH}}(\mathbf{\Theta}^\diamond_{\mathrm{HH}})^{-1}\mathbf{\Theta}^\diamond_{\mathrm{HO}}}_{\mathbf{\Lambda}^*}. \tag{11.56}$$

By our modeling assumptions, the matrix $\mathbf{\Gamma}^* := \mathbf{\Theta}^\diamond_{\mathrm{OO}}$ is sparse, whereas the second component $\mathbf{\Lambda}^* := \mathbf{\Theta}^\diamond_{\mathrm{OH}}(\mathbf{\Theta}^\diamond_{\mathrm{HH}})^{-1}\mathbf{\Theta}^\diamond_{\mathrm{HO}}$ has rank at most $\min\{r, d\}$. Consequently, it has low rank

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

whenever the number of hidden variables r is substantially less than the number of observed variables d. In this way, the addition of hidden variables leads to an inverse covariance matrix that can be decomposed as the sum of a sparse and a low-rank matrix.

Now suppose that we are given n i.i.d. samples $x_i \in \mathbb{R}^d$ from a zero-mean Gaussian with covariance $\mathbf{\Sigma}^*_{OO}$. In the absence of any sparsity in the low-rank component, we require $n > d$ samples to obtain any sort of reasonable estimate (recall our results on covariance estimation from Chapter 6). When $n > d$, then the sample covariance matrix $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ will be invertible with high probability, and hence setting $\mathbf{Y} := (\widehat{\mathbf{\Sigma}})^{-1}$, we can consider an observation model of the form

$$\mathbf{Y} = \mathbf{\Gamma}^* - \mathbf{\Lambda}^* + \mathbf{W}. \tag{11.57}$$

Here $\mathbf{W} \in \mathbb{R}^{d \times d}$ is a stochastic noise matrix, corresponding to the difference between the inverses of the population and sample covariances. This observation model (11.57) is a particular form of additive matrix decomposition, as previously discussed in Section 10.7.

How to estimate the components of this decomposition? In this section, we analyze a very simple two-step estimator, based on first computing a soft-thresholded version of the inverse sample covariance $\mathbf{Y}$ as an estimate of $\mathbf{\Gamma}^*$, and secondly, taking the residual matrix as an estimate of $\mathbf{\Lambda}^*$. In particular, for a threshold $\nu_n > 0$ to be chosen, we define the estimates

$$\widehat{\mathbf{\Gamma}} := T_{\nu_n}((\widehat{\mathbf{\Sigma}})^{-1}) \quad \text{and} \quad \widehat{\mathbf{\Lambda}} := \widehat{\mathbf{\Gamma}} - (\widehat{\mathbf{\Sigma}})^{-1}. \tag{11.58}$$

Here the hard-thresholding operator is given by $T_{\nu_n}(v) = v\, \mathbb{I}[|v| > \nu_n]$.

As discussed in Chapter 10, sparse-plus-low-rank decompositions are unidentifiable unless constraints are imposed on the pair $(\mathbf{\Gamma}^*, \mathbf{\Lambda}^*)$. As with our earlier study of matrix decompositions in Section 10.7, we assume here that the low-rank component satisfies a "spikiness" constraint, meaning that its elementwise max-norm is bounded as $\|\mathbf{\Lambda}^*\|_{\max} \leq \frac{\alpha}{d}$. In addition, we assume that the matrix square root of the true precision matrix $\mathbf{\Theta}^* = \mathbf{\Gamma}^* - \mathbf{\Lambda}^*$ has a bounded ℓ_∞-operator norm, meaning that

$$|||\sqrt{\mathbf{\Theta}^*}|||_\infty = \max_{j=1,\ldots,d} \sum_{k=1}^d |\sqrt{\mathbf{\Theta}^*}|_{jk} \leq \sqrt{M}. \tag{11.59}$$

In terms of the parameters (α, M), we then choose the threshold parameter ν_n in our estimates (11.58) as

$$\nu_n := M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{\alpha}{d} \qquad \text{for some } \delta \in [0,1]. \tag{11.60}$$

Proposition 11.19 *Consider a precision matrix $\mathbf{\Theta}^*$ that can be decomposed as the difference $\mathbf{\Gamma}^* - \mathbf{\Lambda}^*$, where $\mathbf{\Gamma}^*$ has most s non-zero entries per row, and $\mathbf{\Lambda}^*$ is α-spiky. Given $n > d$ i.i.d. samples from the $\mathcal{N}(0, (\mathbf{\Theta}^*)^{-1})$ distribution and any $\delta \in (0,1]$, the estimates $(\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Lambda}})$ satisfy the bounds*

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*\|_{\max} \leq 2M\left(4\sqrt{\frac{\log d}{n}} + \delta\right) + \frac{2\alpha}{d} \tag{11.61a}$$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

and

$$|||\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}^*|||_2 \leq M\left(2\sqrt{\frac{d}{n}} + \delta\right) + s\,|||\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*||_{\max} \tag{11.61b}$$

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

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

$$|||\mathbf{Y} - \mathbf{\Theta}^*|||_2 \leq M\left(2\sqrt{\frac{d}{n}} + \delta\right) \tag{11.62a}$$

and

$$|||\mathbf{Y} - \mathbf{\Theta}^*||_{\max} \leq M\Big(4\sqrt{\frac{\log d}{n}} + \delta\Big) \tag{11.62b}$$

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

To prove the first bound (11.62a), we note that

$$(\widehat{\mathbf{\Sigma}})^{-1} - \mathbf{\Theta}^* = \sqrt{\mathbf{\Theta}^*}\Big\{n^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathbf{I}_d\Big\}\sqrt{\mathbf{\Theta}^*}, \tag{11.63}$$

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

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

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

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

from which the bound (11.62a) follows.

Turning to the bound (11.62b), using the decomposition (11.63) and introducing the shorthand $\widetilde{\mathbf{\Sigma}} = \frac{\mathbf{V}^{\mathrm{T}}\mathbf{V}}{n} - \mathbf{I}_d$, we have

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

Now observe that

$$\max_{j=1,\ldots,d} |||\sqrt{\mathbf{\Theta}^*}e_j|||_1 \leq \max_{|||u|||_1=1} |||\sqrt{\mathbf{\Theta}^*}u|||_1 = \max_{\ell=1,\ldots,d} \sum_{k=1}^{d} |[\sqrt{\mathbf{\Theta}^*}]|_{k\ell} = |||\sqrt{\mathbf{\Theta}^*}|||_\infty,$$

where the final inequality uses the symmetry of $\sqrt{\mathbf{\Theta}^*}$. Putting together the pieces yields that

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

$\|(\widehat{\boldsymbol{\Sigma}})^{-1} - \boldsymbol{\Theta}^*\|_{\max} \leq M\|\widetilde{\boldsymbol{\Sigma}}\|_{\max}$. Since $\widetilde{\boldsymbol{\Sigma}} = \mathbf{V}^{\mathrm{T}}\mathbf{V}/n - I$, where $\mathbf{V} \in \mathbb{R}^{n \times d}$ is a matrix of i.i.d. standard normal variates, we have $\|\widetilde{\boldsymbol{\Sigma}}\|_{\max} \leq 4\sqrt{\frac{\log d}{n}} + \delta$ 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 (11.62b).

Next we establish bounds on the estimates $(\widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Lambda}})$ previously defined in equation (11.58). Recalling our shorthand $\mathbf{Y} = (\widehat{\boldsymbol{\Sigma}})^{-1}$, by the definition of $\widehat{\boldsymbol{\Gamma}}$ and the triangle inequality, we have

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

thereby establishing inequality (11.61a).

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

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

Recall that $\boldsymbol{\Gamma}^*$ 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

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

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

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

Putting together the pieces yields the claimed bound (11.61b).

□

11.5 Bibliographic details and background

Graphical models have a rich history, with parallel developments taking place in statistical physics (Ising, 1925; Bethe, 1935; Baxter, 1982), information and coding theory (Gallager, 1968; Richardson and Urbanke, 2008), artificial intelligence (Pearl, 1988) and image processing (Geman and Geman, 1984), among other areas. See the books (Lauritzen, 1996; Mézard and Montanari, 2008; Wainwright and Jordan, 2008; Koller and Friedman, 2010) for further background. The Ising model from Example 11.4 was first proposed as a model for ferromagnetism in statistical physics (Ising, 1925), and has been extensively studied. The

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

Hammersley–Clifford theorem derives its name from the unpublished manuscript (Hammersley and Clifford, 1971). Grimmett (1973) and Besag (1974) were the first to publish proofs of the result; see Clifford (1990) for further discussion of its history. Lauritzen (1996) provides discussion of how the Markov factorization equivalence can break down when the strict positivity condition is not satisfied. There are a number of connections between the classical theory of exponential families (Barndorff-Nielson, 1978; Brown, 1986) and graphical models; see the monograph (Wainwright and Jordan, 2008) for further details.

The Gaussian graphical Lasso (11.10) has been studied by a large number of researchers (e.g., Friedman et al., 2007; Yuan and Lin, 2007; Banerjee et al., 2008; d'Aspremont et al., 2008; Rothman et al., 2008; Ravikumar et al., 2011), in terms of both its statistical and optimization-related properties. The Frobenius norm bounds in Proposition 11.9 were first proved by Rothman et al. (2008). Ravikumar et al. (2011) proved the model selection results given in Proposition 11.10; they also analyzed the estimator for more general non-Gaussian distributions, and under a variety of tail conditions. There are also related analyses of Gaussian maximum likelihood using various forms of non-convex penalties (e.g., Lam and Fan, 2009; Loh and Wainwright, 2017). Among others, Friedman et al. (2007) and d'Asprémont et al. (2008) have developed efficient algorithms for solving the Gaussian graphical Lasso.

Neighborhood-based methods for graph estimation have their roots in the notion of pseudo-likelihood, as studied in the classical work of Besag (1974; 1975; 1977). Besag (1974) discusses various neighbor-based specifications of graphical models, including the Gaussian graphical model from Example 11.3, the Ising (binary) graphical model from Example 11.4, and the Poisson graphical model from Example 11.14. Meinshausen and Bühlmann (2006) provided the first high-dimensional analysis of the Lasso as a method for neighborhood selection in Gaussian graphical models. Their analysis, and that of related work by Zhao and Yu (2006), was based on assuming that the design matrix itself satisfies the α-incoherence condition, whereas the result given in Theorem 11.12, adapted from Wainwright (2009b), imposes these conditions on the population, and then proves that the sample versions satisfy them with high probability. Whereas we only proved Theorem 11.12 when the maximum degree m is at most $\log d$, the paper (Wainwright, 2009b) provides a proof for the general case.

Meinshausen (2008) discussed the need for stronger incoherence conditions with the Gaussian graphical Lasso (11.10) as opposed to the neighborhood selection method; see also Ravikumar et al. (2011) for further comparison of these types of incoherence conditions. Other neighborhood-based methods have also been studied in the literature, including methods based on the Dantzig selector (Yuan, 2010) and the CLIME-based method (Cai et al., 2011). Exercise 11.4 works through some analysis for the CLIME estimator.

Ravikumar et al. (2010) analyzed the ℓ_1-regularized logistic regression method for Ising model selection using the primal–dual witness method; Theorem 11.15 is adapted from their work. Other authors have studied different methods for graphical model selection in discrete models, including various types of entropy tests, thresholding methods and greedy methods (e.g., Netrapalli et al., 2010; Anandkumar et al., 2012; Bresler et al., 2013; Bresler, 2014). Santhanam and Wainwright (2012) prove lower bounds on the number of samples required for Ising model selection; combined with the improved achievability results of Bento and Montanari (2009), these lower bounds show that ℓ_1-regularized logistic regression is an order-optimal method. It is more natural—as opposed to estimating each neighborhood

separately—to perform a joint estimation of all neighborhoods simultaneously. One way in which to do so is to sum all of the conditional likelihoods associated with each node, and then optimize the sum jointly, ensuring that all edges use the same parameter value in each neighborhood. The resulting procedure is equivalent to the pseudo-likelihood method (Besag, 1975, 1977). Hoefling and Tibshirani (2009) compare the relative efficiency of various pseudo-likelihood-type methods for graph estimation.

The corrected least-squares cost (11.47) is a special case of a more general class of corrected likelihood methods (e.g., Carroll et al., 1995; Iturria et al., 1999; Xu and You, 2007). The corrected non-convex Lasso (11.48) was proposed and analyzed by Loh and Wainwright (2012; 2017). A related corrected form of the Dantzig selector was analyzed by Rosenbaum and Tsybakov (2010). Proposition 11.18 is a special case of more general results on non-convex M-estimators proved in the papers (Loh and Wainwright, 2015, 2017).

The matrix decomposition approach to Gaussian graph selection with hidden variables was pioneered by Chandrasekaran et al. (2012b), who proposed regularizing the global likelihood (log-determinant function) with nuclear and ℓ_1-norms. They provided sufficient conditions for exact recovery of sparsity and rank using the primal–dual witness method, previously used to analyze the standard graphical Lasso (Ravikumar et al., 2011). Ren and Zhou (2012) proposed more direct approaches for estimating such matrix decompositions, such as the simple estimator analyzed in Proposition 11.19. Agarwal et al. (2012) analyzed both a direct approach based on thresholding and truncated SVD, as well as regularization-based methods for more general problems of matrix decomposition. As with other work on matrix decomposition problems (Candès et al., 2011; Chandrasekaran et al., 2011), Chandrasekaran et al. (2012b) performed their analysis under strong incoherence conditions, essentially algebraic conditions that ensure perfect identifiability for the sparse-plus-low-rank problem. The milder constraint, namely of bounding the maximum entry of the low-rank component as in Proposition 11.19, was introduced by Agarwal et al. (2012).

In addition to the undirected graphical models discussed here, there is also a substantial literature on methods for directed graphical models; we refer the reader to the sources (Spirtes et al., 2000; Kalisch and Bühlmann, 2007; Bühlmann and van de Geer, 2011) and references therein for more details. Liu et al. (2009; 2012) propose and study the nonparanormal family, a nonparametric generalization of the Gaussian graphical model. Such models are obtained from Gaussian models by applying a univariate transformation to the random variable at each node. The authors discuss methods for estimating such models; see also Xue and Zou (2012) for related results.

11.6 Exercises

Exercise 11.1 (Properties of log-determinant function) Let $\mathcal{S}^{d\times d}$ denote the set of symmetric matrices, and $\mathcal{S}_+^{d\times d}$ denote the cone of symmetric and strictly positive definite matrices. In this exercise, we study properties of the (negative) log-determinant function $F : \mathcal{S}^{d\times d} \to \mathbb{R}$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

given by

$$F(\mathbf{\Theta}) = \begin{cases} -\sum_{j=1}^{d} \log \gamma_j(\mathbf{\Theta}) & \text{if } \mathbf{\Theta} \in \mathcal{S}_+^{d \times d}, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\gamma_j(\mathbf{\Theta}) > 0$ are the eigenvalues of $\mathbf{\Theta}$.

(a) Show that F is a strictly convex function on its domain $\mathcal{S}_+^{d \times d}$.
(b) For $\mathbf{\Theta} \in \mathcal{S}_+^{d \times d}$, show that $\nabla F(\mathbf{\Theta}) = -\mathbf{\Theta}^{-1}$.
(c) For $\mathbf{\Theta} \in \mathcal{S}_+^{d \times d}$, show that $\nabla F^2(\mathbf{\Theta}) = \mathbf{\Theta}^{-1} \otimes \mathbf{\Theta}^{-1}$.

Exercise 11.2 (Gaussian MLE) Consider the maximum likelihood estimate of the inverse covariance matrix $\mathbf{\Theta}^*$ for a zero-mean Gaussian. Show that it takes the form

$$\widehat{\mathbf{\Theta}}_{\mathrm{MLE}} = \begin{cases} \widehat{\mathbf{\Sigma}}^{-1} & \text{if } \widehat{\mathbf{\Sigma}} \succ 0, \\ \text{not defined} & \text{otherwise,} \end{cases}$$

where $\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^{\mathrm{T}}$ is the empirical covariance matrix for a zero-mean vector. (When $\widehat{\mathbf{\Sigma}}$ is rank-deficient, you need to show explicitly that there exists a sequence of matrices for which the likelihood diverges to infinity.)

Exercise 11.3 (Gaussian neighborhood regression) Let $X \in \mathbb{R}^d$ be a zero-mean jointly Gaussian random vector with strictly positive definite covariance matrix $\mathbf{\Sigma}^*$. Consider the conditioned random variable $Z := (X_j \mid X_{\backslash\{j\}})$, where we use the shorthand $\backslash\{j\} = V \setminus \{j\}$.

(a) Establish the validity of the decomposition (11.21).
(b) Show that $\theta_j^* = (\mathbf{\Sigma}^*_{\backslash\{j\}, \backslash\{j\}})^{-1} \Sigma_{\backslash\{j\}, j}$.
(c) Show that $\theta_{jk}^* = 0$ whenever $k \notin \mathcal{N}(j)$.
Hint: The following elementary fact could be useful: let $\mathbf{A}$ be an invertible matrix, given in the block-partitioned form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}.$$

Then letting $\mathbf{B} = \mathbf{A}^{-1}$, we have (see Horn and Johnson (1985))

$$\mathbf{B}_{22} = (\mathbf{A}_{22} - \mathbf{A}_{21}(\mathbf{A}_{11})^{-1}\mathbf{A}_{12})^{-1} \quad \text{and} \quad \mathbf{B}_{12} = (\mathbf{A}_{11})^{-1}\mathbf{A}_{12}\big[\mathbf{A}_{21}(\mathbf{A}_{11})^{-1}\mathbf{A}_{12} - \mathbf{A}_{22}\big]^{-1}.$$

Exercise 11.4 (Alternative estimator of sparse precision matrix) Consider a d-variate Gaussian random vector with zero mean, and a sparse precision matrix $\mathbf{\Theta}^*$. In this exercise, we analyze the estimator

$$\widehat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta} \in \mathbb{R}^{d \times d}} \{|||\mathbf{\Theta}|||_1\} \qquad \text{such that } |||\widehat{\mathbf{\Sigma}}\mathbf{\Theta} - \mathbf{I}_d|||_{\max} \leq \lambda_n, \tag{11.64}$$

where $\widehat{\mathbf{\Sigma}}$ is the sample covariance based on n i.i.d. samples.

(a) For $j = 1, \ldots, d$, consider the linear program

$$\widehat{\Gamma}_j \in \arg\min_{\Gamma_j \in \mathbb{R}^d} \|\Gamma_j\|_1 \qquad \text{such that } \|\widehat{\mathbf{\Sigma}}\Gamma_j - e_j\|_{\max} \leq \lambda_n, \tag{11.65}$$

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

where $e_j \in \mathbb{R}^d$ is the jth canonical basis vector. Show that $\widehat{\boldsymbol{\Theta}}$ is optimal for the original program (11.64) if and only if its jth column $\widehat{\Theta}_j$ is optimal for the program (11.65).

(b) Show that $\|\widehat{\Gamma}_j\|_1 \leq \|\Theta^*_j\|_1$ for each $j = 1, \ldots, d$ whenever the regularization parameter is lower bounded as $\lambda_n \geq |||\boldsymbol{\Theta}^*|||_1 \|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max}$.

(c) State and prove a high-probability bound on $\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^*\|_{\max}$. (For simplicity, you may assume that $\max_{j=1,\ldots,d} \Sigma^*_{jj} \leq 1$.)

(d) Use the preceding parts to show that, for an appropriate choice of λ_n, there is a universal constant c such that

$$\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_{\max} \leq c \, |||\boldsymbol{\Theta}^*|||_1^2 \sqrt{\frac{\log d}{n}} \tag{11.66}$$

with high probability.

Exercise 11.5 (Special case of general neighborhood regression) Show that the general form of neighborhood regression (11.37) reduces to linear regression (11.22) in the Gaussian case. (*Note*: You may ignore constants, either pre-factors or additive ones, that do not depend on the data.)

Exercise 11.6 (Structure of conditional distribution) Given a density of the form (11.32), show that the conditional likelihood of X_j given $X_{\setminus\{j\}}$ depends only on

$$\boldsymbol{\Theta}_{j+} := \{\Theta_j, \boldsymbol{\Theta}_{jk}, k \in V \setminus \{j\}\}.$$

Prove that $\boldsymbol{\Theta}_{jk} = 0$ whenever $(j, k) \notin E$.

Exercise 11.7 (Conditional distribution for Ising model) For a binary random vector $X \in \{-1, 1\}^d$, consider the family of distributions

$$p_\theta(x_1, \ldots, x_d) = \exp\Big\{ \sum_{(j,k) \in E} \theta_{jk} x_j x_k - \Phi(\theta) \Big\}, \tag{11.67}$$

where E is the edge set of some undirected graph G on the vertices $V = \{1, 2, \ldots, d\}$.

(a) For each edge $(j, k) \in E$, show that $\frac{\partial \Phi(\theta)}{\partial \theta_{jk}} = \mathbb{E}_\theta[X_j X_k]$.

(b) Compute the conditional distribution of X_j given the subvector of random variables $\mathbf{X}_{\setminus\{j\}} := \{X_k, k \in V \setminus \{j\}\}$. Give an expression in terms of the logistic function $f(t) = \log(1 + e^t)$.

Exercise 11.8 (Additive noise and Markov properties) Let $X = (X_1, \ldots, X_d)$ be a zero-mean Gaussian random vector that is Markov with respect to some graph G, and let $Z = X + V$, where $V \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ is an independent Gaussian noise vector. Supposing that $\sigma^2 |||\boldsymbol{\Theta}^*|||_2 < 1$, derive an expression for the inverse covariance of Z in terms of powers of $\sigma^2 \boldsymbol{\Theta}^*$. Interpret this expression in terms of weighted path lengths in the graph.

Exercise 11.9 (Solutions for corrected graphical Lasso) In this exercise, we explore properties of the corrected graphical Lasso from equation (11.44).

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

(a) Defining $\mathbf{\Sigma}_x := \text{cov}(x)$, show that as long as $\lambda_n > \|\widehat{\mathbf{\Gamma}} - \mathbf{\Sigma}_x\|_{\max}$, then the corrected graphical Lasso (11.44) has a unique optimal solution.
(b) Show what can go wrong when this condition is violated. (*Hint:* It suffices to consider a one-dimensional example.)

Exercise 11.10 (Inconsistency of uncorrected Lasso) Consider the linear regression model $y = \mathbf{X}\theta^* + w$, where we observe the response vector $y \in \mathbb{R}^n$ and the corrupted matrix $\mathbf{Z} = \mathbf{X} + \mathbf{V}$. A naive estimator of θ^* is

$$\widetilde{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2n} \|y - \mathbf{Z}\theta\|_2^2 \right\},$$

where we regress y on the corrupted matrix $\mathbf{Z}$. Suppose that each row of $\mathbf{X}$ is drawn i.i.d. from a zero-mean distribution with covariance $\mathbf{\Sigma}$, and that each row of $\mathbf{V}$ is drawn i.i.d. (and independently from $\mathbf{X}$) from a zero-mean distribution with covariance $\sigma^2 I$. Show that $\widetilde{\theta}$ is inconsistent even if the sample size $n \to +\infty$ with the dimension fixed.

Exercise 11.11 (Solutions for corrected Lasso) Show by an example in two dimensions that the corrected Lasso (11.48) may not achieve its global minimum if an ℓ_1-bound of the form $\|\theta\|_1 \leq R$ for some radius R is not imposed.

Exercise 11.12 (Corrected Lasso for additive corruptions) In this exercise, we explore properties of corrected linear regression in the case of additive corruptions (Example 11.16), under the standard model $y = \mathbf{X}\theta^* + w$.

(a) Assuming that $\mathbf{X}$ and $\mathbf{V}$ are independent, show that $\widehat{\mathbf{\Gamma}}$ from equation (11.42) is an unbiased estimate of $\mathbf{\Sigma}_x = \text{cov}(x)$, and that $\widehat{\gamma} = \mathbf{Z}^T y/n$ is an unbiased estimate of $\text{cov}(x, y)$.
(b) Now suppose that in addition both $\mathbf{X}$ and $\mathbf{V}$ are generated with i.i.d. rows from a zero-mean distribution, and that each element X_{ij} and V_{ij} is sub-Gaussian with parameter 1, and that the noise vector w is independent with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. Show that there is a universal constant c such that

$$\|\widehat{\mathbf{\Gamma}}\theta^* - \widehat{\gamma}\|_\infty \leq c(\sigma + \|\theta^*\|_2) \sqrt{\frac{\log d}{n}}$$

with high probability.
(c) In addition to the previous assumptions, suppose that $\mathbf{\Sigma}_v = \nu \mathbf{I}_d$ for some $\nu > 0$. Show that $\widehat{\mathbf{\Gamma}}$ satisfies the RE condition (11.50) with high probability. (*Hint:* The result of Exercise 7.10 may be helpful to you.)

Exercise 11.13 (Corrected Lasso for missing data) In this exercise, we explore properties of corrected linear regression in the case of missing data (Example 11.17). Throughout, we assume that the missing entries are removed completely independently at random, and that $\mathbf{X}$ has zero-mean rows, generated in an i.i.d. fashion from a 1-sub-Gaussian distribution.

(a) Show that the matrix $\widehat{\mathbf{\Gamma}}$ from equation (11.43) is an unbiased estimate of $\mathbf{\Sigma}_x := \text{cov}(x)$, and that the vector $\widehat{\gamma} = \frac{\mathbf{Z}^T y}{n}$ is an unbiased estimate of $\text{cov}(x, y)$.

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press

(b) Assuming that the noise vector $w \in \mathbb{R}^n$ has i.i.d. $\mathcal{N}(0, \sigma^2)$ entries, show there is a universal constant c such that

$$\|\widehat{\mathbf{\Gamma}}\theta^* - \widehat{\gamma}\|_\infty \leq c(\sigma + \|\theta^*\|_2)\sqrt{\frac{\log d}{n}}$$

with high probability.

(c) Show that $\widehat{\mathbf{\Gamma}}$ satisfies the RE condition (11.50) with high probability. (*Hint:* The result of Exercise 7.10 may be helpful to you.)

https://doi.org/10.1017/9781108627771.011 Published online by Cambridge University Press