9

Gaussian Graphical Models

Caroline Uhler
Laboratory for Information and Decision Systems,
Department of Electrical Engineering and Computer Science,
and Institute for Data, Systems and Society,
Massachusetts Institute of Technology

CONTENTS

After the discussion of discrete graphical models in the preceding Chapter 8, we now turn to the continuous setting and introduce Gaussian graphical models. As we will see in this chapter, assuming Gaussianity leads to a rich geometric structure that can be exploited for parameter estimation. However, Gaussianity is not only assumed for mathematical simplicity. Gaussian distributions are commonly used for modeling continuous phenomena. As a consequence of the central limit theorem, physical quantities that are expected to be the sum of many independent contributions often follow approximately a Gaussian distribution. For example, people's height is approximately normally distributed; height is believed to be the sum of many independent contributions from various genetic and environmental factors.

Another reason for assuming normality is that the Gaussian distribution has maximum entropy among all real-valued distributions with a specified mean and covariance. Hence, assuming Gaussianity imposes the least number of structural constraints beyond the first and second moments. So another reason for assuming Gaussianity is that it is the least-informative distribution. In addition, many physical systems tend to move towards maximal entropy configurations over time.

Throughout this chapter, we denote by $\mathbb{S}^p$ the vector space of real symmetric $p \times p$ matrices. This vector space is equipped with the *trace inner product* $\langle A, B\rangle := \mathrm{tr}(AB)$. In addition, we denote by $\mathbb{S}^p_{\succeq 0}$ the convex cone of positive semidefinite matrices. Its interior is the open cone $\mathbb{S}^p_{\succ 0}$ of positive definite matrices.

A random vector $X \in \mathbb{R}^p$ is distributed according to the *multivariate Gaussian distribution* $\mathcal{N}(\mu, \Sigma)$ with parameters $\mu \in \mathbb{R}^p$ (the *mean*) and $\Sigma \in \mathbb{S}^p_{\succ 0}$ (the *covariance matrix*), if it has density function

$$f_{\mu,\Sigma}(x) = (2\pi)^{-p/2}(\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \quad x \in \mathbb{R}^p.$$

In the following, we denote the inverse covariance matrix, also known as the *precision matrix* or the *concentration matrix*, by K. In terms of $K = \Sigma^{-1}$ and using the trace inner product on $\mathbb{S}^p$, the density $f_{\mu,\Sigma}$ can equivalently be formulated as:

$$f_{\mu,K}(x) = \exp\left\{\mu^T K x - \left\langle K, \frac{1}{2}xx^T\right\rangle - \frac{p}{2}\log(2\pi) + \frac{1}{2}\log\det(K) - \frac{1}{2}\mu^T K\mu\right\}.$$

Hence, the Gaussian distribution is an *exponential family* with *canonical parameters* $(-\mu^T K, K)$, *sufficient statistics* $(x, \frac{1}{2}xx^T)$ and *log-partition function* (also known as the *cumulant generating function*) $\frac{p}{2}\log(2\pi) - \frac{1}{2}\log\det(K) + \frac{1}{2}\mu^T K\mu$; see [5, 11] for an introduction to exponential families.

Let $G = (V, E)$ be an undirected graph with vertices $V = [p]$ and edges E, where $[p] = \{1, \ldots, p\}$. A random vector $X \in \mathbb{R}^p$ is said to *satisfy the (undirected) Gaussian graphical model with graph* G, if X has a multivariate Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with

$$\left(\Sigma^{-1}\right)_{i,j} = 0 \quad \text{for all } (i,j) \notin E.$$

Hence, the graph G describes the sparsity pattern of the concentration matrix. This explains why G is also known as the *concentration graph*. As we will see in Section 9.1, missing edges in G also correspond to conditional independence relations in the corresponding Gaussian graphical model. Hence, sparser graphs correspond to simpler models with fewer canonical parameters and more conditional independence relations.

Gaussian graphical models are the continuous counter-piece to the Ising models that were discussed in Section 5.4.1 (Ising models are special versions of the log-linear models from Chapter 8). Like Ising models, Gaussian graphical models are quadratic exponential families. These families only model pairwise interactions between nodes, i.e., interactions are only on the edges of the underlying graph G. But nevertheless, Ising models and Gaussian graphical models are extremely flexible models; in fact, they can capture any pairwise correlation structure that can be constructed for binary or for continuous data.

This chapter discusses maximum likelihood (ML) estimation for Gaussian graphical models. There are two problems of interest in this regard: (1) to estimate the edge weights, i.e. the canonical parameters, given the graph structure, and (2) to learn the underlying graph structure. This chapter is mainly focused with the first problem (Sections 9.2-9.6), while the second problem is only discussed in Section 9.7. The second problem is particularly important in the high-dimensional setting when the number of samples n is smaller than the number of variables p. This problem is the subject of Chapters 12 and 14.

The remainder of this chapter is structured as follows: In Section 9.1, we examine conditional independence relations for Gaussian distributions. Then, in Section 9.2, we introduce the Gaussian likelihood. We show that ML estimation for Gaussian graphical models is a convex optimization problem and we describe its dual optimization problem. In Section 9.3, we analyze this dual optimization problem and explain the close links to positive definite matrix completion problems studied in linear algebra. In Section 9.4, we develop a geometric picture of ML estimation for Gaussian graphical models that complements the point of view of convex optimization. The combination of convex optimization, positive definite matrix completion, and convex geometry allows us to obtain results about the existence of the maximum likelihood estimator (MLE) and algorithms for computing the MLE. These are presented in Section 9.5 and in Section 9.6, respectively. Gaussian graphical models are defined by zero constraints on the concentration matrix K. In Section 9.7, we describe methods for learning the underlying graph, or equivalently, the zero pattern of K. Finally, in Section 9.8, we end with a discussion of other Gaussian models with linear constraints on the concentration matrix or the covariance matrix.

9.1 The Gaussian distribution and conditional independence

We start this section by reviewing some of the extraordinary properties of Gaussian distributions. The following result shows that the Gaussian distribution is closed under marginalization and conditioning. We here only provide proofs that will be useful in later sections of the chapter. A complete proof of the following well-known result can be found for example in [4, 17].

Proposition 9.1.1. *Let $X \in \mathbb{R}^p$ be distributed as $\mathcal{N}(\mu, \Sigma)$ and partition the random vector X into two components $X_A \in \mathbb{R}^a$ and $X_B \in \mathbb{R}^b$ such that $a + b = p$. Let μ and Σ be partitioned accordingly, i.e.,*

$$\mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{A,A} & \Sigma_{A,B} \\ \Sigma_{B,A} & \Sigma_{B,B} \end{pmatrix},$$

where, for example, $\Sigma_{B,B} \in \mathbb{S}^b_{\succ 0}$. Then,

(a) the marginal distribution of X_A is $\mathcal{N}(\mu_A, \Sigma_{A,A})$;

(b) the conditional distribution of $X_A \mid X_B = x_B$ is $\mathcal{N}(\mu_{A|B}, \Sigma_{A|B})$, where

$$\mu_{A|B} = \mu_A + \Sigma_{A,B}\Sigma_{B,B}^{-1}(x_B - \mu_B) \quad \text{and} \quad \Sigma_{A|B} = \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}.$$

Proof. We only prove (b) to demonstrate the importance of Schur complements when working with Gaussian distributions. Fixing x_B, we find by direct calculation that the conditional density $f(x_A \mid x_B)$ is proportional to:

$$\begin{aligned} f(x_A \mid x_B) &\propto \exp\Big\{ -\frac{1}{2}(x_A - \mu_A)^T K_{A,A}(x_A - \mu_A) - (x_A - \mu_A)^T K_{A,B}(x_B - \mu_B) \Big\} \\ &\propto \exp\Big\{ -\frac{1}{2}\big(x_A - \mu_A - K_{A,A}^{-1}K_{A,B}(x_B - \mu_B)\big)^T K_{A,A} \\ &\qquad \times \big(x_A - \mu_A - K_{A,A}^{-1}K_{A,B}(x_B - \mu_B)\big) \Big\}, \end{aligned} \tag{9.1}$$

where we used the same partitioning for K as for Σ. Using Schur complements, we obtain

$$K_{A,A}^{-1} \;=\; \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,B},$$

and hence $K_{A,A} = \Sigma_{A|B}^{-1}$. Similarly, we obtain $K_{A,A}^{-1}K_{A,B} = -\Sigma_{A,B}\Sigma_{B,B}^{-1}$. Combining these two identities with the conditional density in (9.1) completes the proof. □

These basic properties of the multivariate Gaussian distribution have interesting implications with respect to the interpretation of zeros in the covariance and the concentration matrix. Namely, as described in the following corollary, zeros correspond to (*conditional*) *independence relations*. For disjoint subsets $A, B, C \subset [p]$ we denote the statement that X_A is conditionally independent of X_B given X_C by $X_A \perp\!\!\!\perp X_B \mid X_C$. If $C = \emptyset$, then we write $X_A \perp\!\!\!\perp X_B$.

Corollary 9.1.2. *Let $X \in \mathbb{R}^p$ be distributed as $\mathcal{N}(\mu, \Sigma)$ and let $i, j \in [p]$ with $i \neq j$. Then*

(a) $X_i \perp\!\!\!\perp X_j$ if and only if $\Sigma_{i,j} = 0$;

(b) $X_i \perp\!\!\!\perp X_j \mid X_{[p]\setminus\{i,j\}}$ if and only if $K_{i,j} = 0$ if and only if $\det(\Sigma_{[p]\setminus\{i\},[p]\setminus\{j\}}) = 0$.

Proof. Statement (a) follows directly from the expression for the conditional mean in Proposition 9.1.1 (b). From the expression for the conditional covariance in Proposition 9.1.1 (b) it follows that $\Sigma_{\{i,j\}|([p]\setminus\{i,j\})} = (K_{\{i,j\},\{i,j\}})^{-1}$. To prove (b), note that if follows from (a) that $X_i \perp\!\!\!\perp X_j \mid X_{[p]\setminus\{i,j\}}$ if and only if the 2×2 conditional covariance matrix $\Sigma_{\{i,j\}|([p]\setminus\{i,j\})}$ is diagonal. This is the case if and only if $K_{\{i,j\},\{i,j\}}$ is diagonal, or equivalently, $K_{i,j} = 0$. This proves the first equivalence in (b). The second equivalence is a consequence of the cofactor formula for matrix inversion, since

$$K_{i,j} = (\Sigma^{-1})_{i,j} = (-1)^{i+j}\frac{\det(\Sigma_{[p]\setminus\{i\},[p]\setminus\{j\}})}{\det(\Sigma)},$$

which completes the proof. □

Corollary 9.1.2 shows that for undirected Gaussian graphical models a missing edge (i,j) in the underlying graph G (i.e. the concentration graph) corresponds to the conditional independence relation $X_i \perp\!\!\!\perp X_j \mid X_{[p]\setminus\{i,j\}}$. Corollary 9.1.2 can be generalized to an equivalence between any conditional independence relation and the vanishing of a particular almost principal minor of Σ or K. This is shown in the following proposition; see also Proposition 3.3.10.

Proposition 9.1.3. *Let $X \in \mathbb{R}^p$ be distributed as $\mathcal{N}(\mu, \Sigma)$. Let $i, j \in [p]$ with $i \neq j$ and let $S \subseteq [p] \setminus \{i,j\}$. Then the following statements are equivalent:*

(a) $X_i \perp\!\!\!\perp X_j \mid X_S$*;*

(b) $\det(\Sigma_{iS,jS}) = 0$*, where* $iS = \{i\} \cup S$*;*

(c) $\det(K_{iR,jR}) = 0$*, where* $R = [p] \setminus (S \cup \{i,j\})$*.*

Proof. By Proposition 9.1.1 (a), the marginal distribution of $X_{S\cup\{i,j\}}$ is Gaussian with covariance matrix $\Sigma_{ijS,ijS}$. Then Corollary 9.1.2 (b) implies the equivalence between (a) and (b). Next we show the equivalence between (a) and (c): It follows from Proposition 9.1.1 (b) that the inverse of $K_{ijR,ijR}$ is equal to the conditional covariance $\Sigma_{ijR|S}$. Hence by Corollary 9.1.2 (a), the conditional independence statement in (a) is equivalent to $((K_{ijR,ijR})^{-1})_{ij} = 0$, which by the cofactor formula for matrix inversion is equivalent to (c). This completes the proof. □

9.2 The Gaussian likelihood and convex optimization

Given n i.i.d. observations $X^{(1)}, \ldots, X^{(n)}$ from $\mathcal{N}(\mu, \Sigma)$, we define the *sample covariance matrix* as

$$S = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X^{(i)}$ is the *sample mean*. We will see that $\bar{X}$ and S are sufficient statistics for the Gaussian model and hence we can write the log-likelihood function in terms of these quantities. Ignoring the normalizing constant, the Gaussian log-likelihood expressed as a

function of (μ, Σ) is

$$\begin{aligned}\ell(\mu,\Sigma) &\propto -\frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}(X^{(i)}-\mu)^T\Sigma^{-1}(X^{(i)}-\mu)\\ &= -\frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}\Big(\sum_{i=1}^{n}(X^{(i)}-\mu)(X^{(i)}-\mu)^T\Big)\right)\\ &= -\frac{n}{2}\log\det(\Sigma) - \frac{n}{2}\operatorname{tr}(S\Sigma^{-1}) - \frac{n}{2}(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu),\end{aligned}$$

where for the last equality we expanded $X^{(i)} - \mu = (X^{(i)} - \bar{X}) + (\bar{X} - \mu)$ and used the fact that $\sum_{i=1}^{n}(X^{(i)} - \bar{X}) = 0$. Hence, it can easily be seen that in the *saturated* (unconstrained) *model* where $(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}^p_{\succ 0}$, the MLE is given by

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\Sigma} = S,$$

assuming that $S \in \mathbb{S}^p_{\succ 0}$.

ML estimation under general constraints on the parameters (μ, Σ) can be complicated. Since Gaussian graphical models only pose constraints on the covariance matrix, we will restrict ourselves to models where the mean μ is unconstrained, i.e. $(\mu, \Sigma) \in \mathbb{R}^p \times \Theta$, where $\Theta \subseteq \mathbb{S}^p_{\succ 0}$. In this case, $\hat{\mu} = \bar{X}$ and the ML estimation problem for Σ boils down to the optimization problem

$$\begin{aligned}\underset{\Sigma}{\text{maximize}} \quad & -\log\det(\Sigma) - \operatorname{tr}(S\Sigma^{-1})\\ \text{subject to} \quad & \Sigma \in \Theta.\end{aligned} \tag{9.2}$$

While this objective function as a function of the covariance matrix Σ is in general not concave over the whole cone $\mathbb{S}^p_{\succ 0}$, it is concave over a large region of the cone, namely for all $\Sigma \in \mathbb{S}^p_{\succ 0}$ such that $\Sigma - 2S \in \mathbb{S}^p_{\succ 0}$ (see [10, Excercise 7.4]).

Gaussian graphical models are given by linear constraints on K. So it is convenient to write the optimization problem (9.2) in terms of the concentration matrix K:

$$\begin{aligned}\underset{K}{\text{maximize}} \quad & \log\det(K) - \operatorname{tr}(SK)\\ \text{subject to} \quad & K \in \mathcal{K},\end{aligned} \tag{9.3}$$

where $\mathcal{K} = \Theta^{-1}$. In particular, for a Gaussian graphical model with graph $G = (V, E)$ the constraints are given by $K \in \mathcal{K}_G$, where

$$\mathcal{K}_G := \{K \in \mathbb{S}^p_{\succ 0} \mid K_{i,j} = 0 \text{ for all } i \neq j \text{ with } (i,j) \notin E\}.$$

In the following, we show that the objective function in (9.3), i.e. as a function of K, is concave over its full domain $\mathbb{S}^p_{\succ 0}$. Since $\mathcal{K}_G$ is a convex cone, this implies that ML estimation for Gaussian graphical models is a convex optimization problem.

Proposition 9.2.1. *The function* $f(Y) = \log\det(Y) - \operatorname{tr}(SY)$ *is concave on its domain* $\mathbb{S}^p_{\succ 0}$.

Proof. Since $\operatorname{tr}(SY)$ is linear in Y it suffices to prove that the function $\log\det(Y)$ is concave over $\mathbb{S}^p_{\succ 0}$. We prove this by showing that the function is concave on any line in $\mathbb{S}^p_{\succ 0}$. Let $Y \in \mathbb{S}^p_{\succ 0}$ and consider the line $Y + tV$, $V \in \mathbb{S}^p$, that passes through Y. It suffices to prove that $g(t) = \log\det(Y + tV)$ is concave for all $t \in \mathbb{R}$ such that $Y + tV \in \mathbb{S}^p_{\succ 0}$. This can be

seen from the following calculation:

$$\begin{aligned} g(t) &= \log\det(Y + tV) \\ &= \log\det(Y^{1/2}(I + tY^{-1/2}VY^{-1/2})Y^{1/2}) \\ &= \log\det(Y) + \sum_{i=1}^{p} \log(1 + t\lambda_i), \end{aligned}$$

where I denotes the identity matrix and λ_i are the eigenvalues of $Y^{-1/2}VY^{-1/2}$. This completes the proof, since $\log\det(Y)$ is a constant and $\log(1 + t\lambda_i)$ is concave in t. □

As a consequence of Proposition 9.2.1, we can study the dual of (9.3) with $\mathcal{K} = \mathcal{K}_G$. See e.g. [10] for an introduction to convex optimization and duality theory. The Lagrangian of this convex optimization problem is given by:

$$\begin{aligned} \mathcal{L}(K, \nu) &= \log\det(K) - \operatorname{tr}(SK) - 2 \sum_{(i,j)\notin E, i\neq j} \nu_{i,j} K_{i,j} \\ &= \log\det(K) - \sum_{i=1}^{p} S_{i,i}K_{i,i} - 2\sum_{(i,j)\in E} S_{i,j}K_{i,j} - 2 \sum_{(i,j)\notin E,\, i\neq j} \nu_{i,j}K_{i,j}, \end{aligned}$$

where $\nu = (\nu_{i,j})_{(i,j)\notin E}$ are the Lagrangian multipliers. To simplify the calculations, we omit the constraint $K \in \mathbb{S}^p_{\succ 0}$. This can be done, since it is assumed that K is in the domain of $\mathcal{L}$. Maximizing $\mathcal{L}(K, \nu)$ with respect to K gives

$$(\hat{K}^{-1})_{i,j} = \begin{cases} S_{i,j} & \text{if } \ i = j \text{ or } (i,j) \in E \\ \nu_{i,j} & \text{otherwise.} \end{cases}$$

The Lagrange dual function is obtained by plugging in $\hat{K}$ for K in $\mathcal{L}(K, \nu)$, which results in

$$g(\nu) = \log\det(\hat{K}) - \operatorname{tr}(\hat{K}^{-1}\hat{K}) = \log\det(\hat{K}) - p.$$

Hence, the dual optimization problem to ML estimation in Gaussian graphical models is given by

$$\begin{aligned} \underset{\Sigma\in\mathbb{S}^p_{\succ 0}}{\text{minimize}} \quad & -\log\det\Sigma - p \\ \text{subject to} \quad & \Sigma_{i,j} = S_{i,j} \ \text{ for all } \ i = j \text{ or } (i,j) \in E. \end{aligned} \tag{9.4}$$

Note that this optimization problem corresponds to *entropy maximization* for fixed sufficient statistics. In fact, this dual relationship between likelihood maximization and entropy maximization holds more generally for exponential families; see [46].

Sections 9.4 and 9.5 are centered around the existence of the MLE. We say that the MLE does not exist if the likelihood does not attain the global maximum. Note that the identity matrix is a strictly feasible point for (9.3) with $\mathcal{K} = \mathcal{K}_G$. Hence, the MLE does not exist if and only if the likelihood is unbounded. *Slater's constraint qualification* states that the existence of a strictly primal feasible point is sufficient for strong duality to hold for a convex optimization problem (see e.g. [10] for an introduction to convex optimization). Since the identity matrix is a strictly feasible point for (9.3), strong duality holds for the optimization problems (9.3) with $\mathcal{K} = \mathcal{K}_G$ and (9.4), and thus we can equivalently study the dual problem (9.4) to obtain insight into ML estimation for Gaussian graphical models. In particular, the MLE does not exist if and only if there exists no feasible point for the dual optimization problem (9.4). In the next section, we give an algebraic description of this property. A generalization of this characterization for the existence of the MLE holds also more generally for regular exponential families; see [5, 11].

9.3 The MLE as a positive definite completion problem

To simplify notation, we use $E^* = E \cup \{(i,i) \mid i \in V\}$. We introduce the projection on the augmented edge set E^*, namely

$$\pi_G : \mathbb{S}^p_{\succeq 0} \to \mathbb{R}^{|E^*|}, \quad \pi_G(S) = \{S_{i,j} \mid (i,j) \in E^*\}.$$

Note that $\pi_G(S)$ can be seen as a partial matrix, where the entries corresponding to missing edges in the graph G have been removed (or replaced by question marks as shown in (9.5) for the case where G is the 4-cycle). In the following, we use S_G to denote the partial matrix corresponding to $\pi_G(S)$. Using this notation, the constraints in the optimization problem (9.4) become $\Sigma_G = S_G$. Hence, existence of the MLE in a Gaussian graphical model is a *positive definite matrix completion problem*: The MLE exists if and only if the partial matrix S_G can be completed to a positive definite matrix. In that case, the MLE $\hat{\Sigma}$ is the unique positive definite completion that maximizes the determinant. And as a consequence of strong duality, we obtain that $(\hat{\Sigma}^{-1})_{i,j} = 0$ for all $(i,j) \notin E^*$.

Positive definite completion problems have been widely studied in the linear algebra literature [6, 7, 22, 30]. Clearly, if a partial matrix has a positive definite completion, then every specified (i.e., with given entries) principal submatrix is positive definite. Hence, having a positive definite completion imposes some obvious necessary conditions. However, these conditions are in general not sufficient as seen in the following example, where the graph G is the 4-cycle:

$$S_G = \begin{pmatrix} 1 & 0.9 & ? & -0.9 \\ 0.9 & 1 & 0.9 & ? \\ ? & 0.9 & 1 & 0.9 \\ -0.9 & ? & 0.9 & 1 \end{pmatrix}. \tag{9.5}$$

It can easily be checked that this partial matrix does not have a positive definite completion, although all the specified 2×2-minors are positive. Hence, the MLE does not exist for the sufficient statistics given by S_G.

This example leads to the question if there are graphs for which the obvious necessary conditions are also sufficient for the existence of a positive definite matrix completion. The following remarkable theorem proven in [22] answers this question.

Theorem 9.3.1. *For a graph G the following statements are equivalent:*

(a) A G-partial matrix $M_G \in \mathbb{R}^{|E^*|}$ has a positive definite completion if and only if all completely specified submatrices in M_G are positive definite.

(b) G is chordal *(also known as* triangulated*), i.e. every cycle of length 4 or larger has a chord.*

The proof in [22] is constructive. It makes use of the fact that any chordal graph can be turned into a complete graph by adding one edge at a time in such a way, that the resulting graph remains chordal at each step. Following this ordering of edge additions, the partial matrix is completed entry by entry in such a way as to maximize the determinant of the largest complete submatrix that contains the missing entry. Hence the proof in [22] can be turned into an algorithm for finding a positive definite completion for partial matrices on chordal graphs.

We will see in Section 9.5 how to make use of positive definite completion results to determine the minimal number of observations required for existence of the MLE in a Gaussian graphical model.

9.4 ML estimation and convex geometry

After having introduced the connections to positive definite matrix completion problems, we now discuss how convex geometry enters the picture for ML estimation in Gaussian graphical models. We already introduced the set

$$\mathcal{K}_G := \{K \in \mathbb{S}^p_{\succ 0} \mid K_{i,j} = 0 \text{ for all } (i,j) \notin E^*\}.$$

Note that $\mathcal{K}_G$ is a convex cone obtained by intersecting the convex cone $\mathbb{S}^p_{\succ 0}$ with a linear subspace. We call $\mathcal{K}_G$ the *cone of concentration matrices*.

A second convex cone that plays an important role for ML estimation in Gaussian graphical models is the *cone of sufficient statistics* denoted by $\mathcal{S}_G$. It is defined as the projection of the positive semidefinite cone onto the entries E^*, i.e.,

$$\mathcal{S}_G := \pi_G(\mathbb{S}^p_{\succeq 0}).$$

In the following proposition, we show how these two cones are related to each other.

Proposition 9.4.1. *Let G be an undirected graph. Then the cone of sufficient statistics $\mathcal{S}_G$ is the dual cone to the cone of concentration matrices $\mathcal{K}_G$, i.e.*

$$\mathcal{S}_G \;=\; \big\{ S_G \in \mathbb{R}^{|E^*|} \;\mid\; \langle S_G, K\rangle \geq 0 \text{ for all } K \in \mathcal{K}_G \big\}. \tag{9.6}$$

Proof. Let $\mathcal{K}_G^\vee$ denote the dual of $\mathcal{K}_G$, i.e. the right-hand side of (9.6). Let

$$\mathcal{L}_G := \{K \in \mathbb{S}^p \mid K_{i,j} = 0 \text{ for all } (i,j) \notin E^*\}$$

denote the linear subspace defined by the graph G. We denote by $\mathcal{L}_G^\perp$ the orthogonal complement of $\mathcal{L}_G$ in $\mathbb{S}^p$. Using the fact that the dual of the full-dimensional cone $\mathbb{S}^p_{\succ 0}$ is $\mathbb{S}^p_{\succeq 0}$, i.e. $(\mathbb{S}^p_{\succ 0})^\vee = \mathbb{S}^p_{\succeq 0}$, general duality theory for convex cones (see e.g. [8]) implies:

$$\mathcal{K}_G^\vee \;=\; (\mathbb{S}^p_{\succ 0} \cap \mathcal{L}_G)^\vee \;=\; (\mathbb{S}^p_{\succeq 0} + \mathcal{L}_G^\perp)/\mathcal{L}_G^\perp \;=\; \mathcal{S}_G,$$

which completes the proof. □

It is clear from this proof that the geometric picture we have started to draw holds more generally for any Gaussian model that is given by linear constraints on the concentration matrix. We will therefore use $\mathcal{L}$ to denote any linear subspace of $\mathbb{S}^p$ and we assume that $\mathcal{L}$ intersects the interior of $\mathbb{S}^p_{\succeq 0}$. Hence, $\mathcal{L}_G$ is a special case defined by zero constraints given by missing edges in the graph G. Then,

$$\mathcal{K}_\mathcal{L} = \mathcal{L} \cap \mathbb{S}^p_{\succ 0}, \quad \mathcal{S}_\mathcal{L} = \pi_\mathcal{L}(\mathbb{S}^p_{\succeq 0}) = \mathcal{K}_\mathcal{L}^\vee,$$

where $\pi_\mathcal{L} : \mathbb{S}^p \to \mathbb{S}^p/\mathcal{L}^\perp$. Note that given a basis $K_1, \ldots, K_d$ for $\mathcal{L}$, this map can be identified with

$$\pi_\mathcal{L} : \mathbb{S}^p \to \mathbb{R}^d, \quad S \mapsto \big(\langle S, K_1\rangle, \ldots, \langle S, K_d\rangle\big).$$

A *spectrahedron* is a convex set that is defined by linear matrix inequalities. Given a sample covariance matrix S, we define the spectrahedron

$$\text{fiber}_\mathcal{L}(S) \;=\; \big\{\Sigma \in \mathbb{S}^p_{\succ 0} \mid \langle \Sigma, K\rangle = \langle S, K\rangle \text{ for all } K \in \mathcal{L}\big\}.$$

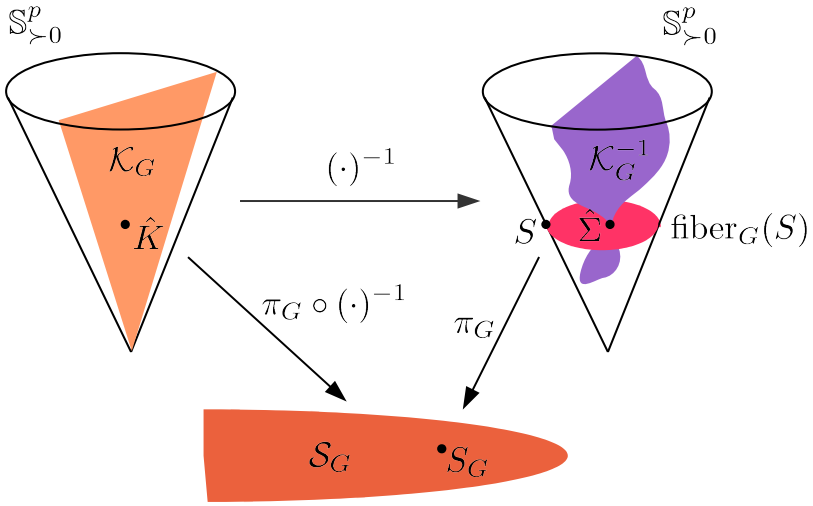


FIGURE 9.1: Geometry of maximum likelihood estimation in Gaussian graphical models. The cone $\mathcal{K}_G$ consists of all concentration matrices in the model and $\mathcal{K}_G^{-1}$ is the corresponding set of covariance matrices. The cone of sufficient statistics $\mathcal{S}_G$ is defined as the projection of $\mathbb{S}^p_{\succ 0}$ onto the (augmented) edge set E^* of G. It is dual and homeomorphic to $\mathcal{K}_G$. Given a sample covariance matrix S, $\mathrm{fiber}_G(S)$ consists of all positive definite completions of the G-partial matrix S_G, and it intersects $\mathcal{K}_G^{-1}$ in at most one point, namely the MLE $\hat{\Sigma}$.

For a Gaussian graphical model with underlying graph G this spectrahedron consists of all positive definite completions of S_G, i.e.

$$\mathrm{fiber}_G(S) \quad = \quad \left\{\Sigma \in \mathbb{S}^p_{\succ 0} \mid \Sigma_G = S_G\right\}.$$

The following theorem combines the point of view of convex optimization developed in Section 9.2, the connection to positive definite matrix completion discussed in Section 9.3, and the link to convex geometry described in this section into a result about the existence of the MLE in Gaussian models with linear constraints on the concentration matrix, which includes Gaussian graphical models as a special case. This result is essentially also given in [22, Theorem 2].

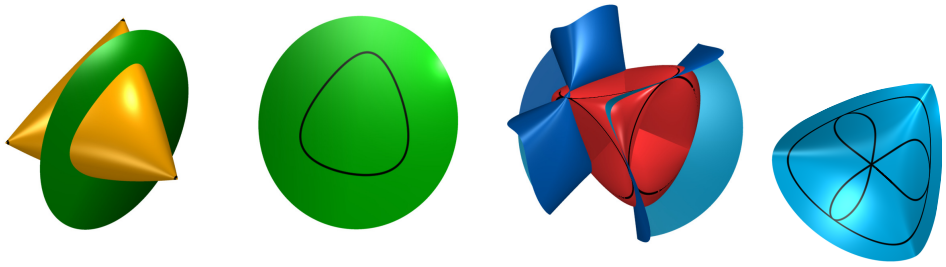
Theorem 9.4.2. *Consider a Gaussian model with linear constraints on the concentration matrix defined by $\mathcal{L}$ with $\mathcal{L} \cap \mathbb{S}^p_{\succ 0} \neq \emptyset$. Then the MLEs $\hat{\Sigma}$ and $\hat{K}$ exist for a given sample covariance matrix S if and only if* $\mathrm{fiber}_{\mathcal{L}}(S)$ *is non-empty, in which case* $\mathrm{fiber}_{\mathcal{L}}(S)$ *intersects $\mathcal{K}_{\mathcal{L}}^{-1}$ in exactly one point, namely the MLE $\hat{\Sigma}$. Equivalently, $\hat{\Sigma}$ is the unique maximizer of the determinant over the spectrahedron* $\mathrm{fiber}_{\mathcal{L}}(S)$.

Proof. This proof is a simple exercise in convex optimization; see [10] for an introduction. The ML estimation problem for Gaussian models with linear constraints on the concentration matrix is given by

$$\begin{aligned} \underset{K}{\text{maximize}} \quad & \log\det K - \mathrm{tr}(SK) \\ \text{subject to} \quad & K \in \mathcal{K}_{\mathcal{L}}. \end{aligned}$$

Its dual is

$$\begin{aligned} \underset{\Sigma}{\text{minimize}} \quad & -\log\det\Sigma - p \\ \text{subject to} \quad & \Sigma \in \mathrm{fiber}_{\mathcal{L}}(S). \end{aligned}$$



(i) $\mathcal{L}_G \cap \mathbb{S}^3_{\succeq 0}$ (ii) $\mathcal{K}_G$ (iii) $(\mathcal{K}_G)^{-1}$ (iv) $\mathcal{S}_G$

FIGURE 9.2: Geometry of Gaussian graphical models for $p = 3$. The tetrahedral-shaped pillow in (a) corresponds to the set of all 3×3 concentration matrices with ones on the diagonal. The linear subspace in (a) is defined by the missing edges in G. The resulting cone of concentration matrices is shown in (b). The corresponding set of covariance matrices is shown in (c), and the cone of sufficient statistics $\mathcal{S}_G$, dual to $\mathcal{K}_G$, is shown in (d).

Since by assumption the primal problem is strictly feasible, strong duality holds by Slater's constraint qualification with the solutions satisfying $\hat{\Sigma} = \hat{K}^{-1}$. The MLE exists, i.e. the global optimum of the two optimization problems is attained, if and only if the dual is feasible, i.e. $\text{fiber}_{\mathcal{L}}(S)$ is non-empty. Let $\Sigma \in \text{fiber}_{\mathcal{L}}(S) \cap \mathcal{K}_{\mathcal{L}}^{-1}$. Then (Σ^{-1}, Σ) satisfies the KKT conditions, namely stationarity, primal and dual feasibility, and complimentary slackness. Hence, this pair is primal and dual optimal. Thus, if $\text{fiber}_{\mathcal{L}}(S)$ is non-empty, then $\text{fiber}_{\mathcal{L}}(S)$ intersects $\mathcal{K}_{\mathcal{L}}^{-1}$ in exactly one point, namely the MLE $\hat{\Sigma}$, which is the dual optimal solution. This completes the proof. □

The geometry of ML estimation in Gaussian models with linear constraints on the concentration matrix is summarized in Figure 9.1 for the special case of Gaussian graphical models. The geometric picture for general linear concentration models is completely analogous. The convex geometry of Gaussian graphical models on 3 nodes is shown in Figure 9.2. Since a general covariance matrix on 3 nodes lives in a 6-dimensional space, we show the picture for correlation matrices instead, which live in the 3-dimensional space.

Theorem 9.4.2 was first proven for Gaussian graphical models by Dempster [15] and later more generally for regular exponential families in [5, 11]. One can show that the map

$$\pi_G \circ (\cdot)^{-1} : \mathcal{K}_G \to \mathcal{S}_G$$

in Figure 9.1 corresponds to the gradient of the log-partition function. To embed this result into the theory of regular exponential families, we denote canonical parameters by θ, minimal sufficient statistics by $t(X)$, and the log-partition function of a regular exponential family by $A(\theta)$. Then the theory of regular exponential families (see e.g. [5, 11]) implies that the gradient of the log-partition function $\nabla A(\cdot)$ defines a homeomorphism between the space of canonical parameters and the relative interior of the convex hull of sufficient statistics, and it is defined by $\nabla A(\theta) = \mathbb{E}_\theta(t(X))$. For Gaussian models we have $A(\theta) = \log\det(\theta)$; the algebraic structure in maximum likelihood estimation for Gaussian graphical models is a consequence of the fact that $\nabla A(\cdot)$ is a rational function.

The geometric results and duality theory that hold for Gaussian graphical models can be extended to all regular exponential families [5, 11]. The algebraic picture can be extended to exponential families where $\nabla A(\cdot)$ is a rational function. This was shown in [36], where it was proven that such exponential families are defined by *hyperbolic polynomials*.

The problem of existence of the MLE can be studied at the level of sufficient statistics,

i.e. in the cone $\mathcal{S}_G$, or at the level of observations. As explained in Section 9.3, the MLE exists if and only if the sufficient statistics S_G lie in the interior of the cone $\mathcal{S}_G$. Hence, analyzing existence of the MLE at the level of sufficient statistics requires analyzing the boundary of the cone $\mathcal{S}_G$. The boundary of $\mathcal{K}_G$ is defined by the hypersurface $\det(K) = 0$ with $K_{i,j} = 0$ for all $(i,j) \notin E^*$. It has been shown in [42] that the boundary of the cone $\mathcal{S}_G$ can be obtained by studying the dual of the variety defined by $\det(K) = 0$. This algebraic analysis results in conditions that characterize existence of the MLE at the level of sufficient statistics.

But perhaps more interesting from a statistical point of view, is a characterization of existence of the MLE at the level of observations. Note that if $\operatorname{rank}(S) < p$ then it can happen that $\operatorname{fiber}_{\mathcal{L}}(S)$ is empty, in which case the MLE does not exist for $(\mathcal{L}, S)$. In the next section, we discuss conditions on the number of observations n, or equivalently on the rank of S, that ensure existence of the MLE with probability 1 for particular classes of graphs.

9.5 Existence of the MLE for various classes of graphs

Since the Gaussian density is strictly positive, $\operatorname{rank}(S) = \min(n, p)$ with probability 1. The *maximum likelihood threshold* of a graph G, denoted $\operatorname{mlt}(G)$, is defined as the minimum number of observations n such that the MLE in the Gaussian graphical model with graph G exists with probability 1. This is equivalent to the smallest integer n such that for all generic positive semidefinite matrices S of rank n there exists a positive definite matrix Σ with $S_G = \Sigma_G$. Although in this section we only consider Gaussian graphical models, note that this definition can easily be extended to general linear Gaussian concentration models.

The maximum likelihood threshold of a graph was introduced by Gross and Sullivant in [23]. Ben-David [9] introduced a related but different notion, the *Gaussian rank* of a graph, namely the smallest n such that the MLE exists for every positive semidefinite matrix S of rank n for which every $n \times n$ principal submatrix is non-singular. Note that with probability 1 every $n \times n$ principal submatrix of a sample covariance matrix based on n i.i.d. samples from a Gaussian distribution is non-singular. Hence, the Gaussian rank of G is an upper bound on $\operatorname{mlt}(G)$. Since a sample covariance matrix of size $p \times p$ based on $n \leq p$ observations from a Gaussian population is of rank n with probability 1, we here concentrate on the maximum likelihood threshold of a graph.

A *clique* in a graph G is a completely connected subgraph of G. We denote by $q(G)$ the maximal clique-size of G. It is clear that the MLE cannot exist if $n < q(G)$, since otherwise the partial matrix S_G would contain a completely specified submatrix that is not positive definite (the submatrix corresponding to the maximal clique). This results in a lower bound for the maximum likelihood threshold of a graph, namely

$$\operatorname{mlt}(G) \geq q(G).$$

For chordal graphs, Theorem 9.3.1 shows that the MLE exists with probability 1 if and only if $n \geq q(G)$. Hence for chordal graphs it holds that $\operatorname{mlt}(G) = q(G)$. However, this is not the case in general as shown by the following example.

Example 9.5.1. Let G be the 4-cycle with edges $(1,2)$, $(2,3)$, $(3,4)$, and $(1,4)$. Then $q(G) = 2$. We define $X \in \mathbb{R}^{4\times 2}$ consisting of 2 samples in $\mathbb{R}^4$ and the corresponding sample

covariance matrix $S = XX^T$ by

$$X = \begin{pmatrix} 1 & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & 1 \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \quad \text{and hence} \quad S = \begin{pmatrix} 1 & \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 1 & \frac{1}{\sqrt{2}} & 0 \\ 0 & \frac{1}{\sqrt{2}} & 1 & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} & 1 \end{pmatrix}.$$

One can check that S_G cannot be completed to a positive definite matrix. In addition, there exists an open ball around X for which the MLE does not exist. This shows that in general for non-chordal graphs $\text{mlt}(G) > q(G)$.

From Theorem 9.3.1 we can determine an upper bound on $\text{mlt}(G)$ for general graphs. For a graph $G = (V, E)$ we denote by $G^+ = (V, E^+)$ a *chordal cover* of G, i.e. a chordal graph satisfying $E \subseteq E^+$. We denote the maximal clique size of G^+ by q^+. A *minimal chordal cover*, denoted by $G^\# = (V, E^\#)$, is a chordal cover of G, whose maximal clique size $q^\#$ achieves $q^\# = min(q^+)$ over all chordal covers of G. The quantity $q^\#(G) - 1$ is also known as the *treewidth* of G. It follows directly from Theorem 9.3.1 that

$$\text{mlt}(G) \leq q^\#(G),$$

since if $S_{G^\#}$ can be completed to a positive definite matrix, so can S_G.

If G is a cycle, then $q(G) = 2$ and $q^\#(G) = 3$. Hence the MLE does not exist for $n = 1$ and it exists with probability 1 for $n = 3$. From Example 9.5.1 we can conclude that for cycles $\text{mlt}(G) = 3$. Buhl [12] shows that for $n = 2$ the MLE exists with probability in $(0, 1)$. More precisely, for $n = 2$ we can view the two samples as vectors $x_1, \ldots, x_p \in \mathbb{R}^2$. We denote by $\ell_1, \ldots, \ell_p$ the lines defined by $x_1, \ldots, x_p$. Then Buhl [12] shows using an intricate trigonometric argument that the MLE for the p-cycle for $n = 2$ exists if and only if the lines $\ell_1, \ldots, \ell_p$ do not occur in one of the two sequences conforming with the ordering in the cycle G as shown in Figure 9.3. In the following, we give an algebraic proof of this result by using an intriguing characterization of positive definiteness for 3×3 symmetric matrices given in [7].

Proposition 9.5.2 (Barrett et al. [7]). *The matrix*

$$\begin{pmatrix} 1 & \cos(\alpha) & \cos(\beta) \\ \cos(\alpha) & 1 & \cos(\gamma) \\ \cos(\beta) & \cos(\gamma) & 1 \end{pmatrix}$$

with $0 < \alpha, \beta, \gamma < \pi$ *is positive definite if and only if*

$$\alpha < \beta + \gamma, \quad \beta < \alpha + \gamma, \quad \gamma < \alpha + \beta, \quad \alpha + \beta + \gamma < 2\pi.$$

Let G denote the p-cycle. Then, as shown in [7], this result can be used to give a characterization for completability of a G-partial matrix to a positive definite matrix through induction on the cycle length p.

Corollary 9.5.3 (Barrett et al. [7]). *Let G be the p-cycle. Then the G-partial matrix*

$$\begin{pmatrix} 1 & \cos(\theta_1) & & & \cos(\theta_p) \\ \cos(\theta_1) & 1 & \cos(\theta_2) & ? & \\ & \cos(\theta_2) & 1 & & \\ & ? & & \ddots & \cos(\theta_{p-1}) \\ \cos(\theta_p) & & & \cos(\theta_{p-1}) & 1 \end{pmatrix}$$

with $0 < \theta_1, \theta_2, \ldots \theta_p < \pi$ *has a positive definite completion if and only if for each* $S \subseteq [p]$ *with* $|S|$ *odd,*

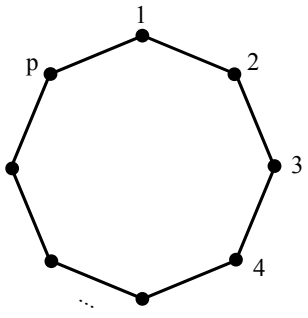
$$\sum_{i \in S} \theta_i < (|S| - 1)\pi + \sum_{j \notin S} \theta_j.$$

Buhl's result [12] can easily be deduced from this algebraic result about the existence of positive definite completions: For $n = 2$ we view the observations as vectors $x_1, \ldots, x_p \in \mathbb{R}^2$. Note that we can rescale and rotate the data vectors $x_1, \ldots, x_p$ (i.e. perform an orthogonal transformation) without changing the problem of existence of the MLE. So without loss of generality we can assume that the vectors $x_1, \ldots, x_p \in \mathbb{R}^2$ have length one, lie in the upper unit half circle, and $x_1 = (1, 0)$. Now we denote by θ_i the angle between x_i and x_{i+1}, where $x_{p+1} := x_1$. One can show that the angular conditions in Corollary 9.5.3 are equivalent to requiring that the vectors $x_1, \ldots, x_p \in \mathbb{R}^2$ do not occur in one of the two sequences conforming with the ordering in the cycle G as shown in Figure 9.3.

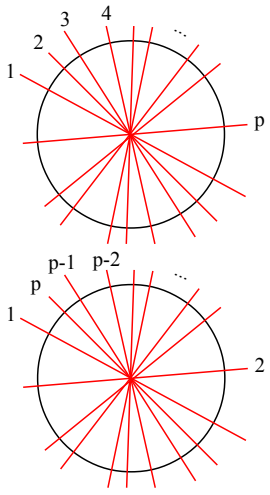
Hence, for a G-partial matrix to be completable to a positive definite matrix, it is necessary that every submatrix corresponding to a clique in the graph is positive definite and every partial submatrix corresponding to a cycle in G satisfies the conditions in Corollary 9.5.3. Barrett et al. [6] characterized the graphs for which these conditions are sufficient for existence of a positive definite completion. They showed that this is the case for graphs that have a chordal cover with no new 4-cliques. Such graphs can be obtained as a *clique sum* of chordal graphs and series-parallel graphs (i.e. graphs G with $q^{\#}(G) \leq 3$) [28]. To be more precise, for such graphs $G = (V, E)$ the vertex set can be decomposed into three disjoint subsets $V = V_1 \cup V_2 \cup V_3$ such that there are no edges between V_1 and V_3, the subgraph induced by V_2 is a clique, and the subgraphs induced by $V_1 \cup V_2$ and $V_2 \cup V_3$ are either chordal or series-parallel graphs or can themselves be decomposed as a clique sum of chordal or series-parallel graphs. For such graphs it follows that

$$\text{mlt}(G) = \max(3, q(G)) = q^{\#}(G).$$

This raises the question whether there exist graphs for which $\text{mlt}(G) < q^{\#}(G)$, i.e., graphs



(i) p-cycle.



(ii) Line configurations for which the MLE does not exist.

FIGURE 9.3: Buhl's geometric criterion [12] for existence of the MLE for $n = 2$ in a Gaussian graphical model on the p-cycle.

for which the MLE exists with probability 1 even if the number of observations is strictly smaller than the maximal clique size in a minimal chordal cover of G. This question has been answered to the positive for 3×3 grids using an algebraic argument in [45] and more generally for grids of size $m \times m$ using a combinatorial argument in [23]. In particular, let G be a grid of size $m \times m$. Then $q^{\#}(G) = m + 1$, but it was shown in [23] that the MLE exists with probability 1 for $n = 3$, independent of the grid size m. Grids are a special class of planar graphs. Gross and Sullivant [23] more generally proved that for any planar graph it holds that $\mathrm{mlt}(G) \leq 4$.

9.6 Algorithms for computing the MLE

After having discussed when the MLE exists, we now turn to the question of how to compute the MLE for Gaussian graphical models. As described in Section 9.2, determining the MLE in a Gaussian model with linear constraints on the inverse covariance matrix is a convex optimization problem. Hence, it can be solved in polynomial time for instance using interior point methods [10]. These are implemented for example in `cvx`, a user-friendly `matlab` software for disciplined convex programming [21].

Although interior point methods run in polynomial time, for very large Gaussian graphical models it is usually more practical to apply coordinate descent algorithms. The idea of using coordinate descent algorithms for computing the MLE in Gaussian graphical models was already present in the original paper by Dempster [15]. Coordinate descent on the entries of Σ was first implemented by Wermuth and Scheidt [47] and is shown in Algorithm 9.1. In this algorithm, we start with $\Sigma^0 = S$ and iteratively update the entries $(i, j) \notin E^*$ by maximizing the log-likelihood in direction $\Sigma_{i,j}$ and keeping all other entries fixed.

Note that step (2) in Algorithm 9.1 can be given in closed-form: Let $A = \{u, v\}$ and $B = V \setminus A$. We now show that the objective function in step (2) of Algorithm 9.1 can be written in terms of the 2×2 Schur complement $\Sigma' = \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}$. To do this, note that $\det(\Sigma) = \det(\Sigma') \det(\Sigma_{B,B})$. Since $\Sigma_{B,B}$ is held constant in the optimization problem, then up to an additive constant it holds that

$$\log\det(\Sigma) = \log\det(\Sigma').$$

Thus, the optimization problem in step (2) of Algorithm 9.1 is equivalent to

$$\begin{aligned} \underset{\Sigma' \succeq 0}{\text{maximize}} \quad & \log\det(\Sigma') \\ \text{subject to} \quad & \Sigma'_{i,i} = \Sigma^0_{i,i} - \Sigma^0_{i,B}(\Sigma^0_{B,B})^{-1}\Sigma^0_{B,i}, \quad i \in A, \end{aligned}$$

and the global maximum is attained by $\Sigma'_{u,v} = 0$. Hence, the solution to the univariate optimization problem in step (2) of Algorithm 9.1 is

$$\Sigma_{u,v} = \Sigma_{u,B}\Sigma_{B,B}^{-1}\Sigma_{B,v},$$

forcing the corresponding entry of Σ^{-1} to be equal to zero.

Dual to this algorithm, one can define an equivalent algorithm that cycles through entries of the concentration matrix corresponding to $(i, j) \in E$, starting in the identity matrix. This procedure is shown in Algorithm 9.2. Similarly as for Algorithm 9.1, the solution to the optimization problem in step (2) can be given in closed-form. Defining as before, $A = \{u, v\}$ and $B = V \setminus A$, then analogously as in the derivation above, one can show that the solution

Algorithm 9.1 Coordinate descent on Σ

Input: Graph $G = (V, E)$, sample covariance matrix S, and precision ϵ.
Output: MLE $\hat{\Sigma}$.

1: Let $\Sigma^0 = S$.
2: Cycle through $(u, v) \notin E^*$ and solve the following optimization problem:

$$\begin{aligned} \underset{\Sigma \succeq 0}{\text{maximize}} \quad & \log\det(\Sigma) \\ \text{subject to} \quad & \Sigma_{i,j} = \Sigma^0_{i,j} \text{ for all } (i,j) \neq (u,v) \end{aligned}$$

and update $\Sigma^1 := \Sigma$.
3: **if** $\|\Sigma^0 - \Sigma^1\|_1 < \epsilon$ **then**
4: let $\hat{\Sigma} := \Sigma^1$
5: **else**
6: let $\Sigma^0 := \Sigma^1$ and return to line 2.
7: **end if**

to the optimization problem in step (2) of Algorithm 9.2 is

$$K_{A,A} = (S_{A,A})^{-1} + K_{A,B}K^{-1}_{B,B}K_{B,A},$$

forcing $\Sigma_{A,A}$ to be equal to $S_{A,A}$. This algorithm, which tries to match the sufficient statistics, is analogous to *iterative proportional scaling* for computing the MLE in contingency tables [24]. Convergence proofs for both algorithms were given by Speed and Kiiveri [41].

In general the MLE must be computed iteratively. However, in some cases estimation can be made in closed form. A trivial case when the MLE of a Gaussian graphical model can be given explicitly is for complete graphs: In this case, assuming that the MLE exists, i.e. S is non-singular, then $\hat{K} = S^{-1}$. In [32, Section 5.3.2], Lauritzen showed that also for chordal graphs the MLE has a closed-form solution. This result is based on the fact that any chordal graph $G = (V, E)$ is a clique sum of cliques, i.e., the vertex set can be decomposed into three disjoint subsets $V = A \cup B \cup C$ such that there are no edges between A and C, the subgraph induced by B is a clique, and the subgraphs induced by $A \cup B$ and $B \cup C$

Algorithm 9.2 Coordinate descent on K

Input: Graph $G = (V, E)$, sample covariance matrix S, and precision ϵ.
Output: MLE $\hat{K}$.

1: Let $K^0 = \mathrm{Id}$.
2: Cycle through $(u, v) \in E$ and solve the following optimization problem:

$$\begin{aligned} \underset{K \succeq 0}{\text{maximize}} \quad & \log\det(K) - \operatorname{trace}(KS) \\ \text{subject to} \quad & K_{i,j} = K^0_{i,j} \text{ for all } (i,j) \in (V \times V) \setminus \{(u,u),(v,v),(u,v)\} \end{aligned}$$

and update $K^1 := K$.
3: **if** $\|K^0 - K^1\|_1 < \epsilon$ **then**
4: let $\hat{K} := K^1$
5: **else**
6: let $K^0 := K^1$ and return to line 2.
7: **end if**

are either cliques or can themselves be decomposed as a clique sum of cliques. In such a decomposition, B is known as a *separator*. In [32, Proposition 5.9], Lauritzen shows that, assuming existence of the MLE, then the MLE for a chordal Gaussian graphical model is given by

$$\hat{K} = \sum_{C \in \mathcal{C}} \left[(S_{C,C})^{-1} \right]^{\text{fill}} - \sum_{B \in \mathcal{B}} \left[(S_{B,B})^{-1} \right]^{\text{fill}}, \tag{9.7}$$

where $\mathcal{C}$ denotes the maximal cliques in G, $\mathcal{B}$ denotes the separators in the clique decomposition of G (with multiplicity, i.e., a clique could appear more than once), and $[A_{HH}]^{\text{fill}}$ denotes a $p \times p$ matrix, where the submatrix corresponding to $H \subset V$ is given by A and all the other entries are filled with zeros.

To gain more insight into the formula (9.7), consider the simple case where the subgraphs corresponding to $A \cup B$ and $B \cup C$ are cliques. Then (9.7) says that the MLE is given by

$$\hat{K} = \left[S_1^{-1} \right]^{\text{fill}} + \left[S_2^{-1} \right]^{\text{fill}} - \left[S_B^{-1} \right]^{\text{fill}}, \tag{9.8}$$

where we simplified notation by setting $S_1 = S_{AB,AB}$, $S_2 = S_{BC,BC}$, and $S_B = S_{B,B}$, also to clarify that we first take the submatrix and then invert it. To prove (9.8), it suffices to show that $\left(\hat{K}^{-1}\right)_G = S_G$, since $\hat{K}_{i,j} = 0$ for all $(i,j) \notin E^*$. We first expand $\hat{K}$ and then use Schur complements to compute its inverse:

$$\hat{K} = \begin{pmatrix} \left(S_1^{-1}\right)_{A,A} & \left(S_1^{-1}\right)_{A,B} & 0 \\ \left(S_1^{-1}\right)_{B,A} & \left(S_1^{-1}\right)_{B,B} + \left(S_2^{-1}\right)_{B,B} - S_B^{-1} & \left(S_2^{-1}\right)_{B,C} \\ 0 & \left(S_2^{-1}\right)_{C,B} & \left(S_2^{-1}\right)_{C,C} \end{pmatrix}. \tag{9.9}$$

Denoting $\hat{K}^{-1}$ by $\hat{\Sigma}$ and using Schur complements, we obtain

$$\hat{\Sigma}_{AB,AB} = \begin{pmatrix} \left(S_1^{-1}\right)_{A,A} & \left(S_1^{-1}\right)_{A,B} \\ \left(S_1^{-1}\right)_{B,A} & \left(S_1^{-1}\right)_{B,B} + \left(S_2^{-1}\right)_{B,B} - S_B^{-1} - \left(S_2^{-1}\right)_{B,C} \left((S_2^{-1})_{C,C}\right)^{-1} \left(S_2^{-1}\right)_{C,B}. \end{pmatrix}^{-1}$$

Note that by using Schur complements once again,

$$\left(S_2^{-1}\right)_{B,B} - \left(S_2^{-1}\right)_{B,C} \left((S_2^{-1})_{C,C}\right)^{-1} \left(S_2^{-1}\right)_{C,B} = S_B^{-1},$$

and hence $\hat{\Sigma}_{AB,AB} = S_1$. Analogously, it follows that $\hat{\Sigma}_{BC,BC} = S_2$, implying that $\hat{\Sigma}_G = S_G$. The more general formula for the MLE of chordal Gaussian graphical models in (9.7) is obtained by induction and repeated use of (9.9).

A stronger property than existence of a closed-form solution for the MLE is to ask which Gaussian graphical models have rational formulas for the MLE in terms of the entries of the sample covariance matrix. An important observation is that the number of critical points to the likelihood equations is constant for generic data, i.e., it is constant with probability 1 (it can be smaller on a measure zero subspace). The number of solutions to the likelihood equations for generic data, or equivalently, the maximum number of solutions to the likelihood equations, is called the *maximum likelihood degree* (*ML degree*). Hence, a model has a rational formula for the MLE if and only if it has ML degree 1. It was shown in [42] that the ML degree of a Gaussian graphical model is 1 if and only if the underlying graph is chordal. The ML degree of the 4-cycle can easily be computed and is known to be 5; see [16, Example 2.1.13] for some code on how to do the computation using the open-source computer algebra system `Singular` [14]. It is conjectured in [16, Section 7.4] that the ML degree of the cycle grows exponentially in the cycle length, namely as $(p-3)2^{p-2}+1$, where $p \geq 3$ is the cycle length.

Since the likelihood function is strictly concave for Gaussian graphical models, this implies that even when the ML degree is larger than 1, there is still a unique local maximum of the likelihood function. As a consequence, while there are multiple complex solutions to the ML equations for non-chordal graphs, there is always a unique solution that is real and results in a positive definite matrix.

9.7 Learning the underlying graph

Until now we have assumed that the underlying graph is given to us. In this section, we present methods for learning the underlying graph. We here only provide a short overview of some of the most prominent methods for model selection in Gaussian graphical models; for more details and for practical examples, see [25].

A popular method for performing model selection is to take a stepwise approach. We start in the empty graph (or in the complete graph) and run a forward search (or a backward search). We cycle through the possible edges and add an edge (or remove an edge) if it decreases some criterion. Alternatively, one can also search for the edge which minimizes some criterion and add (or remove) this edge, but this is considerably slower. Two popular objective functions are the *Akaike information criterion* (AIC) and the *Bayesian information criterion* (BIC) [1, 39]. These criteria are based on penalizing the likelihood according to the model complexity, i.e.

$$-2\ell + \lambda|E|, \tag{9.10}$$

where ℓ is the log-likelihood function, λ is a parameter that penalizes model complexity, and $|E|$ denotes the number of edges, or equivalently, the number of parameters in the model. The AIC is defined by choosing $\lambda = 2$, whereas the BIC is defined by setting $\lambda = \log(n)$ in (9.10).

Alternatively, one can also use significance tests for testing whether a particular partial correlation is zero and removing the corresponding edge accordingly. A hypothesis test for zero partial correlation can be built based on Fisher's z-transform [19]: For testing whether $K_{i,j} = 0$, let $A = \{i, j\}$ and $B = V \setminus A$. In Proposition 9.1.1 we saw that $K_{A,A}^{-1} = \Sigma_{A|B}$. Hence testing whether $K_{i,j} = 0$ is equivalent to testing whether the correlation $\rho_{i,j|B}$ is zero. The sample estimate of $\rho_{i,j|B}$ is given by

$$\hat{\rho}_{i,j|B} = S_{i,j} - S_{i,B}S_{B,B}^{-1}S_{B,j}.$$

Fisher's z-transform is defined by

$$\hat{z}_{i,j|B} = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{i,j|B}}{1-\hat{\rho}_{i,j|B}}\right).$$

Fisher [19] showed that using the test statistic $T_n = \sqrt{n-p+2-3}|\hat{z}_{i,j|B}|$ with a rejection region $R_n = (-\Phi^{-1}(1-\alpha/2), \Phi^{-1}(1-\alpha/2))$, where Φ denotes the cumulative distribution function of $\mathcal{N}(0,1)$, leads to a test of size α.

A problem with stepwise selection strategies is that they are impractical for large problems or only a small part of the relevant search space can be covered during the search. A simple alternative, but a seemingly naive method for model selection in Gaussian graphical models, is to set a specific threshold for the partial correlations and remove all edges corresponding to the partial correlations that are less than the given threshold. This often works well, but a disadvantage is that the resulting estimate of the inverse covariance matrix might not be positive definite.

An alternative is to use the *glasso* algorithm [20]. It is based on maximizing the ℓ_1-penalized log-likelihood function, i.e.

$$\ell_{\text{pen}}(K) = \log\det(K) - \text{tr}(KS) - \lambda|K|_1,$$

where λ is a non-negative parameter that penalizes model complexity and $|K|_1$ is the sum of the absolute values of the off-diagonal elements of the concentration matrix. The use of $|K|_1$ is a convex proxy for the number of non-zero elements of K and allows efficient optimization of the penalized log-likelihood function by convex programming methods such as interior point algorithms or coordinate descent approaches similar to the ones discussed in Section 9.6; see e.g. [34]. A big advantage of using ℓ_1-penalized maximum likelihood estimation for model selection in Gaussian graphical models is that it can also be applied in the high-dimensional setting and comes with structural recovery guarantees [38]. Various alternative methods for learning high-dimensional Gaussian graphical models have been proposed that have similar guarantees, including node-wise regression with the lasso [35], a constrained ℓ_1-minimization approach for inverse matrix estimation (CLIME) [13], and a testing approach with false discovery rate control [33]. Graphical models in the high-dimensional setting are discussed in detail in Chapter 12.

9.8 Other Gaussian models with linear constraints

Gaussian graphical models are Gaussian models with particular equality constraints on the concentration matrix, namely where some of the entries are set to zero. We end this chapter by giving an overview on other Gaussian models with linear constraints.

Gaussian graphical models can be generalized by introducing a vertex and edge coloring: Let $G = (V, E)$ be an undirected graph, where the vertices are colored with s different colors and the edges with t different colors. This leads to a partition of the vertex and edge set into color classes, namely,

$$V = V_1 \cup V2 \cup V_s, \quad s \leq p, \quad \text{and} \quad E = E_1 \cup E_2 \cup \cdots \cup E_t, \quad t \leq |E|.$$

An *RCON* model on G is a Gaussian graphical model on G with some additional equality constraints, namely that $K_{i,i} = K_{j,j}$ if i and j are in the same vertex color class and $K_{i,j} = K_{u,v}$ if (i, j) and (u, v) are in the same edge color class. Hence a Gaussian graphical model on a graph G is an RCON model on G, where each vertex and edge has a separate color.

Determining the MLE for RCON models leads to a convex optimization problem and the corresponding dual optimization problem can be readily computed:

$$\begin{aligned}
\underset{\Sigma \succeq 0}{\text{minimize}} \quad & -\log\det\Sigma - p \\
\text{subject to} \quad & \sum_{\alpha \in V_i} \Sigma_{\alpha,\alpha} = \sum_{\alpha \in V_i} S_{\alpha,\alpha}, \qquad \text{for all } 1 \leq i \leq s, \\
& \sum_{(\alpha,\beta) \in E_j} \Sigma_{\alpha,\beta} = \sum_{(\alpha,\beta) \in E_j} S_{\alpha,\beta}, \quad \text{for all } 1 \leq j \leq t.
\end{aligned}$$

This shows that the constraints for existence of the MLE in an RCON model on a graph G are relaxed as compared to a Gaussian graphical model on G; namely, in an RCON model the constraints are only on the sum of the entries in a color class, whereas in a Gaussian graphical model the constraints are on each entry.

RCON models were introduced by Højsgaard and Lauritzen in [26]. These models are useful for applications, where symmetries in the underlying model can be assumed. Adding symmetries reduces the number of parameters and in some cases also the number of observations needed for existence of the MLE. For example, defining G to be the 4-cycle and having only one vertex color class and one edge color class (i.e., we color each vertex in the same color and each edge in the same color), then one can show that the MLE already exists for 1 observation with probability 1. This is in contrast to the result that $\text{mlt}(G) = 3$ for cycles as shown in Section 9.5. For further examples see [26, 45].

More general Gaussian models with linear equality constraints on the concentration matrix or the covariance matrix were introduced by Anderson [2]. He was motivated by the linear structure of covariance and concentration matrices resulting from various time series models. As pointed out in Section 9.2, the Gaussian likelihood as a function of Σ is not concave over the whole cone of positive definite matrices. Hence maximum likelihood estimation for Gaussian models with linear constraints on the covariance matrix in general does not lead to a convex optimization problem and has many local maxima. Anderson proposed iterative procedures for calculating the MLE for such models, such as the Newton-Raphson method [2] and a scoring method [3].

As mentioned in Section 9.2, while not being concave over the whole cone of positive definite matrices, the Gaussian likelihood as a function of Σ is concave over a large region of $\mathbb{S}^p_{\succ 0}$, namely for all Σ that satisfy $\Sigma - 2S \in \mathbb{S}^p_{\succ 0}$. This is useful, since it was shown in [48] that the MLE for Gaussian models with linear equality constraints on the covariance matrix lies in this region with high probability as long as the sample size is sufficiently large ($n \simeq 14p$). Hence in this regime, maximum likelihood estimation for linear Gaussian covariance models behaves as if it were a convex optimization problem.

Similarly as we posed the question for Gaussian graphical models in Section 9.6, one can ask when the MLE of a linear Gaussian covariance model has a closed form representation. Szatrowski showed in [43, 44] that the MLE for linear Gaussian covariance models has an explicit representation if and only if Σ and Σ^{-1} satisfy the same linear constraints. This is equivalent to requiring that the linear subspace $\mathcal{L}$, which defines the model, forms a Jordan algebra, i.e., if $\Sigma \in \mathcal{L}$ then also $\Sigma^2 \in \mathcal{L}$ [27]. Furthermore, Szatrowski proved that for this model class Anderson's scoring method [3] yields the MLE in one iteration when initiated at any positive definite matrix in the model.

Linear inequality constraints on the concentration matrix also lead to a convex optimization problem for ML estimation. An example of such models are Gaussian distributions that are *multivariate totally positive of order two* (MTP_2). This is a form of positive dependence, which for Gaussian distributions implies that $K_{i,j} \leq 0$ for all $i \neq j$. Gaussian MTP_2 distributions were studied by Karlin and Rinott [29] and more recently in [31, 40] from a machine learning and more applied perspective. It was shown in [18] that MTP_2 distributions have remarkable properties with respect to conditional independence constraints. In addition, for such models the spanning forest of the sample correlation matrix is always a subgraph of the maximum likelihood graph, which can be used to speed up graph learning algorithms [31]. Furthermore, the MLE for MTP_2 Gaussian models exists already for 2 observations with probability 1 [40]. These properties make MTP_2 Gaussian models interesting for the estimation of high-dimensional graphical models.

We end by referring to Pourahmadi [37] for a comprehensive review of covariance estimation in general and a discussion of numerous other specific covariance matrix constraints.

Bibliography

[1] H. Akaike. A new look at the statistical identification problem. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] T. W. Anderson. Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices. In *Essays in Probability and Statistics*, pages 1–24. University of North Carolina Press, Chapel Hill, N.C., 1970.

[3] T. W. Anderson. Asymptotically efficient estimation of covariance matrices with linear structure. *Annals of Statistics*, 1:135–141, 1973.

[4] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, New Jersey, third edition, 2003.

[5] O. Barndorff-Nielsen. *Information and Exponential Families in Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester, 1978.

[6] W. Barrett, C. Johnson, and R. Loewy. The real positive definite completion problem: cycle completability. *Memoirs of the American Mathematical Society*, 584:69, 1996.

[7] W. Barrett, C. Johnson, and P. Tarazaga. The real positive definite completion problem for a simple cycle. *Linear Algebra and its Applications*, 192:3–31, 1993.

[8] A. I. Barvinok. *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, 2002.

[9] E. Ben-David. Sharp lower and upper bounds for the Gaussian rank of a graph. *Journal of Multivariate Analysis*, 139:207–218, 2014.

[10] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[11] L. D. Brown. *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*, volume 9 of *Institute of Mathematical Statistics Lecture Notes—Monograph Series*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[12] S. L. Buhl. On the existence of maximum likelihood estimators for graphical Gaussian models. *Scandinavian Journal of Statistics*, 20:263–270, 1993.

[13] T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607, 2011.

[14] W. Decker, G.-M. Greuel, G. Pfister, and H. Schönemann. SINGULAR 4-0-2 — A computer algebra system for polynomial computations. `http://www.singular.uni-kl.de`, 2015.

[15] A. P. Dempster. Covariance selection. *Biometrics*, 28:157–175, 1972.

[16] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*, volume 39 of *Oberwolfach Seminars*. Springer, 2009.

[17] M. L. Eaton. *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons, New York, 1983.

[18] S. Fallat, S. L. Lauritzen, K. Sadeghi, C. Uhler, N. Wermuth, and P. Zwiernik. Total positivity in Markov structures. *Annals of Statistics*, 45:1152–1184, 2017.

[19] R. A. Fisher. Frequency distribution of the values of the correlation coefficient samples of an indefinitely large population. *Biometrika*, 10:507–521, 1915.

[20] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.

[21] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`, March 2014.

[22] R. Grone, C. R. Johnson, E. M. de Sá, and H. Wolkowicz. Positive definite completions of partial hermitian matrices. *Linear Algebra and its Applications*, 58:109–124, 1984.

[23] E. Gross and S. Sullivant. The maximum likelihood threshold of a graph. To appear in *Bernoulli*, 2014.

[24] S. J. Haberman. *The Analysis of Frequency Data.* Statistical Research Monographs. University of Chicago Press, Chicago, 1974.

[25] S. Hojsgaard, D. Edwards, and S. L. Lauritzen. *Graphical Models with R.* Use R! Springer, New York, 2012.

[26] S. Hojsgaard and S. L. Lauritzen. Graphical Gaussian models with edge and vertex symmetries. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70:1005–1027, 2008.

[27] S. T. Jensen. Covariance hypotheses which are linear in both the covariance and the inverse covariance. *Annals of Statistics*, 16(1):302–322, 1988.

[28] C. R. Johnson and T. A. McKee. Structural conditions for cycle completable graphs. *Discrete Mathematics*, 159:155–160, 1996.

[29] S. Karlin and Y. Rinott. M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, 52:419 – 438, 1983.

[30] M. Laurent. The real positive semidefinite completion problem for series-parallel graphs. *Linear Algebra and its Applications*, 252:347–366, 1997.

[31] S. Lauritzen, C. Uhler, and P. Zwiernik. Maximum likelihood estimation in gaussian models under total positivity. Preprint available at *https://arxiv.org/abs/1702.04031*, 2017.

[32] S. L. Lauritzen. *Graphical Models.* Oxford University Press, 1996.

[33] W. Liu. Gaussian graphical model estimation with false discovery rate control. *Annals of Statistics*, 41:2948–2978, 2013.

[34] R. Mazumder and T. Hastie. The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, 6:2125–2149, 2012.

[35] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[36] M. Michalek, B. Sturmfels, C. Uhler, and P. Zwiernik. Exponential varieties. *Proceedings of the London Mathematical Society*, 112:27–56, 2016.

[37] M. Pourahmadi. Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, 3:369–387, 2011.

[38] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[39] G. Schwarz. Estimating the dimension of a model. *Annals of Mathematical Statistics*, 6:461–464, 1978.

[40] M. Slawski and M. Hein. Estimation of positive definite M-matrices and structure learning for attractive Gaussian Markov random field. *Linear Algebra and its Applications*, 473:145–179, 2015.

[41] T. P. Speed and H. T. Kiiveri. Gaussian Markov distributions over finite graph. *Annals of Statistics*, 14:138–150, 1986.

[42] B. Sturmfels and C. Uhler. Multivariate Gaussians, semidefinite matrix completion, and convex algebraic geometry. *Annals of the Institute of Statistical Mathematics*, 62:603–638, 2010.

[43] T. H. Szatrowski. Necessary and sufficient conditions for explicit solutions in the multivariate normal estimation problem for patterned means and covariances. *Annals of Statistics*, 8:802–810, 1980.

[44] T. H. Szatrowski. Patterned covariances. In *Encyclopedia of Statistical Sciences*, pages 638–641. Wiley, New York, 1985.

[45] C. Uhler. Geometry of maximum likelihood estimation in Gaussian graphical models. *Annals of Statistics*, 40:238–261, 2012.

[46] M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*, volume 1 of *Foundations and Trends in Machine Learning*. 2008.

[47] N. Wermuth and E. Scheidt. Fitting a covariance selection model to a matrix. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26:88–92, 1977.

[48] P. Zwiernik, C. Uhler, and D. Richards. Maximum likelihood estimation for linear Gaussian covariance models. To appear in *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2016.