On perfectness in Gaussian graphical models

Arash A. Amini
University of California, Los Angeles

Bryon Aragam
University of Chicago

Qing Zhou
University of California, Los Angeles

Abstract

Knowing when a graphical model perfectly encodes the conditional independence structure of a distribution is essential in applications, and this is particularly important when performing inference from data. When the model is perfect, there is a one-to-one correspondence between conditional independence statements in the distribution and separation statements in the graph. Previous work has shown that almost all models based on linear directed acyclic graphs as well as Gaussian chain graphs are perfect, the latter of which subsumes Gaussian graphical models (i.e., the undirected Gaussian models) as a special case. In this paper, we directly approach the problem of perfectness for the Gaussian graphical models, and provide a new proof, via a more transparent parametrization, that almost all such models are perfect. Our approach is based on, and substantially extends, a construction of Lněnička and Matúš showing the existence of a perfect Gaussian distribution for any graph. The analysis involves constructing a probability measure on the set of normalized covariance matrices Markov with respect to a graph that may be of independent interest.

1 INTRODUCTION

Graphical models are among the most common approaches to modeling dependencies in multivariate data (Lauritzen, 1996; Koller and Friedman, 2009). They are a foundational object of study in statistics and machine learning, and have found a variety of applications in causal inference, medicine, finance, distributed systems, and climate science. Recently, graphical models have also found applications in interpretable data science owing to their flexibility and natural interpretability (Al-Shedivat et al., 2017; Nguyen et al., 2018; Johnson et al., 2016).

To be concrete, consider a random vector $X = (X_1, \dots, X_d) \in \mathbb{R}^d$. The general idea behind graphical modeling is to represent the conditional independence (CI) statements satisfied by the multivariate distribution $\mathbb{P}$ of X by the *separating sets* in a graph $G = (V, E)$ with nodes $V = \{X_1, \dots, X_d\}$. Whenever graph separation in G implies conditional independence in $\mathbb{P}$, the distribution is said to be *Markov* with respect to (w.r.t.) G and we have a graphical model for $\mathbb{P}$; see Section 1.2 for details. In this paper, we focus on undirected graphs (UGs), in which case G is called a conditional independence graph (CIG) for $\mathbb{P}$.

A question that arises is to what extent such correspondence is possible for a given distribution. A particular case of interest is when the correspondence is exact, that is, the set of CI statements entailed by the distribution is the same as the set of separation statements in the graph. If this desirable property holds, the distribution $\mathbb{P}$ is said to be *perfect* with respect to the graph G. In other words, in the perfect case, both the Markov property above and its reverse implication hold (i.e., CI in $\mathbb{P}$ implies graph separation in G, also known as *faithfulness*). Thus, we can "read off" the CI relations in $\mathbb{P}$ by inspecting the graph G. This in turn makes perfectness a key assumption for learning the structure of graphical models from data.

In previous work (Spirtes and Schienes, 1993; Meek, 1995; Levitz et al., 2001; Peña, 2011), it has been shown that almost all linear directed acyclic graph (DAG) and Gaussian chain graph (CG) models are perfect. In this work, we consider the case of undirected Gaussian graphical models (GGMs), i.e. $X \sim N(0, \Sigma)$, and show that almost all of them are perfect. In other words, almost all Gaussian distributions are capable of being perfectly represented by an undirected graph G. Technically speaking, the results of Levitz et al. (2001); Peña (2011) already show perfectness for almost all Gaussian distributions that factor according to a UG (i.e. as a special case of a CG), however, the constructions and proofs are obscured by the complexity of the

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

CG case. In particular, although showing essentially the same result, the proofs in Peña (2011) and Levitz et al. (2001) use two different indirect parametrizations of the CG-Markov Gaussian distributions. In this paper, we provide a much simpler and more direct parameterization for the undirected case, which should be of independent interest. Our technique is based on an elegant construction of Lněnička and Matúš (2007) which was used to prove the existence of a perfect Gaussian distribution for any given UG. We extend this construction to a full parametrization of the UG-Markov Gaussian distributions and prove the so-called *strong completeness* of this class (i.e. that almost all are perfect).

Contributions Our main contributions can be outlined as follows:

- Our main result is a new, direct proof that almost all Gaussian distributions are capable of being perfectly represented by an undirected graph G (Theorem 1).
- As a matter of independent interest, our proof involves a simpler constructive description of the set of imperfect covariance matrices, which provides useful intuition for understanding perfectness assumption in modeling and estimation with UGs (Theorem 2).
- Finally, as a byproduct of our proof, we construct a probability measure over inverse covariance matrices supported on the edge set of a graph G. This measure may be used as a trial or proposal distribution in Monte Carlo algorithms to simulate from many distributions over positive definite matrices with support restriction (Section 4).

The paper is organized as follows: Section 1.1 reviews related work and Section 1.2 provides some background on graphical modeling and sets up the notation. Section 2 contains the statement of the main result and some discussion. Section 3 provides the details of our parameterization of Markovian distributions, the construction of the null set of imperfect distributions and a more technical version of our main result. Section 4 briefly discusses some applications of this construction. The proof of the main result appears in Section 5 with the proof of some of the technical lemmas deferred to Appendix A.

1.1 Related work

The notion of perfect graphical models has a long history, and we refer the reader to textbooks such as Pearl (1988) and Koller and Friedman (2009) for details. More recently, Sadeghi (2017) has characterized perfect distributions and a related line of work studies the problem of testing whether or not a given graph is perfect for a given distribution (Tatikonda et al., 2014). In this paper, we focus on a related but distinct question:

> *Given a graph G, how likely is it that a random Gaussian distribution that is Markov to G is also perfect with respect to G?*

Making this statement precise requires a bit of care; see Section 2. Similar results are already known for other classes of graphical models. For DAGs, almost-sure perfectness was shown in Spirtes and Schienes (1993); Meek (1995). Using the same techniques, the result was extended to Gaussian distributions that factor according to chain graphs in Levitz et al. (2001); Peña (2011). Chain graphs allow for both directed and undirected edges and the corresponding graphical models extend both the UG and DAG models. There are two equivalent formulations of the Markov property for chain graphs referred to as the Andersson–Madigan–Perlman (AMP) versus the Lauritzen–Wermuth–Frydenberg (LWF) interpretation (Lauritzen, 1996; Andersson et al., 2001; Studeny, 2006). In Levitz et al. (2001) (Section 6), perfectness of almost all Gaussian distributions that are Markov w.r.t. to a CG was shown using the AMP interpretation. A similar result was obtained in Peña (2011) using the LWF interpretation.

Remark 1. A different but equally interesting question arises in the study of Gaussoids, introduced also by Lněnička and Matúš (2007), which are combinatorial structures satisfying certain properties of the CI relations of Gaussian distributions. Here, an important problem is *realizability*: When is a Gaussoid realizable as the CI structure of a Gaussian distribution? Perhaps surprisingly, it turns out that the fraction of realizable Gaussoids vanishes as the number of nodes increases. This provides an interesting contrast to our result, which asks what the fraction of perfect Gaussians is relative to all Markovian Gaussians (i.e. wrt G). See Boege and Kahle (2020) (Remark 3.11) and Boege (2019) (Corollary 3.4) for details.

1.2 Gaussian graphical models

Consider an undirected graph $G = (V, E)$, where $V = [d] := \{1, \dots, d\}$. Two nodes i and j are *adjacent*, or *neighbors*, if $(i, j) \in E$, in which case we write $i \sim j$, otherwise $i \nsim j$. A *path* from i to j is a sequence $i = k_1, k_2, \dots, k_{n-1}, k_n = j \in [d]$ of distinct elements with $(k_\ell, k_{\ell+1}) \in E$ for each $\ell = 1, \dots, n-1$. Given two subsets $A, B \subset [d]$, a path connecting A to B is any path with $k_1 \in A$ and $k_n \in B$. A subset $C \subset [d]$ *separates* A from B, denoted by $A - C - B$, if all paths

connecting A to B intersect C (i.e. $k_\ell \in C$ for some $1 < \ell < n$), otherwise we write $\neg(A - C - B)$. Implicit in this definition is that A, B and C are disjoint.

To simplify the notation, we often identify G with its edge set E, i.e., $G \simeq E$. For example, we also write $|G| := |E|$ to denote the number of edges. We also adopt the following shorthands: $\{i\} = i$ and $\{i, j\} = ij$, $A \cup \{i\} = Ai$, $A \cup B = AB$ and so on, that is, the union of sets is denoted by juxtaposition. In addition, we let $[d]_S = [d] \setminus S = \{1, \dots, d\} \setminus S$. Common uses of these notational conventions are: $[d]_j = [d] \setminus \{j\}$ and $[d]_{ij} = [d] \setminus ij = [d] \setminus \{i, j\}$. For a matrix $\Sigma \in \mathbb{R}^{d \times d}$, and subsets $A, B \subset [d]$, we use $\Sigma_{A,B}$ for the submatrix on rows and columns indexed by A and B, respectively. Single index notation is used for principal submatrices, so that $\Sigma_A = \Sigma_{A,A}$. For example, $\Sigma_{i,j}$ is the (i, j)th element of Σ (using the singleton notation), whereas $\Sigma_{ij} = \Sigma_{ij,ij}$ is the 2×2 submatrix on $\{i, j\}$ and $\{i, j\}$. Similarly, $\Sigma_{Ai,Bj}$ is the submatrix indexed by rows $A \cup \{i\}$ and columns $B \cup \{j\}$.

Now, consider a random vector $X = (X_1, \dots, X_d) \in \mathbb{R}^d$ and a graph $G = (V, E)$ where node i represents the random variable X_i. A random vector X (or its distribution $\mathbb{P}$) is called *Markov* w.r.t. G (and G a CIG for X) if

$$A - C - B \text{ in } G \implies X_A \perp\!\!\!\perp X_B \mid X_C \text{ in } \mathbb{P}. \quad (1)$$

Here, $X_S = \{X_i : i \in S\}$ for any $S \subset [d]$. That is, the separation of the nodes in A and B by the nodes in C implies that X_A is independent of X_B given X_C. The special case where (1) is assumed to hold only for sets of the form $A = \{i\}$, $B = \{j\}$ and $C = [d] \setminus \{i, j\}$ is called the *pairwise Markov property*. This special case implies the full condition (1) if the distribution has a positive and continuous density w.r.t. a product measure on $\mathbb{R}^d$ (Lauritzen, 1996, p. 34).

Even if (1) holds, the converse need not necessarily hold. When the reverse implication of (1) is true, we say the distribution of X is *faithful* to G. When X is both Markov and faithful to G, we say that G is *perfect* for X:

Definition 1. A graph G is *perfect* for X if $A - C - B$ in $G \iff X_A \perp\!\!\!\perp X_B \mid X_C$ in $\mathbb{P}$.

In the Gaussian case, we have $X \sim N(0, \Sigma)$ where $\Sigma = (\Sigma_{i,j}) \in \mathbb{R}^{d \times d}$ is the covariance matrix of X, that is, $\Sigma_{i,j} = \mathbb{E}[X_i X_j]$. Using known results on Gaussian pairwise conditional independence (Lauritzen, 1996, Proposition 5.2), $X_i \perp\!\!\!\perp X_j \mid X_{[d]_{ij}}$ if and only if $[\Sigma^{-1}]_{i,j} = 0$. Thus, letting G be defined for $i \neq j$ by

$$i \nsim j \text{ in } G \iff [\Sigma^{-1}]_{i,j} = 0, \quad (2)$$

we have that X (or $N(0, \Sigma)$ or Σ) satisfies the pairwise Markov property w.r.t. G. Assuming that $\Sigma \succ 0$, it follows that Xsatisfies the (global) Markov property w.r.t. G, hence G is a CIG for X. Throughout, we will assume $\Sigma \succ 0$, i.e. the Gaussian distribution is *regular*.

From the above discussion, in the Gaussian case, Markov properties and CIGs can be equivalently characterized by the covariance matrix Σ. Thus, we can equivalently talk about *perfectness of a covariance matrix*. The corresponding graph is uniquely implied in this case, given by the support of Σ^{-1}, i.e., $\operatorname{supp}(\Sigma^{-1}) := \{(i, j) : (\Sigma^{-1})_{i,j} \neq 0,\ i \leq j\}$. We caution the reader that while the graph G has $|G|$ edges by definition, the support of Σ^{-1} has $|G| + d$ elements. We will write G° for the graph G with self-loops added, i.e., edges of the form (i, i) for all $i \in [d]$. Then we have $|G^\circ| = |\operatorname{supp}(\Sigma^{-1})| = |G| + d$. The above discussion is summarized in the following definition:

Definition 2. We say that a positive definite matrix Σ is G-Markov if $\operatorname{supp}(\Sigma^{-1}) = G^\circ$. We say that it is G-perfect if $N(0, \Sigma)$ is perfect w.r.t. G.

2 MAIN RESULT

In Lněnička and Matúš (2007), it was shown that for any graph G, there exists a regular Gaussian distribution which is perfect w.r.t. G. As discussed in Section 1.2, given any positive definite matrix Σ, we can ask whether it is perfect or not, with the graph of G being implicit from the support of Σ^{-1}. This is the language that we will use throughout. The result of Lněnička and Matúš (2007) can be restated as follows: for any potential CIG, there is at least one covariance matrix Σ which is perfect w.r.t. it. Here, we extend the argument in Lněnička and Matúš (2007) to show that *almost all* covariance matrices are perfect.

Theorem 1. *For any undirected graph G on $[d]$, the set of positive definite matrices $A \in \mathbb{R}^{d \times d}$ for which $\Sigma = A^{-1}$ is G-Markov but not G-perfect has Lebesgue measure zero.*

In Theorem 1 (and its corollary below), the Lebesgue measure is of dimension $|G^\circ| = |G| + d$.

According to the discussion in Section 1.2, Σ is G-Markov if Σ^{-1} is supported on G°. It follows that the set of matrices $A \in \mathbb{R}^{d \times d}$ for which $\Sigma = A^{-1}$ is G-Markov can be identified with a set in $\mathbb{R}^{|G^\circ|}$ of positive Lebesgue measure. Theorem 1 then states that those A in this set whose inverse is not perfect occupy a null subset. An equivalent restatement of this result in terms of probability distributions is the following:

Corollary 1. *Let G be an undirected graph on $[d]$, and let $A \in \mathbb{R}^{d \times d}$ be drawn from a continuous distribution (w.r.t. the Lebesgue measure) on positive definite*

matrices with support G°. Then with probability one, $\Sigma = A^{-1}$ is G-perfect.

Theorem 1 is a consequence of a more technical result, Theorem 2, which is discussed in Section 3.3 and could be of independent interest. One needs a fair amount of technical work to make the notion of "almost all" precise. This is done in Section 3.2 by constructing appropriate measures on a suitable parametrization of the set of covariance matrices that are G-Markov. Once done, the same techniques in Lněnička and Matúš (2007) can be extended to show the stronger result as illustrated in the proof of Theorem 2. In addition, Theorem 1 further strengthens this result by showing that the notion of "almost all" is independent of the particular parametrization of Theorem 2.

3 CONSTRUCTION OF NULL SETS

We begin by setting up notation to refer to paths in a graph. Next, we discuss how to parametrize the space of G-Markov covariance matrices. We then characterize the subset of perfect covariance matrices in Theorem 2, a result that is interesting in its own right.

3.1 Path notation

An *ij-path* on $[d]$, of length $t+1$, is an ordered sequence $i_0 \to i_1 \to i_2 \cdots \to i_t \to i_{t+1}$, where $i_j, j = 0, \ldots, t+1$ are distinct elements of $[d]$, $i_0 = i$ and $i_{t+1} = j$. We represent such a path as an ordered subset $\Pi = \{i_0, i_1 \ldots, i_{t+1}\}$ of $[d]$. An *i_0-cycle* on $[d]$, of length $t+1$, is an i_0i_0-path; that is, an ordered sequence of the form $i_0 \to i_1 \to i_2 \cdots \to i_t \to i_0$, where $i_j, j = 0, \ldots, t$ are distinct elements of $[d]$. We will represent such a cycle as an ordered subset $C = \{i_0, i_1 \ldots, i_t\}$ of $[d]$. Ultimately, the ij-paths and i_0-cycles will be used to represent non-intersecting paths and cycles on a graph G on nodes $[d]$.

From here on, we consider the edges of a graph G to be directed, i.e., ordered pairs of nodes. An undirected edge $ij \in G$ is interpreted as bidirected, i.e. $\{i, j\} \in G$ and $\{j, i\} \in G$. We say that an ij-path $\Pi = \{i = i_0, i_1 \ldots, i_{t+1} = j\}$ belongs to G, denoted as $\Pi \in G$, if all the edges in the path belong to G, that is, $i_j i_{j+1} \in G$ for all $j = 0, \ldots, t$. The set of ij-paths that belong to G is denoted as $\mathcal{P}^{ij}(G)$. With some abuse of notation, we let $\mathcal{P}^{ij} = \mathcal{P}^{ij}([d])$ denote the set of all ij-paths on $[d]$ in the complete graph. The set of all ij-paths of G of length $t+1$ is denoted as $\mathcal{P}_t^{ij}(G)$, that is

$$\mathcal{P}_t^{ij}(G) := \{\Pi \in \mathcal{P}^{ij} : \Pi \in G,\ |\Pi| = t+1\}.$$

We let $\mathcal{P}_t(G) := \bigcup_{i,j\in[d]} \mathcal{P}_t^{ij}(G)$, the set of all paths of length $t+1$ in G. The parallel notations for i-cycles, namely

$$\mathcal{C}^i(G),\ \ \mathcal{C}^i = \mathcal{C}^i([d]),\ \ \mathcal{C}_t^i(G),\ \text{ and }\ \mathcal{C}_t(G) \qquad (3)$$

are defined similarly (by setting $i = j$ in the corresponding definitions for paths). Note that in our notation, an undirected edge $ij \in G$, with $i \neq j$, is considered a valid cycle $\{i, j\}$ of length 2, since both $i \to j$ and $j \to i$ are in G.

For an ij-path $\Pi = \{i_0 = i, i_1, \ldots, i_t, j = i_{t+1}\} \in \mathcal{P}^{ij}$ and a matrix $B = (b_{i,j}) \in \mathbb{R}^{d\times d}$, let

$$b_\Pi = \prod_{j=0}^{t} b_{i_j, i_{j+1}}. \qquad (4)$$

A similar notation, namely b_C, is well-defined when C is an i_0-cycle. (In this case, $i_{t+1} = i_0$ in (4).)

3.2 A parametrization of G-Markov covariance matrices

We now give a parametrization of the G-Markov covariance matrices that provides a simple way of putting distributions on them. It also allows us to explicitly construct the set of imperfect covariance matrices from pieces that are all Lebesgue null sets.

Let $\mathbb{S}^d$ be the set of symmetric $d \times d$ matrices, $\mathbb{S}^d_{++}$ the set of $d \times d$ positive definite matrices, and define

$$\mathbb{S}^d_{++,1} = \{\Gamma \in \mathbb{S}^d_{++} : \Gamma_{i,i} = 1,\ i \in [d]\}.$$

Matrices in $\mathbb{S}^d_{++,1}$ are often called correlation matrices. Since we use this normalization mainly for precision matrices, to avoid confusion, we call elements of $\mathbb{S}^d_{++,1}$ normPrc matrices. It is not hard to see that for any diagonal matrix $D \in \mathbb{S}^d_{++}$, the two matrices $D\Sigma D$ and Σ have the same Markov properties. Thus, it is enough to focus on the case where $(\Sigma^{-1})_{i,i} = 1$ for all $i \in [d]$. We will make the following shorthand:

Definition 3. A matrix Σ is called a normCov matrix if $\Sigma^{-1} \in \mathbb{S}^d_{++,1}$, i.e., its inverse is a normalized precision matrix.

Given any graph G on nodes $[d]$, our first step is to construct a probability measure (mutually absolutely continuous w.r.t. the uniform probability measure), over all normCov matrices that are G-Markov. We then show that with probability one, such normCov matrices are perfect. Later, we will show how the result extends to all covariance matrices G-Markov (see Step 2 in the proof of Theorem 1). The class of normCov matrices that are G-Markov can be written as

$$\Psi_G^{-1} := \{\Gamma^{-1} : \Gamma \in \Psi_G\}, \quad \text{where}$$
$$\Psi_G := \{\Gamma \in \mathbb{S}^d_{++,1} : \Gamma_{i,j} \neq 0 \iff ij \in G\}.$$

Ψ_G is just the set of normPrc matrices with support G. The first step in our approach is to put a distribution on Ψ_G^{-1} as the push-forward of a distribution constructed on Ψ_G. Although our construction is not uniform w.r.t. the Lebesgue measure, in Corollary 1 we extend the result to any distribution on Ψ_G which is absolutely continuous w.r.t. the Lebesgue measure.

Before describing our construction of a random normCov matrix, let us set up some more notation. We let $\mathcal{L}^n$ and $\mathcal{H}^s$ ($s > 0$) denote, respectively, the Lebesgue measure and the s-dimensional Hausdorff measure on $\mathbb{R}^n$. The dimension of the ambient space of $\mathcal{H}^s$ will be clear from the context. For the graph G, let $g = |G|$ be the number of edges. We often identify Ψ_G with a subset of $\mathbb{R}^G$, and often identify $\mathbb{R}^G$ with $\mathbb{R}^g$, after ordering the edges, the particular order being unimportant. For example, if G is $1 - 2 - 3$, with $g = |G| = 2$, and $\Gamma \in \Psi_G$ is

$$\Gamma = \begin{pmatrix} 1 & \delta_{12} & 0 \\ \delta_{12} & 1 & \delta_{23} \\ 0 & \delta_{23} & 1 \end{pmatrix},$$

we either view Γ as $\{\delta_{12}, \delta_{23}\} = (\delta_{ij}, ij \in G)$, as an element of $\mathbb{R}^G$, or as the ordered pair $(\delta_{12}, \delta_{23})$ as an element of $\mathbb{R}^g = \mathbb{R}^2$.

For $\delta = (\delta_{ij},\ ij \in G) \in \mathbb{R}^G$ and $\varepsilon > 0$, define $A^{G,\delta,\varepsilon} = (a_{ij}^{G,\delta,\varepsilon}) \in \mathbb{R}^{d\times d}$ by setting

$$a_{ij}^{G,\delta,\varepsilon} = \begin{cases} \delta_{ij}\,\varepsilon & ij \in G \\ 1 & i = j \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

For a fixed $\delta \in \mathbb{R}^G$, let $\varepsilon_G(\delta)$ be the largest $\varepsilon > 0$ such that $A^{G,\delta,\varepsilon}$ is positive definite, that is,

$$\varepsilon_G(\delta) := \sup\{\varepsilon > 0 :\ A^{G,\delta,\varepsilon} \in \mathbb{S}^d_{++}\}. \tag{6}$$

Then $A^{G,\delta,\varepsilon}$ is positive definite for all $\varepsilon \in [0, \varepsilon_G(\delta))$, due to the convexity of $\mathbb{S}^d_{++}$. Let $[-1,1]_* := [-1,1] \setminus \{0\}$ and consider

$$\mathcal{M}^G := \big\{\, (\delta,\varepsilon) :\ \ \delta \in [-1,1]_*^G,\ \ \varepsilon \in (0, \varepsilon_G(\delta))\,\big\}.$$

The set $\{(A^{G,\delta,\varepsilon})^{-1} :\ (\delta,\varepsilon) \in \mathcal{M}^G\}$ is a parametrization of the set of all normCov matrices that are G-Markov. In other words, with the map $\zeta : \mathcal{M}^G \to \mathbb{R}^{d\times d}$

$$\zeta(\delta,\varepsilon) = (A^{G,\delta,\varepsilon})^{-1}, \tag{7}$$

we have $\zeta(\mathcal{M}^G) = \Psi_G^{-1}$. We note that $\mathcal{M}^G$ is a subset of $[-1,1]^G \times (0,\infty) \subset \mathbb{R}^G \times \mathbb{R} \simeq \mathbb{R}^{g+1}$. We will equip $\mathcal{M}^G$ with the Lebesgue measure (i.e., $\mathcal{L}^{g+1}$).

The map ζ overparametrizes the set Ψ_G^{-1} since $\zeta(c\delta, \varepsilon/c)$ is the same for all $c > 0$, i.e. it defines the same normCov matrix for all $c > 0$. To remove this ambiguity (and to avoid unnecessary complications in working with equivalence classes), without loss of generality, we focus on the subset of $\mathcal{M}^G$ for which δ has unit ℓ_∞ norm. Let $\mathbb{S}^G_\infty := \{\delta \in \mathbb{R}^G :\ \|\delta\|_\infty = 1\}$, $\mathbb{S}^G_{\infty,*} = [-1,1]^G_* \cap \mathbb{S}^G_\infty$, and

$$\begin{aligned}\mathcal{M}^G_\infty &:= \mathcal{M}^G \cap (\mathbb{S}^G_\infty \times \mathbb{R}) \\ &= \big\{\, (\delta,\varepsilon) :\ \delta \in \mathbb{S}^G_{\infty,*},\ \varepsilon \in (0, \varepsilon_G(\delta))\,\big\}.\end{aligned}$$

The function ε_G, restricted to $\mathbb{S}^G_{\infty,*}$, is continuous and bounded. In fact, $\sup \varepsilon_G(\mathbb{S}^G_{\infty,*}) = 1$ so that $\mathcal{M}^G_\infty \subset [-1,1]^G \times (0,1]$. Hence $\mathcal{M}^G_\infty$ has finite and positive $\mathcal{H}^g$-measure (on $\mathbb{R}^{g+1}$), where $g := |G|$.

Definition 4. We equip $\mathcal{M}^G_\infty$ with the measure μ_G defined as follows: Pick δ' by drawing each entry uniformly from $[-1,1]_*$, and given δ', set $\delta = \delta'/\|\delta'\|_\infty$ and draw ε uniformly from $[0, \varepsilon_G(\delta)]$; the vector (δ,ε) has the desired distribution μ_G. See Figure 1.

Measure μ_G defined above is equivalent to (i.e., mutually absolutely continuous w.r.t.) the normalized $\mathcal{H}^g$-measure on $\mathcal{M}^G_\infty$. The latter defines a uniform probability distribution on $\mathcal{M}^G_\infty$. We note that μ_G has density proportional to $(\varepsilon,\delta) \mapsto 1/\varepsilon_G(\delta)$ relative to this uniform distribution.

Restricted to $\mathcal{M}^G_\infty$, the map ζ defined earlier is well-behaved: It is one-to-one and onto Ψ_G^{-1}, that is, $\zeta : \mathcal{M}^G_\infty \to \Psi_G^{-1}$ is a bijection. We can now put a distribution on normCov matrices, Ψ_G^{-1}, as the push-forward of μ_G by ζ.

3.3 Characterizing imperfect covariance matrices

Let us now consider the subclass of Ψ_G^{-1} which is imperfect. It is enough to work with the corresponding subsets in $\mathcal{M}^G$ and $\mathcal{M}^G_\infty$:

$$\begin{aligned}\mathcal{N}^G &= \{(\delta,\varepsilon) \in \mathcal{M}^G :\ (A^{G,\delta,\varepsilon})^{-1} \text{ is not perfect}\}, \\ \mathcal{N}^G_\infty &= \{(\delta,\varepsilon) \in \mathcal{M}^G_\infty :\ (A^{G,\delta,\varepsilon})^{-1} \text{ is not perfect}\} \\ &= \mathcal{N}^G \cap (\mathbb{S}^G_\infty \times \mathbb{R}).\end{aligned}$$

For any $\delta \in \mathbb{R}^G$ and any path $\Pi \in \mathcal{P}_t(G)$ in G (of length $t+1$), the quantity δ_Π is well-defined using (4) even though δ is undefined outside G. We define:

$$\begin{aligned}\mathcal{D}_G := \Big\{\delta \in [-1,1]^G_* :\ &\sum_{\Pi \in \mathcal{P}_t^{ij}(G)} \delta_\Pi \neq 0, \\ &\text{for all nonempty } \mathcal{P}_t^{ij}(G),\ ij \in G,\ 0 \le t < d\Big\}.\end{aligned} \tag{8}$$

Note that by definition, $\delta_{i,j} = \delta_{j,i}$ for $\delta \in \mathbb{R}^G$.

The following result is the key component of Theorem 1:

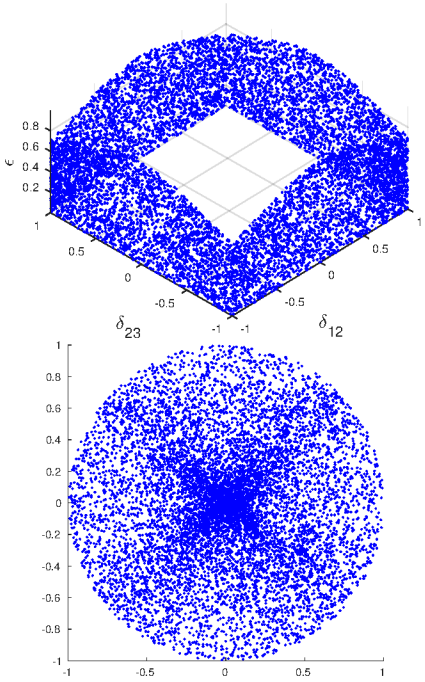

Figure 1: (Top) The plot of the samples from the uniform distribution μ_G on $\mathcal{M}_\infty^G$ for the graph $G = 1-2-3$. We have $\varepsilon_G((\delta_{12}, \pm 1)) = 1/\sqrt{1+\delta_{12}^2}$ and similarly for $\varepsilon_G((\pm 1, \delta_{23}))$. Note that the distribution is supported on $\mathbb{S}_\infty^2 \times [0,1]$. It is singular w.r.t. $\mathcal{L}^3$ but absolutely continuous w.r.t. $\mathcal{H}^2$. (Bottom) Pushforward of μ_G by the map F defined in Step 1 in the proof of Theorem 1 (cf. Section 5.1).

Theorem 2. *Let G be a graph with $g := |G| \geq 2$. Let*

$$B_\delta := \{\varepsilon : \ (\delta, \varepsilon) \in \mathcal{N}^G\},$$

for any $\delta \in [-1,1]_*^G$ and let $\mathcal{D}_G$ be as given in (8). Then, the following hold:

(a) $\mathcal{D}_G^c$ is an $\mathcal{L}^g$-null set,

(b) $\mathbb{S}_\infty^G \cap \mathcal{D}_G^c$ is an $\mathcal{H}^{g-1}$-null set, and

(c) for every $\delta \in \mathcal{D}_G$, B_δ is finite.

In particular, (d) the set $\mathcal{N}^G$ is a $\mathcal{L}^{g+1}$-null set, and $\mathcal{N}_\infty^G$ is a $\mathcal{H}^g$-null set, i.e. $\mu_G(\mathcal{N}_\infty^G) = 0$.

Theorem 2 is proved in Section 5.2. To gain some intuition for this result, consider the example $G = 1-2-3$ illustrated in Figure 1. Theorem 2 says that there is a "good" set $\mathcal{D}_G$ of $(\delta_{12}, \delta_{23}) \in [-1,1]_*^2$ which has full 2-dimensional measure; its boundary (i.e., the intersection with the perimeter of the square $[-1,1]_*^2$) has full 1-dimensional measure. Moreover, for any $(\delta_{12}, \delta_{23}) \in \mathcal{D}_G$, at most finitely many ε are problematic, that is, they lead to a imperfect covariance matrix via (7).

Theorem 2 also provides an explicit construction of the "bad set" $\mathcal{D}_G^c$ containing the bulk of non-perfect covariance matrices. This set is identified to be an algebraic variety in $\mathbb{R}^g$, the union of roots of polynomials of the form $\sum_{\Pi \in \mathcal{P}_t^{ij}(G)} \delta_\Pi$. In addition, the measure μ_G constructed in the course of the proof is itself interesting with many practical uses as discussed in Section 4.

Comparison with the literature The main advantage of our approach compared to the existing literature (Levitz et al., 2001; Peña, 2011) is the simpler and more direct parametrization of the class of G-Markov undirected models. Technically, our parametrization (δ, ε) decouples the positive semidefinite constraints (on ε) that need to be satisfied by any Gaussian distribution from the perfectness constraints (on δ); see Eq. (5)–(6). The main parameter δ is free to take any values in $[-1,1]^d$ and the only constraints are those imposed by perfectness or lack thereof. Moreover, our explicit identification of the bulk of non-perfect matrices via $\mathcal{D}_G^c$, provides insights about how perfectness can be violated in terms of the behavior of the coefficients along paths in G.

4 PRACTICAL APPLICATIONS

The construction of the measure μ_G (cf. Definition 4) has many possible uses in practice. Specifically, using μ_G as a trial or proposal distribution in rejection sampling or Markov chain Monte Carlo, we have a device to generate a random sample from any continuous distribution over (properly normalized) G-Markov covariance matrices, for any graph G. By the results of this paper, the resulting sample is also guaranteed to be G-perfect. Simulating from μ_G itself is quite straightforward. The only computational burden is that of computing $\varepsilon_G(\delta)$ which is a convex optimization problem that can be solved efficiently in practice. Here we briefly highlight two possible uses for such sampling schemes.

4.1 Uncertainty quantification

Consider the problem of uncertainty quantification when estimating the structure of a graphical model G. For simplicity, let us assume μ_G specifies the distribution of Σ^{-1} given a graph G. Then, by generating samples from μ_G, one can perform simulation-based uncertainty quantification given an estimated graph G. This is essentially a version of the *parametric bootstrap* for graphical models, where we view G as the discrete parameter. It allows practitioners to explicitly quantify the uncertainty of particular edges and even the entire network. We can also assess the variability of the regularization paths for methods that use a regularization

parameter. If one chooses to use distributions other than μ_G for Σ^{-1}, we can easily implement that by using importance weights with μ_G as the trial distribution.

These ideas are illustrated in Figure 2 where we have compared the variability of the estimates obtained by CLIME (Cai et al., 2011) and the graphical lasso (Friedman et al., 2008), both popular methods for estimating sparse undirected graphical models. For a random graph G on $d = 10$ nodes, with edge density ≈ 0.2, we have sampled 100 precision matrices Γ from μ_G and then sampled $n = 50$ data points from each $N(0, \Gamma^{-1})$, and ran the two methods on each dataset. Figure 2 (top) shows the variability of the regularization paths for the two methods, namely, the relative operator norm error between the true and estimated precision matrices versus the regularization parameter λ. The solid lines are the mean paths showing that $\lambda \approx 0.12$ is optimal for both methods. For this λ, and $n = 50, 100$ and 200, Figure 2 (bottom) shows the pairwise normalized Hamming distance between each pair of the 100 graphs estimated by each method, a measure of structure variability. The code is available at GitHub repository `aaamini/GMarkov-sampling` (Amini et al., 2022) and uses CVXR (Fu et al., 2020).

4.2 Bayesian inference

Consider the case where we are given a graph G and would like to estimate a Gaussian model that is G-Markov. In Bayesian analysis, one puts a prior on the precision matrix, which is often a Wishart prior. Restricting the domain of this prior to G, we obtain a support-restricted Wishart distribution. Sampling from the resulting posterior is difficult due to the support restriction, however, one can use μ_G as a trial distribution in the Metropolis-Hastings algorithm to efficiently sample from the posterior. We can also use μ_G as a trial to sample from other—not necessarily conjugate—priors (e.g., a uniform prior on the set of normalized G-Markov precision matrices).

5 PROOF OF MAIN RESULTS

In this section, we provide the proofs of the two main theorems. We first show how Theorem 2 implies Theorem 1. The rest of the section is then focused on proving Theorem 2. The main component is Lemma 1 whose proof is deferred to Section A.1.

5.1 Proof of Theorem 1

Recall the identification of Ψ_G with a subset of $\mathbb{R}^G \simeq \mathbb{R}^g$. The cases $g = 0$ and $g = 1$ are trivial, so we will assume $g \geq 2$. We proceed in two steps:

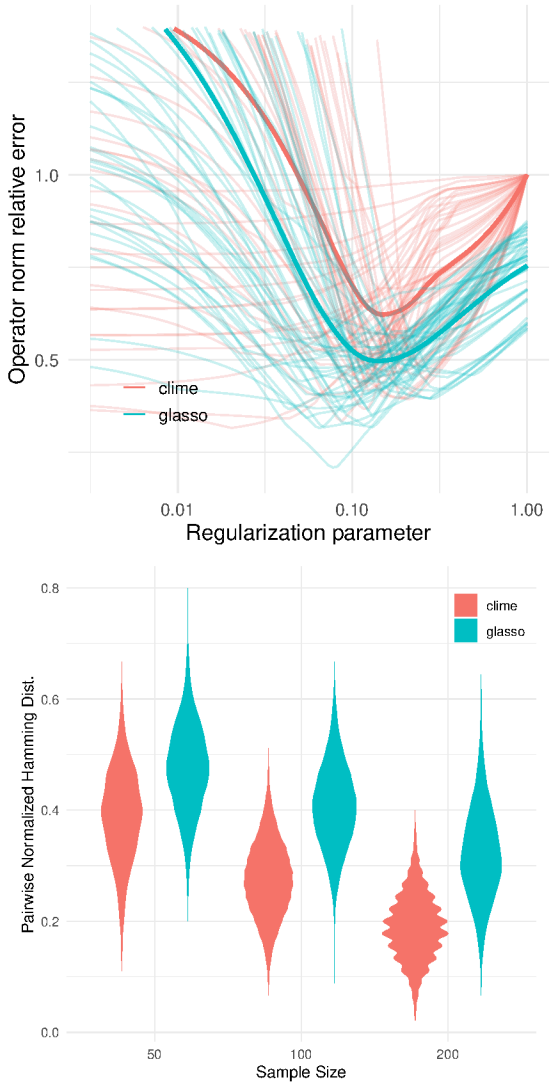

Figure 2: (Top) Operator norm relative error vs. λ. (Bottom) Hamming distance vs. sample size.

Step 1. The first step is to show that almost all normPrc matrices supported on G° lead to perfect covariance matrices after inversion. Consider the following subset of normPrc matrices,

$$N = \{\Gamma \in \Psi_G : \Gamma^{-1} \text{ is not perfect}\}. \tag{9}$$

We wish to show that $\mathcal{L}^g(N) = \mathcal{H}^g(N) = 0$. Let $F : \mathbb{R}^G \times \mathbb{R} \to \mathbb{R}^G$ be the map given by $F(\delta, \varepsilon) = (\delta_{ij}\varepsilon,\ ij \in G)$. Since F is a C^1 map, it is locally Lipschitz over $\mathbb{R}^{|G|+1}$, hence Lipschitz over $\mathcal{M}_\infty^G$. It is well-known that a Lipschitz map (between two metric spaces) maps $\mathcal{H}^s$-null sets to $\mathcal{H}^s$-null sets, for any $s > 0$; see for example (Krantz and Parks, 2008, Proposition 2.4.7). Since $\mathcal{H}^g(\mathcal{N}_\infty^G) = 0$ according to Theorem 2, it follows that $\mathcal{H}^g(F(\mathcal{N}_\infty^G)) = 0$. Recalling the identification of Ψ_G with a subset of $\mathbb{R}^G$, and that $F : \mathcal{M}_\infty^G \to \Psi_G$ is a bijection, we have $F(\mathcal{N}_\infty^G) \simeq N$, that is, $\mathcal{H}^g(N) = 0$.

Step 2. The second step is to extend the previous result for normPrc matrices to general positive definite matrices. Let $\mathbb{R}_{++}$ be the set of positive reals and let $\xi : \mathbb{R}^d_{++} \times \mathbb{R}^G \to \mathbb{R}^G$ be given by

$$\xi : \big(x_k, k \in [d];\, y_{ij},\ ij \in G\big) \mapsto \Big(\frac{y_{ij}}{\sqrt{x_i x_j}},\, ij \in G\Big). \tag{10}$$

This map should be thought of as mapping a general positive definite (PD) matrix, with support G°, to its corresponding normPrc matrix (ignoring the diagonal of all ones). We claim that the push-forward of $\mathcal{L}^{g+d}$ by ξ is absolutely continuous w.r.t. to $\mathcal{H}^g = \mathcal{L}^g$ on $\mathbb{R}^G \simeq \mathbb{R}^g$; see Lemma 4 in Section A.2 for a proof. Combined with the result that $\mathcal{H}^g(F(\mathcal{N}^G_\infty)) = 0$ of Step 1, this implies $\mathcal{L}^{g+d}(\xi^{-1}(N)) = 0$. But $\xi^{-1}(N)$ is the set of all PD matrices with support G° that are not perfect. The proof is complete.

5.2 Proof of Theorem 2

Let us introduce some notation, most of which is borrowed from Lněnička and Matúš (2007) with minor modifications. Recall also our subsetting and indexing notations from Section 1.2 and path notation from Section 3.1.

The proof of Theorem 2 relies on the following key technical lemma, which is as an extension of Lemma 4 in Lněnička and Matúš (2007) and is proven in Section A.1. Here, we treat node 1 specially, hence the emphasis on the collection of 1-cycles (cycles which begin and end on node 1) of a given length $t+1$, namely $\mathcal{C}^1_t(G)$. The special role given to 1 becomes clear in the proof of Theorem 3 below, where in dealing with an ij-path of G, we identify the endpoints with node 1 of a new graph H, hence obtaining a 1-cycle of H.

In the sequel, $\mathbb{C}[x]$ is the set of polynomials in the indeterminate variable x with complex coefficients. For $p \in \mathbb{C}[x]$, we say that $p = 0$ in $\mathbb{C}[x]$, or $p(x) = 0$ in $\mathbb{C}[x]$, if p is the zero polynomial (i.e., all its coefficients are zero). For a square matrix B, $|B|$ denotes its determinant.

Lemma 1. *Consider a (directed) graph H on $[r]$ with no self-loops on any node except possibly node 1. Let $\delta_{i,j}$ for $i, j \in [r]$ be a collection of nonzero real numbers. Define a matrix $B(x) = (b_{i,j}(x)) \in \mathbb{R}^{r\times r}$ by*

$$b_{i,i} = 1,\ \forall i > 1, \qquad \text{and}$$
$$b_{i,j}(x) = \delta_{i,j}\, x\, 1\big\{\{i,j\} \in H\big\},\ \text{ for } i \neq j \text{ and } i = j = 1,$$

treating $b_{i,j}(x)$ as a polynomial in $\mathbb{C}[x]$. The following two statements hold:

(a) $|B(x)| = 0$ in $\mathbb{C}[x]$ if $\mathcal{C}^1_t(H) = \emptyset$ for all $t \geq 0$.

(b) Assume further that for any $0 \leq t < r$,

$$\sum_{C\,\in\,\mathcal{C}^1_t(H)} \delta_C \neq 0 \quad \text{whenever } \mathcal{C}^1_t(H) \text{ is nonempty.} \tag{11}$$

Then, $|B(x)| = 0$ in $\mathbb{C}[x]$ implies $\mathcal{C}^1_t(H) = \emptyset$ for all $t \geq 0$.

Note that in the case $t = 0$, $\mathcal{C}^1_t(H)$ is nonempty only if H has a self-loop on node 1. Following Lněnička and Matúš (2007), let $\mathcal{N} = [d]$ and $\mathcal{R}(\mathcal{N})$ be the set of all couples $(ij|K)$ such that i and j are distinct singletons of $\mathcal{N}$ and $K \subset \mathcal{N} \setminus ij$. Subsets of $\mathcal{R}(\mathcal{N})$ are called relations. To simplify notation, unless otherwise stated, couples of the form $(ij|K)$ are always assumed to belong to $\mathcal{R}(\mathcal{N})$.

Definition 5. The dual couple of $(ij|K)$ is $(ij|\mathcal{N} \setminus ijK)$. For a relation $\mathcal{L} \subset \mathcal{R}(\mathcal{N})$, the *dual relation* $\mathcal{L}^\rceil$ is defined as the relation containing all the dual couples of the elements of $\mathcal{L}$.

For any matrix $A \in \mathbb{R}^{d\times d}$, let

$$\langle\!\langle A \rangle\!\rangle := \big\{(ij|K) :\ \ |A_{iK,jK}| = 0\big\}.$$

By Lemma 1 in Lněnička and Matúš (2007), for an invertible matrix A, we have $\langle\!\langle A \rangle\!\rangle^\rceil = \langle\!\langle A^{-1} \rangle\!\rangle$. For a simple undirected graph G with vertex set $\mathcal{N}$, let

$$\langle G \rangle := \big\{(ij|K) :\ \ K \text{ separates } i \text{ and } j \text{ in } G\big\}.$$

Recalling the notation $\mathcal{P}^{ij}_t(G)$ from Section 3.1, let

$$\mathcal{P}^{ij}_t(G; K) := \{\Pi \in \mathcal{P}^{ij}_t(G) :\ \Pi \subset ijK\}$$

denote the set of ij-paths in G of length $t+1$ that pass entirely through K. We also recall the definition of $\mathcal{D}_G$ from (8).

Theorem 3. *Let G be a simple undirected graph with vertex set $\mathcal{N}$. Then, for any $\delta \in \mathcal{D}_G$, there are finitely many $\varepsilon \in \mathbb{C}$ for which $\langle G \rangle^\rceil = \langle\!\langle A^{G,\delta,\varepsilon} \rangle\!\rangle$ fails, where $A^{G,\delta,\varepsilon}$ is defined in (5).*

Proof. Let $A^{G,\delta,x}$ be the matrix with elements in $\mathbb{C}[x]$ obtained by replacing ε in $A^{G,\delta,\varepsilon}$ by indeterminate x. Consider

$$\langle\!\langle A^{G,\delta,x} \rangle\!\rangle_{\mathbb{C}[x]} := \{(ij|K) :\ |A^{G,\delta,x}_{iK,jK}| = 0 \text{ in } \mathbb{C}[x]\}.$$

Step 1. Fix $\delta \in \mathcal{D}_G$. We show that $\langle G \rangle^\rceil = \langle\!\langle A^{G,\delta,x} \rangle\!\rangle_{\mathbb{C}[x]}$. Consider the matrix $A^{G,\delta,x}_{iK,jK}$, and assume that its first row and column correspond to the ith row and jth column of $A^{G,\delta,x}$, by swapping rows and columns if necessary, noting that such operations do not change

the determinant $|A^{G,\delta,x}_{iK,jK}|$. We identify (i,j) with element $(1,1)$ in Lemma 1. Let H be the subgraph of G induced on nodes ijK, with nodes i and j identified together and renamed node 1. The only edges in H that can be directed are those incident on node 1: $\{1,k\} \in H$ iff $\{i,k\} \in G$, and $\{k,1\} \in H$ iff $\{k,j\} \in G$. All the undirected edges in H_{KK} are considered bidirected. In other words, the support of $A_{iK,jK}$ is the adjacency matrix of H, which can be asymmetric and thus correspond to a directed graph. A path from i to j in G that lies entirely in ijK corresponds to a cycle in H starting at node 1, that is, we can identify $\mathcal{P}^{ij}_t(G;K)$ with $\mathcal{C}^1_t(H)$. A possible edge between i and j in G will be a self-loop on node 1 in H, i.e., $\delta_{ij}x$ plays the role of $\delta_{11}x$ in Lemma 1. Since $\delta \in \mathcal{D}_G$, it follows from (8) that assumption (11) holds for H whenever $(ij|K) \in \mathcal{R}(\mathcal{N})$. It follows from Lemma 1 that $|A^{G,\delta,x}_{iK,jK}| = 0$ if and only if $\mathcal{P}^{ij}_t(G;K)$ is empty for all $t \geq 0$. Hence, i and j are separated in G by $\mathcal{N} \setminus ijK$, or in symbols $(ij|K) \in \langle G \rangle^\urcorner$, if and only if $|A^{G,\delta,x}_{iK,jK}| = 0$.

Step 2. Fix $\delta \in \mathbb{R}^G$. Then $|A^{G,\delta,x}_{iK,jK}| = 0$ in $\mathbb{C}[x]$ implies $|A^{G,\delta,\varepsilon}_{iK,jK}| = 0$ for all $\varepsilon \in \mathbb{C}$. That is,

$$\langle\!\langle A^{G,\delta,x} \rangle\!\rangle_{\mathbb{C}[x]} \subset \langle\!\langle A^{G,\delta,\varepsilon} \rangle\!\rangle, \quad \forall \varepsilon \in \mathbb{C}. \tag{12}$$

The inclusion is strict if and only if there is $(ij|K)$ such that

$$p_{ijK}(x,\delta) := |A^{G,\delta,x}_{iK,jK}|$$

is a nonzero polynomial (in $\mathbb{C}[x]$) with root ε. Since any such polynomial has a finite number of roots, we have $\langle\!\langle A^{G,\delta,x} \rangle\!\rangle_{\mathbb{C}[x]} = \langle\!\langle A^{G,\delta,\varepsilon} \rangle\!\rangle$, for all but finitely many $\varepsilon \in \mathbb{C}$. Combined with Step 1, the assertion follows. □

Remark 2. The above proof contains the key intuition for defining $\mathcal{D}_G$ as in (8): For any $\delta \in \mathcal{D}_G$ and all but a finite number of ε, $\langle G \rangle^\urcorner = \langle\!\langle A^{G,\delta,\varepsilon} \rangle\!\rangle$ and thus, $\langle G \rangle = \langle\!\langle A^{G,\delta,\varepsilon} \rangle\!\rangle^\urcorner = \langle\!\langle \Sigma \rangle\!\rangle$, where $\Sigma = (A^{G,\delta,\varepsilon})^{-1}$ is the covariance matrix of X. This implies that $i - K - j$ in G if and only if $|\Sigma_{iK,jK}| = 0$, which is equivalent to $X_i \perp\!\!\!\perp X_j \mid X_K$ by Lemma 2 below. See the proof of Lemma 3 for a rigorous argument.

The following lemma is straightforward (see for example Amini et al. (2017)):

Lemma 2. *Suppose $X \sim N(0,\Sigma)$ and $\Sigma \succ 0$. Then, $|\Sigma_{Si,Sj}| = 0$ is equivalent to $X_i \perp\!\!\!\perp X_j \mid X_S$ for all i,j and $S \subset [d]_{ij}$.*

Lemma 3. *If $\langle G \rangle^\urcorner = \langle\!\langle \Sigma^{-1} \rangle\!\rangle$, then Σ is G-perfect.*

Proof. First we note that by Lemma 1 in Lněnička and Matúš (2007), $\langle G \rangle = \langle G \rangle^{\urcorner\urcorner} = \langle\!\langle \Sigma^{-1} \rangle\!\rangle^\urcorner = \langle\!\langle \Sigma \rangle\!\rangle$. Recall $\mathcal{N} := [d]$. Assume that Σ is not perfect. Then, there exist nonempty disjoint sets $A, B \subset \mathcal{N}$ and $K \subset \mathcal{N} \setminus AB$ such that $X_A \perp\!\!\!\perp X_B \mid X_K$, and K does not separate A and B. Then, $\exists i \in A, j \in B$ such that $\neg(i - K - j)$ and clearly $K \subset \mathcal{N} \setminus ij$ (i.e., $(ij|K) \in \mathcal{R}(\mathcal{N})$). We also have $X_i \perp\!\!\!\perp X_j \mid X_K$, hence $|\Sigma_{Ki,Kj}| = 0$ by Lemma 2. That is, $(ij|K) \in \langle\!\langle \Sigma \rangle\!\rangle$, hence we should have $(ij|K) \in \langle G \rangle$, contradicting $\neg(i - K - j)$. The proof is complete. □

Proof of Theorem 2. Part (c) of Theorem 2, with $\mathcal{D}_G$ defined by (8), follows from Theorem 3 and Lemma 3 and the relation $\Sigma^{-1} = A^{G,\delta,\varepsilon}$. For part (a), we note that $\mathcal{D}^c_G := \{\delta \in [-1,1]^G_* : \delta \notin \mathcal{D}_G\}$ is the finite union of the zero sets of nontrivial polynomials, hence of $\mathcal{L}^g$-measure zero in $[-1,1]^G_*$ (as a subset of $\mathbb{R}^G \simeq \mathbb{R}^g$). For part (b), let $\mathbb{S}^G_\infty = \bigcup_{ij}(F^+_{ij} \cup F^-_{ij})$ be the decomposition of $\mathbb{S}^G_\infty$ into its $(g-1)$-dimensional faces: $F^\pm_{ij} = \{\delta : \delta_{ij} = \pm 1\}$. It is enough to show, for example, that $F^+_{ij} \cap \mathcal{D}^c_G$ has $\mathcal{H}^{g-1}$-measure zero. Let G' be G with edge ij removed. By fixing $\delta_{ij} = 1$, we can view $F^+_{ij} \cap \mathcal{D}^c_G$ as a subset of $F^+_{ij} \subset \mathbb{R}^{G'} \simeq \mathbb{R}^{g-1}$. Recalling the definition of $\mathcal{D}_G$, (8), we observe, as before, that $F^+_{ij} \cap \mathcal{D}^c_G$ as a subset of $\mathbb{R}^{g-1}$ has $\mathcal{L}^{g-1}$-measure zero as a finite union of the zero sets of nontrivial polynomials in $g-1$ variables $\delta_{G'} = (\delta_{rs}, rs \in G')$. Since $\mathcal{L}^{g-1} = \mathcal{H}^{g-1}$ on $\mathbb{R}^{g-1}$, the assertion follows.

For part (d), both $\mathcal{L}^{g+1}(\mathcal{N}^G) = 0$ and $\mathcal{H}^g(\mathcal{N}^G_\infty) = 0$ follow from the Fubini theorem for the Lebesgue measure. For example, consider the latter assertion. It is enough to show $\mathcal{H}^g(\mathcal{N}^G \cap (F^+_{ij} \times \mathbb{R})) = 0$. Viewing $\mathcal{N}^G \cap (F^+_{ij} \times \mathbb{R})$ as a subset of $\mathbb{R}^{g-1} \times \mathbb{R}$, as above, and using the decomposition of the Lebesgue measure $\mathcal{L}^g = \mathcal{L}^{g-1} \times \mathcal{L}^1$, Fubini theorem gives

$$\begin{aligned}
\mathcal{H}^g\big(\mathcal{N}^G \cap (F^+_{ij} \times \mathbb{R})\big) &= \int_{F^+_{ij}} \mathcal{L}^1(B_\delta)\, d\mathcal{H}^{g-1}(\delta) \\
&= \int_{F^+_{ij} \cap \mathcal{D}^c_G} \mathcal{L}^1(B_\delta)\, d\mathcal{H}^{g-1}(\delta) \\
&\quad + \int_{F^+_{ij} \cap \mathcal{D}_G} \mathcal{L}^1(B_\delta)\, d\mathcal{H}^{g-1}(\delta).
\end{aligned}$$

Both integrals are zero, the first since $\mathcal{H}^{g-1}(F^+_{ij} \cap \mathcal{D}^c_G) = 0$ by part (b), and the second since B_δ has finitely many elements hence $\mathcal{L}^1(B_\delta) = 0$, by part (c). The proof is complete. □

Acknowledgements

A.A. acknowledges the support of NSF DMS-1945667. B.A. acknowledges the support of NSF IIS-1956330, NIH R01GM140467, and the Robert H. Topel Faculty Research Fund at the University of Chicago Booth School of Business. Q.Z. acknowledges the support of NSF DMS-1952929 and NSF IIS-1546098.

References

M. Al-Shedivat, A. Dubey, and E. P. Xing. Contextual explanation networks. *arXiv preprint arXiv:1705.10301*, 2017.

A. A. Amini, B. Aragam, and Q. Zhou. The neighborhood lattice for encoding partial correlations in a hilbert space. *arXiv preprint arXiv:1711.00991*, 2017.

A. A. Amini, B. Aragam, and Q. Zhou. G-Markov Sampling. 2022. doi: 10.5281/zenodo.6147844. URL `https://github.com/aaamini/GMarkov-sampling`.

S. A. Andersson, D. Madigan, and M. D. Perlman. Alternative markov properties for chain graphs. *Scandinavian journal of statistics*, 28(1):33–85, 2001.

T. Boege. Construction methods for gaussoids. Master's thesis, Otto-von-Guericke-Universität Magdeburg, 2019.

T. Boege and T. Kahle. Construction methods for gaussoids. *Kybernetika*, 56(6):1045–1062, 2020.

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 (494):594–607, 2011.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

A. Fu, B. Narasimhan, and S. Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020. doi: 10.18637/jss.v094.i14.

M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2946–2954. Curran Associates, Inc., 2016.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*, volume 2009. 2009. ISBN 0262013193. doi: 10.1016/j.ccl.2010.07.006.

S. G. Krantz and H. R. Parks. *Geometric integration theory*. Springer Science & Business Media, 2008.

S. L. Lauritzen. *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA, 1996. ISBN 0198522193.

M. Levitz, M. D. Perlman, and D. Madigan. Separation and completeness properties for amp chain graph markov models. *Ann. Stat.*, 29(6):1751–1784, Dec. 2001.

R. Lněnička and F. Matúš. On Gaussian condititional independence structures. *Kybernetika*, 43(3):327–342, 2007.

C. Meek. Strong completeness and faithfulness in bayesian networks. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligenc*, 1995.

A. T. Nguyen, A. Kharosekar, M. Lease, and B. Wallace. An interpretable joint graphical model for fact-checking from crowds. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

J. M. Peña. Faithfulness in chain graphs: the gaussian case. In *International Conference on Artificial Intelligence and Statistics*, pages 588–599, 2011.

K. Sadeghi. Faithfulness of probability distributions and graphs. *The Journal of Machine Learning Research*, 18(1):5429–5457, 2017.

G. P. Spirtes and R. Schienes. Causation, prediction, and search. 1993.

R. P. Stanley. Enumerative combinatorics (volume 1). *Cambridge studies in advanced mathematics*, 1997.

M. Studeny. *Probabilistic conditional independence structures*. Springer Science & Business Media, 2006.

S. C. Tatikonda et al. Testing unfaithful gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 2681–2689, 2014.

Supplementary Material: On perfectness in Gaussian graphical models

A PROOFS OF AUXILIARY RESULTS

We recall the following notational conventions: For a matrix $\Sigma \in \mathbb{R}^{d\times d}$, and subsets $A, B \subset [d]$, we use $\Sigma_{A,B}$ for the submatrix on rows and columns indexed by A and B, respectively. Single index notation is used for principal submatrices, so that $\Sigma_A = \Sigma_{A,A}$. For example, $\Sigma_{i,j}$ is the (i,j)th element of Σ (using the singleton notation), whereas $\Sigma_{ij} = \Sigma_{ij,ij}$ is the 2×2 submatrix on $\{i,j\}$ and $\{i,j\}$.

A.1 Proof of Lemma 1

Recall the definition of the i_0-cycle (of $[d]$ or a graph G) from Section 5.2. In proving Lemma 1, we will use the term *cycle* to also refer to cycles of a permutation. The necessary background on cycle decomposition is briefly reviewed below. The two notions of cycle (graph versus permutation) are related in our arguments, and the distinction in each occurrence should be clear from the context.

Recall that every permutation π on $[d]$, that is, a bijective map $\pi : [d] \to [d]$, has a unique *cycle decomposition*, once we agree on a particular order within cycles and among them (Stanley, 1997, Section 1.3). For example, representing $\pi = (142)(35)$ means that π has two cycles $C_1 = \{1,4,2\}$ and $C_2 = \{3,5\}$. C_1 being a cycle means that π maps 1 to 4, 4 to 2, and 2 back to 1, and similarly for C_2. We treat the cycles of π as ordered sets with the smallest element written first, and the rest of the order determined by the action of π. (That is, if $C = \{i_0, i_1, \dots, i_t\}$ is a cycle of π, we have $i_0 < i_j$ and $\pi(i_{j-1}) = i_j$ for $j = 1, \dots, t$.) Thus, permutation cycles are also graph cycles in the sense of Section 5.2. The (unordered) collection of cycles of π will be denoted as $\mathcal{S}_\pi$. In the example, $\mathcal{S}_\pi = \{C_1, C_2\}$. The ordering among the cycles is unimportant. In forming $\mathcal{S}_\pi$, we disregard trivial cycles, those containing a single element, except for the cycle containing 1. We often talk about "single cycle" permutations: for example, $\pi' = (142)(3)(5)$ has a single cycle $C_1 = \{1,4,2\}$ in our convention, while $\pi'' = (1)(42)(3)(5)$ has two cycles $C_1 = \{1\}$ and $C_2 = \{42\}$. Similarly, the identity permutation has a single cycle in our convention.

For matrix $B = (b_{i,j}) \in \mathbb{R}^{d\times d}$ and permutation π on $[d]$, we write

$$b_\pi := \prod_{i\in[d]} b_{i,\pi(i)} = \prod_{C\,\in\,\mathcal{S}_\pi} b_C, \tag{13}$$

where b_C is as defined[1] in (4). Since $b_{\{i\}} = b_{ii} = 1$ for $i \neq 1$, dropping single cycles $\{i\}$, for $i \neq 1$, from $\mathcal{S}_\pi$ does not affect (13). For the example above, the two expressions are

$$b_\pi = b_{1,4}b_{2,1}b_{3,5}b_{4,2}b_{5,3} = (b_{1,4}b_{4,2}b_{2,1})(b_{3,5}b_{5,3}).$$

For any permutation π, let C_π be its 1-cycle, i.e., its cycle that contains 1 and let $t_\pi = |C_\pi \setminus \{1\}| = |C_\pi| - 1$. Note that $b_{C_\pi} = \prod_{i\in C_\pi} b_{i,\pi(i)}$ is a factor of b_π.

Proof of Lemma 1. For simplicity, we will drop the explicit dependence on x and write $B = (b_{i,j})$. It is well-known that

$$|B| = \sum_\pi \operatorname{sign}(\pi) b_\pi.$$

[1]The notation b_π is also consistent with the definition of b_C in Section 5.2 due to the following connection: Every (graph) cycle C can be viewed as a permutation that leaves elements outside C intact.

First, consider the part (a). Assume $\mathcal{C}_t^1(H) = \emptyset$ for all $t \geq 0$. The case $t = 0$ gives $\{1,1\} \notin H$, hence $b_{i,\pi(i)} = b_{1,1} = 0$ whenever $C_\pi = \{1\}$. Similarly, for any C_π with $|C_\pi| > 1$, there are $i, j \in C_\pi$ with $i \neq j = \pi(i)$, such that $\{i,j\} \notin H$, hence $b_{i,\pi(i)} = 0$. Thus, $b_\pi = 0$ for all π, giving $|B| = 0$ and proving part (a).

Now assume $|B| = 0$. We start by showing that $b_{C_\pi} = 0$ for all π. We proceed by induction on $t_\pi = |C_\pi| - 1$. Fix $0 \leq t < r$. It suffices to show that if $b_{C_\pi} = 0$ for all π with $t_\pi < t$, then $b_{C_\pi} = 0$ for all π with $t_\pi = t$. The same argument below, with $t = 0$, establishes the initial step of the induction. For any cycle C,

$$b_C = \delta_C x^{|C|} 1\{C \in H\}, \tag{14}$$

that is, b_C is equal to 0 or $\delta_C x^{|C|}$, the latter if and only if $C \in H$. Here, δ_C is defined similar to b_C.

By the induction assumption, it follows that $b_\pi = 0$ for all π for which $t_\pi < t$ since b_{C_π} is a factor of b_π. It follows that $0 = |B| = \sum_{\pi:\, t_\pi \geq t} \mathrm{sign}(\pi) b_\pi$. There are three types of terms in this expansion: (Below, $\mathcal{S}_\pi$ is the cycle decomposition of π, using the convention discussed earlier.)

(a) $|\mathcal{S}_\pi| = 1, t_\pi = t$: These have a cycle C_π of length $t+1$ containing 1, and every other cycle is trivial. All of these permutations have the same sign, and we have

$$b_\pi = b_{C_\pi} = \delta_{C_\pi} x^{t+1} 1\{C_\pi \in H\}. \tag{15}$$

The first equality is since $b_{i,i} = 1$ for all $i \neq 1$. As π varies over the permutations in this category, C_π runs over all $\mathcal{C}_t^1$, i.e., cycles of length $t+1$ over $[r]$ containing 1. That is,

$$\{C_\pi : t_\pi = t\} = \mathcal{C}_t^1.$$

(Note that the correspondence also holds for $t = 1$ since the edges as considered directed. E.g., the permutation cycle $C_\pi = (12)$ corresponds to the graph cycle $1 \to 2$ and $2 \to 1$ in $\mathcal{C}_1^1$. In this case, we have $b_\pi = \delta_{12}\delta_{21} x^2 \big\{\{1,2\} \in H, \{2,1\} \in H\big\}$.)

However, only the subset $\mathcal{C}_t^1(H)$ of $\mathcal{C}_t^1$ contributes to $|B|$ due to the indicator $1\{C_\pi \in H\}$ in (15). There are two possible cases:

(i) $\mathcal{C}_t^1(H) = \emptyset$; then $b_{C_\pi} = 0$ for all π such that $|\mathcal{S}_\pi| = 1$ and $t_\pi = t$.
(ii) $\mathcal{C}_t^1(H) \neq \emptyset$; then, these permutations contribute to $|B|$, a term $\pm\big(\sum_{C \in \mathcal{C}_t^1(H)} \delta_C\big) x^{t+1}$.

(b) $|\mathcal{S}_\pi| \geq 2, t_\pi = t$: Any such permutation has at least a cycle C of size $\nu \geq 2$ in $[r] \setminus C_\pi$. Hence, b_π has a factor of the form

$$b_{C_\pi} b_C = \delta_{C_\pi} \delta_C \, x^{t+\nu+1} 1\{C_\pi, C \in H\}$$

Thus, any such b_π, if nonzero, contributes a polynomial of degree at least $t+3$.

(c) $t_\pi \geq t+1$: In this case, b_π has a factor of

$$b_{C_\pi} = \delta_{C_\pi} x^{t_\pi + 1} 1\{C_\pi \in H\}$$

and as the previous case contributes a polynomial of degree at least $t+2$, if nonzero.

Thus, the coefficient of x^{t+1} in $|B|$ is determined only by permutations of type (a). But, this coefficient is zero by the assumption that $|B| = 0$. We conclude that case (ii) above cannot occur, since then $\sum_{C \in \mathcal{C}_t^1(H)} \delta_C = 0$ for some nonempty $\mathcal{C}_t^1(H)$ with $t \in \{0, \ldots, r-1\}$, contradicting assumption (11).

This in turn implies that for any permutation π of type (a), we have $b_{C_\pi} = 0$, by (15) and that H cannot contain any cycle of size $t+1$. But this proves the induction claim: For any permutation π' with $t_{\pi'} = t$, there is permutation π of type (a) such that $C_\pi = C_{\pi'}$ (i.e. break all the cycles of π', other than $C_{\pi'}$, into trivial ones).

As a byproduct of establishing the induction claim, we also obtain $\mathcal{C}_t^1(H) = \emptyset$ for all $t \geq 0$ which is the desired result. (In particular, with $t = 0$, it means that H cannot have a self-loop on node 1 if $|B| = 0$.) The proof is complete. □

A.2 Auxiliary lemmas

The following lemma is used in the proof of Theorem 1. The notation $\xi_* \mu$ denotes the push-forward of measure μ by map ξ.

Lemma 4. *With $\xi : \mathbb{R}^d_{++} \times \mathbb{R}^g \to \mathbb{R}^g$ defined as in (10), we have $\xi_* \mathcal{L}^{d+g} \ll \mathcal{L}^g$, that is, $\mathcal{L}^g(A) = 0$ implies $\mathcal{L}^{g+d}(\xi^{-1}(A)) = 0$.*

Proof. Let $\Omega := \mathbb{R}^d_{++} \times \mathbb{R}^g$ be a subset of $\mathbb{R}^{d+g}$. Let $x = (x_k,\ k \in [d])$ and $y = (y_{ij},\ ij \in G)$. Consider the function $F_1 : \Omega \to \Omega$ defined by

$$F_1(x, y) = \Big(x, \frac{y_{ij}}{\sqrt{x_i x_j}},\ ij \in G\Big).$$

F_1 is a C^∞ diffiomorphism of Ω onto itself, that is, $F_1 : \Omega \to \Omega$ is a bijection and both F_1 and its inverse $F_2 := F_1^{-1}$ belong to class C^∞. This implies that F_1 and F_2 are locally Lipschitz (i.e., Lipschitz when restricted to any compact subset of Ω), hence they both preserve $\mathcal{L}^{g+d}$-null sets (i.e., map null sets to null sets).

Let $\pi : \mathbb{R}^{d+g} \to \mathbb{R}^g$ be the projection $\pi(x, y) = y$. We can write $\xi = \pi \circ F_1$. We first show that $\pi_* \mathcal{L}^{d+g} \ll \mathcal{L}^g$. This follows from Fubini theorem: Let $A \subset \mathbb{R}^g$ be such that $\mathcal{L}^g(A) = 0$. We have $\pi^{-1}(A) = \mathbb{R}^d \times A$. Hence, $\mathcal{L}^{d+g}(\pi^{-1}(A)) = \mathcal{L}^d(\mathbb{R}^d) \cdot \mathcal{L}^g(A) = 0$ since the Lebesgue measure is σ-finite.

Now assuming that $\mathcal{L}^g(A) = 0$, we thus have $\mathcal{L}^{g+d}(\pi^{-1}(A)) = 0$. But then $\mathcal{L}^{g+d}(F_2 \circ \pi^{-1}(A)) = 0$, due to the diffiomorphic nature of F_2. Noting that $\xi^{-1} = (\pi \circ F_1)^{-1} = F_1^{-1} \circ \pi^{-1} = F_2 \circ \pi^{-1}$, we have the desired result. The proof is complete. □