

CENTRE DE RECERCA MATEMÀTICA
BIMR 2026



Maximum likelihood thresholds for colored Gaussian graphical models: why uniform sampling struggles near the boundary

Project: *Can algebra tell us whether we have enough data?*

Olaya González Campos

Supervised by: Roser Homs Pons

Barcelona Introduction to Mathematical Research 2026
Barcelona, 2026

Abstract

Gaussian graphical models are statistical models in which a graph encodes conditional independence relations among the random variables represented by the nodes of the graph. They are used in a variety of applications, especially in biological studies, where data tends to be scarce because of the high cost of obtaining it or data protection issues. The minimal number of observations that we need to ensure the existence of the maximum likelihood estimator with probability one is called the maximum likelihood threshold, and it can be studied geometrically through the cone of sufficient statistics $\mathcal{C}_G$.

In this project we first introduce colored Gaussian graphical models and the geometry of the existence of the maximum likelihood estimator. Then, we study the approach based on topological data analysis (persistent homology) proposed in [HKS26], and, in particular, the sampling challenge pointed out there: the reference point cloud, obtained by projecting a uniform sample of positive definite matrices, is sparse near the boundary of $\mathcal{C}_G$, which is exactly where the sample covariance matrices that we want to classify lie. We compute this phenomenon explicitly for the colored 2-cycle, we state the general results that this example suggests, whose proofs are left for future work, and we propose some directions for future research.

Contents

Abstract	i
1 Graphical models	1
1.1 Colored graphs and conditional independence	1
1.2 The cone of concentration matrices	5
2 Existence of the maximum likelihood estimator	7
2.1 The MLE and the sufficient statistics	7
2.2 The cone of sufficient statistics and the fiber	8
2.3 Maximum likelihood thresholds	9
3 The TDA approach and its sampling issue	10
3.1 The TDA approach	10
3.2 Trace normalization	11
3.3 The problem we are focusing on	12
3.4 Why $\pi_G(M)$ is not uniform: the fiber-volume density	12
4 Future research	17
Bibliography	18

Chapter 1

Graphical models

In this chapter we introduce colored graphs and the Gaussian graphical models associated to them, following [HKS26] and [Sul18].

1.1 Colored graphs and conditional independence

We write $\text{Sym}(m)$ for the vector space of $m \times m$ real symmetric matrices, and $\text{PD}(m) \subseteq \text{PSD}(m) \subseteq \text{Sym}(m)$ for the subsets of positive definite and positive semidefinite matrices. For $n < m$, we also consider the matrices of rank at most n : $\text{Sym}(m, n)$ among the symmetric ones and $\text{PSD}(m, n)$ among the positive semidefinite ones. The latter will be the natural place for sample covariance matrices of $n < m$ observations.

Graphical interpretation: $\text{PD}(m)$ and $\text{PSD}(m)$ are convex cones. In particular, $\text{PD}(m)$ is an open convex cone and $\text{PSD}(m)$ is its closure, i.e., $\text{cl}(\text{PD}(m)) = \text{PSD}(m)$ or, equivalently, $\text{PD}(m) = \text{Int}(\text{PSD}(m))$. The boundary of the cone, $\text{PSD}(m) \setminus \text{PD}(m)$, consists of the positive semidefinite matrices M with $\det M = 0$.

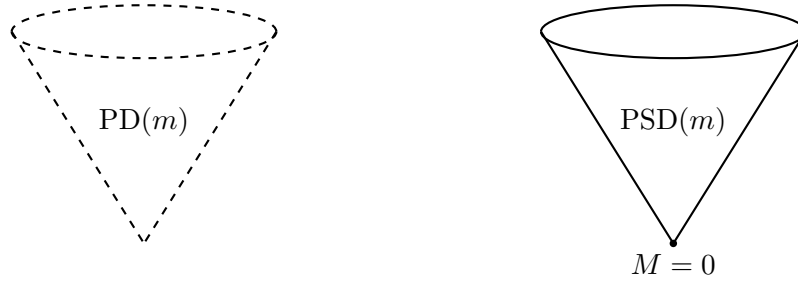


Figure 1.1: Left: convex cone of positive definite matrices (open cone). Right: convex cone of positive semidefinite matrices (closed cone).

More specifically, given a real vector space V , a subset $C \subseteq V$ is a convex cone if $\lambda x + \mu y \in C$ for every scalars $\lambda, \mu \in \mathbb{R}^+$ s.t. at least one of them is strictly positive, and for every elements $x, y \in C$. Note that, taking $\lambda, \mu \in \mathbb{R}^+$ such that $\lambda + \mu = 1$, we have that C verifies the definition of convex set. With this in mind, let us consider $A, B \in \text{PSD}(m)$ and $\lambda, \mu \in \mathbb{R}^+$. Then, for $x \in \mathbb{R}^m$:

$$x^\top (\lambda A + \mu B) x = \lambda (x^\top A x) + \mu (x^\top B x).$$

It is clear that, since both λ and μ are non-negative constants, $\lambda (x^\top A x) + \mu (x^\top B x) \geq 0$, and, consequently, $\lambda A + \mu B \in \text{PSD}(m)$. Furthermore, if $A, B \in \text{PD}(m)$ and $x \neq 0$, then $x^\top A x > 0$ and $x^\top B x > 0$ and, because λ and μ cannot be both zero, it follows that $\lambda (x^\top A x) + \mu (x^\top B x) > 0$, which implies that $\lambda A + \mu B \in \text{PD}(m)$.

Definition 1.1 ([HKS26, Definition 2.1]). The tuple $G = (V, E, \lambda)$ is a *colored graph* if it is an undirected graph on the vertex set $V = [m] = \{1, \dots, m\}$ and the edge set E together with a coloring function $\lambda : V \sqcup E \rightarrow [d]$ such that $\lambda(V) \cap \lambda(E) = \emptyset$.

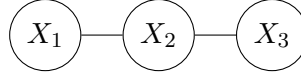
The d colors of a colored graph will be interpreted as symmetries in the concentration matrix $K = \Sigma^{-1}$, and the condition $\lambda(V) \cap \lambda(E) = \emptyset$ ensures that the colors of the vertices and the edges do not coincide.

Where does this covariance matrix Σ come from? To each vertex of the graph, we assign a random variable. These m random variables (there are m vertices) constitute an m -dimensional random vector:

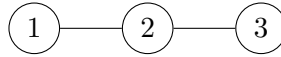
$$X = (X_1, \dots, X_m) \sim \mathcal{N}(0, \Sigma), \quad X_i \in \mathbb{R},$$

where $\mathcal{N}(0, \Sigma)$ denotes the multivariate Gaussian distribution with mean vector 0 and covariance matrix Σ . The next natural question is: ‘What is the intuition behind this correspondence between vertices and random variables?’ The best way to understand it is with examples.

Example 1.2. The graph



which will be denoted by



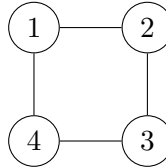
for simplicity, can be interpreted as

$$X_1 \perp\!\!\!\perp X_3 \mid X_2 \quad (\text{Markov property}),$$

which means that X_1 and X_3 are conditionally independent given X_2 . This is called a *conditional independence statement*. $\blacklozenge$

From now on, for simplicity, the random variables will not be explicitly displayed in the vertices of all considered graphs, and the vertices will always be identified with random variables throughout this work.

Example 1.3. The graph



can be interpreted as

$$X_1 \perp\!\!\!\perp X_3 \mid \{X_2, X_4\},$$

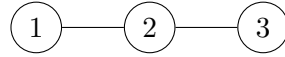
which means that X_1 and X_3 are conditionally independent given both X_2 and X_4 . Symmetrically,

$$X_2 \perp\!\!\!\perp X_4 \mid \{X_1, X_3\}.$$

$\blacklozenge$

Now, it will be seen how the two previous examples can be expressed in terms of the concentration matrix (*concentration models*).

Example 1.4. The graph



can be expressed in terms of the concentration matrix as

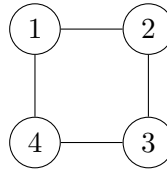
$$K = \begin{pmatrix} \lambda_1 & \lambda_4 & 0 \\ \lambda_4 & \lambda_2 & \lambda_5 \\ 0 & \lambda_5 & \lambda_3 \end{pmatrix},$$

where the zeros represent the conditional independence statement

$$X_1 \perp\!\!\!\perp X_3 \mid X_2.$$

◆

Example 1.5. The graph



can be expressed in terms of the concentration matrix as

$$K = \begin{pmatrix} \lambda_1 & \lambda_5 & 0 & \lambda_8 \\ \lambda_5 & \lambda_2 & \lambda_6 & 0 \\ 0 & \lambda_6 & \lambda_3 & \lambda_7 \\ \lambda_8 & 0 & \lambda_7 & \lambda_4 \end{pmatrix},$$

where the zeros represent the conditional independence statements

$$X_1 \perp\!\!\!\perp X_3 \mid \{X_2, X_4\}, \quad X_2 \perp\!\!\!\perp X_4 \mid \{X_1, X_3\}.$$

◆

In these examples, the zeros of the concentration matrix correspond to conditional independence statements; this is true in general. Let us state this result formally.

Proposition 1.6 ([Sul18, Propositions 6.3.2 and 13.1.5]). *Let $X \sim \mathcal{N}(0, \Sigma)$ belong to the Gaussian graphical model of G , that is, its concentration matrix $K = \Sigma^{-1}$ satisfies $k_{ij} = 0$ whenever $(i, j) \notin E$. Given $i, j \in V$ with $i \neq j$, the following hold:*

$$(i, j) \notin E \implies k_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j \mid X_{[m] \setminus \{i, j\}}.$$

Analogously, the zeros of the covariance matrix correspond to *independence statements*. Let us see it with an example.

Example 1.7. The graph



can be interpreted as

$$X_1 \perp\!\!\!\perp X_2, \quad X_1 \perp\!\!\!\perp X_3,$$

which are *independence statements* that, in particular, mean that X_1 is independent from both X_2 and X_3 . Furthermore, in terms of the covariance matrix, this is translated to

$$\Sigma = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_{33} \end{pmatrix}, \quad \sigma_{ij} = \text{Cov}(X_i, X_j), \quad \sigma_{ii} = \text{Cov}(X_i, X_i) = \text{Var}(X_i).$$

◆

Remark 1.8. It is interesting to note that, since we are working with a multivariate Gaussian distribution, independence is equivalent to zero covariance (which, in general, is a weaker condition). $\diamond$

The following proposition makes the intuition around independence statements and their relation with the covariance matrix explicit.

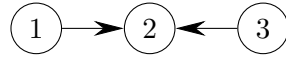
Proposition 1.9 ([Sul18, Proposition 2.4.4 and Section 13.1]). *Given $i, j \in V$ with $i \neq j$, the following equivalence holds:*

$$\sigma_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j.$$

In particular, if X belongs to the Gaussian graphical model of G and there is no path between i and j in G , then $\sigma_{ij} = 0$.

Next, we are going to see how the independence and conditional independence statements behave when considering directed graphs.

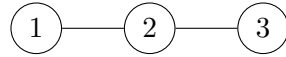
Example 1.10. The graph



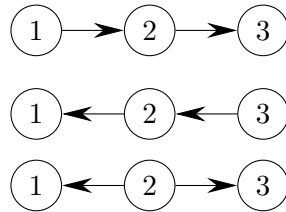
implies that X_1 is independent from X_3 , that is, $X_1 \perp\!\!\!\perp X_3$, or equivalently, $\sigma_{13} = 0$. $\blacklozenge$

As the next example shows, with directed graphs several configurations can be equivalent regarding conditional independence statements.

Example 1.11. The undirected graph



and each of the 3 directed graphs



imply $X_1 \perp\!\!\!\perp X_3 \mid X_2$, or equivalently, $k_{13} = 0$. $\blacklozenge$

How can we obtain equations from the conditional independence statements?

Consider a model with $m = 3$ vertices,

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}, \quad K = \Sigma^{-1}.$$

In this setting, the conditional independence statement $X_1 \perp\!\!\!\perp X_3 \mid X_2$ is equivalent to $k_{13} = 0$, which, by using the adjoint matrix to compute that entry of the inverse, is equivalent to the equation

$$\sigma_{12}\sigma_{23} - \sigma_{13}\sigma_{22} = 0.$$

1.2 The cone of concentration matrices

The model can be defined by describing the type of concentration matrices that we are going to ‘accept’. Let us consider the linear vector subspace

$$\begin{aligned}\mathcal{L}_G &= \{K = (k_{ij}) \in \text{Sym}(m) \mid k_{ij} = 0 \text{ if } (i, j) \notin E, \ i \neq j, \\ &\quad k_{ii} = k_{jj} \text{ if } \lambda(i) = \lambda(j), \\ &\quad k_{ij} = k_{lw} \text{ if } \lambda(i, j) = \lambda(l, w) \text{ for } (i, j), (l, w) \in E\},\end{aligned}$$

which has dimension $\dim(\mathcal{L}_G) = d$ (see [HKS26, Definition 2.2]). A statistical interpretation of this set can be found in [HL08, Section 3].

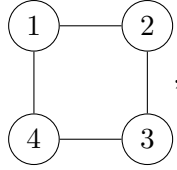
To be a covariance matrix, Σ must be symmetric and positive definite, which necessarily implies that $K = \Sigma^{-1} \in \text{PD}(m)$. Therefore, we are going to work with the set

$$\mathcal{K}_G = \mathcal{L}_G \cap \text{PD}(m),$$

which will be called the *cone of concentration matrices*.

Let $K_1, \dots, K_d$ be a basis of $\mathcal{L}_G$, that is, $\mathcal{L}_G = \langle K_1, \dots, K_d \rangle$. Thus, for each $K \in \mathcal{L}_G$, there exists a unique vector $(\lambda_1, \dots, \lambda_d) \in \mathbb{R}^d$ such that $K = \sum_{i=1}^d \lambda_i K_i$. Consequently, the set $\mathcal{K}_G$ can be identified with the set $\{(\lambda_1, \dots, \lambda_d) \in \mathbb{R}^d \mid \sum_{i=1}^d \lambda_i K_i \in \text{PD}(m)\}$ (see [HKS26, Remark 2.3]).

Example 1.12. Let G be the uncolored 4-cycle



then,

$$\mathcal{L}_G = \left\{ \begin{pmatrix} \lambda_1 & \lambda_5 & 0 & \lambda_8 \\ \lambda_5 & \lambda_2 & \lambda_6 & 0 \\ 0 & \lambda_6 & \lambda_3 & \lambda_7 \\ \lambda_8 & 0 & \lambda_7 & \lambda_4 \end{pmatrix} : \lambda_i \in \mathbb{R}, \ i = 1, \dots, 8 \right\}, \quad d = \dim(\mathcal{L}_G) = 8,$$

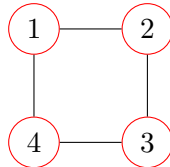
and the basis elements are

$$\begin{aligned}K_1 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & K_2 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & K_3 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & K_4 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \\ K_5 &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & K_6 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & K_7 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, & K_8 &= \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.\end{aligned}$$

◆

In the next example it can be seen that adding colors reduces the dimension of $\mathcal{L}_G$.

Example 1.13. Let G be the colored 4-cycle with all vertices of the same color and all edges of different colors (see also [HKS26, Example 2.6])



then

$$\mathcal{L}_G = \left\{ \begin{pmatrix} \lambda_1 & \lambda_2 & 0 & \lambda_5 \\ \lambda_2 & \lambda_1 & \lambda_3 & 0 \\ 0 & \lambda_3 & \lambda_1 & \lambda_4 \\ \lambda_5 & 0 & \lambda_4 & \lambda_1 \end{pmatrix} : \lambda_i \in \mathbb{R}, i = 1, \dots, 5 \right\}, \quad d = \dim(\mathcal{L}_G) = 5,$$

and the basis elements are

$$\begin{aligned} K_1 &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, & K_2 &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, & K_3 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \\ K_4 &= \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, & K_5 &= \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \end{aligned}$$

◆

Finally, the *colored Gaussian graphical model* associated to G is the set of covariance matrices

$$\mathcal{M}_G = \mathcal{K}_G^{-1} = \{\Sigma \in \text{PD}(m) : \Sigma^{-1} \in \mathcal{L}_G\}$$

(see [HKS26, Definition 2.4]). Note that, since $\mathcal{L}_G$ is a linear subspace of $\text{Sym}(m)$ and $\text{PD}(m)$ is a convex cone, $\mathcal{K}_G = \mathcal{L}_G \cap \text{PD}(m)$ is a convex cone. Also, since $\text{PD}(m)$ is open, $\mathcal{K}_G$ is open inside $\mathcal{L}_G$ (with the Euclidean topology).

Chapter 2

Existence of the maximum likelihood estimator

2.1 The MLE and the sufficient statistics

Let $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^m$ be an i.i.d. sample of n observations drawn from $\mathcal{N}(0, \Sigma)$, with $\Sigma \in \text{PD}(m)$. The *sample covariance matrix* is

$$S = \frac{1}{n} \sum_{i=1}^n X^{(i)} X^{(i)\top} = \frac{1}{n} X X^\top \in \text{PSD}(m), \quad X = (X^{(1)} | \dots | X^{(n)}) \in \mathbb{R}^{m \times n}.$$

Since $1/n$ is a positive constant, we will sometimes omit it and write $S = X X^\top$. Note that X is in general not a square matrix, because we are interested in the case $n < m$.

Where does the log-likelihood come from? The density of $\mathcal{N}(0, \Sigma)$ at $x \in \mathbb{R}^m$ is $(2\pi)^{-m/2} (\det \Sigma)^{-1/2} \exp(-\frac{1}{2} x^\top \Sigma^{-1} x)$. Hence, writing $K = \Sigma^{-1}$ and using that $x^\top K x = \text{tr}(K x x^\top)$, the log-likelihood of the sample is

$$\sum_{i=1}^n \left(-\frac{m}{2} \log(2\pi) + \frac{1}{2} \log \det K - \frac{1}{2} X^{(i)\top} K X^{(i)} \right) = -\frac{nm}{2} \log(2\pi) + \frac{n}{2} (\log \det K - \text{tr}(SK)).$$

Dropping the additive constant and the positive factor $n/2$, which do not change where the maximum is attained, maximizing the likelihood is equivalent to maximizing

$$\ell_S(K) = \log \det K - \langle S, K \rangle,$$

where $\langle S, K \rangle = \text{tr}(SK) = \sum_{1 \leq i, j \leq m} s_{ij} k_{ij}$ is the trace (or Frobenius) inner product, which is the standard inner product on $\text{Sym}(m)$.

Why do we work with K instead of Σ ? In terms of Σ , the MLE is the solution of

$$\max_{\Sigma \in \mathcal{M}_G} -\log \det \Sigma - \langle S, \Sigma^{-1} \rangle,$$

whose objective function is not concave in general. Since $\Sigma \mapsto \Sigma^{-1}$ is a bijection between $\mathcal{M}_G$ and $\mathcal{K}_G$, this problem is equivalent to

$$\max_{K \in \mathcal{K}_G} \log \det K - \langle S, K \rangle,$$

whose objective function ℓ_S is concave on $\text{PD}(m)$ [Uhl17, Proposition 3.1] and whose feasible set $\mathcal{K}_G$ is a convex cone. Therefore, the second one is a convex optimization problem: if it has a maximum $\hat{K}$, it is unique, and it is the MLE of the concentration matrix. The MLE of the covariance matrix is then $\hat{\Sigma} = \hat{K}^{-1}$ (see [Uhl17, Section 3]).

It is important to note that, in $\ell_S(K) = \log \det K - \langle S, K \rangle$, the only part that depends on the data (that is, on the sample) is $\langle S, K \rangle$. Given the basis $K_1, \dots, K_d$ of $\mathcal{L}_G$, if $K = \sum_{i=1}^d \lambda_i K_i$, then

$$\langle S, K \rangle = \sum_{i=1}^d \lambda_i \langle S, K_i \rangle.$$

Therefore, the optimization problem does not see the whole matrix S , but only the d numbers $\langle S, K_1 \rangle, \dots, \langle S, K_d \rangle$: two sample covariance matrices with the same values of these numbers define the same log-likelihood on $\mathcal{L}_G$, and hence the same MLE. This motivates the *projection map*

$$\pi_G : \text{Sym}(m) \longrightarrow \mathbb{R}^d, \quad \pi_G(S) = (\langle S, K_1 \rangle, \dots, \langle S, K_d \rangle).$$

The components $\langle S, K_i \rangle$ are called the *sufficient statistics*, because they are the only thing we need to compute $\langle S, K \rangle$.

Example 2.1. For the uncolored 4-cycle G of Example 1.12, with $m = 4$ and $d = 8$, the basis $K_1, \dots, K_8$ given there yields

$$\pi_G : \text{Sym}(4) \longrightarrow \mathbb{R}^8, \quad S \longmapsto (s_{11}, s_{22}, s_{33}, s_{44}, 2s_{12}, 2s_{23}, 2s_{34}, 2s_{14}).$$

Thus, the sufficient statistics are the diagonal entries of S and (twice) the entries corresponding to the edges, while the entries s_{13} and s_{24} , which correspond to the missing edges, are not needed to compute the log-likelihood. If, instead, all vertices have the same color (Example 1.13), the four diagonal statistics merge into $s_{11} + s_{22} + s_{33} + s_{44} = \text{tr } S$, and π_G takes values in $\mathbb{R}^5$. $\blacklozenge$

2.2 The cone of sufficient statistics and the fiber

Write $D = \binom{m+1}{2} = \dim \text{Sym}(m)$. Since $K_1, \dots, K_d$ are linearly independent, π_G is surjective and $\ker \pi_G = \mathcal{L}_G^\perp$, the orthogonal complement of $\mathcal{L}_G$ in $\text{Sym}(m)$ with respect to the trace inner product.

Definition 2.2 ([HKS26, Definition 2.8]). The *fiber* of $S \in \text{Sym}(m)$ with respect to G is

$$\text{fiber}_G(S) = \{S' \in \text{Sym}(m) : \pi_G(S') = \pi_G(S)\}.$$

Note that $\text{fiber}_G(S) = S + \ker \pi_G = S + \mathcal{L}_G^\perp$ is an affine translate of $\mathcal{L}_G^\perp$ of dimension

$$k = \dim \ker \pi_G = D - d.$$

For a point $t \in \mathbb{R}^d$ we write $\pi_G^{-1}(t)$ for the corresponding affine subspace of $\text{Sym}(m)$; if $t = \pi_G(S)$ then $\pi_G^{-1}(t) = \text{fiber}_G(S)$. The fiber collects all the matrices that are indistinguishable from S for the MLE: if $S' \in \text{fiber}_G(S)$, then $S' - S \in \mathcal{L}_G^\perp$, so $\langle S', K \rangle - \langle S, K \rangle = \langle S' - S, K \rangle = 0$ for every $K \in \mathcal{L}_G$, that is, $\ell_{S'} = \ell_S$ on $\mathcal{L}_G$.

We are going to focus on the map π_G restricted to those matrices that could be sample covariance matrices, that is, $S \in \text{PSD}(m, n)$. The set that we are going to study is the following.

Definition 2.3 ([HKS26, Definition 2.9]). The *cone of sufficient statistics* is the open convex cone $\mathcal{C}_G = \pi_G(\text{PD}(m))$ in $\mathbb{R}^d$.

Since $\text{PSD}(m) = \text{cl}(\text{PD}(m))$ and π_G is continuous, $\pi_G(\text{PSD}(m)) \subseteq \text{cl}(\mathcal{C}_G)$; the reverse inclusion also holds [SU10], so $\text{cl}(\mathcal{C}_G) = \pi_G(\text{PSD}(m))$. The key result that connects the existence of the MLE with this geometry is the following.

Theorem 2.4 (Characterization of the MLE, [Uhl17, Theorem 5.2]). *The MLE $\hat{\Sigma}$ exists if and only if $\text{fiber}_G(S) \cap \text{PD}(m) \neq \emptyset$. If it exists, $\hat{\Sigma}$ is the unique matrix in $\text{fiber}_G(S)$ satisfying $\hat{\Sigma}^{-1} \in \mathcal{K}_G$, that is, $\{\hat{\Sigma}\} = \text{fiber}_G(S) \cap \mathcal{M}_G$.*

Equivalently, the MLE exists for S if and only if $\pi_G(S) \in \mathcal{C}_G$. Indeed,

$$\begin{aligned} \exists \text{ MLE} &\iff \exists \tilde{S} \in \text{fiber}_G(S) \cap \text{PD}(m) \iff \exists \tilde{S} \in \text{PD}(m) \text{ s.t. } \pi_G(\tilde{S}) = \pi_G(S) \\ &\iff \pi_G(S) \in \pi_G(\text{PD}(m)) = \mathcal{C}_G. \end{aligned}$$

Thus, the whole question of MLE existence is the membership question

$$\text{the MLE exists for } S \iff \pi_G(S) \in \mathcal{C}_G \iff \text{fiber}_G(S) \cap \text{PD}(m) \neq \emptyset.$$

- When $S \in \text{PD}(m)$, $\pi_G(S) \in \pi_G(\text{PD}(m)) = \mathcal{C}_G$, so the MLE always exists.
- When $S \in \text{PSD}(m)$, $\pi_G(S) \in \pi_G(\text{PSD}(m)) = \text{cl}(\mathcal{C}_G)$, so $\pi_G(S)$ could be in $\mathcal{C}_G$ (the interior) or in the boundary $\partial\mathcal{C}_G$ of $\mathcal{C}_G$. There are no more options.

When is S positive definite? For every $x \in \mathbb{R}^m$,

$$x^\top S x = \frac{1}{n} x^\top X X^\top x = \frac{1}{n} \|X^\top x\|^2 \geq 0,$$

so S is always positive semidefinite, and $S \in \text{PD}(m)$ if and only if $X^\top x \neq 0$ for every $x \neq 0$, that is, if and only if $\text{rank}(X) = m$. Since $X \in \mathbb{R}^{m \times n}$, we have $\text{rank}(X) \leq \min\{n, m\}$, and the sample covariance matrix of n observations has rank $\min\{n, m\}$ with probability one [HKS26, Remark 2.14]. This gives two cases:

- If $n \geq m$ (there are at least as many observations as components), then $\text{rank}(X) = m$ with probability one, so $S \in \text{PD}(m)$ and $\pi_G(S) \in \pi_G(\text{PD}(m)) = \mathcal{C}_G$ with probability one. Consequently, the MLE exists with probability one.
- If $n < m$ (there are fewer observations than components), then $\text{rank}(X) \leq n < m$, so $S \in \text{PSD}(m, n) \setminus \text{PD}(m)$ always. Hence $\pi_G(S) \in \pi_G(\text{PSD}(m)) = \text{cl}(\mathcal{C}_G)$, but we do not know whether $\pi_G(S)$ lies in $\mathcal{C}_G$ or in $\partial\mathcal{C}_G$.

Therefore, the case that we have to study is $n < m$.

2.3 Maximum likelihood thresholds

We are not interested in the behavior of an isolated S , but in the general behavior of $S \in \text{PSD}(m, n)$.

Definition 2.5 ([HKS26, Definition 2.13]). Let S be the sample covariance matrix of n generic observations. The *maximum likelihood threshold* $\text{mlt}(G)$ is the smallest n such that $P(\text{the MLE exists for } S) = 1$, and the *weak maximum likelihood threshold* $\text{wmlt}(G)$ is the smallest n such that $P(\text{the MLE exists for } S) > 0$.

Geometrically [HKS26, Section 2.3],

$$\begin{aligned} \text{mlt}(G) &= \min\{n : \pi_G(S) \in \mathcal{C}_G \text{ for almost every } S \in \text{PSD}(m, n)\}, \\ \text{wmlt}(G) &= \min\{n : \pi_G(S) \in \mathcal{C}_G \text{ for some } S \in \text{PSD}(m, n)\}. \end{aligned}$$

The first is a statement of existence with probability $p = 1$, and the second one with probability $p > 0$. It may seem surprising that a single S is enough for the second one, but it follows from $\mathcal{C}_G$ being open: if $\pi_G(X_0 X_0^\top) \in \mathcal{C}_G$ for some $X_0 \in \mathbb{R}^{m \times n}$, then, by continuity, $\pi_G(X X^\top) \in \mathcal{C}_G$ for every X in an open neighborhood of X_0 , which has positive probability (see also [HKS26, Remark 2.15]).

Chapter 3

The TDA approach and its sampling issue

This chapter studies the sampling difficulty that limits the Topological Data Analysis (TDA) approach to estimating $\text{mlt}(G)$ proposed in [HKS26, Section 4]. The first two sections recall the approach and fix the setting. The last two contain our contribution: we state the difficulty precisely as two problems; for the colored 2-cycle, we compute explicitly the distribution of the projected reference cloud; and we state general results, together with the conditions on G under which they hold, that describe the behavior of the reference cloud near the boundary.

3.1 The TDA approach

Given $X^{(1)}, \dots, X^{(n)} \in \mathbb{R}^m$, we want to know whether the MLE exists for this n . For this, it suffices to check whether $\pi_G(S) \in \mathcal{C}_G$, where S is the sample covariance matrix. We know the explicit formula of the linear map π_G , so where is the difficulty? We do not have an explicit formula of $\mathcal{C}_G = \pi_G(\text{PD}(m))$ in general, so it is not trivial to know whether $t = \pi_G(S) \in \mathcal{C}_G$ or $t \in \partial\mathcal{C}_G$.

The idea of [HKS26, Section 4] is to sample a set $M \subseteq \text{PD}(m)$ and consider that the point cloud $\pi_G(M)$ describes $\mathcal{C}_G$ well enough, and then to use persistent homology of annular neighborhoods [Sto+20] to see whether $\pi_G(S)$ is an interior point of $\mathcal{C}_G$ or a boundary point. Intuitively, if $\pi_G(S)$ is an interior point, the points of $\pi_G(M)$ at distance between R_1 and R_2 from $\pi_G(S)$ surround it in every direction, and form (approximately) a sphere; if $\pi_G(S)$ is a boundary point, there are regions of the annulus with no points of $\pi_G(M)$. More precisely:

- *Algorithm 2* of [HKS26] receives the graph G , a candidate n for $\text{mlt}(G)$ and the number MC of Monte Carlo replicates. It computes M , an i.i.d. set of positive definite matrices with trace 1 sampled uniformly [MMJ12]; it generates MC samples $X^{(i)} = (X^{(i,1)} | \dots | X^{(i,n)})$, whose columns are i.i.d. uniform on $[-1, 1]^m$; it computes $S^{(i)} = X^{(i)} X^{(i)\top}$, rescales its trace, and projects it to obtain the query points $\pi_G(S^{(i)})$; it classifies them with Algorithm 3, and returns the number of interior points.
- *Algorithm 3* of [HKS26] receives a finite point set and radii $0 < R_1 < R_2$. For each point y it finds the annulus A_y of points at distance between R_1 and R_2 from y and, if A_y contains at least 100 points, computes the 1- and 2-dimensional persistence barcodes of 100 of them, and classifies y as an interior point if some bar is longer than $R_2 - R_1$, and as a boundary point otherwise.

The approach is tested in [HKS26, Figure 4] on the graphs G_3 , G_6 and G_{18} of [Uhl12, Table 2] (see also [HKS26, Table 1]), which are 4-cycles ($m = 4$), G_{18} being the uncolored one, for $n = 1, 2, 3$ observations. Theoretically, $\text{wmlt}(G_3) = \text{mlt}(G_3) = 1$ and $\text{wmlt}(G_6) = \text{mlt}(G_6) = 1$, so for these graphs the projection of the sample covariance matrix lies in the interior of $\mathcal{C}_G$ with probability one for $n = 1, 2, 3$; and $\text{wmlt}(G_{18}) = 2$, $\text{mlt}(G_{18}) = 3$. As pointed out in [HKS26, Section 4.1.2], the approach captures the relative trends but does not provide sharp decision boundaries: for $n = 1$, it classifies around 20% of the query points of G_3 and 40% of those of G_6 as boundary points [HKS26, Figure 4], although, theoretically, all of them are interior points with probability one. The rest of this chapter aims to explain why.

3.2 Trace normalization

Recall that $\mathcal{L}_G \subseteq \text{Sym}(m)$ has dimension d and basis $K_1, \dots, K_d$, that $D = \binom{m+1}{2} = \dim \text{Sym}(m)$, and that $k = D - d$ is the dimension of the fibers.

Remark 3.1 (Normalizing the trace). The vertex colors of G partition $V = [m]$ into classes $V_1, \dots, V_s$, and the corresponding basis elements of $\mathcal{L}_G$ are the diagonal matrices $K_i = \sum_{j \in V_i} E_{jj}$ for $i = 1, \dots, s$ (without loss of generality, we order the d basis elements of $\mathcal{L}_G$ such that the first s correspond to the vertex color classes and the remaining $d - s$ to the edge color classes). Note that, for an uncolored graph, each vertex is its own class and $s = m$. Since $\langle S, K_i \rangle = \sum_{j \in V_i} s_{jj}$ and the classes partition $[m]$,

$$\sum_{i=1}^s \langle S, K_i \rangle = \sum_{i=1}^s \sum_{j \in V_i} s_{jj} = \text{tr } S,$$

so the trace of S is determined by the first s components of $\pi_G(S)$, namely $\text{tr } S = t_1 + \dots + t_s$, where $t = (t_1, \dots, t_d) = \pi_G(S)$. Hence the normalization $\text{tr } S = 1$ used when sampling corresponds, in the target $\mathbb{R}^d$, to the affine hyperplane

$$H = \{t \in \mathbb{R}^d : t_1 + \dots + t_s = 1\}.$$

Two consequences are used repeatedly below.

1. *We sample in a bounded slice.* The set $\{S \in \text{PSD}(m) : \text{tr } S = 1\}$ is compact and π_G is linear, so $\pi_G(\text{PSD}(m) \cap \{\text{tr} = 1\}) = \pi_G(\text{PSD}(m)) \cap H = \text{cl}(\mathcal{C}_G) \cap H$ is compact. Since $\mathcal{C}_G$ is a convex cone, it is unbounded and we cannot sample directly from it; instead, we sample the slice $\mathcal{C}_G \cap H$, which is convex (intersection of the convex $\mathcal{C}_G$ with the affine H) and bounded (by the compactness above). Moreover, it has dimension $d - 1$, so we are working in a lower-dimensional space.
2. *Nothing is lost.* Let $S \in \text{PSD}(m)$ with $S \neq 0$, so $\text{tr } S > 0$. Since π_G is linear,

$$\pi_G(S) \in \mathcal{C}_G \iff \pi_G(S / \text{tr } S) \in \mathcal{C}_G \cap H.$$

Thus, deciding membership for a trace-normalized matrix is the same problem as in the non-normalized setting.

◇

3.3 The problem we are focusing on

The TDA approach (Algorithms 2 and 3 of [HKS26]) decides whether a point $\pi_G(S)$ belongs to $\mathcal{C}_G$ or to $\partial\mathcal{C}_G$ by computing the persistent homology of an annular neighborhood of $\pi_G(S)$ inside a finite point cloud that is meant to be a discrete proxy of $\mathcal{C}_G$. Its accuracy depends on how well that point cloud represents $\mathcal{C}_G$ near the *query point* $t = \pi_G(S)$. Two difficulties arise.

Problem 3.2 (Scarcity of cloud points near $\partial\mathcal{C}_G$). Let M be a uniform sample of $\text{PD}(m) \cap \{\text{tr} = 1\}$. Then, for the graphs that we consider (see Corollary 3.7), $\pi_G(M)$ is not a uniform sample of $\mathcal{C}_G \cap H$, and its density decays towards the boundary $\partial\mathcal{C}_G$. This is also observed empirically in [HKS26, Section 4.2].

Problem 3.3 (Concentration of the query points near $\partial\mathcal{C}_G$). Consider the matrix $X = (X^{(1)} | \dots | X^{(n)}) \in \mathbb{R}^{m \times n}$, where $X^{(1)}, \dots, X^{(n)}$ are i.i.d. realizations uniform on $[-1, 1]^m$ with $n < m$. Consider the empirical covariance matrix $S = XX^\top \in \text{PSD}(m, n)$ and rescale so that $\text{tr} S = 1$. If we replicate this MC times, obtaining MC covariance matrices, their projections $\pi_G(S)$'s, which are the points we want to classify as interior or boundary points (the query points), concentrate near $\partial\mathcal{C}_G$ (see [HKS26, Section 4.2]).

Putting together both Problem 3.3 and Problem 3.2, we see that the reference cloud $\pi_G(M)$ is dense where there are few query points and sparse exactly where most of them are. Therefore, when using persistent homology to determine whether the points $\pi_G(S)$ are interior against the cloud $\pi_G(M)$, some interior points are classified as boundary points, as observed in [HKS26, Figure 4]. On the one hand, Problem 3.3 needs to be explicitly stated to understand the difficulty we face, but the generation of the covariance matrices S must not be changed; they are precisely the sample covariance matrices whose MLE existence we want to decide, so their concentration near $\partial\mathcal{C}_G$ is part of the problem, not an artifact of the method. On the other hand, regarding Problem 3.2, whether the non-uniformity of $\pi_G(M)$ is structural or an artifact that can be overcome is addressed in the next section.

3.4 Why $\pi_G(M)$ is not uniform: the fiber-volume density

Volumes. We regard $\text{Sym}(m)$ as a D -dimensional Euclidean space under the Frobenius inner product $\langle A, B \rangle = \text{tr}(AB)$. Any affine subspace $W \subseteq \text{Sym}(m)$ of dimension j inherits this inner product, and we write vol_j for the j -dimensional Lebesgue measure on W (for $j = 1$ it is the usual length, for $j = 2$ the usual area, and for $j \geq 3$ the usual j -dimensional volume). We use the same symbol vol_{d-1} for the Lebesgue measure on the affine hyperplane $H \subseteq \mathbb{R}^d$ of dimension $d - 1$.

Consider $\Omega := \text{PD}(m) \cap \{\text{tr} = 1\}$, a bounded, open, nonempty convex subset of the hyperplane $\{\text{tr} = 1\} = \{S \in \text{Sym}(m) : \text{tr} S = 1\} \subseteq \text{Sym}(m)$. Its closure is the compact set $\text{PSD}(m) \cap \{\text{tr} = 1\}$, from which it differs by a vol_{D-1} -null set. It holds that $0 < \text{vol}_{D-1}(\Omega) < \infty$, and consequently, the *uniform* probability measure μ on Ω , which is given by

$$\mu(A) = \frac{\text{vol}_{D-1}(A)}{\text{vol}_{D-1}(\Omega)} \quad \text{for measurable } A \subseteq \Omega,$$

is well defined. For $t \in H$, since π_G is surjective there exists $\tilde{S} \in \text{Sym}(m)$ with $\pi_G(\tilde{S}) = t$, so

$$\pi_G^{-1}(t) = \text{fiber}_G(\tilde{S}) = \tilde{S} + \mathcal{L}_G^\perp$$

is an affine translate of $\mathcal{L}_G^\perp$ of dimension $k = \dim \mathcal{L}_G^\perp = D - d$. We write

$$\Phi(t) := \pi_G^{-1}(t) \cap \text{PSD}(m)$$

for the positive semidefinite part of this fiber.

Densities. Given two σ -finite measures ν and ρ on the same measurable space, ν is *absolutely continuous* with respect to ρ , written $\nu \ll \rho$, if every ρ -null set is also ν -null. If $\nu \ll \rho$, by the Radon–Nikodym theorem there exists a nonnegative function f , unique up to a ρ -null set, with

$$\nu(B) = \int_B f d\rho \quad \text{for every measurable } B,$$

called the *density* (or Radon–Nikodym derivative) of ν with respect to ρ , and written $f = d\nu/d\rho$. It measures how concentrated ν is relative to the reference ρ .

The following proposition, which we state without proof (see Chapter 4), describes the probability measure that π_G induces on H when considering the uniform measure μ on the source Ω .

Proposition 3.4 (Push-forward density). *The push-forward $\nu := (\pi_G)_*\mu$ is a probability measure on H , supported on $\mathcal{C}_G \cap H$, absolutely continuous with respect to vol_{d-1} , and its density $f = d\nu/d\text{vol}_{d-1}$ satisfies*

$$f(t) \propto \text{vol}_k(\Phi(t)), \quad \text{for every } t \in H.$$

Example 3.5 (The colored 2-cycle: a semicircular density). Take $G = C_2$ (both vertices one color, the edge its own), so there are $m = 2$ vertices, and $D = \dim \text{Sym}(2) = 3$. With the basis $K_1 = I$, $K_2 = E_{12} + E_{21}$ of $\mathcal{L}_G$ we have $d = 2$, $k = D - d = 1$, and

$$t := \pi_G(S) = (\langle S, K_1 \rangle, \langle S, K_2 \rangle) = (\text{tr } S, 2s_{12}).$$

The cone $\mathcal{C}_G$ and its slice. Let us determine $\mathcal{C}_G = \pi_G(\text{PD}(2))$. Consider $S \in \text{PD}(2)$, that is $S \in \text{Sym}(2)$ such that $s_{11} > 0$ and $\det S = s_{11}s_{22} - s_{12}^2 > 0$. In terms of its image $\pi_G(S) = t = (t_1, t_2) = (s_{11} + s_{22}, 2s_{12})$ this forces $t_1 > 0$ and, by the arithmetic–geometric mean inequality,

$$s_{12}^2 < s_{11}s_{22} \leq \left(\frac{s_{11}+s_{22}}{2}\right)^2 = \frac{t_1^2}{4}, \quad \text{i.e. } |t_2| < t_1.$$

Conversely, for any $t = (t_1, t_2)$ with $|t_2| < t_1$ there exists a positive definite matrix S whose image by π_G yields t ; for instance $S = \frac{1}{2} \begin{pmatrix} t_1 & t_2 \\ t_2 & t_1 \end{pmatrix}$, which has $s_{11} = t_1/2 > 0$ and $\det S = \frac{1}{4}(t_1^2 - t_2^2) > 0$, satisfies $\pi_G(S) = t$. Hence,

$$\mathcal{C}_G = \{(t_1, t_2) \in \mathbb{R}^2 : |t_2| < t_1\},$$

the open sector between the rays $t_2 = t_1$ and $t_2 = -t_1$.

On the slice $H = \{t_1 = 1\}$ this gives

$$\mathcal{C}_G \cap H = \{(1, t_2) \in \mathbb{R}^2 : -1 < t_2 < 1\},$$

a segment whose two endpoints constitute $\partial\mathcal{C}_G \cap H = \{(1, -1), (1, 1)\}$.

The positive semidefinite part of the fibers and their volume. Fix $t = (1, t_2) \in \mathcal{C}_G \cap H$. Then, its preimage consists of the symmetric matrices with $\text{tr } S = 1$ and $s_{12} = t_2/2$, that is

$$\pi_G^{-1}(t) = \left\{ S_c = \begin{pmatrix} \frac{1}{2} + c & t_2/2 \\ t_2/2 & \frac{1}{2} - c \end{pmatrix}, \quad c \in \mathbb{R} \right\}.$$

Note that $S_c \succeq 0$ iff $\frac{1}{2} \pm c \geq 0$ and $\det S_c = \frac{1}{4}(1 - t_2^2) - c^2 \geq 0$ which, since $|t_2| < 1$, is equivalent to $|c| \leq \frac{1}{2}\sqrt{1 - t_2^2}$. Therefore,

$$\Phi(t) = \pi_G^{-1}(t) \cap \text{PSD}(2) = \{S_c : S_c \succeq 0\} = \left\{ S_c : |c| \leq \frac{1}{2}\sqrt{1 - t_2^2} \right\}.$$

Taking $c_+ := \frac{1}{2}\sqrt{1 - t_2^2}$ and $c_- := -c_+$,

$$\begin{aligned} \text{vol}_1(\Phi(t)) &= \text{dist}(S_{c_+}, S_{c_-}) = \|S_{c_+} - S_{c_-}\| = \left\| \begin{pmatrix} c_+ - c_- & 0 \\ 0 & -(c_+ - c_-) \end{pmatrix} \right\| \\ &= \sqrt{2(c_+ - c_-)^2} = \sqrt{2} \sqrt{1 - t_2^2}. \end{aligned}$$

The push-forward density. We now derive explicitly the push-forward, under π_G , of the uniform distribution on the source. Consider $\Omega := \text{PD}(2) \cap \{\text{tr } S = 1\}$, an open subset of the affine plane $P := \text{Sym}(2) \cap \{\text{tr } S = 1\}$, and let μ be the *uniform* probability measure on Ω . For a measurable set $A \subseteq \Omega$,

$$\mu(A) = \frac{\text{vol}_2(A)}{\text{vol}_2(\Omega)},$$

where vol_2 is the 2-dimensional Lebesgue measure on P induced by the Frobenius inner product $\langle A, B \rangle = \text{tr}(AB)$; intuitively, it measures the area of A . Let $\nu := (\pi_G)_* \mu$ be its push-forward, which is a probability measure defined by $\nu(B) = \mu(\pi_G^{-1}(B))$ for each $B \subseteq \pi_G(\Omega) = \mathcal{C}_G \cap H$.

The segment $\mathcal{C}_G \cap H = \{(1, t_2) : |t_2| < 1\} = \{1\} \times (-1, 1)$ can be identified with the interval $(-1, 1)$ via the isometry

$$\iota : (-1, 1) \longrightarrow \pi_G(\Omega), \quad \iota(t_2) = (1, t_2),$$

which does not distort the lengths. Let $\tilde{\nu}(B) := \nu(\iota(B))$ be the reading of ν on $(-1, 1)$, and let λ be the Lebesgue measure (length) on $(-1, 1)$. The density of $\tilde{\nu}$ with respect to λ is the nonnegative function f on $(-1, 1)$ that satisfies

$$\tilde{\nu}(B) = \int_B f d\lambda \quad \text{for every measurable } B \subseteq (-1, 1),$$

that is, $f = d\tilde{\nu}/d\lambda$, which is also known as the Radon–Nikodym derivative of $\tilde{\nu}$ with respect to λ . It is unique up to a set of measure (length) zero, and its existence is exhibited by its explicit computation, which will be performed below.

Unpacking the definitions, f is characterized by

$$\begin{aligned} \mu(\{S \in \Omega : a < 2s_{12} < b\}) &= \mu(\{S \in \Omega : a < \iota^{-1}(\pi_G(S)) < b\}) = \mu(\pi_G^{-1}(\iota((a, b)))) \\ &\stackrel{(i)}{=} \nu(\iota((a, b))) \stackrel{(ii)}{=} \tilde{\nu}((a, b)) \stackrel{(iii)}{=} \int_{(a, b)} f d\lambda = \int_a^b f(t_2) dt_2, \end{aligned} \quad (3.1)$$

for every $(a, b) \subseteq (-1, 1)$, where (i) is the definition of the push-forward; (ii) is the definition of $\tilde{\nu}$; and (iii) is the definition of f .

Parametrization of P and Ω . Using that $\text{Sym}(2) = \mathcal{L}_G \oplus \mathcal{L}_G^\perp$, every $S \in P = \text{Sym}(2) \cap \{\text{tr } S = 1\}$ can be parametrized as

$$S = S(x, y) = \frac{1}{2}I + xu + yv = \begin{pmatrix} \frac{1}{2} + \frac{y}{\sqrt{2}} & \frac{x}{\sqrt{2}} \\ \frac{x}{\sqrt{2}} & \frac{1}{2} - \frac{y}{\sqrt{2}} \end{pmatrix},$$

$$u = \frac{1}{\sqrt{2}}(E_{12} + E_{21}) \in \mathcal{L}_G, \quad v = \frac{1}{\sqrt{2}}\text{diag}(1, -1) \in \mathcal{L}_G^\perp,$$

where $(x, y) \in \mathbb{R}^2$ are the coordinates that determine S uniquely. Because $v \in \mathcal{L}_G^\perp = \ker \pi_G$, the image $\pi_G(S)$ is not affected by moving y and depends only on x . Specifically, $\pi_G(S) = (1, \sqrt{2}x)$, so the only free coordinate is $t_2 = 2s_{12} = \sqrt{2}x$.

Now, let us determine $\Omega = \text{PD}(2) \cap \{\text{tr } S = 1\} \subseteq P$ in terms of the coordinates (x, y) . Because $\text{tr } S = 1 > 0$ and $\det S = \frac{1}{4} - \frac{1}{2}(x^2 + y^2)$, we have $S \succ 0 \iff \det S > 0 \iff x^2 + y^2 < \frac{1}{2}$, so Ω corresponds to the open disk D of radius $\frac{1}{\sqrt{2}}$:

$$\Omega = \{S(x, y) \in P : (x, y) \in D\}, \quad D = \left\{ (x, y) \in \mathbb{R}^2 : x^2 + y^2 < \frac{1}{2} \right\}.$$

Since u, v are orthonormal, the parametrization

$$\begin{aligned} \mathbb{R}^2 &\rightarrow P \\ (x, y) &\mapsto S := S(x, y) \end{aligned}$$

is an isometry, so vol_2 agrees with the ordinary area $dx dy$. Hence $\text{vol}_2(\Omega)$ equals the area of the disk D , that is $\text{vol}_2(\Omega) = \pi (1/\sqrt{2})^2 = \pi/2$, and, being uniform, μ has constant density $1/\text{vol}_2(\Omega) = 2/\pi$ with respect to $dx dy$.

Computation of f . Fix $(a, b) \subseteq (-1, 1)$. Using Equation (3.1) and $2s_{12} = \sqrt{2}x$, we have

$$\begin{aligned} \int_a^b f(t_2) dt_2 &= \mu(\{S \in \Omega : a < 2s_{12} < b\}) = \frac{\text{vol}_2(\{S \in \Omega : a < 2s_{12} < b\})}{\text{vol}_2(\Omega)} \\ &= \frac{2}{\pi} \text{vol}_2(\{S \in \Omega : a < 2s_{12} < b\}) = \frac{2}{\pi} \text{area} \left(\left\{ (x, y) \in D : \frac{a}{\sqrt{2}} < x < \frac{b}{\sqrt{2}} \right\} \right) \\ &= \frac{2}{\pi} \int_{a/\sqrt{2}}^{b/\sqrt{2}} 2\sqrt{\frac{1}{2} - x^2} dx. \end{aligned}$$

Substituting $x = t_2/\sqrt{2}$ (so $x : \frac{a}{\sqrt{2}} \rightarrow \frac{b}{\sqrt{2}}$ becomes $t_2 : a \rightarrow b$ and $dx = \frac{1}{\sqrt{2}} dt_2$),

$$\int_a^b f(t_2) dt_2 = \int_a^b \frac{2}{\pi} \cdot 2\sqrt{\frac{1}{2} - \frac{t_2^2}{2}} \cdot \frac{1}{\sqrt{2}} dt_2 = \int_a^b \frac{2}{\pi} \sqrt{1 - t_2^2} dt_2.$$

Consequently,

$$f(t_2) = \frac{2}{\pi} \sqrt{1 - t_2^2}, \quad |t_2| < 1,$$

which is a density function on $(-1, 1)$ as it is nonnegative and integrates to 1 over its domain, that is, $\int_{-1}^1 f(t_2) dt_2 = 1$.

Interpretation. The density f is the semicircle density on $(-1, 1)$. It attains its maximum at $t_2 = 0$, that is, at $\pi_G(\frac{1}{2}I)$, where $\frac{1}{2}I = S(0, 0)$ corresponds to the center $(x, y) = (0, 0)$ of the disk D , the point of D farthest from ∂D . On the other hand, $f(t_2) \rightarrow 0$ as $t_2 \rightarrow \pm 1$, which correspond to the two points $(1, \pm 1)$ of $\partial \mathcal{C}_G \cap H$. Thus, the uniform distribution μ in $\text{PD}(2) \cap \{\text{tr } S = 1\}$ yields a push-forward distribution ν on the slice $\mathcal{C}_G \cap H$ of the cone of

sufficient statistics that corresponds, under the identification $\iota : (-1, 1) \rightarrow \mathcal{C}_G \cap H$, to the density f , which attains its maximum at the center and decays to zero at the boundary.

Finally, observe that the density $f(t_2)$ is proportional to the length $\text{vol}_1(\Phi(t)) = \sqrt{2}\sqrt{1-t_2^2}$ computed at the beginning of the example. Indeed,

$$f(t_2) = \frac{\sqrt{2}}{\pi} \text{vol}_1(\Phi(t)) \propto \text{vol}_1(\Phi(t)).$$

This proportionality lets us deduce that f vanishes at $t_2 = \pm 1$ from the size of $\Phi(t) = \pi_G^{-1}(t) \cap \text{PSD}(2)$ alone, without computing f explicitly. Recall that $\Phi(t) = \{S_c : |c| \leq \frac{1}{2}\sqrt{1-t_2^2}\}$, so for $|t_2| < 1$ it is a segment, but as $t_2 \rightarrow \pm 1$ the admissible range of c shrinks to $\{0\}$, so $\Phi(t)$ reduces to the single point S_0 and $\text{vol}_1(\Phi(t)) \rightarrow 0$. Hence $f(t_2) \rightarrow 0$. $\blacklozenge$

Remark 3.6 (The role of f in the sampling). The density f governs the distribution of the projected sample. In Example 3.5, if $S_1, \dots, S_N$ are drawn independently and uniformly from $\Omega = \text{PD}(2) \cap \{\text{tr } S = 1\}$, then their projections $\pi_G(S_1), \dots, \pi_G(S_N)$ are independent with distribution ν , that is, with density f under the identification ι . Consequently, the expected fraction of projected points landing in a subset $\{1\} \times (a, b)$ of $\mathcal{C}_G \cap H$ (where $(a, b) \subseteq (-1, 1)$) is $\int_a^b f(t_2) dt_2$.

Since f attains its maximum at the center $t_2 = 0$ and decays to zero as $t_2 \rightarrow \pm 1$, the projected cloud $\pi_G(\{S_1, \dots, S_N\})$ is dense in the middle of $\mathcal{C}_G \cap H$ and increasingly sparse near $\partial(\mathcal{C}_G \cap H)$. Increasing N adds points everywhere in the same proportion, so the region near the boundary remains comparatively depleted no matter how large N is. Therefore, in this example, this depletion is a structural feature of the projection, not a consequence of having too few samples; for more general graphs G , this is what Proposition 3.4 and Corollary 3.7 would imply. Note that here N is the number of matrices in the reference sample M , not the number MC of Monte Carlo replicates that produce the query points. $\blacklozenge$

The following result, which we also state without proof, shows that this behavior is not particular to the colored 2-cycle.

Corollary 3.7 (Vanishing of the fiber volume on the boundary). *Assume that for every vertex $i \in V$ there exists $j \in V \setminus \{i\}$ such that $(i, j) \notin E$ or $\lambda(i) = \lambda(j)$. If $t \in \partial\mathcal{C}_G$, then $\text{vol}_k(\Phi(t)) = 0$. In particular, by Proposition 3.4, the density f vanishes on $\partial(\mathcal{C}_G \cap H)$.*

The assumption of Corollary 3.7 holds in many cases. For instance, it holds for every graph in which no vertex is adjacent to all the other vertices, whatever the coloring, such as every m -cycle with $m \geq 4$ and, in particular, all the colored 4-cycles of [Uhl12, Table 2], among them G_3 , G_6 and G_{18} . It also holds for every graph in which each vertex shares its color with some other vertex, such as every graph with all vertices of the same color, among them C_2 . On the other hand, it fails for the uncolored complete graph, for which $k = 0$ and π_G is a bijection, so the projection of a uniform sample is again uniform.

Chapter 4

Future research

The results of Chapter 3 are a first step towards one of the directions for future work proposed in [HKS26, Section 5]: understanding the projection onto $\mathcal{C}_G$ of the uniform distribution on $\text{PD}(m) \cap \{\text{tr} = 1\}$.

The first natural step is to prove Proposition 3.4 and Corollary 3.7, which are stated without proof. We expect both of them to hold, since they agree with the explicit computation of Example 3.5 and with the empirical observations of [HKS26, Section 4.2]; nevertheless, they have to be proved before they can be used as established results. For Proposition 3.4, a possible way is to generalize the computation of Example 3.5: parametrize the hyperplane $\{\text{tr} = 1\}$ by orthonormal coordinates adapted to the decomposition $\text{Sym}(m) = \mathcal{L}_G \oplus \mathcal{L}_G^\perp$, so that π_G only depends on the coordinates in $\mathcal{L}_G$, and integrate out the coordinates in $\mathcal{L}_G^\perp$, which describe the fiber. For Corollary 3.7, the starting point is that, if $t \in \partial\mathcal{C}_G$, then $\pi_G^{-1}(t)$ contains no positive definite matrix, so every matrix in $\Phi(t)$ is singular. Once these results are established, it would also be interesting to quantify how fast f decays near $\partial\mathcal{C}_G$: in Example 3.5, $f(t_2) = \frac{2}{\pi} \sqrt{(1-t_2)(1+t_2)}$ decays as the square root of the distance to the boundary.

Finally, note that improving the reference cloud may not be enough on its own. Even with a uniform reference cloud on $\mathcal{C}_G \cap H$, the query points close to $\partial\mathcal{C}_G$ would remain difficult to classify with annuli of fixed radii $0 < R_1 < R_2$: if $y \in \mathcal{C}_G \cap H$ is at distance less than R_2 from $\partial\mathcal{C}_G$, then, by convexity of $\text{cl}(\mathcal{C}_G) \cap H$, part of the annulus of y lies outside $\text{cl}(\mathcal{C}_G) \cap H$, so y cannot be surrounded by cloud points in every direction. Since the query points concentrate near $\partial\mathcal{C}_G$ (Problem 3.3), this suggests combining a better reference cloud with neighborhoods that adapt to the data, for instance with a dynamic choice of radius, as proposed in [HKS26, Section 5].

Bibliography

- [HL08] Søren Højsgaard and Steffen L. Lauritzen. “Graphical Gaussian models with edge and vertex symmetries”. In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 70.5 (2008), pp. 1005–1027.
- [HKS26] Roser Homs, Olga Kuznetsova, and Bernadette J. Stolz. *A computational approach to maximum likelihood thresholds for colored Gaussian graphical models*. 2026. arXiv: [2609.02382](#) [stat.ML].
- [MMJ12] Martin Mittelbach, Bho Matthiesen, and Eduard A. Jorswieck. “Sampling uniformly from the set of positive definite matrices with trace constraint”. In: *IEEE Transactions on Signal Processing* 60.5 (2012), pp. 2167–2179.
- [Sto+20] Bernadette J. Stolz et al. “Geometric anomaly detection in data”. In: *Proceedings of the National Academy of Sciences* 117.33 (2020), pp. 19664–19669.
- [SU10] Bernd Sturmfels and Caroline Uhler. “Multivariate Gaussians, semidefinite matrix completion, and convex algebraic geometry”. In: *Annals of the Institute of Statistical Mathematics* 62 (2010), pp. 603–638. DOI: [10.1007/s10463-010-0295-4](#).
- [Sul18] Seth Sullivant. *Algebraic Statistics*. Vol. 194. Graduate Studies in Mathematics. American Mathematical Society, 2018.
- [Uhl12] Caroline Uhler. “Geometry of maximum likelihood estimation in Gaussian graphical models”. In: *The Annals of Statistics* 40.1 (2012), pp. 238–261. DOI: [10.1214/11-AOS957](#).
- [Uhl17] Caroline Uhler. “Gaussian Graphical Models: An Algebraic and Geometric Perspective”. In: *arXiv preprint arXiv:1707.04345* (2017).