Computing logarithmic Voronoi cells

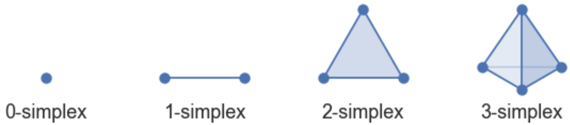
Yulia Alexandr (UC Berkeley)

Joint Math Meeting 2023
Boston, MA

Logarithmic Voronoi cells

- A *probability simplex* is defined as

$$\Delta_{n-1} = \{(p_1, \ldots, p_n) \,:\, p_1 + \cdots + p_n = 1, p_i \geq 0 \text{ for } i \in [n]\}.$$



- A *statistical model* is a subset of Δ_{n-1}. An *algebraic statistical model* is a subset $\mathcal{M} = \mathcal{V} \cap \Delta_{n-1}$ for some variety $\mathcal{V} \subseteq \mathbb{C}^n$.
- For an empirical data point $u = (u_1, ..., u_n) \in \Delta_{n-1}$, the *log-likelihood function* defined by u assuming distribution $p = (p_1, ..., p_n) \in \mathcal{M}$ is

$$\boxed{\ell_u(p) = u_1 \log p_1 + u_2 \log p_2 + \cdots + u_n \log p_n}.$$

Maximum likelihood estimation

Fix an algebraic statistical model $\mathcal{M} \subseteq \Delta_{n-1}$

1. The maximum likelihood estimation problem (MLE):

Given a sampled empirical distribution $u \in \Delta_{n-1}$, which point $p \in \mathcal{M}$ did it most likely come from? In other words, we wish to maximize $\ell_u(p)$ over all points $p \in \mathcal{M}$.

Maximum likelihood estimation

Fix an algebraic statistical model $\mathcal{M} \subseteq \Delta_{n-1}$

1. The maximum likelihood estimation problem (MLE):

Given a sampled empirical distribution $u \in \Delta_{n-1}$, which point $p \in \mathcal{M}$ did it most likely come from? In other words, we wish to maximize $\ell_u(p)$ over all points $p \in \mathcal{M}$.

2. Computing logarithmic Voronoi cells:

Given a point $q \in \mathcal{M}$, what is the set of all points $u \in \Delta_{n-1}$ that have q as a global maximum when optimizing the function $\ell_u(p)$ over $\mathcal{M}$?

The set of all such elements $u \in \Delta_{n-1}$ is the *logarithmic Voronoi cell* at q.

Proposition (A., Heaton)

Logarithmic Voronoi cells are convex sets.

The *log-normal space* at q is the space of possible data points $u \in \mathbb{R}^n$ for which q is a critical point of $\ell_u(p)$. It is a *linear* space.

Intersecting this space with the simplex Δ_{n-1}, we obtain a polytope, which we call the *log-normal polytope* at q.

The log-normal polytope at q contains the logarithmic Voronoi cell at q.

Proposition (A., Heaton)

Logarithmic Voronoi cells are convex sets.

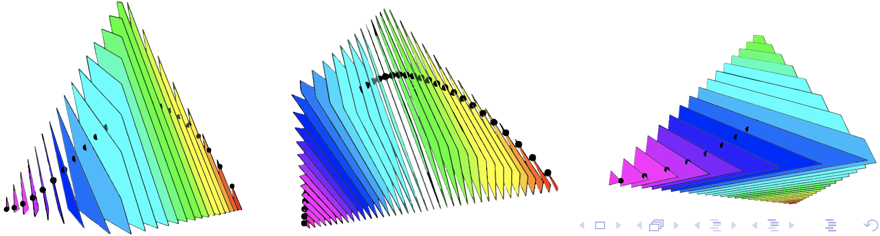
The *log-normal space* at q is the space of possible data points $u \in \mathbb{R}^n$ for which q is a critical point of $\ell_u(p)$. It is a *linear* space.

Intersecting this space with the simplex Δ_{n-1}, we obtain a polytope, which we call the *log-normal polytope* at q.

The log-normal polytope at q contains the logarithmic Voronoi cell at q.

Example (The twisted cubic.)

The curve is given by $p \mapsto \left(p^3, 3p^2(1-p), 3p(1-p)^2, (1-p)^3\right)$.



The Hardy-Weinberg curve

Consider a model parametrized by

$$p \mapsto \left(p^2, 2p(1-p), (1-p)^2\right).$$

Performing implicitization, we find that the model $\mathcal{M} = \mathcal{V}(f)$ where

$$f = \begin{bmatrix} 4x_1x_3 - x_2^2 \\ x_1 + x_2 + x_3 - 1 \end{bmatrix}.$$

The augmented Jacobian is given by:

$$A = \begin{bmatrix} 4x_3 & -2x_2 & 4x_1 \\ 1 & 1 & 1 \\ u_1/x_1 & u_2/x_2 & u_3/x_3 \end{bmatrix}.$$

Fix a point $q \in \mathcal{M}$ and substitute x_i for q_i in A. All points $u \in \mathbb{R}^3$ at which the determinant vanishes define the log-normal space at q.

The Hardy-Weinberg curve

$$\det A = 4u_1 - 4u_3 - 4u_2 \cdot \frac{x_1}{x_2} + 2u_1 \cdot \frac{x_2}{x_1} - 2u_3 \cdot \frac{x_2}{x_3} + 4u_2 \cdot \frac{x_3}{x_2}$$

For example, at $p = 0.2$, we get a point $q = (0.04, 0.32, 0.64) \in \mathcal{M}$. The log-normal space at q is the plane

$$20u_1 + 7.5u_2 - 5u_3 = 0.$$

Sampling more points, we get the following pictures:

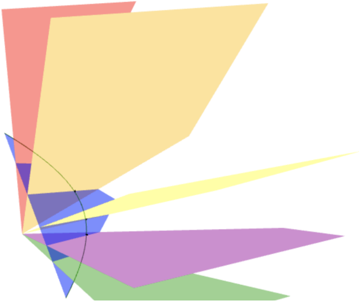
The Hardy-Weinberg curve

$$\det A = 4u_1 - 4u_3 - 4u_2 \cdot \frac{x_1}{x_2} + 2u_1 \cdot \frac{x_2}{x_1} - 2u_3 \cdot \frac{x_2}{x_3} + 4u_2 \cdot \frac{x_3}{x_2}$$

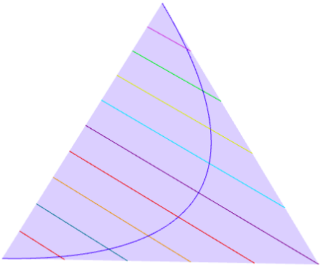
For example, at $p = 0.2$, we get a point $q = (0.04, 0.32, 0.64) \in \mathcal{M}$. The log-normal space at q is the plane

$$20u_1 + 7.5u_2 - 5u_3 = 0.$$

Sampling more points, we get the following pictures:



Log-normal spaces



Log-normal polytopes = Log-Voronoi cells

Polytopal cells

The *maximum likelihood degree* (ML degree) of $\mathcal{M}$ is the number of complex critical points when optimizing $\ell_u(x)$ over $\mathcal{M}$ for generic data u.

Theorem (A., Heaton)

If $\mathcal{M}$ is a finite model, a linear model, a toric model, or a model of ML degree 1, the logarithmic Voronoi cell at any point $p \in \mathcal{M}$ is equal to the log-normal polytope at p.

Polytopal cells

The *maximum likelihood degree* (ML degree) of $\mathcal{M}$ is the number of complex critical points when optimizing $\ell_u(x)$ over $\mathcal{M}$ for generic data u.

Theorem (A., Heaton)

If $\mathcal{M}$ is a finite model, a linear model, a toric model, or a model of ML degree 1, the logarithmic Voronoi cell at any point $p \in \mathcal{M}$ is equal to the log-normal polytope at p.

What about logarithmic Voronoi cells that are **not** polytopes?

Numerical algebraic geometry

Logarithmic Voronoi cells that are not polytopes are mysterious. Are they semialgebraic sets? How to describe their boundary?

Conjecture

The boundary of logarithmic Voronoi cells are, in general, not algebraic.

Numerical algebraic geometry provides an alternative approach to logarithmic Voronoi cells for more complicated models.

Example (Mixture model)

Consider the 3-dimensional model in Δ_5 parametrized as:

$$(a, s, t) \mapsto \left(as^5 + (1-a)t^5, \ldots, a(1-s)^5 + (1-a)(1-t)^5\right)$$

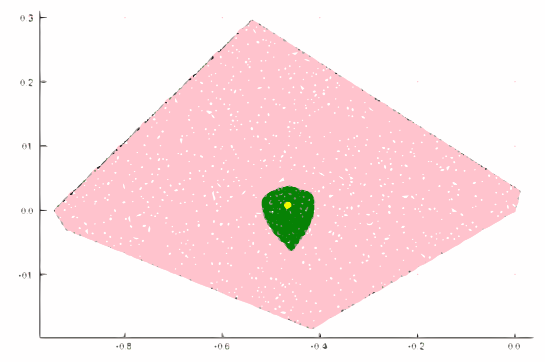
Mixture model, continued

Implicitly, it is described by

$$
\begin{aligned}
f_0(x) &= x_1 + x_2 + x_3 + x_4 + x_5 + x_6 - 1\\
f_1(x) &= 20x_1x_3x_5 - 10x_1x_4^2 - 8x_2^2x_5 + 4x_2x_3x_4 - x_3^3\\
f_2(x) &= 100x_1x_3x_6 - 20x_1x_4x_5 - 40x_2^2x_6 + 4x_2x_3x_5 + 2x_2x_4^2 - x_3^2x_4\\
f_3(x) &= 100x_1x_4x_6 - 40x_1x_5^2 - 20x_2x_3x_6 + 4x_2x_4x_5 + 2x_3^2x_5 - x_3x_4^2\\
f_4(x) &= 20x_2x_4x_6 - 8x_2x_5^2 - 10x_3^2x_6 + 4x_3x_4x_5 - x_4^3.
\end{aligned}
$$

- Pick a point $p_{\text{fav}} = \left(\frac{518}{9375}, \frac{124}{625}, \frac{192}{625}, \frac{168}{625}, \frac{86}{625}, \frac{307}{9375}\right) \in \mathcal{M}$.
- The log-normal space at p_{fav} is 3-dimensional.
- Intersecting with Δ_5, we get the 2-dimensional log-normal polytope at p_{fav}, which is a hexagon.

Mixture model, continued



Mixture model, continued

- Sample 60000 points u_i from within the log-normal hexagon.
- We will color each point green if p_{fav} is its MLE and pink otherwise.
- Let $df = [\nabla f_0 \, \nabla f_1 \, \dots \, \nabla f_4]^T$ be a 5×6 matrix.
- Multiply by a random 3×5 matrix A on the left, then $A \cdot df$ is a 3×6 matrix with the same rowspace. This is dimension reduction.
- Add the 4th row $\nabla \ell_u = (u_i/p_i)$.
- Introducing new variables $\lambda_1, \lambda_2, \lambda_3$ we will require that some linear combination of the first 3 rows is equal to the last row.
- This gives 6 equations:

$$G \quad := \quad [\lambda_1 \ \lambda_2 \ \lambda_3 - 1] \begin{bmatrix} A \cdot df \\ \nabla \ell_u \end{bmatrix} \mathrm{diag}(p_i) = 0.$$

- In addition, we add the original 5 equations f_0, f_1, f_2, f_3, f_4, so we have 11 equations in 9 variables (rather than 225 minors in 6 variables).

HomotopyContinuation.jl and monodromy

We could start with a random point u_{start}, solve the system, and use these starting solutions to track solutions for the 60000 points u_i. But solve(G) reports that we have to track 110592 paths... Use monodromy!

HomotopyContinuation.jl and monodromy

We could start with a random point u_{start}, solve the system, and use these starting solutions to track solutions for the 60000 points u_i. But solve(G) reports that we have to track 110592 paths... Use monodromy!

- The ML degree of our model is 39.
- Start with a random u_0 in the hexagon; we know one solution: p_{fav}!
- Perturb our system of equations around a loop in parameter space, tracking this solution along the way.
- When we return to the original system, the original solution likely has moved to another one of the 39 solutions. Now we have two!
- Repeating this process, we discover all 39 solutions.
- For the 60000 other parameter points u_i, perturb the system of equations at u_0, and track all solutions along the way. This way we can find the 39 solutions for each sample point in the hexagon.
- We check which one of them maximizes log-likelihood. If it is p_{fav}, we color it green. If not, we color it pink.

In general

Theorem (A., Heaton)

Let $\mathcal{M}$ be an irreducible algebraic model in Δ_{n-1} described implicitly by m polynomials f. Let $u \in \Delta_{n-1}$ be fixed and generic. With probability 1, all points $p \in \mathcal{M}$ such that $u \in \log N_p\mathcal{M}$ are among the finitely many isolated solutions to the square system

$$\underbrace{\begin{bmatrix} [\lambda - 1] \begin{bmatrix} A \cdot df \\ \nabla \ell_u \end{bmatrix} & f \end{bmatrix}}_{1\times(n+m)} \begin{bmatrix} I_{n+c} \\ B \end{bmatrix} = \underbrace{[0 \cdots 0]}_{1\times(n+c)}.$$

Thanks!