Tropical Sufficient Statistics for Persistent Homology

Anthea Monod

Columbia University

am4691@cumc.columbia.edu

Joint work with Sara Kališnik (Max Planck Institute),
Juan Ángel Patiño-Galindo (Columbia) & Lorin Crawford (Brown)

Many thanks to Steve Oudot (INRIA Saclay) for helpful discussions

22 February 2018

Topological Signatures & Summary Statistics

Summary Statistic: For a given data sample, calculate a quantity to summarize it (= *feature selection*)

$$f : \text{Data} \rightarrow \text{"Nice" Space}$$

Desired Properties:

- Injectivity
- Ability to define probabilistic models in the transformed space
- Amenability to existing statistical methodology & ML algorithms
- Computable distances

Topologically,

- Persistent Homology Transform (Turner, Mukherjee, Boyer; 2014)
- Smooth Euler Characteristic Transform (Crawford, M., Chen, Mukherjee, Rabadán; 2017)
- Persistence Landscapes (Bubenik; 2015)
- etc. (mentioned by Uli & Wolfgang)

From Summary Statistics to Sufficient Statistics

Idea: Sufficient statistics allow for a lower dimensional or less complex representation of data *without the loss of information*

- Sufficiency for a parameter that defines a distribution
  e.g. $\bar{x}$ for μ in $\mathcal{N}(\mu, \sigma^2)$
- Sufficiency for a family/class of distributions via a statistic
  e.g. Exponential family, distributions on spaces, order statistics
  $\Longrightarrow$ Measure-theoretic interpretation (Halmos & Savage, 1949; Diaconis, 1992)

Sufficient statistics are summary statistics that are injective and measurable, and map between two well-defined probability spaces

M., Kališnik, Patiño-Galindo, Crawford (2017)

Sufficient statistics for persistent homology, constructed via tropical geometry, exist

$\Longrightarrow$ Allows for parametric analysis of recombination in phylogenetics

Statistical Sufficiency & The Factorization Criterion

Definition

Let X be a vector of observations of size n with $X_i \sim f_\vartheta$ *i.i.d.*
A statistic $T(X)$ is *sufficient* for ϑ if

$$\mathbb{P}\big(X = x | T(X) = t, \vartheta\big) = \mathbb{P}\big(X = x | T(X) = t\big)$$

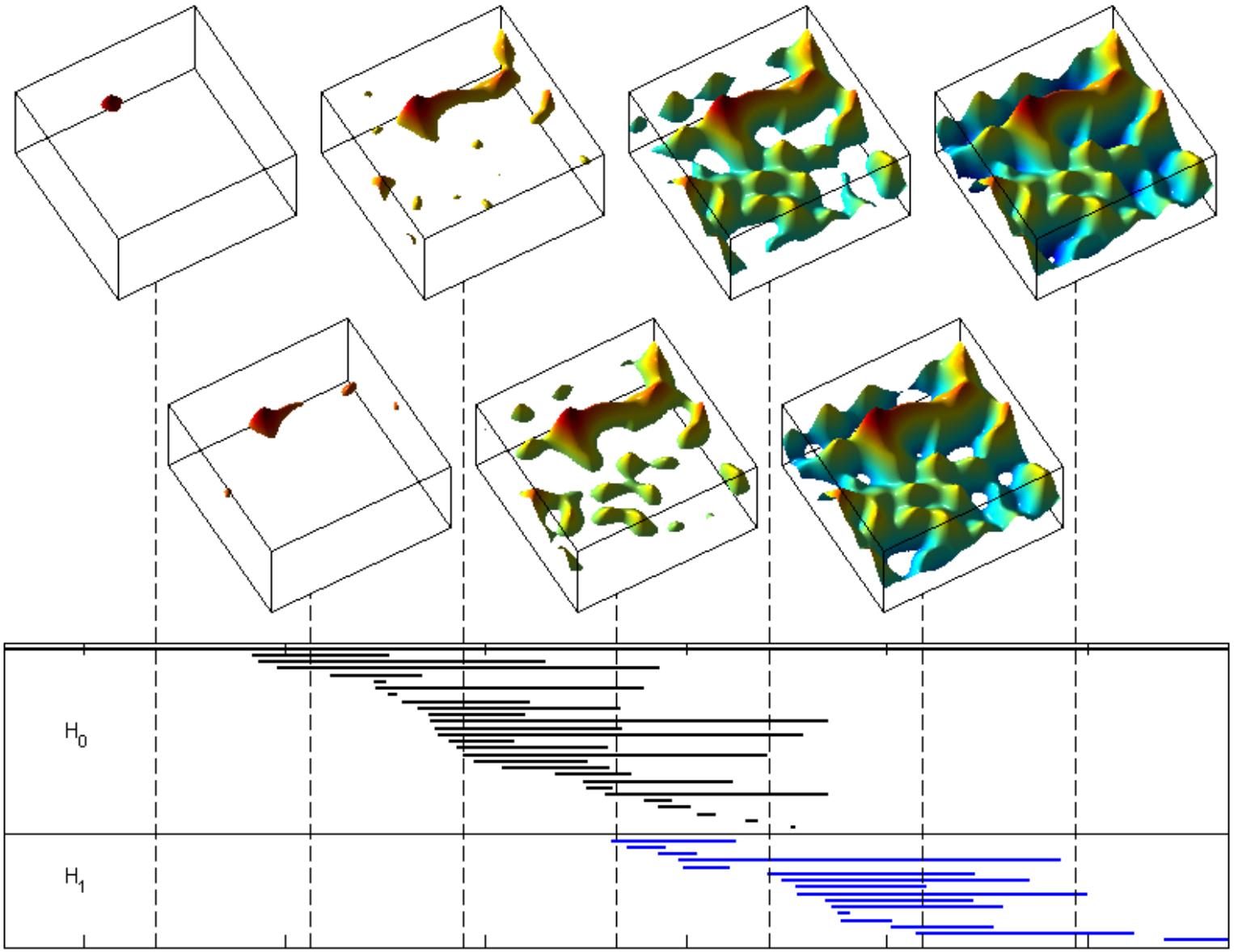
Theorem (Neyman–Fisher, 1922, 1935)

If the pdf for the observed data is $f(x;\vartheta)$, then the statistic $T = T(x)$ is sufficient for $\vartheta \in \Theta$ if and only if $f(x;\vartheta) = h(x)g\big(T(x);\vartheta\big)$

Theorem (Halmos–Savage, 1949)

A necessary and sufficient condition that the statistic $T(\cdot)$ be sufficient for a dominated set $\mathcal{M}$ of measures on a σ-algebra $\mathbf{S}$ is that for every $\mu \in \mathcal{M}$, the density $f_\mu := \mathrm{d}\mu/\,\mathrm{d}\lambda$ admits the factorization $f_\mu(x) = h(x)g_\mu\big(T(x)\big)$

Persistent Homology in 2 Dimensions



Barcode Space

$(x_1, d_1, x_2, d_2, \ldots, x_n, d_n)$; $x_i =$ birth; $d_i =$ length; $x_i \geq 0$
$B_n =$ Orbit space of the action of the symmetric group S_n on n letters on the product $\left([0,\infty) \times [0,\infty)\right)^n$, given by permuting the coordinates

Definition

The barcode space $\mathcal{B}_{\leq n}$ consisting of barcodes with at most n intervals is the quotient

$$\coprod_{n \in \mathbf{N}_{\leq n}} B_n / \sim$$

where $\sim$ is generated by the following equivalences whenever $d_n = 0$:

$$\left\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\right\} \sim \left\{(x_1, d_1), (x_2, d_2), \ldots, (x_{n-1}, d_{n-1})\right\}$$

Regularizing Subsets of Barcode Space: For fixed $m > 0$, denote by $\mathcal{B}^m_{\leq n}$ the subset of $\mathcal{B}_{\leq n}$ that consist of those $(x_1, d_1, \ldots, x_n, d_n)$ with $d_i > 0$ for all $i = 1, \ldots, n$ such that $x_i \leq m d_i$

Fundamentals of Tropical/Arctic Geometry

Tropical geometry = "Skeletonized" version of algebraic geometry

Tropical/Min-plus Semiring:
$(\mathbf{R} \cup \{+\infty\}, \oplus, \odot)$ with $a \oplus b := \min(a, b)$ and $a \odot b := a + b$
Arctic/Max-plus Semiring:
$(\mathbf{R} \cup \{-\infty\}, \boxplus, \odot)$ with $a \boxplus b := \max(a, b)$ and $a \odot b := a + b$

- Commutative
- Associative
- Distributive Law: $a \odot (b \oplus c) = a \odot b \oplus a \odot c$
- Frobenius Identity in Tropical Arithmetic:

$$(a \oplus b)^n = a^n \oplus b^n$$

b^{-1} is the inverse of b w.r.t. $\odot = -b$ in ordinary arithmetic

Tropical/Arctic Functions

Let $x_1, x_2, \ldots, x_n$ be variables of elements in the tropical/arctic semiring

- **Tropical/Arctic Monomial:** Any product or quotient of $x_1, x_2, \ldots, x_n$; repetition is allowed
- **Tropical Polynomial:**

$$p(x_1, x_2, \ldots, x_n) = a_1 \odot x_1^{a_1^1} x_2^{a_2^1} \cdots x_n^{a_n^1} \oplus a_2 \odot x_1^{a_1^2} x_2^{a_2^2} \cdots x_n^{a_n^2} \oplus$$
$$\oplus \cdots \oplus a_m \odot x_1^{a_1^m} x_2^{a_2^m} \cdots x_n^{a_n^m}$$

Each tropical/arctic polynomial is a continuous piecewise linear function
The passage from tropical/arctic polynomials to functions is not 1-1, e.g.:

$$\begin{aligned} x_1^2 \boxplus x_2^2 &= 2x_1 \boxplus 2x_2 \\ &= \max\{2x_1, 2x_2\} \\ x_1^2 \boxplus x_2^2 \boxplus x_1 x_2 &= 2x_1 \boxplus 2x_2 \boxplus (x_1 + x_2) \\ &= \max\{2x_1, 2x_2, x_1 + x_2\} \\ &= \max\{2x_1, 2x_2\} \end{aligned}$$

Functional Equivalence & Semirings of Equivalence Classes

Functional equivalence, $p(x_1, x_2, \ldots, x_n) = q(x_1, x_2, \ldots, x_n)$, denoted by $\sim$, is an equivalence relation on the set of all max-plus polynomial expressions

We want to study functions, so look at the expressions that define the same functions $\Longrightarrow$ Max-plus polynomials are the semiring of equivalence classes of max-plus polynomial expressions w.r.t. $\sim$

We will use this semiring to assign vectors (functions) to barcodes (coordinatize barcode space)

Very related: "The Ring of Algebraic Functions on Persistence Barcodes" — Adcock, Carlsson & Carlsson (2016)
Careful: These functions are not Lipschitz w.r.t. Wasserstein and bottleneck distances...

Identifying Tropical Functions for Barcodes

Fix n and let S_n act on $X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix}$ by left multiplication

$$\mathscr{E}_n = \left\{ \begin{pmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \\ \vdots & \vdots \\ e_{n,1} & e_{n,2} \end{pmatrix} \neq [0]_n^2 : e_{i,j} \in \{0,1\} \text{ for } i = 1,2,\ldots,n; j = 1,2 \right\}$$

Every matrix $E \in \mathscr{E}_n$ determines a max-plus monomial from X by

$$P(E) = x_{1,1}^{e_{1,1}} x_{1,2}^{e_{1,2}} \cdots x_{n,1}^{e_{n,1}} x_{n,2}^{e_{n,2}}$$

The orbits $E_i \in \mathscr{E}_n/S_n$ under the row permutation action on $\mathscr{E}_n$ determine max-plus polynomials by max-plus multiplication over row permutations:

$$E_{(e_{11},e_{12}),(e_{21},e_{22}),\ldots,(e_{n1},e_{n2})} := P(E_1) \boxplus P(E_2) \boxplus \cdots \boxplus P(E_m)$$

Tropical Coordinates on Barcode Space

Proposition

Let $\left[(x_1, d_1, \ldots, x_n, d_n)\right]$ *and* $\left[(x'_1, d'_1, \ldots, x'_n, d'_n)\right]$ *be two orbits under the row permutation action on* $\mathbf{R}^{2n}$*. If*

$$E_{(0,1)^i(1,1)^j}\left[(x_1, d_1, \ldots, x_n, d_n)\right] = E_{(0,1)^i(1,1)^j}\left[(x'_1, d'_1, \ldots, x'_n, d'_n)\right]$$

for all $i, j \leq n$*, then* $\left[(x_1, d_1, \ldots, x_n, d_n)\right] = \left[(x'_1, d'_1, \ldots, x'_n, d'_n)\right]$.

Therefore,

$$E_{m,(1,1)^i,(0,1)^j(x_1,d_1,\ldots,x_n,d_n)} := E_{(1,1)^i,(0,1)^j}(x_1 \oplus d_1^m, d_1, \ldots, x_n \oplus d_n^m, d_n)$$

induces an injective map on $\mathcal{B}^m_{\leq n}$ and separates nonequivalent barcodes

Tropical Sufficient Statistics for Persistent Homology

Theorem (Kališnik (2016); M., Kališnik, Patiño-Galindo, Crawford (2017))

The following collection of tropical polynomials

$$T : \mathcal{B}^m_{\leq n} \to \mathbf{R}^d$$
$$\mathscr{B} \mapsto \left(E_{m,(1,1)^i,(0,1)^j}(x_1, d_1, \ldots, x_n, d_n)\right)_{i+j \in \mathbf{N}_{\leq n}}(\mathscr{B})$$

- *induces a map on $\mathcal{B}^m_{\leq n}$, thereby mapping from barcode space to Euclidean space*
- *are Lipschitz-continuous with respect to the Wasserstein and bottleneck distances*
- *are injective*
- *are measurable via Borel σ-algebras*
- *are sufficient statistics for the family of probability measures $\mathcal{P}$ on the subset of persistence barcodes $\mathcal{B}^m_{\leq n}$*

An Example, $n = 2$

Fix $n = 2 \Longrightarrow$ The set of orbits under the S_2 action is

$$\mathscr{E}_2/S_2 = \left\{ \begin{array}{c} \left[\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}\right], \\ \left[\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\right] \end{array} \right\}$$

We only need a subcollection of all orbits to map barcodes injectively
$\Longrightarrow$ Take the orbits with rows $(1,1)$ and $(0,1)$:

$$\left[\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\right]$$

We need $d = n + \frac{n(n+1)}{2}$ many orbits

Suppose we have two barcodes $\mathscr{B}_1 = \{(1,2),(3,1)\}$ and $\mathscr{B}_2 = \{(2,2)\}$; $\mathscr{B}_1, \mathscr{B}_2 \in \mathcal{B}_{\leq 2}$

1. Compute m: For intervals $(1,2),(3,1),(2,2)$, find the smallest m such that $x_i \leq m d_i \Longrightarrow$ The quotients are $\frac{1}{2}, \frac{3}{1}, 1$, so take $m = 3$, so $\mathscr{B}_1, \mathscr{B}_2 \in \mathcal{B}^3_{\leq 2}$
2. Determine the 2-symmetric max-plus polynomials

$$E_{(1,1)^i,(0,1)^j}(x_1 \oplus d_1^m, d_1, \ldots, x_n \oplus d_n^m, d_n)$$

from

$$\left[\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\right], \left[\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\right] :$$

$$\begin{aligned} E_{3,(0,1),(0,0)}(x_1, d_1, x_2, d_2) &= d_1 \boxplus d_2 \\ &= \max(d_1, d_2) \\ E_{3,(0,1),(0,1)}(x_1, d_1, x_2, d_2) &= d_1 d_2 \\ &= d_1 + d_2 \end{aligned}$$

$$
\begin{aligned}
E_{3,(0,0),(1,1)}(x_1, d_1, x_2, d_2) &= (x_2 \oplus d_2^3)d_2 \boxplus (x_1 \oplus d_1^3)d_1 \\
&= \max\left\{ \min(x_2, 3d_2) + d_2, \min(x_1, 3d_1) + d_1 \right\} \\
E_{3,(1,1),(0,1)}(x_1, d_1, x_2, d_2) &= (x_1 \oplus d_1^3)d_1 d_2 \boxplus (x_2 \oplus d_2^3)d_2 d_1 \\
&= \max\left\{ \min(x_1, 3d_1) + d_1 + d_2, \right. \\
&\qquad \left. \min(x_2, 3d_2) + d_2 + d_1 \right\} \\
E_{3,(1,1),(1,1)}(x_1, d_1, x_2, d_2) &= (x_1 \oplus d_1^3)d_1 (x_2 \oplus d_2^3)d_2 \\
&= \min(x_1, 3d_1) + d_1 + \min(x_2, 3d_2) + d_2
\end{aligned}
$$

3. Evaluate on $\mathscr{B}_1$:

$$
\begin{aligned}
\max(2, 1) &= 2 \\
2 + 1 &= 3 \\
\max\left\{ \min(1, 6) + 2, \min(3, 3) + 1 \right\} &= \max\{1 + 2, 3 + 1\} = 4 \\
\max\left\{ \min(1, 6) + 2 + 1, \min(3, 3) + 2 + 1 \right\} &= \max\{4, 6\} = 6 \\
\min(1, 6) + 2 + \min(3, 3) + 1 &= 7
\end{aligned}
$$

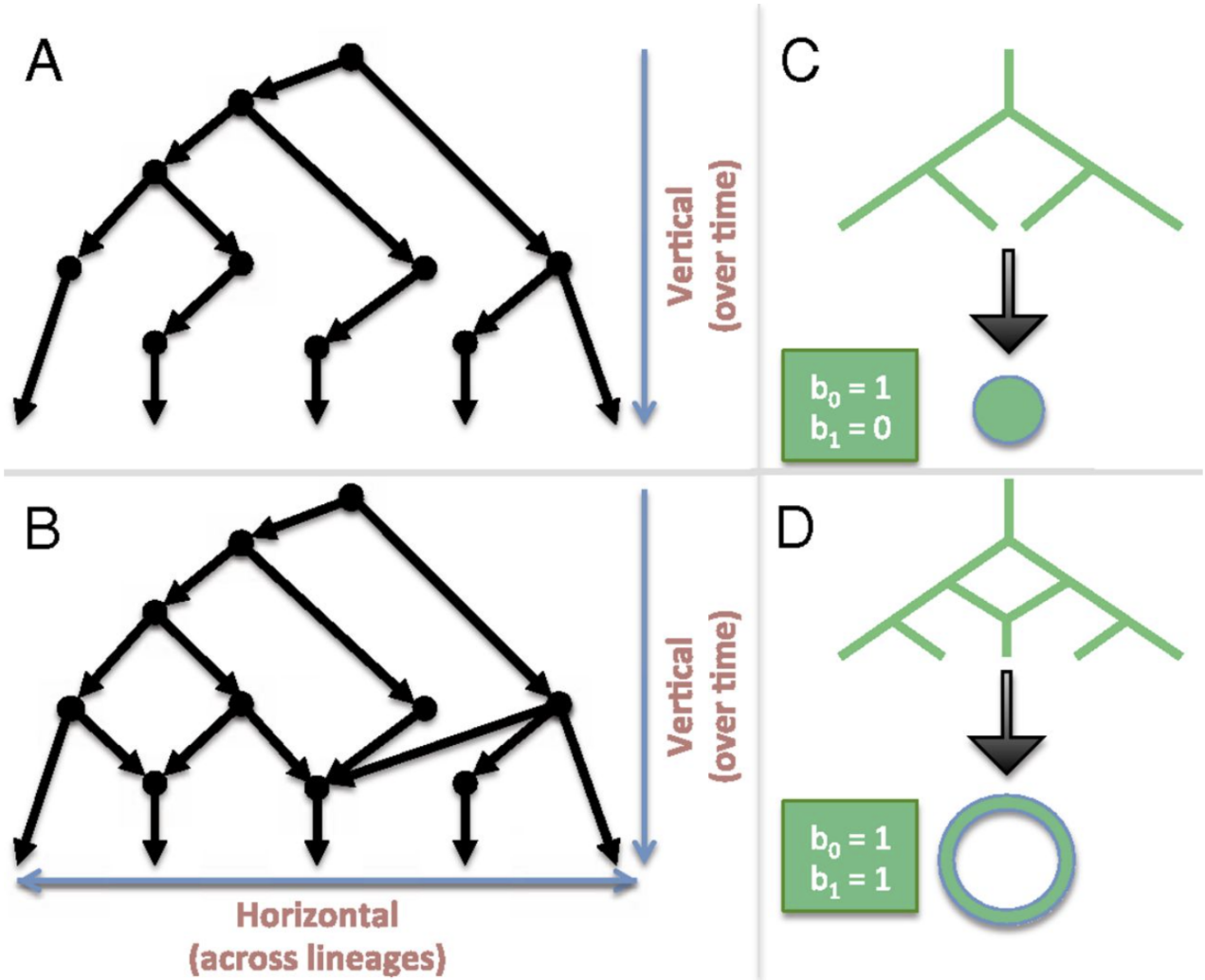
4. Evaluate on $\mathscr{B}_2$:

$$\begin{aligned}
\max(2,2) &= 2 \\
2+0 &= 2 \\
\max\left\{\min(2,6)+2\right\} &= 4 \\
\max\left\{\min(2,6)+2\right\} &= 4 \\
\min(2,6)+2 &= 4
\end{aligned}$$

The Euclidean-space vector representation of $\mathscr{B}_1$ is $(2,3,4,6,7)$, and of $\mathscr{B}_2$ is $(2,2,4,4,4)$

Evolutionary Phylogenetics & Recombination Events

Linking algebraic topology to evolution.



Motivation: Recombination in RNA Viruses

- Horizontal recombination is an important event that causes mutation in RNA viruses (e.g. HIV, avian, swine influenza)
- Molecular phylogenetic analysis to extract and analyze diversification history is extremely tedious and computationally costly
- Applying persistent homology significantly improves computational efficiency:

  Dimension 1 persistence intervals provide explicit information on the genetic divergence of the sequences involved in the recombination event (Chan, Carlsson, Rabadán; PNAS 2013)

  ...but is hard to work with statistically

Application: Analyzing Intra- & Intersubtype Recombination in Avian Influenza

The influenza virus presents a genome with 8 segments (RNA molecules)

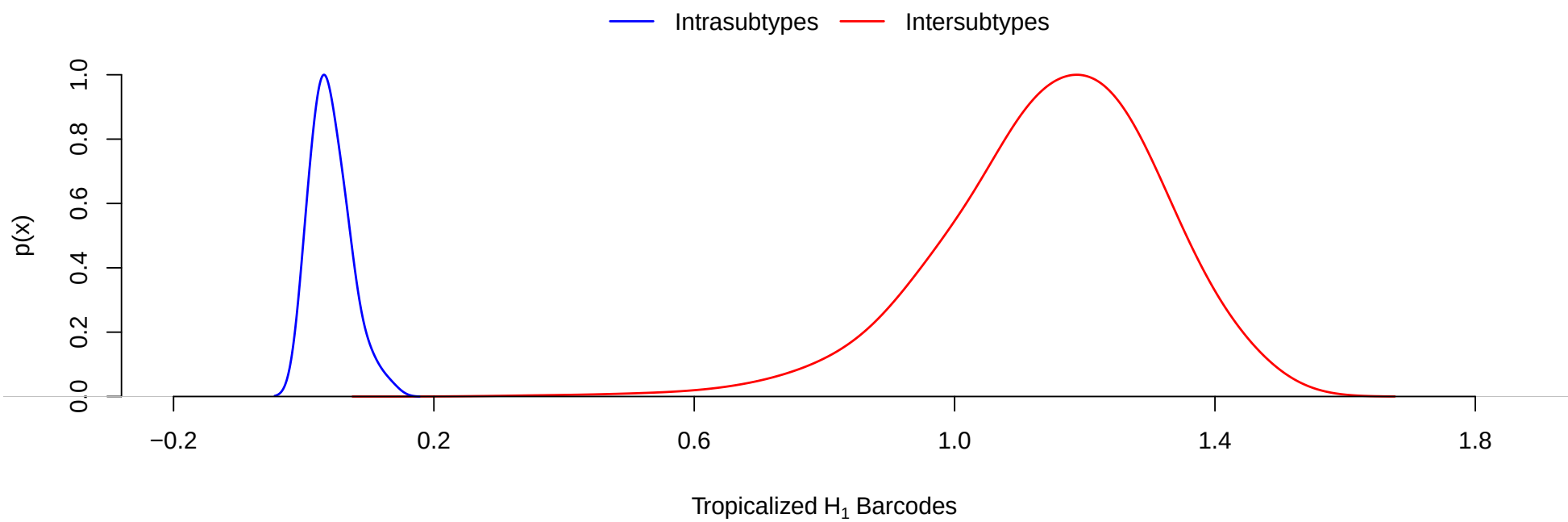
Genetic Recombination:

- *Intrasubtype* — Between viruses of the same subtype
- *Intersubtype* — Between viruses of different subtypes

$\Longrightarrow$ Lengths of PH_1 intrasubtype recombination barcodes will be shorter than those of intersubtype recombination

Detecting gene reassortment is key to understanding mutations within the evolutionary dynamics of viruses

Marginal Distribution of Intra- & Intersubtype Recombination in Avian Influenza



Hellinger Distance

f-divergences measure distances between probability distributions

Definition

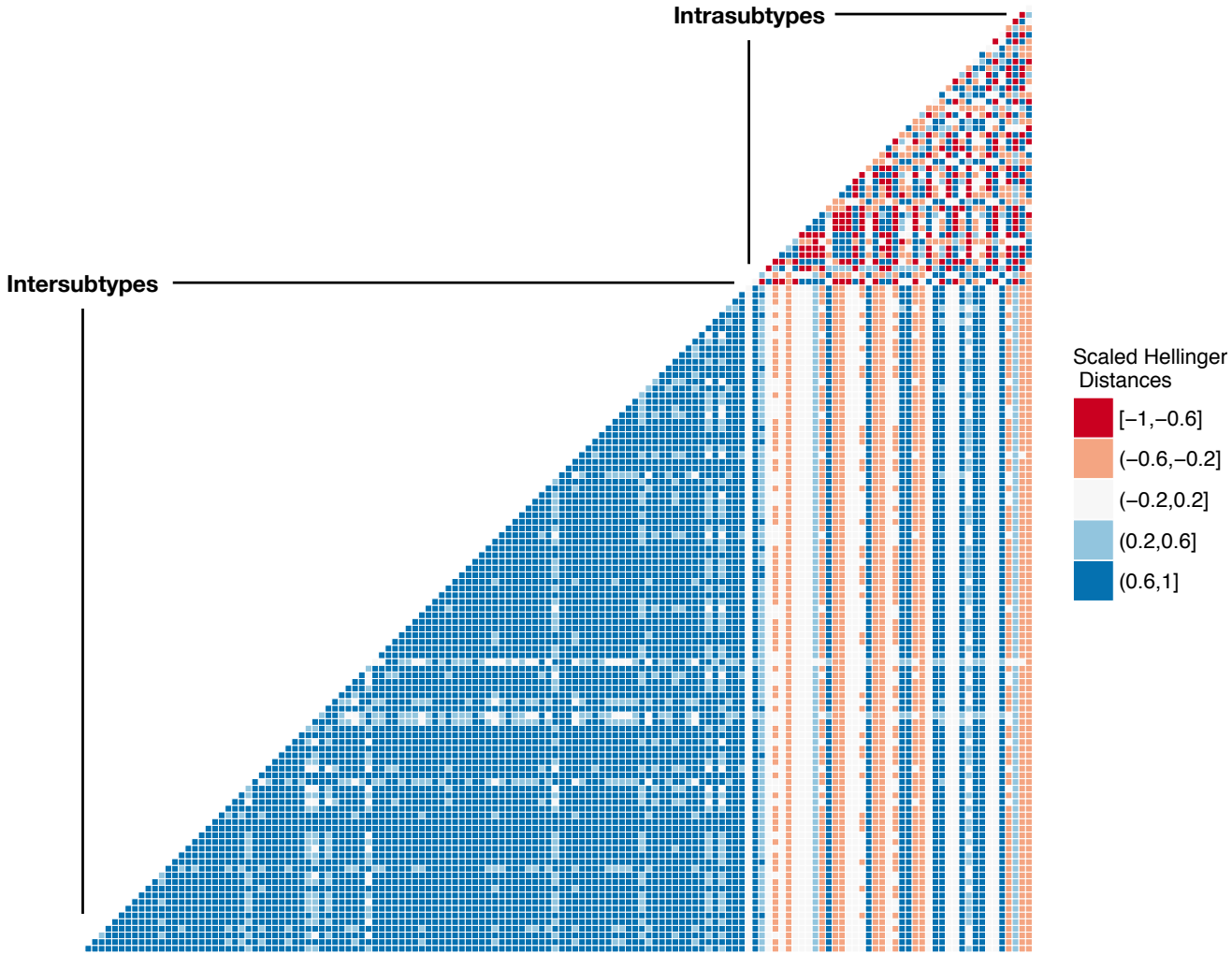
Assume that $T(\mathscr{B}_i)$ and $T(\mathscr{B}_j)$ are probability measures that are absolutely continuous with respect to λ. The *Hellinger distance* is

$$H^2\big(T(\mathscr{B}_i), T(\mathscr{B}_j)\big) = \frac{1}{2}\int\left(\sqrt{\frac{\mathrm{d}\,T(\mathscr{B}_i)}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\,T(\mathscr{B}_j)}{\mathrm{d}\lambda}}\right)^2 \mathrm{d}\lambda$$

For two r.v. $T(\mathscr{B}_i) \sim N(\mu_i, \sigma_i^2)$ and $T(\mathscr{B}_j) \sim N(\mu_j, \sigma_j^2)$, we have:

$$H^2\big(T(\mathscr{B}_i), T(\mathscr{B}_j)\big) = 1 - \sqrt{\frac{2\sigma_i\sigma_j}{\sigma_i^2+\sigma_j^2}}\exp\left\{-\frac{(\mu_i-\mu_j)^2}{4(\sigma_i^2+\sigma_j^2)}\right\}$$

Scaled Hellinger Distances: $\mathbf{H}^* = \mathbf{1}\mathbf{1}^\intercal - \mathbf{H}$



Current/Future Work: Towards Parametric Probability Distributions for Barcodes

Open problem since 2008 (Adler/Taylor, Carlsson, Blumberg et al., Mileyko/Mukherjee/Harer, etc.): Find explicit, parametric probability distributions for barcodes

Challenges: Barcode space is equipped with Alexandrov topology
$\Longrightarrow$ Arbitrarily highly curved; geodesics are not even locally unique

Work in progress (with L. Crawford, S. Kališnik, T. Sudijono):

- Compute inverse
- Bi-Lipschitz?
- Pull back exponential family distributions onto barcode space: Theory + Simulation
- Is "Gaussianity" preserved?

Resources & References

- Fully reproducible research
- Data publicly available from GenBank, the HIV Sequence Database (Los Alamos National Secruity) & NCBI Influenza Virus Database
- Code available at `https://github.com/lorinanthony/Tropix`
- More details can be found in Monod, A., Kališnik Verovšek, S., Patiño-Galindo, J.Á., Crawford, L. (2017). *Tropical Sufficient Statistics for Persistent Homology.* `https://arxiv.org/abs/1709.02647`

Thank You!

Funding: National Institutes of General Medical Sciences at the National Institutes of Health (NIGMS–NIH), Award No. R01GM117591.