Tropical Sufficient Statistics for Persistent Homology

Anthea Monod

Columbia University

am4691@cumc.columbia.edu

Joint work with Sara Kališnik (Wesleyan University),
Juan Ángel Patiño-Galindo (Columbia) & Lorin Crawford (Brown)

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

9 August 2018
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.
**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 $\mu$ 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
  $\implies$ 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

$\implies$ 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 $\vartheta$ if

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

**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 $f(x; \vartheta) = h(x)g(T(x); \vartheta)$*

**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 $\sigma$-algebra $\mathcal{S}$ is that for every $\mu \in \mathcal{M}$, the density $f_\mu := \frac{d\mu}{d\lambda}$ admits the factorization $f_\mu(x) = h(x)g_{T(x)}(T(x))$*
Persistent Homology in 2 Dimensions

- Persistent Homology in 2 Dimensions

- Diagram showing the evolution of homology groups $H_0$ and $H_1$ over time.

- $H_0$: The zeroth homology group captures connected components.

- $H_1$: The first homology group captures cycles that are not boundaries.

- Diagram illustrates the persistence of homological features across different scales or resolutions.

- Anthea Monod (Columbia)

- ICERM Summer Bootcamp

- 9 August 2018
Barcode Space

\[(x_1, d_1, x_2, d_2, \ldots, x_n, d_n); \quad x_i = \text{birth}; \quad d_i = \text{length}; \quad x_i \geq 0\]

\[B_n = \text{Orbit space of the action of the symmetric group } S_n \text{ on } n \text{ letters on the product } ([0, \infty) \times [0, \infty))^{n}, \text{ given by permuting the coordinates}\]

**Definition**

The barcode space \(B_{\leq n}\) consisting of barcodes with at most \(n\) intervals is the quotient

\[\coprod_{n \in \mathbb{N}_{\leq n}} B_n / \sim\]

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

\[\{(x_1, d_1), (x_2, d_2), \ldots, (x_n, d_n)\} \sim \{(x_1, d_1), (x_2, d_2), \ldots, (x_{n-1}, d_{n-1})\}\]

**Regularizing Subsets of Barcode Space:** For fixed \(m > 0\), denote by \(B_{\leq n}^m\) the subset of \(B_{\leq n}\) 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 \(x_i \leq md_i\)
Tropical geometry = "Skeletonized" version of algebraic geometry

**Tropical/Min-plus Semiring:**
\[(\mathbb{R} \cup \{+\infty\}, \oplus, \odot) \quad \text{with} \quad a \oplus b := \min(a, b) \quad \text{and} \quad a \odot b := a + b\]

**Arctic/Max-plus Semiring:**
\[(\mathbb{R} \cup \{-\infty\}, \boxplus, \odot) \quad \text{with} \quad a \boxplus b := \max(a, b) \quad \text{and} \quad a \odot b := a + b\]

- 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:**

\[
p(x_1, x_2, \ldots, x_n) = a_1 \odot x_1^{a_1^1} x_2^{a_1^2} \cdots x_n^{a_1^n} \oplus a_2 \odot x_1^{a_2^1} x_2^{a_2^2} \cdots x_n^{a_2^n} \oplus \cdots \oplus a_m \odot x_1^{a_m^1} x_2^{a_m^2} \cdots x_n^{a_m^n}
\]

Each tropical/arctic polynomial is a continuous piecewise linear function

The passage from tropical/arctic polynomials to functions is not 1-1, e.g.:

\[
x_1^2 \oplus x_2^2 = 2x_1 \oplus 2x_2 = \max\{2x_1, 2x_2\}
\]

\[
x_1^2 \oplus x_2^2 \oplus x_1 x_2 = 2x_1 \oplus 2x_2 \oplus (x_1 + x_2) = \max\{2x_1, 2x_2, x_1 + x_2\} = \max\{2x_1, 2x_2\}
\]
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 $\implies$ 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).


**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.

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

Every matrix $E \in E_n$ 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 $E_i \in E_n/S_n$ under the row permutation action on $E_n$ determine max-plus polynomials by max-plus multiplication over row permutations:

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

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

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

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

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 \(B^m_{\leq n}\) and separates nonequivalent barcodes.
Theorem (Kališnik (2016); M., Kališnik, Patiño-Galindo, Crawford (2017))

The following collection of tropical polynomials

\[ T : \mathcal{B}^m_{\leq n} \to \mathbb{R}^d \]

\[ \mathcal{B} \mapsto (E_{m,(1,1)^i,(0,1)^j} \langle x_1, d_1, \ldots, x_n, d_n \rangle)_{i+j \in \mathbb{N}_{\leq n}}(\mathcal{B}) \]

- induc\es 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 \( \sigma \)-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 $n = 2 \implies$ The set of orbits under the $S_2$ action is

$$\mathcal{E}_2/S_2 = \left\{ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\}$$

We only need a subcollection of all orbits to map barcodes injectively

$\implies$ Take the orbits with rows $(1, 1)$ and $(0, 1)$:

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

We need $d = n + \frac{n(n+1)}{2}$ many orbits
Suppose we have two barcodes $\mathcal{B}_1 = \{(1, 2), (3, 1)\}$ and $\mathcal{B}_2 = \{(2, 2)\}$; $\mathcal{B}_1, \mathcal{B}_2 \in \mathcal{B}_{\leq 2}$

1. Compute $m$: For intervals $(1, 2), (3, 1), (2, 2)$, find the smallest $m$ such that $x_i \leq md_i$. The quotients are $\frac{1}{2}, \frac{3}{1}, 1$, so take $m = 3$, so $\mathcal{B}_1, \mathcal{B}_2 \in \mathcal{B}_{\leq 2}^3$.

2. Determine the $2$-symmetric max-plus polynomials

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

from

$$\begin{bmatrix} (0 & 1) \\ (0 & 0) \end{bmatrix}, \begin{bmatrix} (0 & 1) \\ (0 & 1) \end{bmatrix}, \begin{bmatrix} (0 & 0) \\ (1 & 1) \end{bmatrix}, \begin{bmatrix} (1 & 1) \\ (0 & 1) \end{bmatrix}, \begin{bmatrix} (1 & 1) \\ (1 & 1) \end{bmatrix}.$$ 

$$E_{3,(0,1),(0,0)}(x_1, d_1, x_2, d_2) = d_1 \boxplus d_2$$
$$= \max(d_1, d_2)$$

$$E_{3,(0,1),(0,1)}(x_1, d_1, x_2, d_2) = d_1 d_2$$
$$= d_1 + d_2$$
\[ E_{3,(0,0),(1,1)}(x_1, d_1, x_2, d_2) = (x_2 \oplus d_2^3)d_2 \boxplus (x_1 \oplus d_1^3)d_1 \]
\[ = \max \left\{ \min(x_2, 3d_2) + d_2, \min(x_1, 3d_1) + d_1 \right\} \]

\[ E_{3,(1,1),(0,1)}(x_1, d_1, x_2, d_2) = (x_1 \oplus d_1^3)d_1d_2 \boxplus (x_2 \oplus d_2^3)d_2d_1 \]
\[ = \max \left\{ \min(x_1, 3d_1) + d_1 + d_2, \right. \]
\[ \left. \min(x_2, 3d_2) + d_2 + d_1 \right\} \]

\[ E_{3,(1,1),(1,1)}(x_1, d_1, x_2, d_2) = (x_1 \oplus d_1^3)d_1(x_2 \oplus d_2^3)d_2 \]
\[ = \min(x_1, 3d_1) + d_1 + \min(x_2, 3d_2) + d_2 \]

3. Evaluate on \( \mathcal{B}_1 \):

\[ \max(2, 1) = 2 \]
\[ 2 + 1 = 3 \]
\[ \max \left\{ \min(1, 6) + 2, \min(3, 3) + 1 \right\} = \max\{1 + 2, 3 + 1\} = 4 \]
\[ \max \left\{ \min(1, 6) + 2 + 1, \min(3, 3) + 2 + 1 \right\} = \max\{4, 6\} = 6 \]
\[ \min(1, 6) + 2 + \min(3, 3) + 1 = 7 \]
4. Evaluate on $B_2$:

\[
\begin{align*}
\max(2, 2) &= 2 \\
2 + 0 &= 2 \\
\max \{ \min(2, 6) + 2 \} &= 4 \\
\max \{ \min(2, 6) + 2 \} &= 4 \\
\min(2, 6) + 2 &= 4
\end{align*}
\]

The Euclidean-space vector representation of $B_1$ is $(2, 3, 4, 6, 7)$, and of $B_2$ is $(2, 2, 4, 4, 4)$. 
Linking algebraic topology to evolution.

A

B

C

D

\[
\begin{align*}
\text{Vertical (over time)} & \\
\text{Horizontal (across lineages)} & \\
\end{align*}
\]

\[
\begin{align*}
b_0 &= 1 \\
b_1 &= 0 \\
\end{align*}
\]

\[
\begin{align*}
b_0 &= 1 \\
b_1 &= 1 \\
\end{align*}
\]
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
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

⇒ Lengths of PH₁ 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 $T(B_i)$ and $T(B_j)$ are probability measures that are absolutely continuous with respect to $\lambda$. The *Hellinger distance* is

$$H^2(T(B_i), T(B_j)) = \frac{1}{2} \int \left( \sqrt{\frac{dT(B_i)}{d\lambda}} - \sqrt{\frac{dT(B_j)}{d\lambda}} \right)^2 d\lambda$$

- Probabilistic analog of Euclidean distance between two distributions
- For two r.v. $T(B_i) \sim N(\mu_i, \sigma_i^2)$ and $T(B_j) \sim N(\mu_j, \sigma_j^2)$, we have

$$H^2(T(B_i), T(B_j)) = 1 - \sqrt{\frac{2\sigma_i \sigma_j}{\sigma_i^2 + \sigma_j^2}} \exp \left\{ -\frac{(\mu_i - \mu_j)^2}{4(\sigma_i^2 + \sigma_j^2)} \right\}$$
Scaled Hellinger Distances: $H^* = 11^T - H$
Kullback–Leibler Divergence

Definition

Assume that $T(\mathcal{B}_i)$ and $T(\mathcal{B}_j)$ are probability measures that are absolutely continuous with respect to $\lambda$. The Kullback–Leibler divergence is

$$KLD(T(\mathcal{B}_i), T(\mathcal{B}_j)) = \int \frac{dT(\mathcal{B}_i)}{d\lambda} \log \left( \frac{dT(\mathcal{B}_j)}{d\lambda} \right) d\lambda$$

- Intuitively, KLD measures the relative entropy (or information gain) when comparing inference models.
- For two r.v. $T(\mathcal{B}_i) \sim N(\mu_i, \sigma_i^2)$ and $T(\mathcal{B}_j) \sim N(\mu_j, \sigma_j^2)$, we have

$$KLD(T(\mathcal{B}_i), T(\mathcal{B}_j)) = \log \frac{\sigma_j}{\sigma_i} + \left[ \frac{\sigma_i^2 + (\mu_i - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2} \right]$$
Scaled Kullback–Leibler Distances: $K^* = 11^T - K$
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 an Alexandrov space

⇒ Arbitrarily highly curved; geodesics are not even locally unique
- Fully reproducible research
- Data publicly available from GenBank, the HIV Sequence Database (Los Alamos National Security) & NCBI Influenza Virus Database
- Code available at https://github.com/lorinanthony/Tropix
Thank You!

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

![NIH Logo](image-url)