Principal Observable Analysis

FSU DEPARTMENT OF MATHEMATICS

Ece Karacam   Washington Mio   Osman Berat Okutan

Introduction

Many datasets can be modeled as a metric space (X, d):

- Networks, shapes, images, point clouds.

When data density is also important we equip (X, d) with a Borel probability measure μ, yielding a *metric measure space* (X, d, μ).
This work present a method to study the structure of (X, d, μ) where (X, d) is a compact metric space.

Metric Observables

Probe the geometry of (X, d, μ) using *metric observables* — 1-Lipschitz functions $f : X \to \mathbb{R}$.

Notation:

- $\mathrm{Lip}(X, d)$ = vector space of Lipschitz functions with sup norm $\|f\|_\infty = \sup_{x\in X} |f(x)|$.
- $O(X) \subseteq \mathrm{Lip}(X, d)$ = closed, convex subset of all 1-Lipschitz functions (observables).

Key Objects on $O(X)$

Observable mean: It assings to an observable f its expected value.

$$M_\mu(f) := \mathbb{E}_\mu[f] = \int_X f(x)\, d\mu(x)$$

Observable covariance: For any two μ-centered observables f and g, the observable covariance $\Sigma_\mu : O^c_\mu \times O^c_\mu \to \mathbb{R}$ is given by

$$\Sigma_\mu : (f, g) \mapsto \int_X f(x)\, g(x)\, d\mu(x).$$

Stability Results

Both the observable mean and covariance operators are stable with respect to the Wasserstein 1-distance (w_1).

- **Theorem 5.1 (Mean Stability):** $\|M_\mu - M_\nu\|_\infty \le w_1(\mu, \nu)$
- **Theorem 5.4 (Covariance Stability):** $d_H(\Sigma_\mu, \Sigma_\nu) \le C_X \cdot w_1(\mu, \nu)$ where $C_X = \max\{1, 4D_X\}$.

Stability for Heterogeneous Data. Consider (X, d, μ) and (X', d', μ'). The Kantorovich–Sturm distance, which extends the Wasserstein distance to measures on different domains by optimizing over probabilistic couplings h and metric couplings δ, is

$$d_{KS,p}(\mathcal{M}, \mathcal{M}') = \inf_{h,\delta} \left(\int_{X \times X'} \delta^p(x, y)\, dh(x, y) \right)^{1/p}.$$

- **Theorem 6.3 (Observable Mean Stability):** $d_{GH}(M_{\mu,p}, M_{\mu',p}) \le d_{KS,p}(\mu, \mu')$.
- **Theorem 6.6 (Observable Covariance Stability):** $d_{GH}(\Sigma_{p,\mu}, \Sigma_{p,\mu'}) \le C(X, X')\, d_{KS,p}(\mu, \mu')$, where $C(X, X') = \max\{2,\ 2(D_X + D_{X'})\}$.

Principal Observable Analysis

We define **Principal Observable Analysis (POA)** following a maximization-of-variance principle, analogous to classical PCA.

- **First Principal Observable (ϕ_1):** The centered observable of maximal nonzero variance:

$$\phi_1 := \arg\max_{f \in O^c_\mu} \sigma^2(f) = \arg\max_{f \in O^c_\mu} \Sigma_\mu(f, f).$$

- **Higher Principal Observables (ϕ_n):** For $n \ge 2$, ϕ_n maximizes variance subject to μ-orthogonality to $\phi_1, \ldots, \phi_{n-1}$:

$$\phi_n := \arg\max_{f \in O^c_\mu} \sigma^2(f),$$

$$\text{subject to} \int_X f(x)\, \phi_i(x)\, d\mu(x) = 0, \qquad 1 \le i \le n-1.$$

Implementation. Principal observable computation is a convex maximization problem (the variance $\sigma^2(f)$ is a quadratic form over the convex domain of 1-Lipschitz functions), solved via disciplined convex-concave programming (DCCP). Shen et al. (2016), *Disciplined Convex-Concave Programming.*

Computational Efficiency for Sparse Graphs. Checking $|f(x_i) - f(x_j)| \le d(x_i, x_j)$ for all pairs costs $O(n^2)$. On a graph with shortest-path distance, it suffices to check edge constraints $|f(v_i) - f(v_j)| \le w_{ij}$, reducing complexity to $O(|E|)$.

POA Embedding and Dimension Reduction

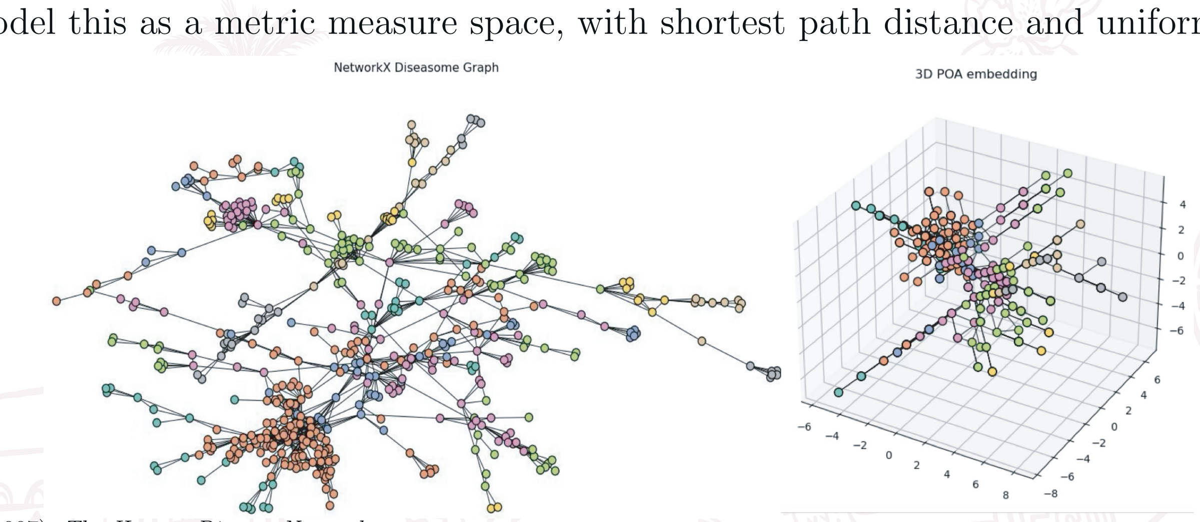
Let $\phi_1, \ldots, \phi_k$ be the first k principal observables. The **k-dimensional POA embedding** $\iota_k : X \to \mathbb{R}^k$ is defined by

$$\iota_k(x) = (\phi_1(x), \ldots, \phi_k(x)).$$

- ι_k provides vectorization and dimension reduction.
- Equip $\mathbb{R}^k$ with L_∞ distance to ensure ι_k is 1-Lipschitz.

Diseasome Network Application. This figure shows POA applied to a diseasome network:

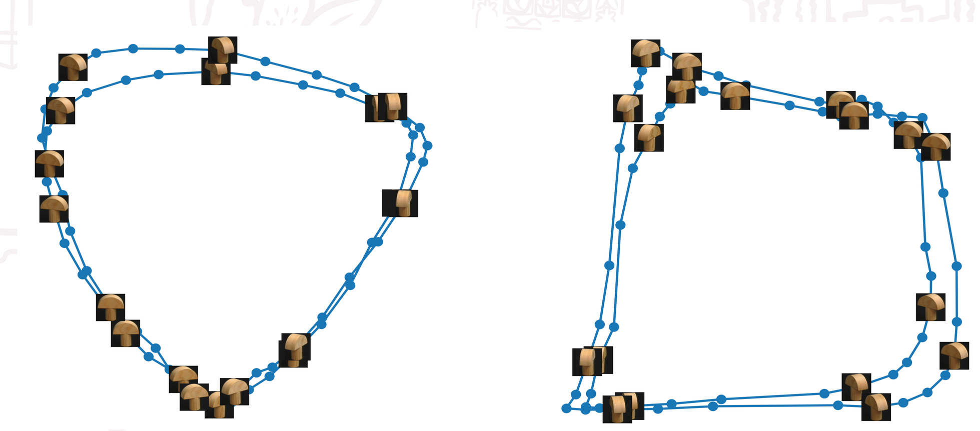
- $N = 516$ nodes representing diseases,
- Edges representing shared genetic origins or biological relationships.
- Model this as a metric measure space, with shortest path distance and uniform measure.



Goh et al. (2007), *The Human Disease Network.*

Rotating Object Images. PCA and POA applied to 72 grayscale images of a rotating T-shaped object (COIL–100).

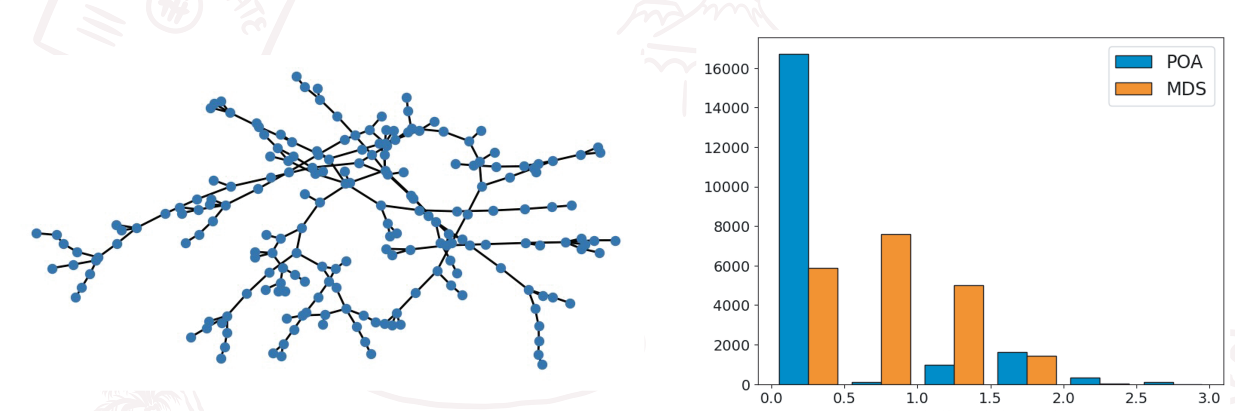
- Each 128×128 image is converted into a vector in $\mathbb{R}^{16384}$.
- POA representation is calculated on a finite metric space comprising 72 data points, using pairwise distances.
- Both methods recover the loop structure.



Nene et al. (1996), *COIL-100.*

Distortion Comparison. A 200-node tree is embedded into $\mathbb{R}^3$ using POA (ι_1) and MDS (ι_2).

- For POA, the tree is treated as a metric measure space (X, d, μ), where d is the shortest-path distance and μ is the normalized counting measure.
- Distortion for POA: $\delta_1(v, w) = \big| d(v, w) - \|\iota_1(v) - \iota_1(w)\|_\infty \big|$.
- Distortion for MDS : $\delta_2(v, w) = \big| d(v, w) - \|\iota_2(v) - \iota_2(w)\|_2 \big|$.
- As shown in the histogram, POA has a much larger number of pairs for which the distortion is small.



Representation of Signals in Observable Domain

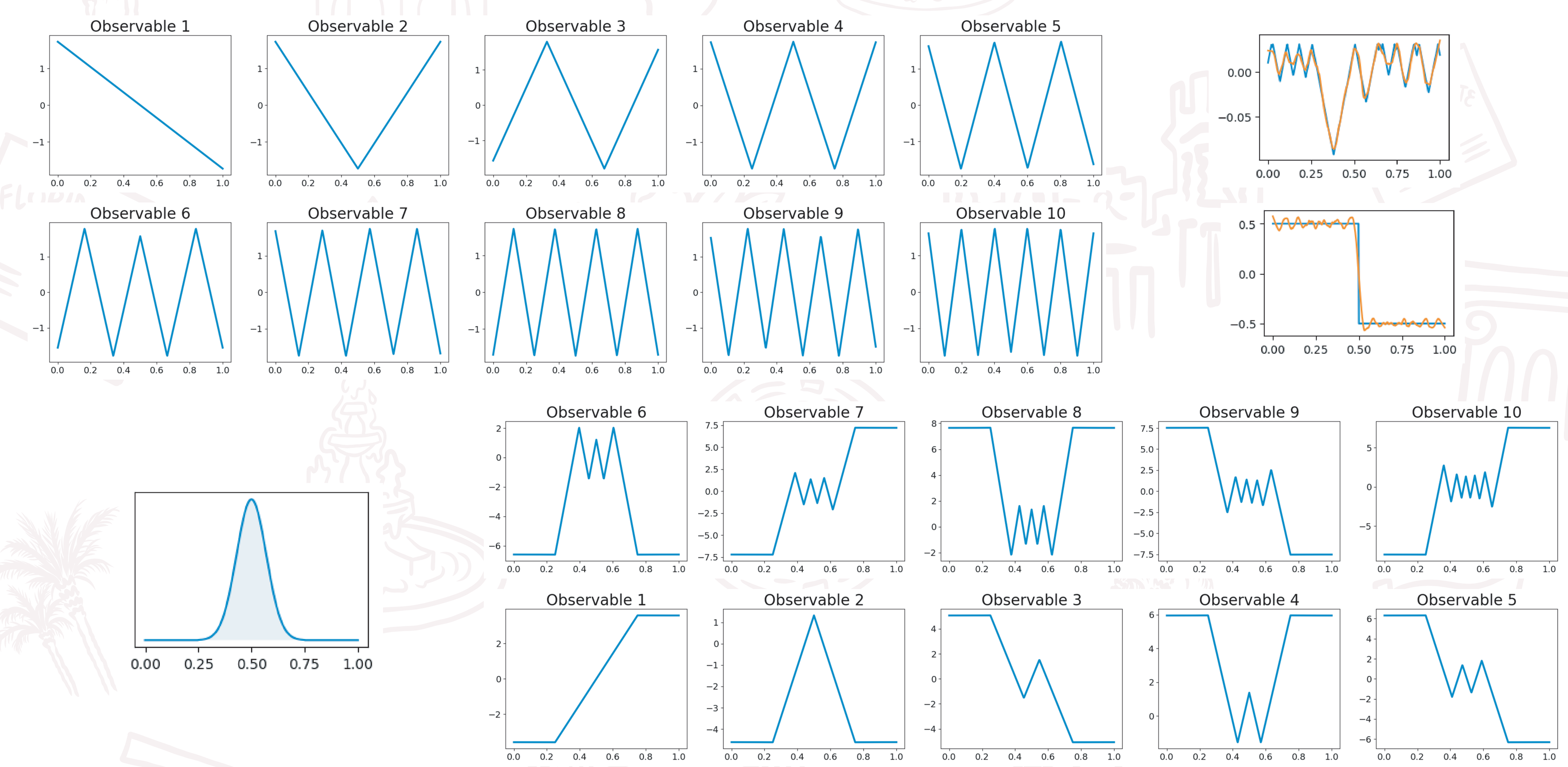
Let $\phi_1, \ldots, \phi_k$ be the first k-POs and set $\phi_0 \equiv 1$.

- Define **Observable Domain** as the span of the μ-orthonormal family of normalized principal observables: $u_i := \frac{\phi_i}{\|\phi_i\|_{2,\mu}}$ for $i = 0, \ldots, k$, where $\|\cdot\|_{2,\mu}$ denotes the $\mathbb{L}_2$-norm with respect to μ.
- Represent signal $f : X \to \mathbb{R}$ in observable domain by $f \approx \sum_{i=0}^k a_i\, u_i$ where the observable-domain coefficients are

$$a_i := \int_X f(x)\, u_i(x)\, d\mu(x), \qquad i = 0, \ldots, k.$$

Line Graph Example. Consider a line graph L with 501 nodes forming a uniform grid on $[0, 1]$ (edges of length $1/500$).

- Using the first 25 normalized observables $\{u_i\}$, we reconstruct signals by $f \approx \sum_{i=1}^{25} a_i u_i$,
- The reconstruction (orange) closely matches the original signals (blue).
- For a localized bell-shaped distribution, the principal observables display similar oscillatory patterns but restricted to the support of the measure.



Reeb Graph Application. Using the first principal observable, we construct a Reeb graph representation for the Western U.S. high-voltage power grid. The Watts–Strogatz (1998) network is modeled as an unweighted, undirected graph X with:

- $N = 4941$ nodes (generators, substations, transformers),
- $E = 6594$ edges (transmission lines).

Equip X with the shortest-path distance and node-degree weights for the measure.

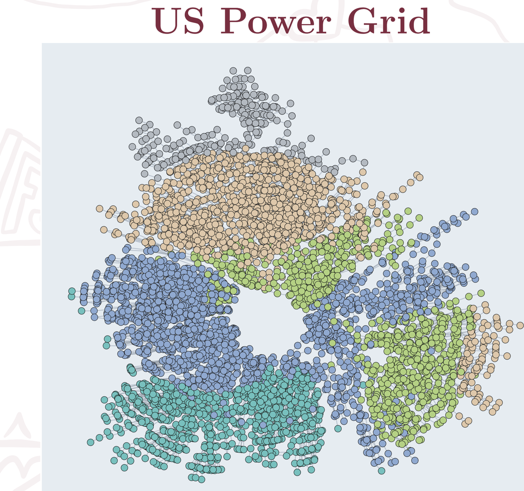
- Compute the first principal observable $f : X \to \mathbb{R}$.
- Bin the values of f into B intervals to define the piecewise-constant function $\bar{f}$.
- Construct the Reeb graph as the quotient $X/\sim$, where for $x, y \in X$,

$$x \sim y \iff \bar{f}(x) = \bar{f}(y) = c \text{ and } x, y \text{ lie in the same connected component of } \bar{f}^{-1}(c).$$

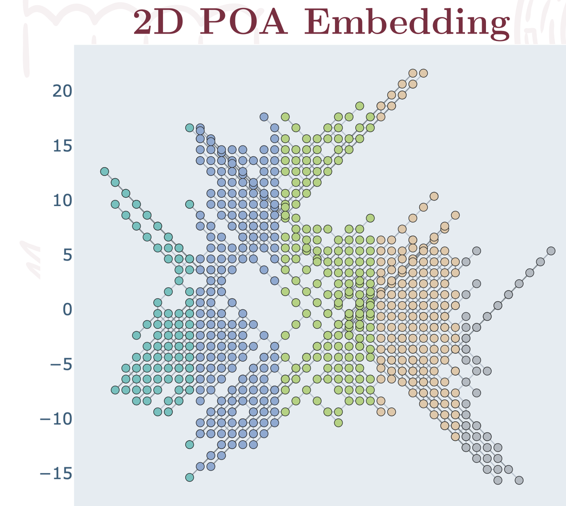
Figures below show:

- **Top-left:** The power grid network X. Nodes with the same color share the same binned PO1 value.
- **Bottom-left:** The 2D POA embedding (ϕ_1, ϕ_2) of X, colored by $\bar{f}$. Here, $B = 5$.
- **Right:** The Reeb graph of X.

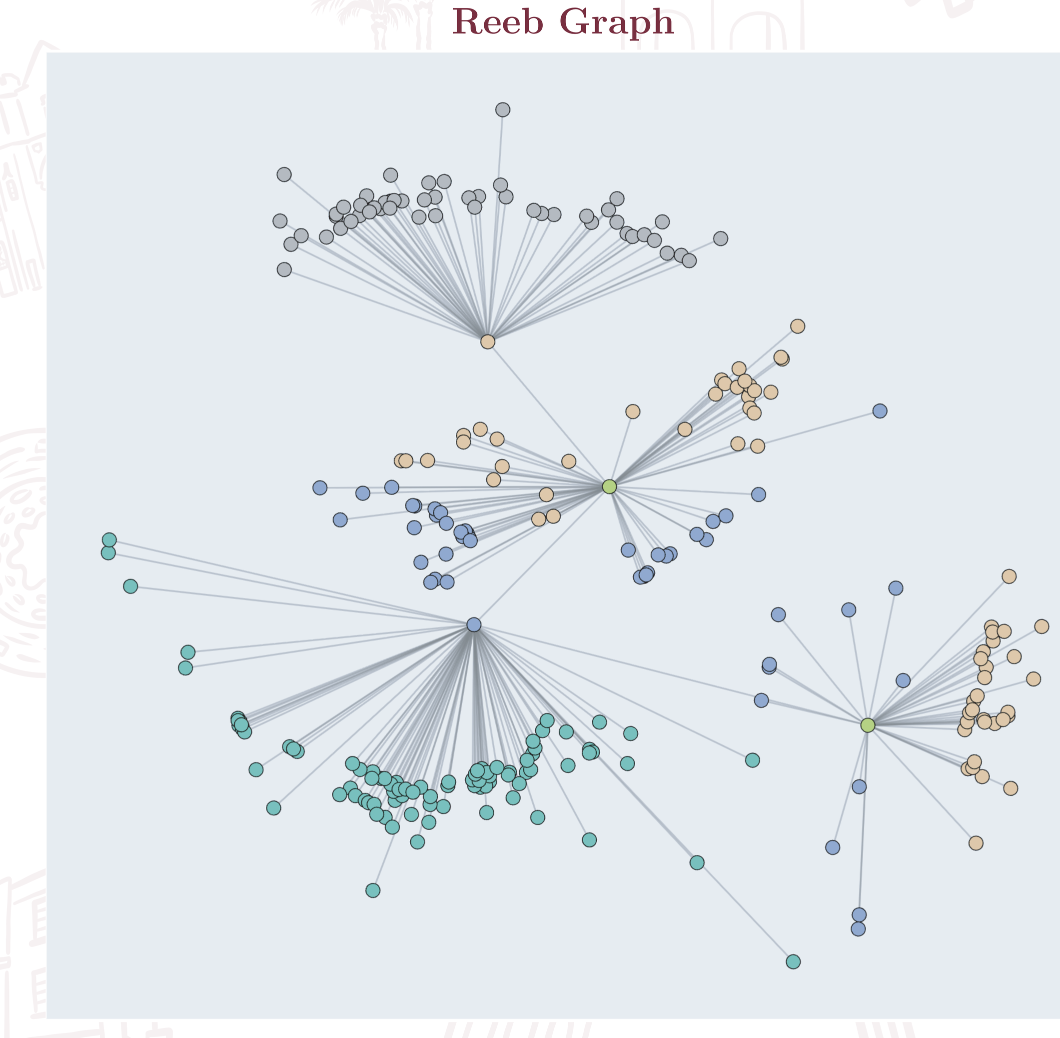
US Power Grid



2D POA Embedding



Reeb Graph



Watts & Strogatz (1998), *US Power Grid Network.*