Measure of shape for object data

J. Virta

University of Turku

ICS and Related Methods
Helsinki, 26.-28. May 2026

Reference

This talk is based on the following article:

- **Virta, J.**. *Measure of shape for object data*, Journal of Nonparametric Statistics, 1-21, 2025.

These slides are available at the speaker's website
`https://users.utu.fi/jomivi/talks/`

Table of Contents

Object data

Setting

- A sample of data $X_1, \ldots X_n$ in some metric space $(\mathcal{X}, d)$.
- The only thing we are allowed to compute are distances:

$$d(X_i, X_j) = \text{“how different } X_i \text{ and } X_j \text{ are”}$$

Descriptive statistics

- Let $X_1, X_2, \ldots$ be i.i.d. random objects from a distribution P_X.

Descriptive statistics

- Let $X_1, X_2, \ldots$ be i.i.d. random objects from a distribution P_X.

Frechét mean (location)

$$\mu(P_X) := \operatorname{argmin}_\mu \mathrm{E}\{d^2(X_1, \mu)\}$$

Descriptive statistics

- Let $X_1, X_2, \ldots$ be i.i.d. random objects from a distribution P_X.

Frechét mean (location)

$$\mu(P_X) := \operatorname{argmin}_\mu \mathrm{E}\{d^2(X_1, \mu)\}$$

Frechét variance (spread)

$$\sigma^2(P_X) := \frac{1}{2}\mathrm{E}\{d^2(X_1, X_2)\}$$

Descriptive statistics

- Let $X_1, X_2, \ldots$ be i.i.d. random objects from a distribution P_X.

Frechét mean (location)

$$\mu(P_X) := \operatorname{argmin}_\mu \mathrm{E}\{d^2(X_1, \mu)\}$$

Frechét variance (spread)

$$\sigma^2(P_X) := \frac{1}{2}\mathrm{E}\{d^2(X_1, X_2)\}$$

Distance variance (spread)

$$\theta^2(P_X) := \mathrm{E}\{d^2(X_1, X_2)\} - 2\mathrm{E}\{d(X_1, X_2)d(X_1, X_3)\} + [\mathrm{E}\{d(X_1, X_2)\}]^2$$

Beyond location and spread

Metric skewness?

Beyond location and spread

Metric skewness?

Metric shape!

Table of Contents

Main concept

Metric shape

$$S(P_X) = \frac{\mathrm{E}\{d(X_1, X_2)^2 d(X_1, X_3)^2\}}{\mathrm{E}\{d(X_1, X_2)^4\}},$$

- If $(\mathcal{X}, d)$ is the Euclidean space $\mathbb{R}^p$, then $S(P_X) = S(P_{aX+b})$ for any $a \in \mathbb{R}$ and $b \in \mathbb{R}^p$.

$\Rightarrow$ measures shape!

Sample version

- Easy to estimate in practice with complexity $\mathcal{O}(n^2)$ as

$$S(P_n) = \frac{\frac{1}{n^3}\sum_{i=1}^{n}\{\sum_{j=1}^{n} d(X_i, X_j)^2\}^2}{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d(X_i, X_j)^4}.$$

Main result

Property I

$$S(P_X) \in \left[\frac{1}{2}, 1\right].$$

Main result

Property I

$$S(P_X) \in \left[\tfrac{1}{2}, 1\right].$$

Property II

$$S(P_X) = \tfrac{1}{2} \quad \Leftrightarrow \quad P(X_1, X_2, X_3 \text{ reside on a } d\text{-line}) = 1.$$

Main result

Property I

$$S(P_X) \in \left[\tfrac{1}{2}, 1\right].$$

Property II

$$S(P_X) = \tfrac{1}{2} \quad \Leftrightarrow \quad P(X_1, X_2, X_3 \text{ reside on a } d\text{-line}) = 1.$$

"Distribution looks like a line"

Main result

Property I

$$S(P_X) \in \left[\tfrac{1}{2}, 1\right].$$

Property II

$$S(P_X) = \tfrac{1}{2} \quad \Leftrightarrow \quad P(X_1, X_2, X_3 \text{ reside on a } d\text{-line}) = 1.$$

"Distribution looks like a line"

Property III

$$S(P_X) = 1 \quad \Leftrightarrow \quad d(X_1, X_2) = \text{constant, a.s.}$$

Main result

Property I

$$S(P_X) \in \left[\tfrac{1}{2}, 1\right].$$

Property II

$$S(P_X) = \tfrac{1}{2} \quad \Leftrightarrow \quad P(X_1, X_2, X_3 \text{ reside on a } d\text{-line}) = 1.$$

"Distribution looks like a line"

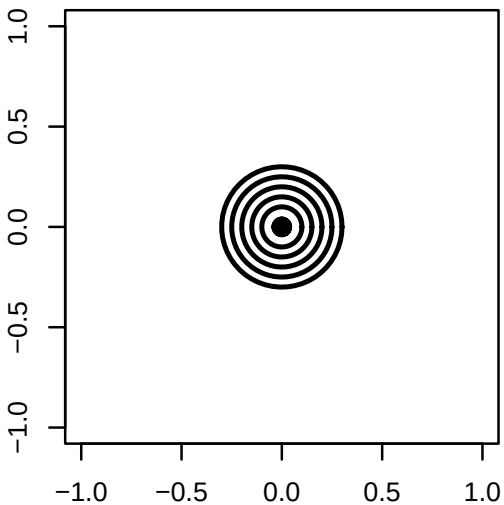
Property III

$$S(P_X) = 1 \quad \Leftrightarrow \quad d(X_1, X_2) = \text{constant, a.s.}$$

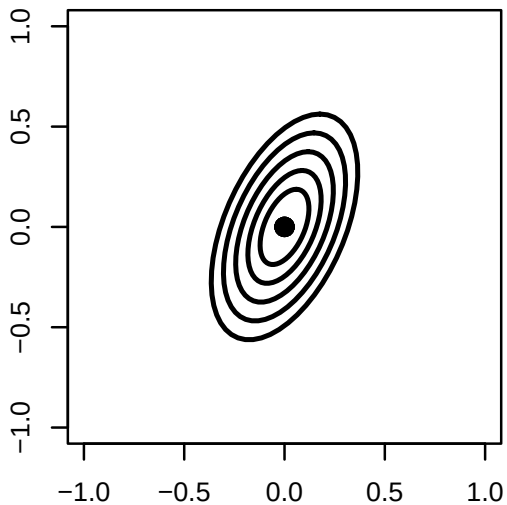
"Distribution is extremely uniform in every direction"

Multivariate data

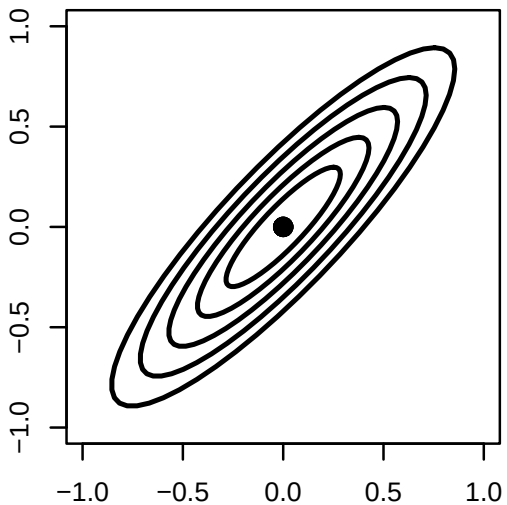
- Space $= (\mathbb{R}^2, \mathrm{d}_{\mathrm{Euc}})$
- Line $=$ Line in the usual sense



$S(P_X) = 0.600$



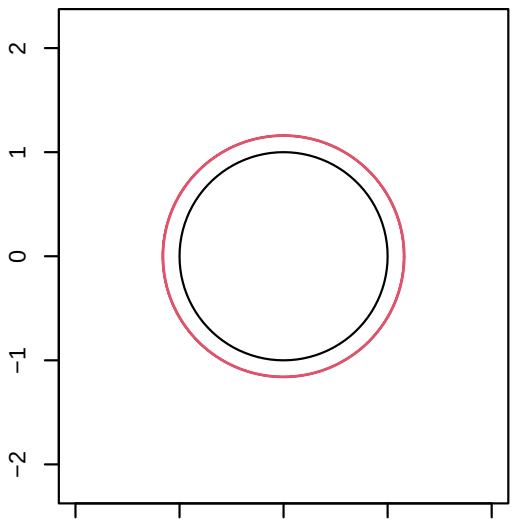
$S(P_X) = 0.554$



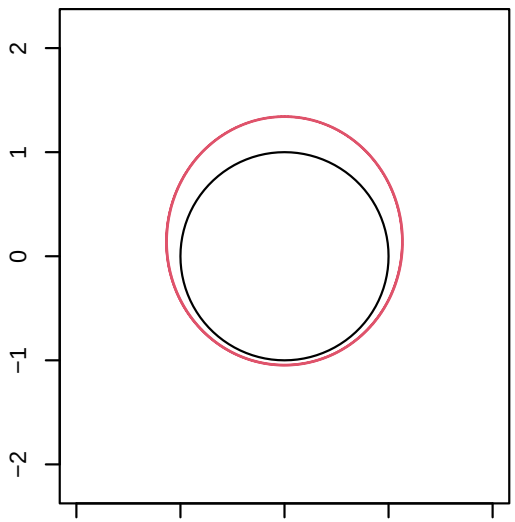
$S(P_X) = 0.512.$

Circular data

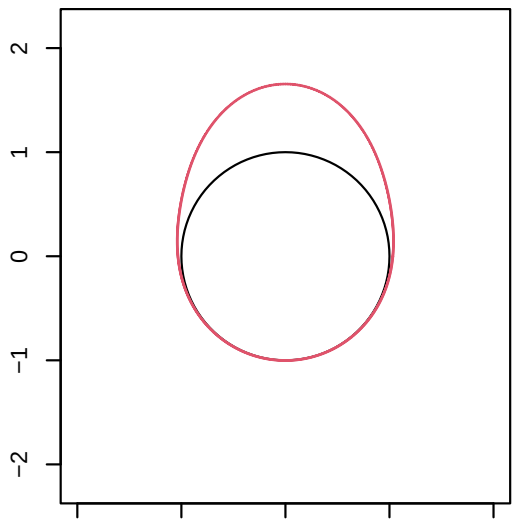
- Space $= (\mathcal{S}^1, \mathrm{d}_{\mathrm{arc}})$
- Line $=$ Semicircle



$S(P_X) = 0.555$



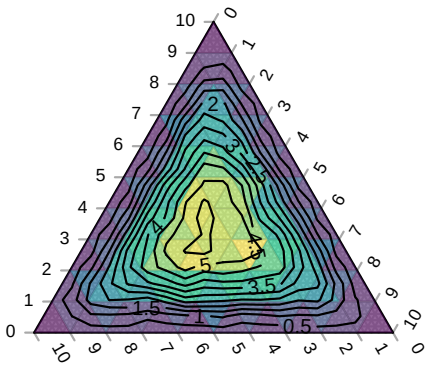
$S(P_X) = 0.543$



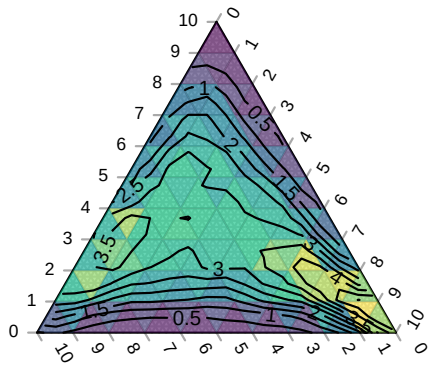
$S(P_X) = 0.508.$

Compositional data

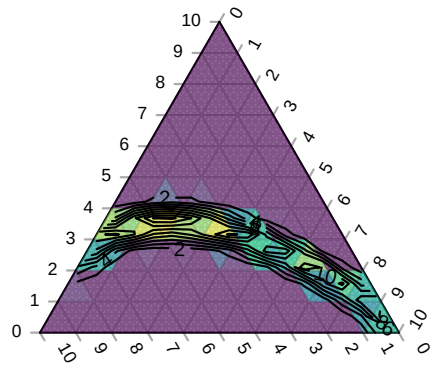
- Space $= (\Delta^3, \mathrm{d}_{\mathrm{Aitchison}})$
- Line $=$ Aitchison geodesic



$S(P_X) = 0.625$



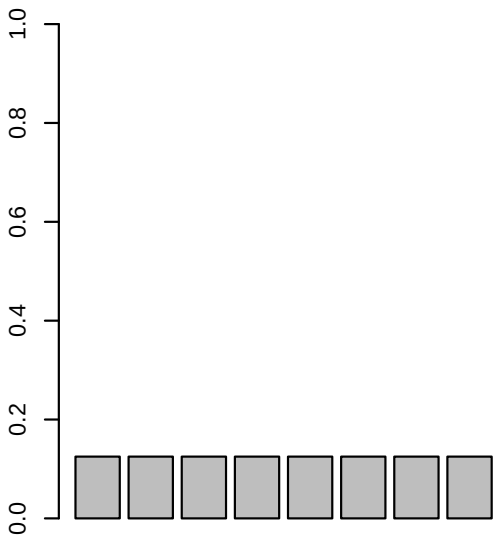
$S(P_X) = 0.584$



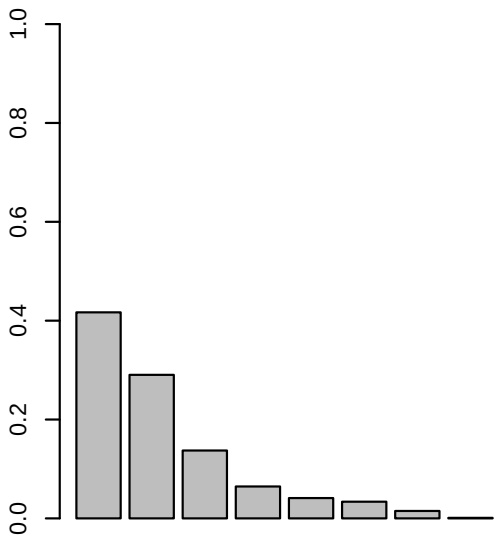
$S(P_X) = 0.502.$

Discrete data

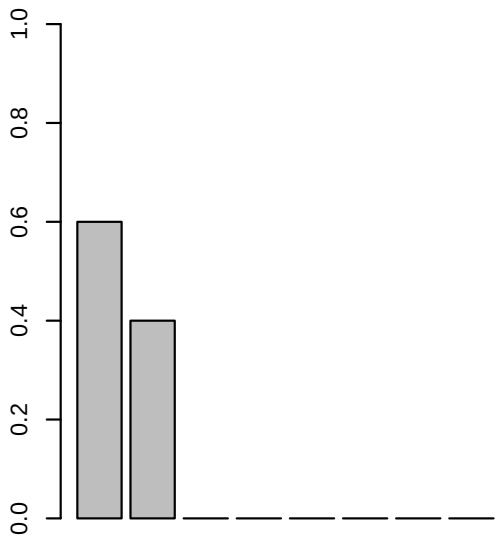
- Space $= (\{1, \ldots, K\}, \mathrm{d}_{\mathrm{discrete}})$
- Line $=$ Pair of points



$S(P_X) = 0.875$



$S(P_X) = 0.743$



$S(P_X) = 0.500.$

Summary

- $S(P_X)$ acts as a universal concept of "shape".
  - Large values = distribution fills the full space in a symmetric manner.
  - Small values = distribution is concentrated on a "line".

Summary

- $S(P_X)$ acts as a universal concept of "shape".
  - Large values = distribution fills the full space in a symmetric manner.
  - Small values = distribution is concentrated on a "line".
- Unifies several existing concepts of uniformity/symmetry.

Table of Contents

Minimization peeling

Greedy algorithm

1. Start with a sample of data $X_1, \ldots, X_n$.
2. Discard the observation i such that $S(P_{-i})$ is minimized.
3. Iteratively repeat step 2.

Minimization peeling

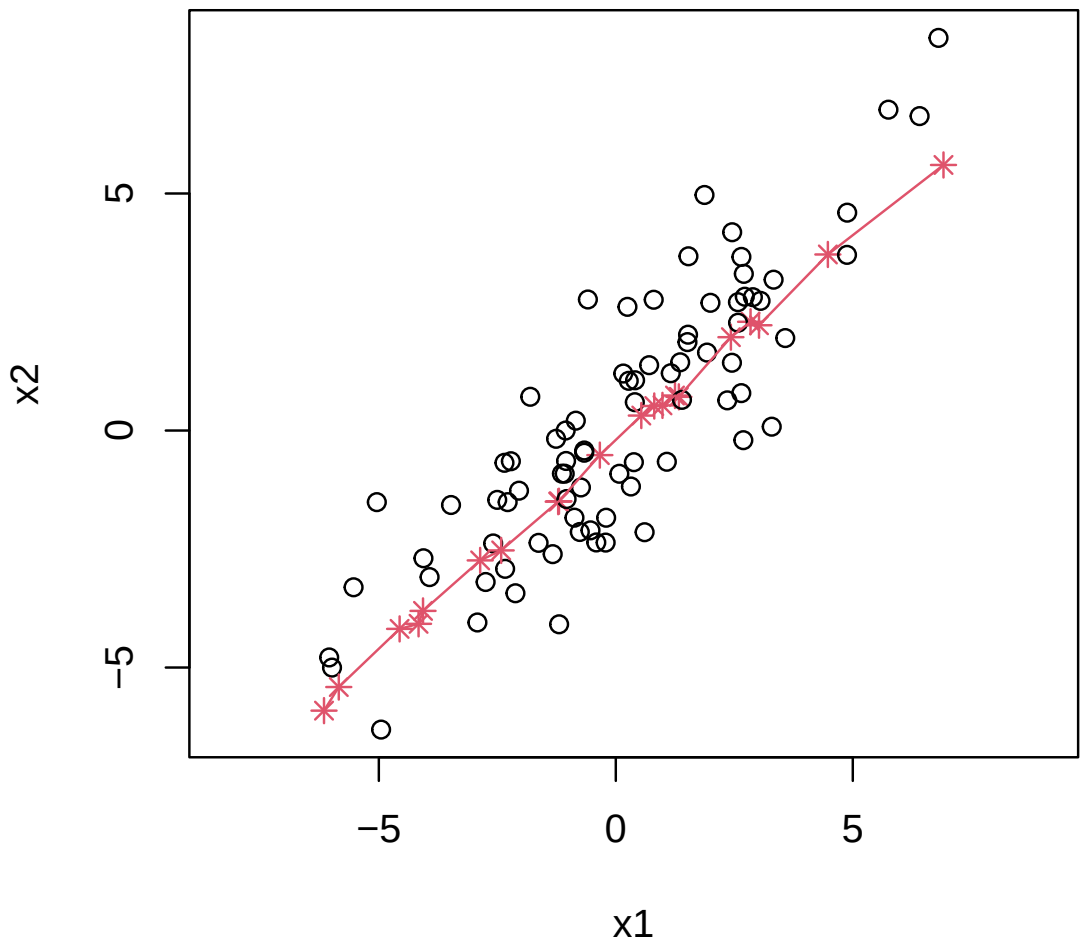
Greedy algorithm

1. Start with a sample of data $X_1, \ldots, X_n$.
2. Discard the observation i such that $S(P_{-i})$ is minimized.
3. Iteratively repeat step 2.

Finds the first "principal component"!

Euclidean example

- Sample of size $n = 100$ is peeled to 20 points.

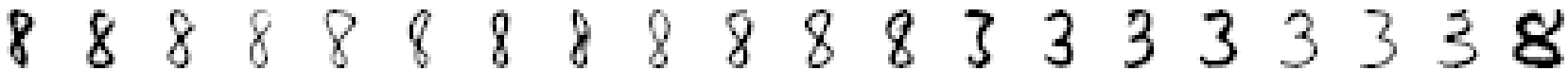


Peeling, image data

- A sample of $n = 100$ images of hand-drawn digits three and eight, sized 16×16.
- Space $= (\mathbb{R}^{256}, \ell_1)$

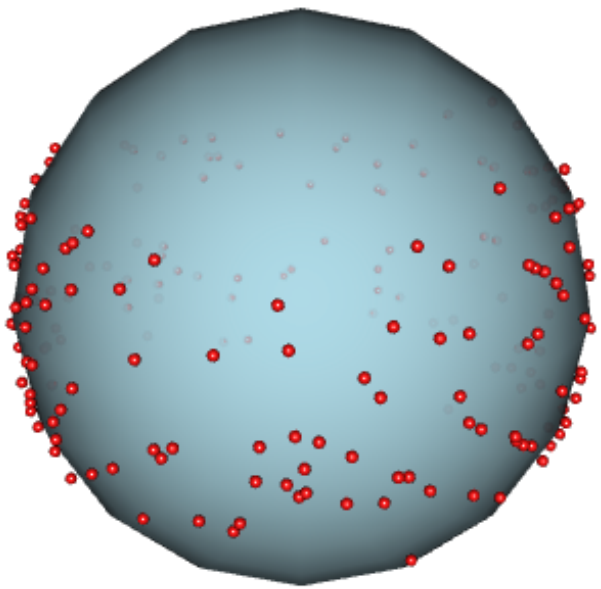
Peeling, image data

- We peel to 20 images:



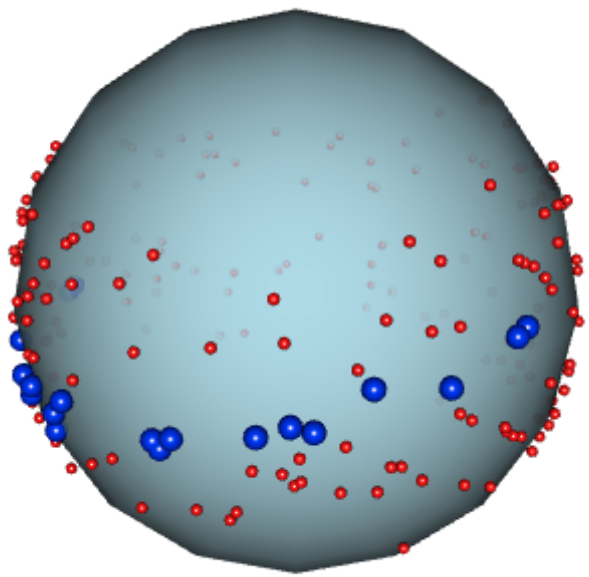
Peeling, spherical data

- Locations of $n = 200$ sunspots on the surface of the sun.
- Space $= (\mathcal{S}^2, \mathrm{d}_{\mathrm{arc}})$



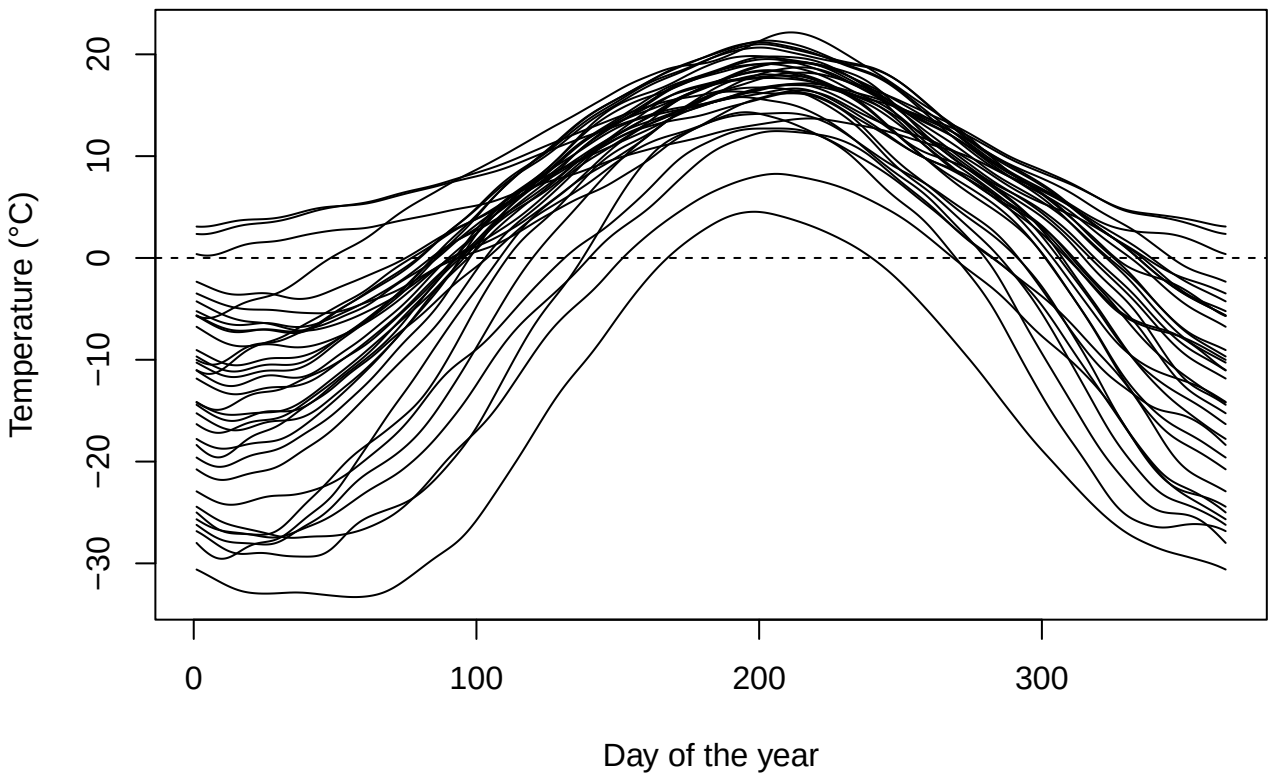
Peeling, spherical data

- We peel to 20 sunspots:



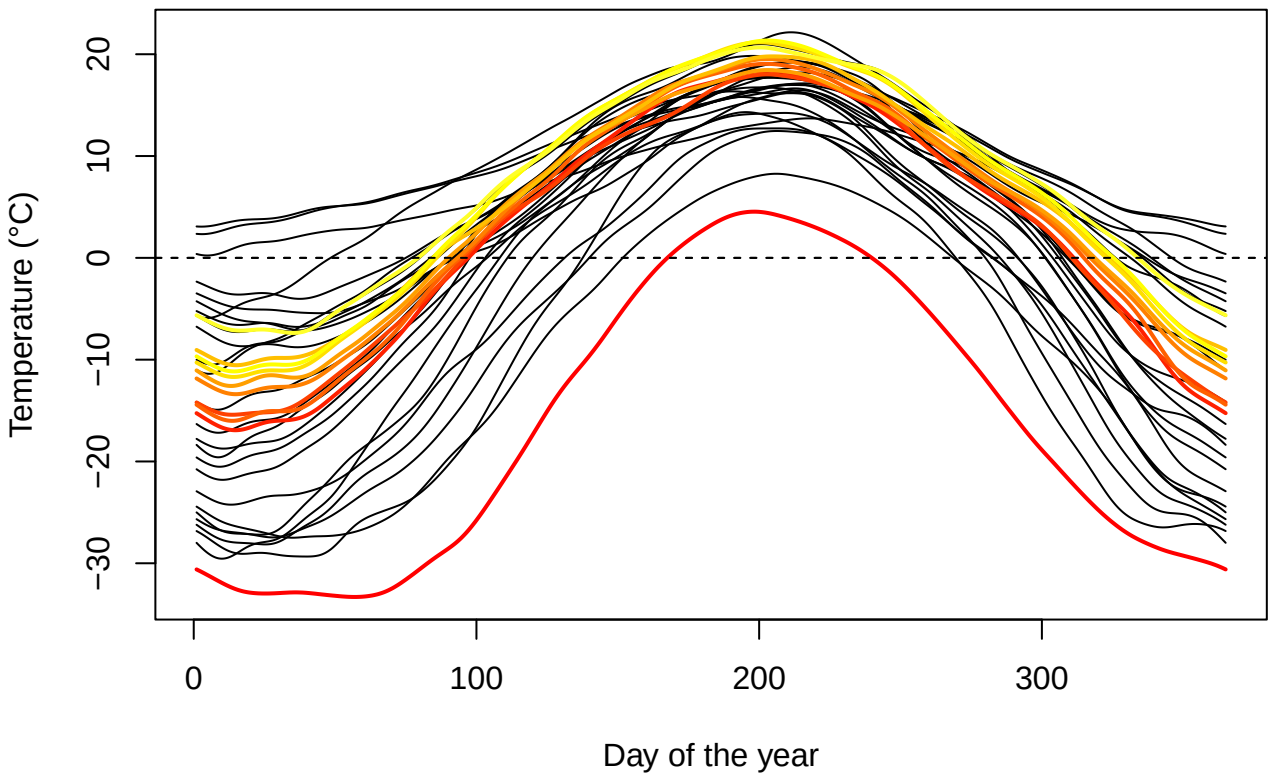
Peeling, functional data

- Yearly temperature curves of $n = 35$ Canadian locations.
- Space $= L_2([0, 365])$



Peeling, functional data

- We peel to 10 locations:



Summary

- Same procedure for finding PC1 applies in every metric space!
- There is no general approach ("loadings") for interpreting the found component.
- High computational cost of order $\mathcal{O}(n^4)$!
- Ordering the representatives is a traveling salesman problem.

Table of Contents

Future research topics

- Developing general tests for "uniformity" based on $S(P_X)$.
  - How much efficiency do we lose compared to parametric tests?
- Extracting further principal components.

Thank you for your attention!