Functional Data Analysis of Populations of Tree-structured Objects

Haonan Wang
Department of Statistics

December 11, 2001

Shape

- interesting
- useful

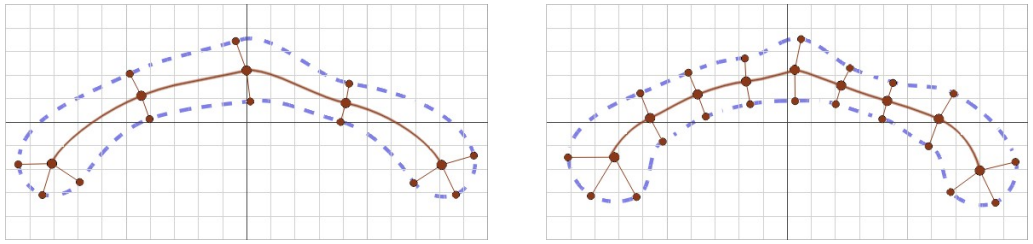
An example of one member of a population of shapes of interest (from Paul Yushkevich).

- bendings at the two ends
- one bump in the middle

M-rep — developed by S. M. Pizer.

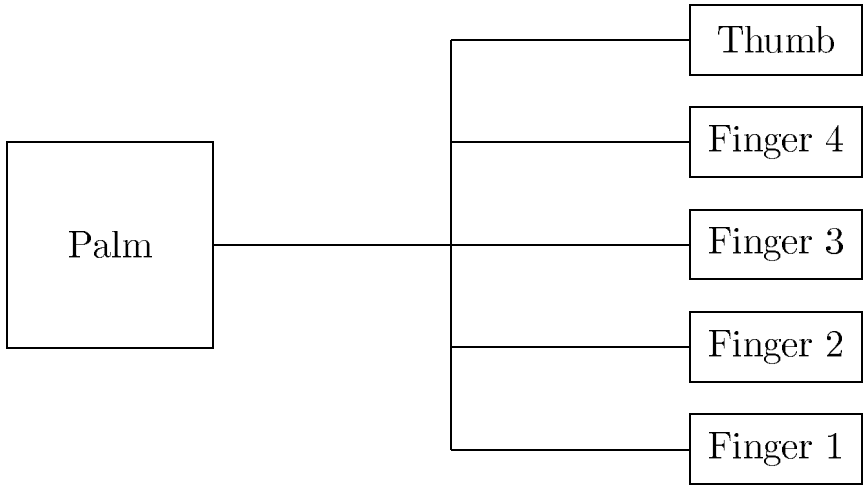
- convenient
- powerful

Coarse scale M-rep and fine scale M-rep



Statistical analysis of multifigural objects

Eg: hand



- Structures of the shapes are all the same —restrictive assumption.

Shape space — Euclidean space

- The feature vectors are not all the same — complicated.

Problem: Understand "structure" of population of multi-figural objects

- Center Point???
- Variation ???
- Analog of PCA ???

Finding a data structure

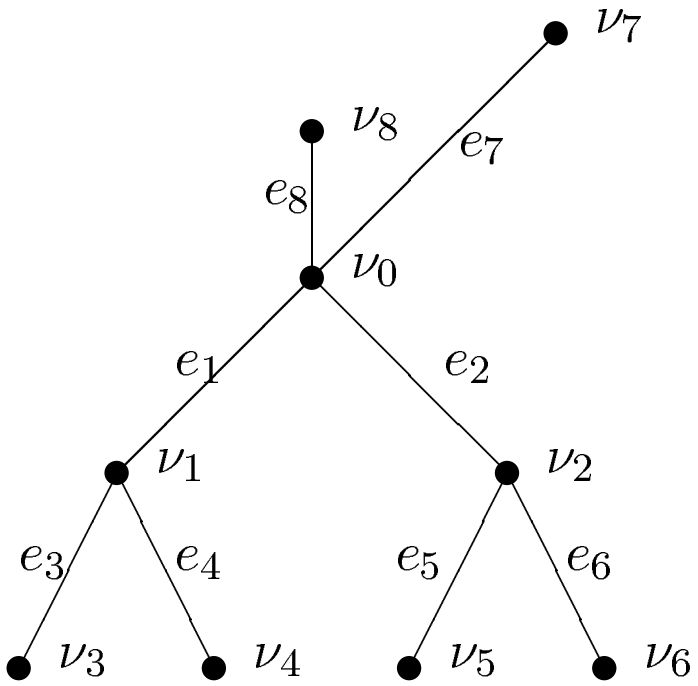
- trees

- figures at each node

tree

- collection of nodes and edges
- unique path (a set of edges) between every pair
- root — one designated node
- level of a node — the length (number of edges) of the path to the root

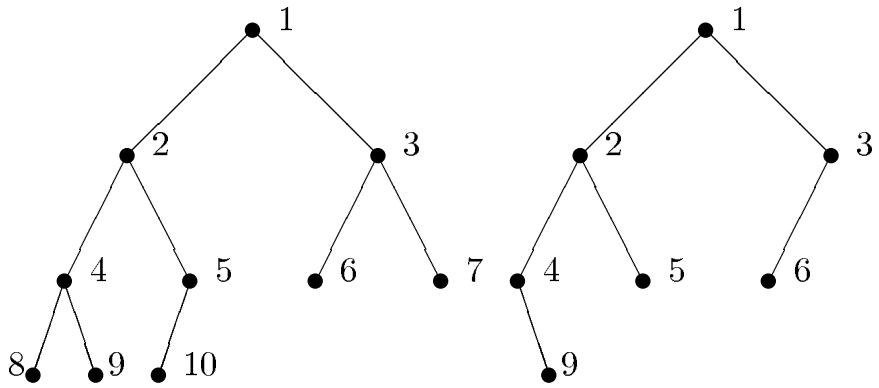
An example of tree t



The tree t has 9 nodes and 8 edges. Let ν_0 be the root of tree t. Note that $\{\nu_1, \nu_2, \nu_7, \nu_8\}$ have level 1, and $\{\nu_3, \nu_4, \nu_5, \nu_6\}$ have level 2. Thus, the level of the tree t is 2.

A special case — Binary tree

Each node has at most two children.



level-order index: lable of the node shown above

binary tree

- Each node has at most 2 children;
- left child and right child

union and intersection of the binary trees

- union or intersection of the level-order index sets

"Information": feature vector for a node;

Eg: collection of M-rep parameters

Q: How to deal with the information?

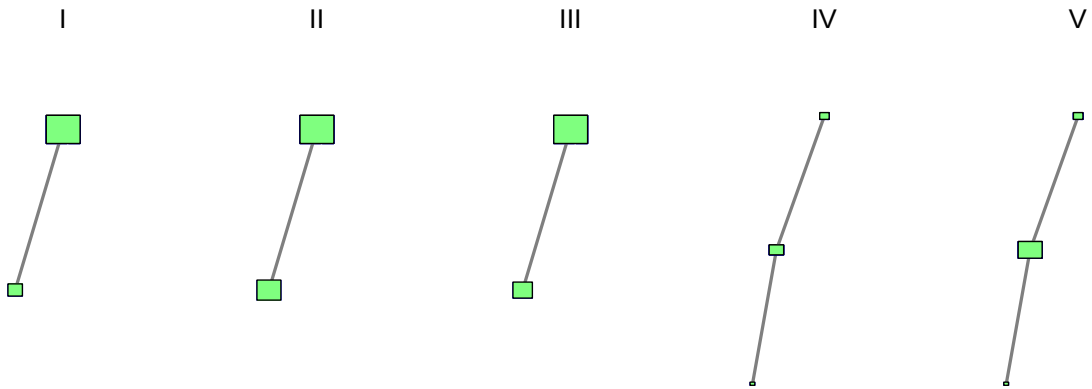
Simplifying assumptions (for toy examples):

1. Only two values for each node;

2. Each feature bounded by 0 and $\frac{\sqrt{2}}{2}$.

A toy sample of binary trees:

Let T be a sample of binary trees with sample size $n = 5m$. There are five types of trees and each has m elements:



Nodal information of the five representatives:

level-order index	I	II	III	IV	V
1	(0.7,0.7)	(0.7,0.7)	(0.7,0.7)	(0.2,0.2)	(0.2,0.2)
2	(0.3,0.3)	(0.5,0.5)	(0.4,0.4)	(0.3,0.3)	(0.5,0.5)
4	n/a	n/a	n/a	(0.1,0.1)	(0.1,0.1)

Where is the "center point"?

Characterization of mean of $x_1, x_2, \ldots, x_n$:

$$\text{argmin} \sum_{i=1}^{n} (x - x_i)^2$$

I.e., minimum of sum of squared **distances** to all the data points

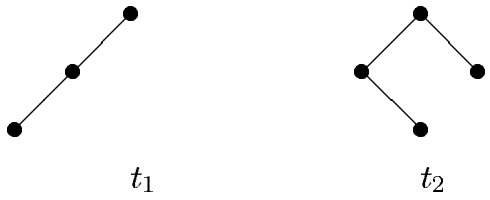
Motivates finding metric on tree space

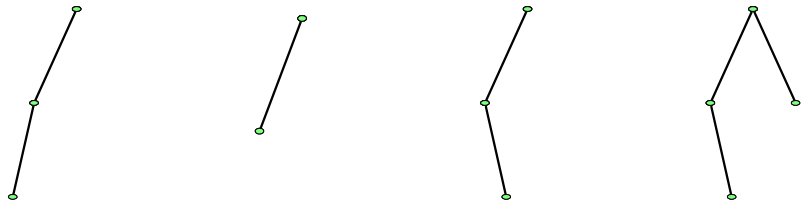
Finding a suitable metric — structure and information

- Integer part metric — tree structures

- Fractional part metric — nodal information

Motivation of Integer part metric





The smallest number of addition and deletion of nodes from one tree to the other is 3.

Define the integer part metric d_I:

$$d_I = \#(Ind(t_1) \triangle Ind(t_2))$$

In the previous example,

$$Ind(t_1) = \{1, 2, 4\} \; Ind(t_2) = \{1, 2, 3, 5\}$$

$$Ind(t_1) \triangle Ind(t_2) = \{3, 4, 5\}$$

Therefore,

$$d_I(t_1, t_2) = 3$$

Given "weights" $\{\alpha_k\}$, fractional part metric f_δ

$$f_\delta(s,t) = [\sum_{k \in Ind(s) \cap Ind(t)} \alpha_k((x_{sk} - x_{tk})^2 + (y_{sk} - y_{tk})^2)$$
$$+ \sum_{k \in Ind(s) \setminus Ind(t)} \alpha_k((x_{sk} - 0)^2 + (y_{sk} - 0)^2)$$
$$+ \sum_{k \in Ind(t) \setminus Ind(s)} \alpha_k((0 - x_{tk})^2 + (0 - y_{tk})^2)]^{\frac{1}{2}}$$

where $\alpha_k > 0$ and $\sum_k \alpha_k = 1$.

Comment:

1. "f" means "fractional", i.e., $f_\delta \leq 1$.

2. "f"— weighted Euclidean distance.

3. "f"— ordinary Euclidean distance between weighted vectors.

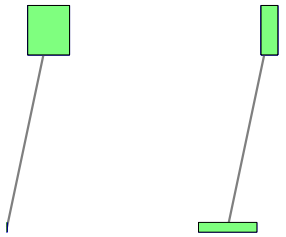
4. Some often used weight sequence:

   - Uniform weights — finite level trees, ρ;

   - power weights — $\frac{1}{2^{(2i+1)}}$ for a node on the i^{th} level.

Example: t_1 and t_2 are two trees with nodal information listed below.

level-order index	t_1	t_2
1	(0.5,0.5)	(0.2,0.5)
2	(0,0.1)	(0.7,0.1)



$$f_\delta(t_1,t_2) = \sqrt{\underbrace{\frac{1}{2}((0.5-0.2)^2+(0.5-0.5)^2)}_{k=1} + \underbrace{\frac{1}{2^3}((0-0.7)^2+(0.1-0.1)^2)}_{k=2}}$$
$$= 0.3260$$

Metric in the binary tree space — δ

$$\delta(t_1, t_2) = d_I(t_1, t_2) + f_\delta(t_1, t_2)$$

In the previous example,

$$\begin{aligned}\delta(t_1, t_2) &= d_I(t_1, t_2) + f_\delta(t_1, t_2) \\ &= 0 + 0.3260 = 0.3260\end{aligned}$$

Finding "center point" of a sample of trees — Median-mean binary tree

Recall that, in classical statistics, two often-used measures of "center" of a group of data $\{x_1, x_2, \ldots, x_n\}$:

- Sample Mean — average $(\frac{\sum x_i}{n})$

- Sample Median — middle point in the ordered data set

Q: In the binary tree space, where is the center point?

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a sample of trees.

First, let's define the variation function

$$V_\delta(t_1, t_2) = d_I(t_1, t_2) + f_\delta^2(t_1, t_2)$$

For a "center" tree m, The total variation is

$$\sum_{i=1}^{n} V_\delta(t_i, m).$$

minimizer of the sum above —**center point**.

Case 1. Without nodal information

In this case, the total variation is

$$\sum_{i=1}^{n} d_I(t_i, m).$$

Characterization of the minimizer is the **majority rule**.

Majority Rule

$$m = A \cup B$$

where
A= {all nodes appearing more than $\frac{n}{2}$ times}

$B \subset$ {nodes that appear exactly $\frac{n}{2}$ times}

Note: This is why we called it "median".

Case 2. With nodal information.

median-mean tree

- median — metric d_I
- mean — metric f_δ.

A tree is called a **median-mean tree** for a sample T, denoted by m_δ, if it minimizes

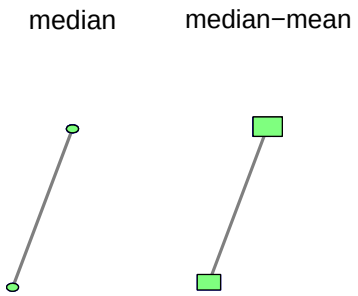
$$\sum_{i=1}^{n} d_I(t, t_i) \tag{1}$$

over all trees $t \in T$ and has nodal information

$$x_{m_\delta k} = \frac{\sum_{i=1}^{n} x_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}} \tag{2}$$

$$y_{m_\delta k} = \frac{\sum_{i=1}^{n} y_{t_i k} 1\{k \in Ind(t_i)\}}{\sum_{i=1}^{n} 1\{k \in Ind(t_i)\}} \tag{3}$$

Back to the example on page 13



The nodal information of the median-mean tree is

level-order index	m_δ
1	(0.5,0.5)
2	(0.4,0.4)
4	n/a

Principal Component Analysis

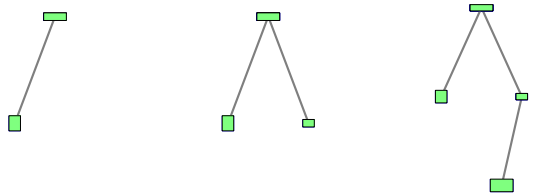
PCA on Euclidean space — linear space

- decompose the total variation
- line of greatest variability — one-dimensional subspace

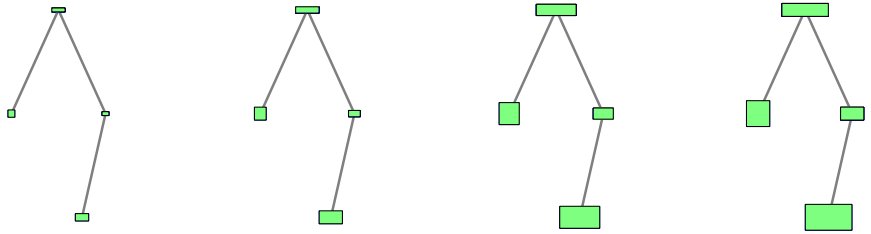
Binary tree space — nonlinear

- Analog of "subspace" of $dim = 1$, i.e., "line"???
- Decomposition of variation??? Orthogonality???

Structure treeline (s-treeline):



Information treeline (i-treeline):



structure treeline (s-treeline)

- nested tree sequence, one node
- nodes with the same level-order index have same nodal information

information treeline (i-treeline)

- same tree structure
- feature vectors — one-dimensional subspace in Euclidean space

In the previous example,

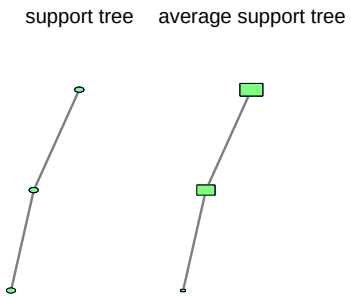
$$\text{feature vector} = \vec{0} + \lambda\vec{v}$$

$\lambda = 1.0, 1.75, 3.0, 3.5$ and
$\vec{v} = [0.2, 0.1, 0.1, 0.2, 0.1, 0.1, 0.2, 0.2]'$.

average support tree of a sample T,

- set of all nodes which appear in the sample
- nodal information is average of corresponding nodal information

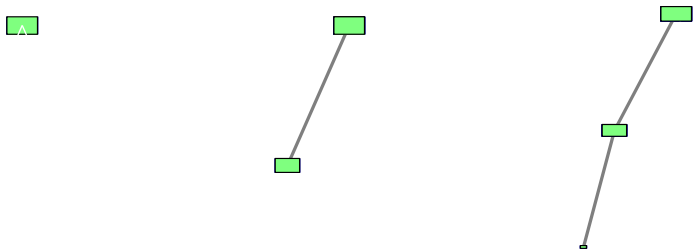
Example of average support of the tree sample on page 13:



Comment:

1. Each member of s-treeline is a subtree of the average support tree.

The unique s-treeline for the sample on Page 13:



Treeline — one dimensional representation of tree data

Projection — The closest element on the treeline

Q: Is the projection unique?

A: yes!

Q: Orthogonality?

A: No!

Pythagorean Theorem — fundamental theorem to variation decomposition

Tree version Pythagorean Theorem

Part I:
Let l be an i-treeline passing through a tree u in the tree space $\mathcal{T}$. Then, for any $t \in \mathcal{T}$,

$$V_\rho(t,u) = V_\rho(t,P_l(t)) + V_\rho(P_l(t),u) \tag{4}$$

Part II:

Let $T = \{t_1, t_2, \ldots, t_n\}$ be a sample of finite level trees. Let l be an s-treeline where every element is a subtree of the average support tree of T. Then, for any $u \in l$,

$$\sum_{i=1}^{n} V_\rho(t_i, u) = \sum_{i=1}^{n} V_\rho(t, P_l(t_i)) + \sum_{i=1}^{n} V_\rho(P_l(t_i), u) \qquad (5)$$

Comment:

For each single tree in the sample, Part II may be false!!!

Principal Component Analysis on tree space:

- Finding principal one-dimensional structure representation;
- Finding principal components on nodal information vectors.

one-dimensional principal structure representation:

- An s-treeline that minimizes the sum

$$\sum_{i=1}^{n} V_\rho(t_i, P_l(t_i)) \tag{6}$$

over all binary s-treelines l passing through the minimal median-mean tree μ_ρ in the sample T.

Remark:

ρ is a special version of the metric δ with equal weights (α_k) on finite level tree space.

Back to the example on page 13:

The total variation is:

$$\sum_{i=1}^{5m} V_\rho(t_i, m_\rho) = 2.18m \tag{7}$$

$$\sum_{i=1}^{5m} V_\rho(P_l(t_i), m_\rho) = 2.01m$$

and

$$\sum_{i=1}^{5m} V_\rho(P_l(t_i), t_i) = 0.17m$$

Proportion of variation explained by the one-dimensional structure representation is

$$\frac{\sum_{i=1}^{5m} V_\rho(P_l(t_i), m_\rho)}{\sum_{i=1}^{5m} V_\rho(t_i, m_\rho)} = \frac{2.01}{2.18} = 92.2\%.$$

Decompose the variability in the nodal information:

Two principal componet:

$$\vec{p_1} = [1, 1, 0, 0, 0, 0, 0, 0]'$$

and

$$\vec{p_2} = [0, 0, 1, 1, 0, 0, 0, 0]'.$$

First one explains $0.15m$ and second one explains $0.02m$.