Dissimilarities Approximation Issues

D. Fortin INRIA-Rocquencourt

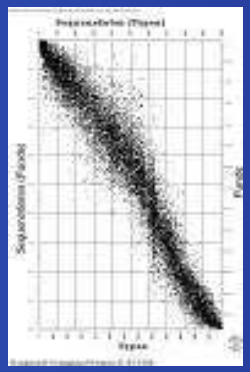
P. Préa LIF-Marseille

Plan

- Dissimilarities $D : d_{ij} \geq 0, d_{ii} = 0$
- Problem Taxonomy
  - Seriation
  - Phylogeny
  - Clustering
- Bottleneck Properties
- Approximations

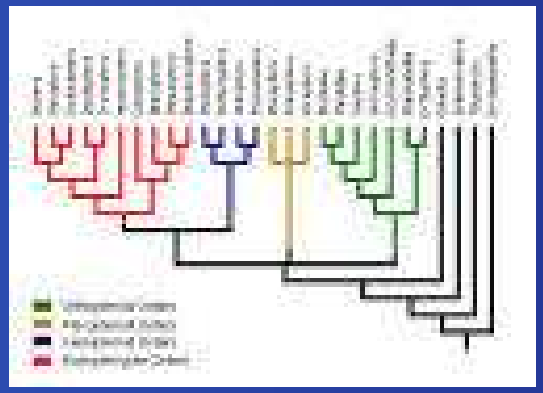
Seriation

- $D \in \mathbf{R}^{n \times n} \longmapsto$ hamiltonian path
- assumption *innovation* propagates along shortest path
-



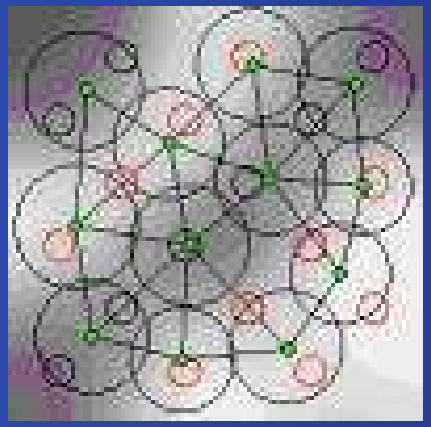
Phylogeny

- $D \in \mathbf{R}^{n \times n} \longmapsto$ tree metrics
- hereditary relationship
-

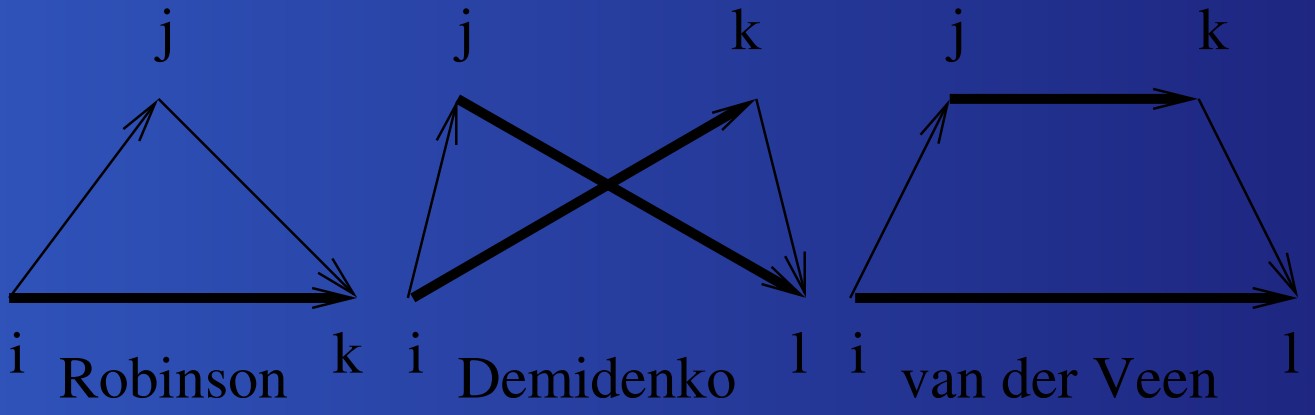


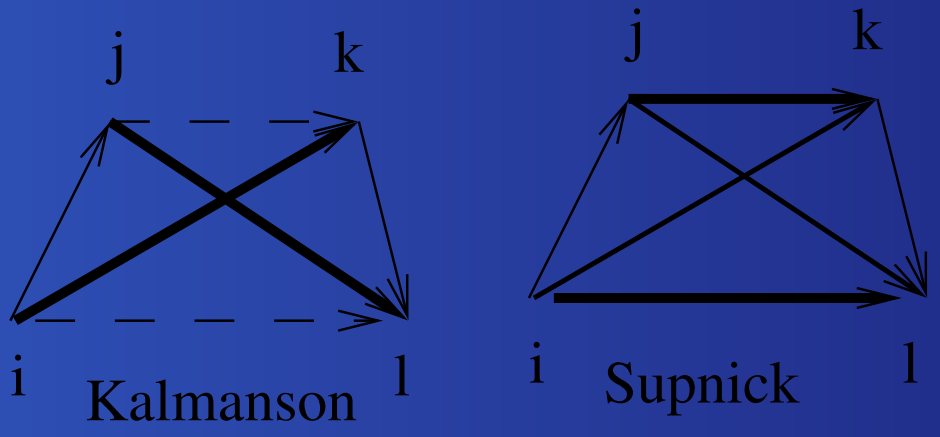
Clustering

- $D \in \mathbf{R}^{n \times n} \longmapsto$ graph
- classes are dense (cliques)
- classes are separable (stables)
-



Bottleneck Properties





Easy Recognition

- ultrametrics $d_{ik} \leq \max(d_{ij}, d_{jk})$ $o(n^2 \log n)$
- treemetrics (phylogeny)
  $d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}$ $o(n^2 \log n)$
- Robinson (seriation, clustering)
  $d_{ik} \geq \max(d_{ij}, d_{jk}), i < j < k$ $o(n^2 \log n)$

Characterization $i < j < k < l$

- Kalmanson (sum case: master tour)

$$d_{ik} \oplus d_{jl} \geq \begin{cases} d_{ij} \oplus d_{kl} \\ d_{il} \oplus d_{jk} \end{cases}$$

- Supnick (sum case: pyramidal tour)

$$d_{ij} \oplus d_{kl} \leq d_{ik} \oplus d_{jl} \leq d_{il} \oplus d_{jk}$$

$$(1, 3, 5 \ldots n, n-1, \ldots 6, 4, 2)$$

Conic Approximations

- definitions

$$\begin{aligned}
\text{vec}(.) \ : \ & \mathbf{R}^{n\times n} \longmapsto \mathbf{R}^{n^2} \text{(matrix vectorization)} \\
\text{Mat}(.) = \ & \text{vec}(.)^{-1} \text{(vector matrixization)} \\
\langle X, Y\rangle = \ & \langle \text{vec}(X), \text{vec}(Y)\rangle \text{(inner product)} \\
\mathcal{K}^0_{\hat{X}} = \ & \{Y \mid \langle Y, X - \hat{X}\rangle \leq 0 \text{ for all } X \in \mathcal{K}\} \text{(normal cone)} \\
\mathcal{K}^\circ = \ & \{Y, y \in \mathbf{R} \mid \langle X, Y\rangle \leq y \text{ for all } X \in \mathcal{K}\} \text{(polar cone)} \\
& \mathcal{K} \text{ cone} \Rightarrow \mathcal{K}^\circ = \mathcal{K}^0_0
\end{aligned}$$

Dissimilarity Modeling

- $\min \max \mapsto \| \ . \ \|_\infty$
- $\min \sum \mapsto \| \ . \ \|_1$
- subdominant $\min(.,.,\ldots,.)$ multicriterion

$$\min \max \mathsf{vec}(X_1), \min \langle E, X_1 \rangle, \min \mathsf{vec}(X_1)$$
$$\text{s.t.} \quad X_1 + X_2 = D$$
$$X_1 \in \overline{\mathcal{N}}, X_2 \in \overline{\mathcal{K}} = \overline{\mathcal{N}} \cap \mathcal{K}$$

- $\mathsf{vec}([X_1, X_2]) = \begin{bmatrix} \mathsf{vec}(X_1) \\ \mathsf{vec}(X_2) \end{bmatrix}$

Dissimilarity Approximations

- primal

$$\begin{aligned} \min \quad & \langle [E, O], X \rangle \\ \text{s.t.} \quad & [I_{n^2}, I_{n^2}]\text{vec}(X) = D \\ & X \in \overline{\mathcal{N}} \times \overline{\mathcal{K}} \end{aligned}$$

- dual

$$\begin{aligned} \max \quad & \langle D, Y_1 \rangle \\ \text{s.t.} \quad & Y_1 + Y_2 = E \\ & Y_2 \in (\overline{\mathcal{N}} \times \overline{\mathcal{K}})^0 \end{aligned}$$

- strong duality (primal=dual) iff $\text{int}(\overline{\mathcal{N}} \times \overline{\mathcal{K}})^0 \neq \emptyset$

Clustering Modeling

- $x_{ic} = 1$ if $i \in c, 0$ otherwise
- intracluster $\min \sum_{i,j,c} d_{ij} x_{ic} x_{jc}$
- intercluster $\max \sum_{i,j,c} d_{ij} x_{ic} (1 - x_{jc})$
- overlapping $1 \leq \sum_{c} x_{ic} \leq k$
  - partition $k = 1$
  - hierarchy $k = 2 \quad \supset$ (D is Robinson)

Clustering Approximation

- $(\min\langle DX, X\rangle, \max\langle DX, E - X\rangle)$
- $e \leq Xe \leq ke, X$ $0-1$ matrix

$$\begin{cases} \min & \langle DX, Y\rangle \\ & e \leq Ye \leq ke \\ & Y = X \end{cases} \qquad \begin{cases} \min & \langle -DE, X\rangle \\ & e \leq Xe \leq ke \end{cases}$$

- lagrangian decomposition $\simeq$ bimatrix game

Intracluster Approximation schemes

- k-Clustering $\min \sum_c \sum_{ij \in c} d_{ij}$
- k-Median $\min \sum_c \sum_{i \in c} d_{ir_c}$ to representative r_c
- k-Center $\min_c \max_{i \in c} d_{ir_c}$ to representative r_c
- metrics (facility location $\longrightarrow$ Guha)
- $\mathbf{R}^d : d(x,y) = \| x - y \|^2$ ($\longrightarrow$ Kenyon)