



Learning from (dis)similarity data

Nathalie Vialaneix

► To cite this version:

Nathalie Vialaneix. Learning from (dis)similarity data. European R Users Meeting (eRum 2018), May 2019, Budapest, Hungary. 70 p. hal-02785273

HAL Id: hal-02785273

<https://hal.inrae.fr/hal-02785273>

Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Learning from (dis)similarity data

Nathalie Villa-Vialaneix

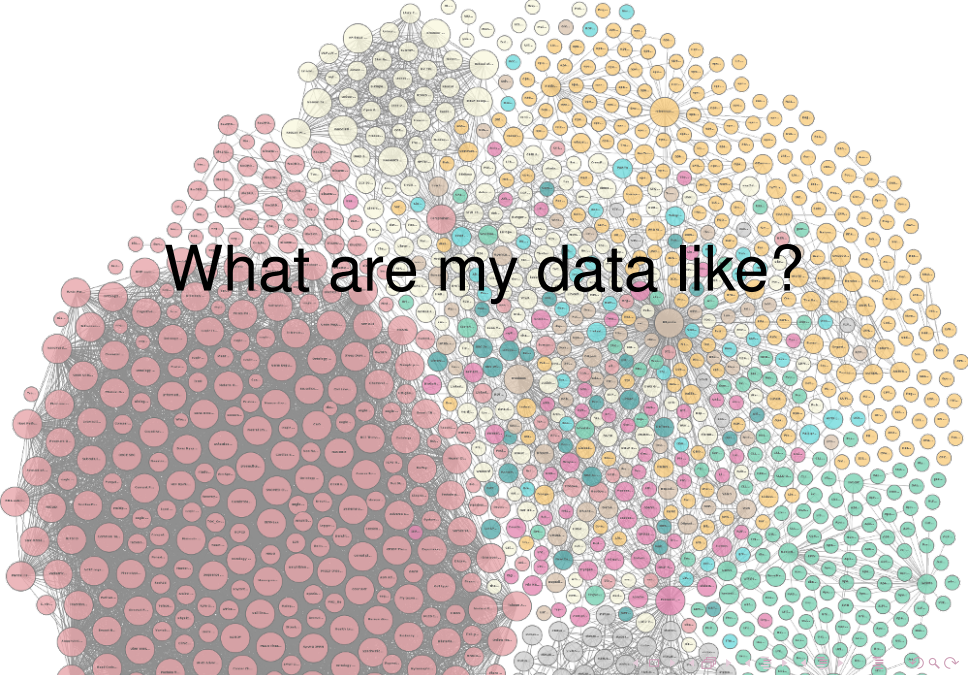
nathalie.villa-vialaneix@inra.fr

<http://www.nathalievilla.org>



eRum 2018

May 15th, 2018 - Budapest, Hungary



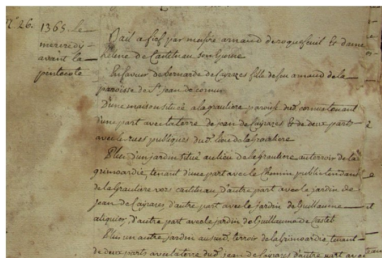
What are my data like?



A medieval social network [Boulet et al., 2008, Rossi et al., 2013]



corpus with more than 6,000
transactions, 3 centuries, all
related to
Castelnaud Montriat



AD 46 48 J6 page 37, acte 26 (analyse détaillée id_acte=72, id_transaction=142)

références documentaires

1365, le mercredi avant la Pentecôte

date

paroisse

tenancier, acteur de l'acte

Bail à fief par messire Amaud de Roquefeuil et Dame Hélène de Castelnaud son épouse en faveur de Bernarde Catrazes, fille de feu Amaud de la paroisse de St Jean de Cornus, d'une maison située à La Graulière, paroisse de Cornus, tenant d'une part avec la terre de Jean Catrazes et de deux parts avec les rues publiques du dit lieu de La Graulière.

tenancier confront

[...] (7 autres transactions pour deux jardins un pré et 4 pièces de terre)

sous la redevance de 6 d cahorsin d'acapte à mutation de seigneur et de 3 (4 quartes) mesures d'avoine et 1 poule à notre Dame en septembre.

Jean de Combeleau, notaire et commissaire d'actes de monsieur l'officiel de Cahors.

notaire

seigneurs, acteurs de l'acte

A medieval social network [Boulet et al., 2008, Rossi et al., 2013]



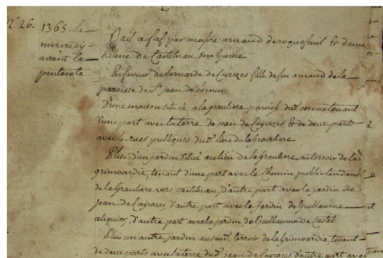
corpus with more than 6,000 transactions, 3 centuries, all related to Castelnaud Montrastier

Individual
Transaction



bipartite network with more than 17,000 nodes (~ 10,000 individuals)

What can we learn from the French medieval society?



AD 46 48 J6 page 37, acte 26 (analyse_détailée_id_acte=72, id_transaction=142)

références documentaires

1365, le mercredi avant la Pentecôte

date

paroisse

tenancier, acteur de l'acte

Bail à fief par messire Amaud de Roquefeuil et Dame Hélène de Castelnaud son épouse en faveur de Bernard Cairaizes, fille de feu Amaud de la paroisse de St Jean de Cornus, d'une maison située à La Graulière, paroisse de Cornus, tenant d'une part avec la terre de Jean Cairaizes et de deux parts avec les rues publiques du dit lieu de La Graulière.

tenancier confront

[...] (7 autres transactions pour deux jardins un pré et 4 pièces de terre)

sous la redevance de 6 d cahorsin d'acapte à mutation de seigneur et de 3 (4 quartes) mesures d'avoine et 1 poule à notre Dame en septembre.

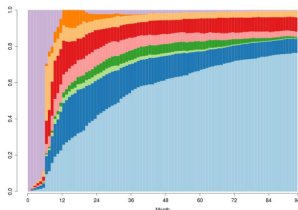
Jean de Combeleau, notaire et commissaire d'actes de monsieur l'officiel de Cahors.

notaire

seigneurs, acteurs de l'acte

Career paths [Olteanu and Villa-Vialaneix, 2015a]

Survey “Génération 98”: labor market status (9 categories) on more than 16,000 people having graduated in 1998 during 94 months.¹



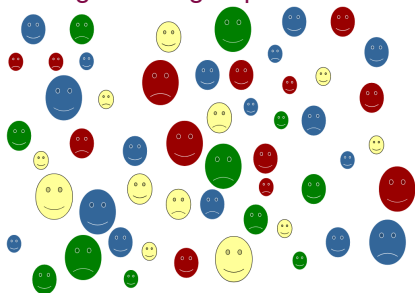
1. Available thanks to Génération 1998 à 7 ans - 2005, [producer] CEREQ, [diffusion] Centre Maurice Halbwachs (CMH)



Career paths [Olteanu and Villa-Vialaneix, 2015a]

Survey “Génération 98”: labor market status (9 categories) on more than 16,000 people having graduated in 1998 during 94 months.¹

How to cluster career paths into homogeneous groups?



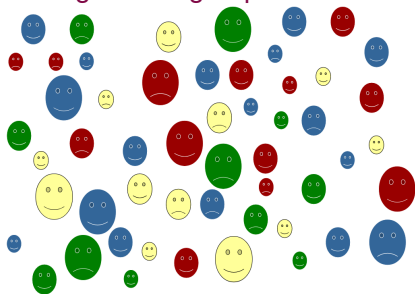
1. Available thanks to Génération 1998 à 7 ans - 2005, [producer] CEREQ, [diffusion] Centre Maurice Halbwachs (CMH)



Career paths [Olteanu and Villa-Vialaneix, 2015a]

Survey “Génération 98”: labor market status (9 categories) on more than 16,000 people having graduated in 1998 during 94 months.¹

How to cluster career paths into homogeneous groups?



It is **all about distance...**

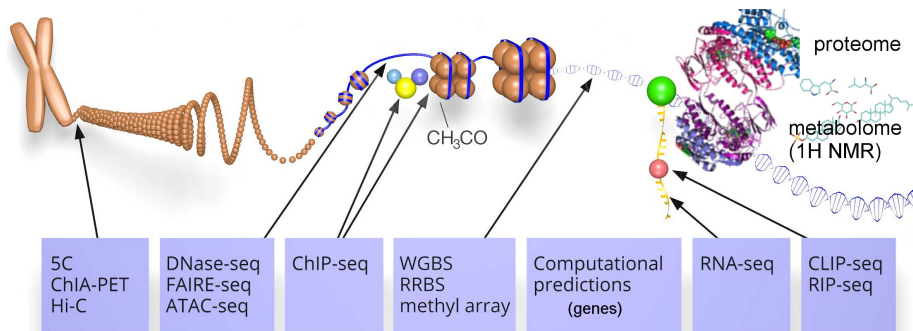
- χ^2 dissimilarity emphasizes the contemporary identical situations
- Optimal-matching dissimilarities is more focused on the sequences similarities

[Needleman and Wunsch, 1970]
(or “edit distance”, “Levenshtein distance”)

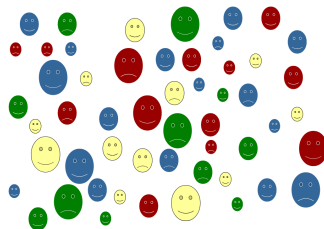
1. Available thanks to Génération 1998 à 7 ans - 2005, [producer] CEREQ, [diffusion] Centre Maurice Halbwachs (CMH)



and then I went into NGS data...



and again...
distances are everywhere



a collection of NGS data...

DNA barcoding

Astraptes fulgerator

optimal matching
(edit) distances to
differentiate species



a collection of NGS data...

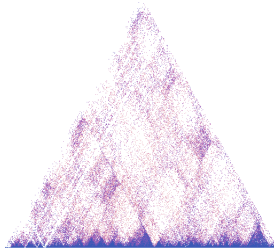
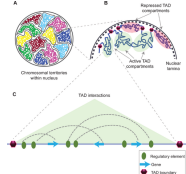
DNA barcoding

Astraptes fulgerator

optimal matching
(edit) distances to
differentiate species



Hi-C data



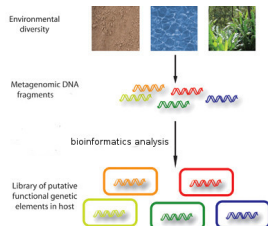
pairwise measure (similarity) related to
the physical 3D distance between loci in
the cell, at genome scale

a collection of NGS data...

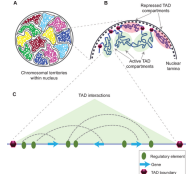
DNA barcoding

Astraptes fulgerator

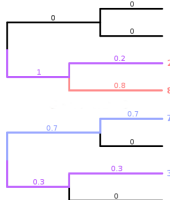
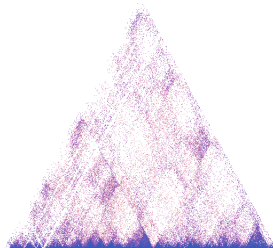
optimal matching
(edit) distances to
differentiate species



Hi-C data



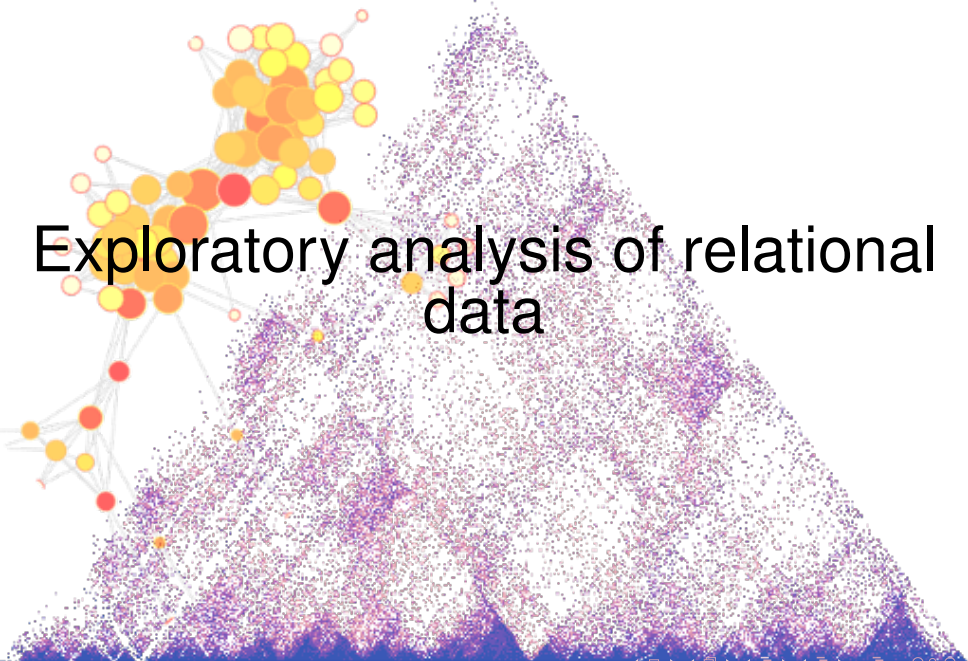
pairwise measure (similarity) related to
the physical 3D distance between loci in
the cell, at genome scale



Metagenomics

dissemblance between
samples is better
captured when
phylogeny between
species is taken into
account (unifrac
distances)



The background features a complex visualization of relational data. On the left, a network graph is shown with nodes of varying sizes in shades of yellow, orange, and red, connected by thin grey lines. To the right, a large, dense triangular cluster of small purple and blue dots is visible. The title 'Exploratory analysis of relational data' is centered over these visualizations.

Exploratory analysis of relational data



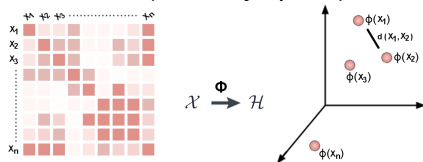
Formally, relational data are:

Euclidean distances or (non
Euclidean) dissimilarities between n
entities: symmetric $(n \times n)$ -matrix **D**
with positive entries and null
diagonal

Formally, relational data are:

Euclidean distances or (non Euclidean) dissimilarities between n entities: symmetric $(n \times n)$ -matrix $\mathbf{D}$ with positive entries and null diagonal

kernels: a symmetric and positive definite $(n \times n)$ -matrix $\mathbf{K}$ that measures a “relation” between n entities in $\mathcal{X}$ (arbitrary space)

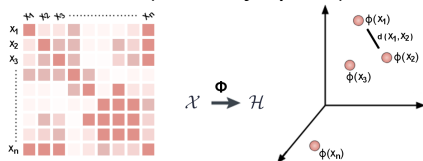


$$\mathbf{K}(x, x') = \langle \phi(x), \phi(x') \rangle$$

Formally, relational data are:

Euclidean distances or (non Euclidean) dissimilarities between n entities: symmetric $(n \times n)$ -matrix $\mathbf{D}$ with positive entries and null diagonal

kernels: a symmetric and positive definite $(n \times n)$ -matrix $\mathbf{K}$ that measures a “relation” between n entities in $\mathcal{X}$ (arbitrary space)



$$\mathbf{K}(x, x') = \langle \phi(x), \phi(x') \rangle$$

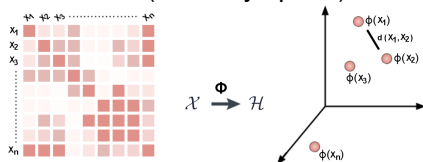
networks/graphs: groups of n entities (nodes/vertices) linked by a (potentially weighted) relation (edges)

$\Rightarrow$ symmetric $(n \times n)$ -matrix with positive entries and null diagonal $\mathbf{W}$

Formally, relational data are:

Euclidean distances or (non **Euclidean**) **dissimilarities** between n entities: symmetric $(n \times n)$ -matrix **D** with positive entries and null diagonal

kernels: a symmetric and positive definite $(n \times n)$ -matrix **K** that measures a “relation” between n entities in $\mathcal{X}$ (arbitrary space)



$$\mathbf{K}(x, x') = \langle \phi(x), \phi(x') \rangle$$

networks/graphs: groups of n entities (nodes/vertices) linked by a (potentially weighted) relation (edges)

$\Rightarrow$ symmetric $(n \times n)$ -matrix with positive entries and null diagonal **W**

Similarities between n entities: symmetric $(n \times n)$ -matrix **S** (with usually positive entries) but not necessarily definite positive

Different relational data types are related to each others

- a kernel is equivalent to an Euclidean distance:

$$\mathbf{D}(x, x') := \sqrt{\mathbf{K}(x, x) + \mathbf{K}(x', x') - 2\mathbf{K}(x, x')}$$

- from a dissimilarity, similarities can be computed:

$$\mathbf{S}(x, x) := a(x) \text{ (arbitrary)}, \mathbf{S}(x, x') = \frac{1}{2} (a(x) + a(x') - \mathbf{D}^2(x, x'))$$

- various kernels have been proposed for graphs (e.g., based on the graph Laplacian): [[Kondor and Lafferty, 2002](#)]

Different relational data types are related to each others

- a kernel is equivalent to an Euclidean distance:

$$\mathbf{D}(x, x') := \sqrt{\mathbf{K}(x, x) + \mathbf{K}(x', x') - 2\mathbf{K}(x, x')}$$

- from a dissimilarity, similarities can be computed:

$$\mathbf{S}(x, x) := a(x) \text{ (arbitrary)}, \mathbf{S}(x, x') = \frac{1}{2} (a(x) + a(x') - \mathbf{D}^2(x, x'))$$

- various kernels have been proposed for graphs (e.g., based on the graph Laplacian): [[Kondor and Lafferty, 2002](#)]

in summary

useful simplification: “is the framework Euclidean or not?” (e.g., kernel vs non Euclidean dissimilarity)

Principles for learning from relational data

Euclidean case (kernel $\mathbf{K}$)

rewrite all quantities using:

- $\mathbf{K}$ to compute distances and dot products
- linear or convex combinations of $(\phi(x_i))_i$ to describe all unobserved elements (centers of gravity and so on...)

Principles for learning from relational data

Euclidean case (kernel $\mathbf{K}$)

rewrite all quantities using:

- $\mathbf{K}$ to compute distances and dot products
- linear or convex combinations of $(\phi(x_i))_i$ to describe all unobserved elements (centers of gravity and so on...)

Works for: PCA, k -means, linear regression, ...

Principles for learning from relational data

Euclidean case (kernel **K**)
rewrite all quantities using:

- **K** to compute distances and dot products
- linear or convex combinations of $(\phi(x_i))_i$ to describe all unobserved elements (centers of gravity and so on...)

Works for: PCA, *k*-means, linear regression, ...

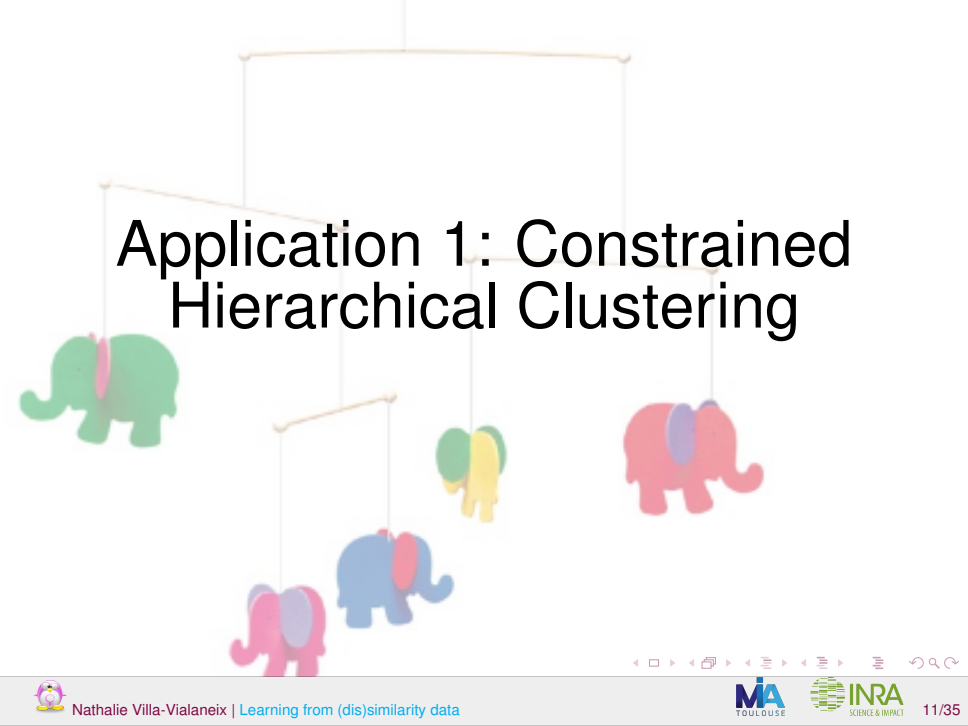
non Euclidean case (non Euclidean dissimilarity **D**): do almost the same using a pseudo-Euclidean framework

[Goldfarb, 1984]

$\exists$ two Euclidean spaces $\mathcal{E}_+$ and $\mathcal{E}_-$ and two mappings ϕ_+ and ϕ_- st:

$$\mathbf{D}(x, x') = \|\phi_+(x) - \phi_+(x')\|_{\mathcal{E}_+}^2 - \|\phi_-(x) - \phi_-(x')\|_{\mathcal{E}_-}^2$$

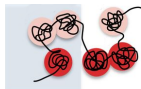
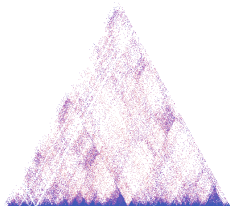




Application 1: Constrained Hierarchical Clustering



Constrained clustering for genomic data



Hi-C data: **S**

- segmentation (or contiguous clustering) of the chromosome
⇔ functional domains (TAD)
- hierarchical clustering is relevant

Other similar problems in biology:
Haplotypes based on LD between SNPs (groups of genomic positions inherited together)



adjclust

<https://cran.r-project.org/package=adjclust>

Features:

- constrained hierarchical clustering for arbitrary similarities (or kernels) or dissimilarities (extends *e.g.*, **rioja**)

adjclust

<https://cran.r-project.org/package=adjclust>

Features:

- constrained hierarchical clustering for arbitrary similarities (or kernels) or dissimilarities (extends *e.g.*, **rioja**)
- can be used for large scale (*e.g.*, genomic) datasets: fast implementation based on sparsity of **S** [Dehman, 2015]
complexity:
 - ▶ original method: $O(n^2)$ (time) and $O(n^2)$ (space)
 - ▶ **adjclust**: $O(nh + n \log n)$ (time) and $O(nh)$ (space) with h the non sparse band around the diagonal

adjclust

<https://cran.r-project.org/package=adjclust>

Features:

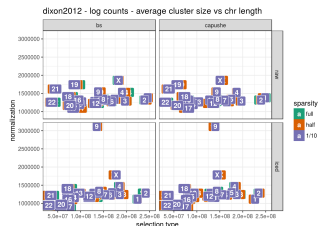
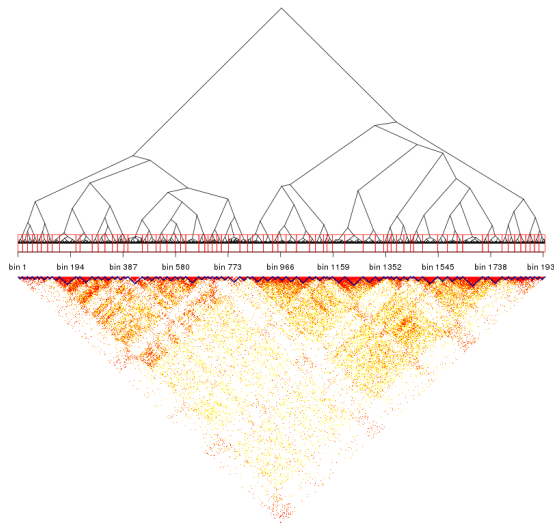
- constrained hierarchical clustering for arbitrary similarities (or kernels) or dissimilarities (extends e.g., **rioja**)
- can be used for large scale (e.g., genomic) datasets: fast implementation based on sparsity of **S** [Dehman, 2015]
complexity:
 - ▶ original method: $O(n^2)$ (time) and $O(n^2)$ (space)
 - ▶ **adjclust**: $O(nh + n \log n)$ (time) and $O(nh)$ (space) with h the non sparse band around the diagonal

Icing on the cake:

- wrappers for Hi-C datasets and LD datasets
- model selection methods (broken stick and slope heuristic)
- corrected dendrogram to avoid reversals [Grimm, 1987]
- ... and other nice plots to compare data with clustering

Application to Hi-C data

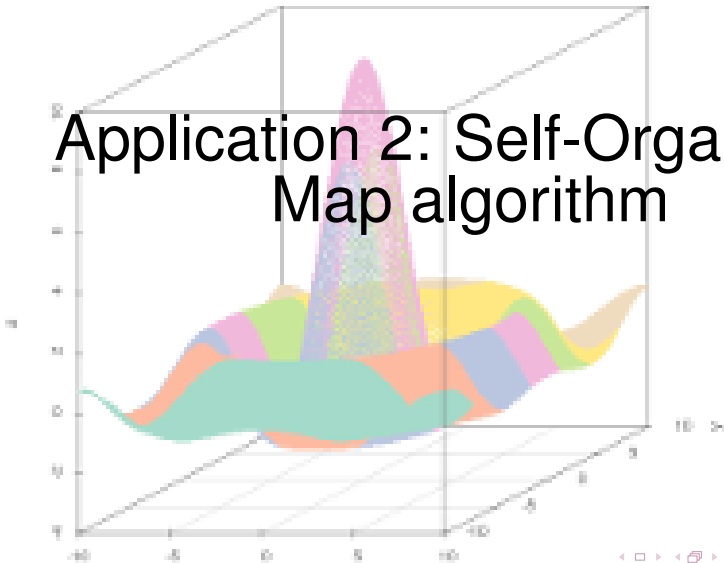
with data from [Dixon et al., 2012]



- constant average TAD size whatever the chromosome length
- similar results for broken stick and slope heuristic
- similar results for full and sparse (half - 1/10) versions

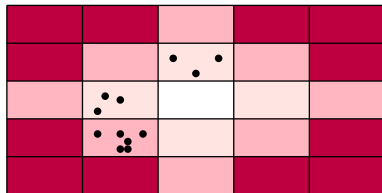
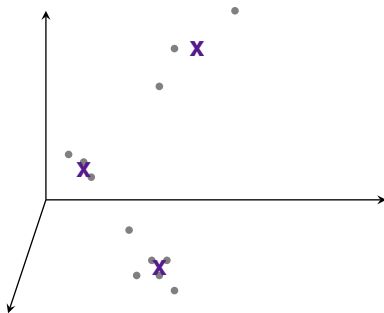


Application 2: Self-Organizing Map algorithm



Basics on (standard) stochastic SOM

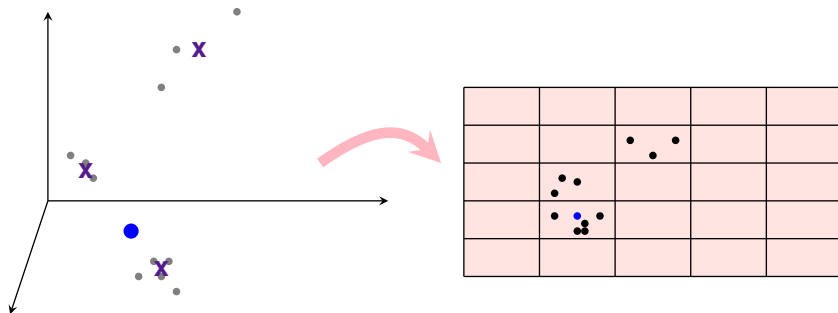
[Kohonen, 2001]



- $(x_i)_{i=1,\dots,n} \subset \mathbb{R}^d$ are affected to a unit $f(x_i) \in \{1, \dots, U\}$
- the grid is equipped with a “distance” between units: $d(u, u')$ and observations affected to close units are close in $\mathbb{R}^d$
- every unit u corresponds to a **prototype**, $p_u(\mathbf{x})$ in $\mathbb{R}^d$

Basics on (standard) stochastic SOM

[Kohonen, 2001]

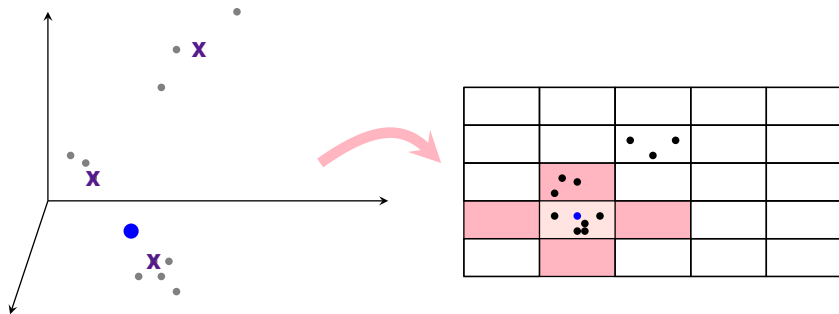


Iterative learning (assignment step): x_i is picked at random within $(x_k)_k$ and affected to *best matching unit*:

$$f^t(x_i) = \arg \min_u \|x_i - p_u^t\|^2$$

Basics on (standard) stochastic SOM

[Kohonen, 2001]



Iterative learning (representation step): all prototypes in neighboring units are updated with a gradient descent like step:

$$p_u^{t+1} \leftarrow p_u^t + \mu(t) H^t(d(f(x_i), u))(x_i - p_u^t)$$

Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathbb{R}^d$

1: **Initialization:**

randomly set $p_1^0, \dots, p_U^0$ in $\mathbb{R}^d$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: **Assignment**

$$f^t(x_i) = \arg \min_{u=1,\dots,U} \|x_i - p_u^t\|^2$$

5: **for all** $u = 1 \rightarrow U$ **do Representation**

6:

$$p_u^{t+1} = p_u^t + \mu(t) H^t(d(f^t(x_i), u)) (x_i - p_u^t)$$

7: **end for**

8: **end for**

Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathcal{X}$

1: Initialization:

randomly set $p_1^0, \dots, p_U^0$ in $\mathbb{R}^d$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: Assignment

$$f^t(x_i) = \arg \min_{u=1,\dots,U} \|x_i - p_u^t\|^2$$

5: **for all** $u = 1 \rightarrow U$ **do** Representation

6:

$$p_u^{t+1} = p_u^t + \mu(t) H^t(d(f^t(x_i), u)) (x_i - p_u^t)$$

7: **end for**

8: **end for**

Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathcal{X}$

1: Initialization:

$$p_u^0 = \sum_{i=1}^n \beta_{ui}^0 \phi(x_i) \text{ (convex combination)}$$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: **Assignment**

$$f^t(x_i) = \arg \min_{u=1,\dots,U} \|x_i - p_u^t\|^2$$

5: **for all** $u = 1 \rightarrow U$ **do Representation**

6:

$$p_u^{t+1} = p_u^t + \mu(t) H^t(d(f^t(x_i), u)) (x_i - p_u^t)$$

7: **end for**

8: **end for**

Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathcal{X}$

1: Initialization:

$$p_u^0 = \sum_{i=1}^n \beta_{ui}^0 \phi(x_i) \text{ (convex combination)}$$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: **Assignment**

$$f^t(x_i) = \arg \min_{u=1,\dots,U} \|\phi(x_i) - p_u^t\|_{\mathcal{H}}^2$$

5: **for all** $u = 1 \rightarrow U$ **do Representation**

6:

$$p_u^{t+1} = p_u^t + \mu(t) H^t(d(f^t(x_i), u)) (x_i - p_u^t)$$

7: **end for**

8: **end for**

Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathcal{X}$

1: Initialization:

$$p_u^0 = \sum_{i=1}^n \beta_{ui}^0 \phi(x_i) \text{ (convex combination)}$$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: **Assignment**

$$f^t(x_i) = \arg \min_{u=1,\dots,U} \|\phi(x_i) - p_u^t\|_{\mathcal{H}}^2$$

5: **for all** $u = 1 \rightarrow U$ **do Representation**

6:

$$p_u^{t+1} = p_u^t + \mu(t) H^t(d(f^t(x_i), u)) (\phi(x_i) - p_u^t)$$

7: **end for**

8: **end for**

Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathcal{X}$

1: Initialization:

$$p_u^0 = \sum_{i=1}^n \beta_{ui}^0 \phi(x_i) \text{ (convex combination)}$$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: **Assignment**

$$f^t(x_i) = \arg \min_{u=1,\dots,U} (\beta_u^t)^\top \mathbf{K} \beta_u^t - 2(\beta_u^t)^\top \mathbf{K}(\cdot, x_i)$$

5: **for all** $u = 1 \rightarrow U$ **do Representation**

6:

$$\beta_u^{t+1} = \beta_u^t + \mu(t) H^t(d(f^t(x_i), u)) (\mathbf{1}_i - \beta_u^t)$$

7: **end for**

8: **end for**



Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathcal{X}$

1: Initialization:

$$p_u^0 \sim \sum_{i=1}^n \beta_{ui}^0 x_i \text{ (convex combination)}$$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: **Assignment**

$$f^t(x_i) = \arg \min_{u=1,\dots,U} \mathbf{D}(p_u^t, x_i)$$

5: **for all** $u = 1 \rightarrow U$ **do Representation**

6:

$$p_u^{t+1} = p_u^t + \mu(t) H^t(d(f^t(x_i), u)) (\sim x_i - p_u^t)$$

7: **end for**

8: **end for**

Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

Data: $(x_i)_{i=1,\dots,n} \in \mathcal{X}$

1: Initialization:

$$p_u^0 \sim \sum_{i=1}^n \beta_{ui}^0 x_i \text{ (convex combination)}$$

2: **for** $t = 1 \rightarrow T$ **do**

3: pick at random $i \in \{1, \dots, n\}$

4: **Assignment**

$$f^t(x_i) = \arg \min_{u=1,\dots,U} (\beta_u^t)^\top \mathbf{D}(\cdot, x_i) - \frac{1}{2} (\beta_u^t)^\top \mathbf{D} \beta_u^t$$

5: **for all** $u = 1 \rightarrow U$ **do Representation**

6:

$$\beta_u^{t+1} = \beta_u^t + \mu(t) H^t(d(f^t(x_i), u)) (\mathbf{1}_i - \beta_u^t)$$

7: **end for**

8: **end for**



SOMbrero

[Villa-Vialaneix, 2017], <https://cran.r-project.org/package=SOMbrero>

- stochastic variants of SOM (standard, KORRESP and relational) with a large number of diagnostic plots
- specific functions to **use with graphs** and obtain simplified representations [Olteanu and Villa-Vialaneix, 2015b]

SOMbrero

[Villa-Vialaneix, 2017], <https://cran.r-project.org/package=SOMbrero>

- stochastic variants of SOM (standard, KORRESP and relational) with a large number of diagnostic plots
- specific functions to **use with graphs** and obtain simplified representations [Olteanu and Villa-Vialaneix, 2015b]
- contains comprehensive **vignettes** illustrated on **3 datasets** corresponding to the three algorithms (iris, presidentielles2002 and lesmis, a graph from “Les Misérables”)

SOMbrero

[Villa-Vialaneix, 2017], <https://cran.r-project.org/package=SOMbrero>

- stochastic variants of SOM (standard, KORRESP and relational) with a large number of diagnostic plots
- specific functions to **use with graphs** and obtain simplified representations [Olteanu and Villa-Vialaneix, 2015b]
- contains comprehensive **vignettes** illustrated on **3 datasets** corresponding to the three algorithms (iris, presidentielles2002 and lesmis, a graph from “Les Misérables”)
- **Web User Interface** (made with **shiny**) with `sombreroGUI()`

Tested on and approved by an historian!

The screenshot displays the SOMbrero Web User Interface (v0.1) in a browser window. The interface has a light gray header with the title 'SOMbrero Web User Interface (v0.1)'. Below the header, there are several tabs: 'Import Data', 'Self-Organize', 'Plot Map', 'Superclasses', 'Combine with external information', and 'Help'. The 'Plot Map' tab is currently selected. The main content area is titled 'Third step: plot the self-organizing map' and contains the following elements:

- A sub-header: 'In this panel and the next ones you can visualize the computed self-organizing map. This panel contains the standard plots used to analyze the map.'
- A section titled 'Options' with two dropdown menus:
 - 'Plot what?' with 'Prototypes' selected.
 - 'Type of plot:' with 'polygon distances' selected.
- A 3D plot on the left side showing a self-organizing map with a colorful surface and a grid of points.
- A welcome message at the bottom left: 'Welcome to SOMbrero, the open-source on-line interface for self-organization maps (SOM)'.

Navigation icons (back, forward, search, etc.) are visible in the bottom right corner of the interface.



Note on drawbacks of RSOM

Two main drawbacks:

- For $T \sim \gamma n$ iterations, complexity of RSOM is $O(\gamma n^3 U)$ (compared to $O(\gamma U d n)$ for numeric) [Rossi, 2014]

Note on drawbacks of RSOM

Two main drawbacks:

- For $T \sim \gamma n$ iterations, complexity of RSOM is $O(\gamma n^3 U)$ (compared to $O(\gamma U d n)$ for numeric) [Rossi, 2014]

Exact solution proposed in [Mariette et al., 2017] to reduce the complexity to $O(\gamma n^2 U)$ with additional storage memory of $O(Un)$ (implemented in **SOMbrero**)

Note on drawbacks of RSOM

Two main drawbacks:

- For $T \sim \gamma n$ iterations, complexity of RSOM is $O(\gamma n^3 U)$ (compared to $O(\gamma U d n)$ for numeric) [Rossi, 2014]

Exact solution proposed in [Mariette et al., 2017] to reduce the complexity to $O(\gamma n^2 U)$ with additional storage memory of $O(Un)$ (implemented in **SOMbrero**)

- For the non Euclidean case, the learning algorithm can be very unstable (saddle points)

Note on drawbacks of RSOM

Two main drawbacks:

- For $T \sim \gamma n$ iterations, complexity of RSOM is $O(\gamma n^3 U)$ (compared to $O(\gamma U d n)$ for numeric) [Rossi, 2014]

Exact solution proposed in [Mariette et al., 2017] to reduce the complexity to $O(\gamma n^2 U)$ with additional storage memory of $O(Un)$ (implemented in **SOMbrero**)

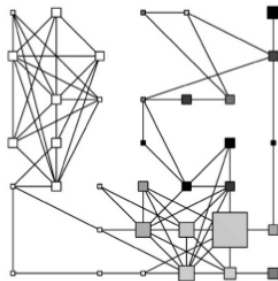
- For the non Euclidean case, the learning algorithm can be very unstable (saddle points)

clip or flip? [Chen et al., 2009]

RSOM for mining a medieval social network

with the heat kernel

● Individual
● Transaction



Graph induced by clusters:

- has nice relations with space and time
- emphasizes leading people
- has helped to identify problems in the database (namesakes)

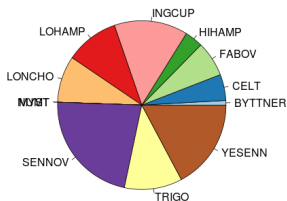
But: biggest communities are still very complex

[Boulet et al., 2008]



RSOM for typology of *Astraptes fulgerator* from DNA barcoding

Edit distances between DNA sequences [Olteanu and Villa-Vialaneix, 2015a]

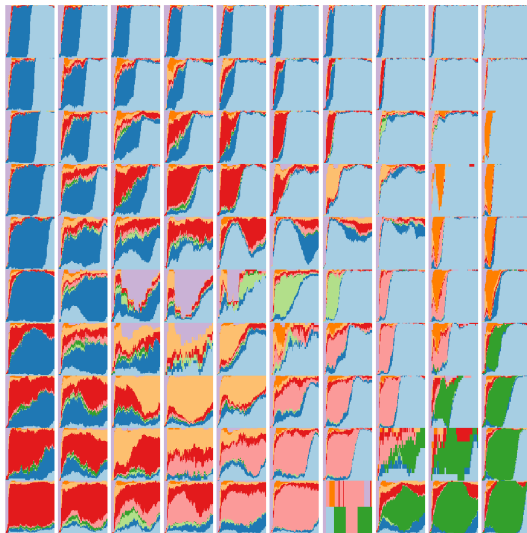


Almost perfect clustering (identifying a possible label error on one sample) with (in addition) **information on relations between species.**

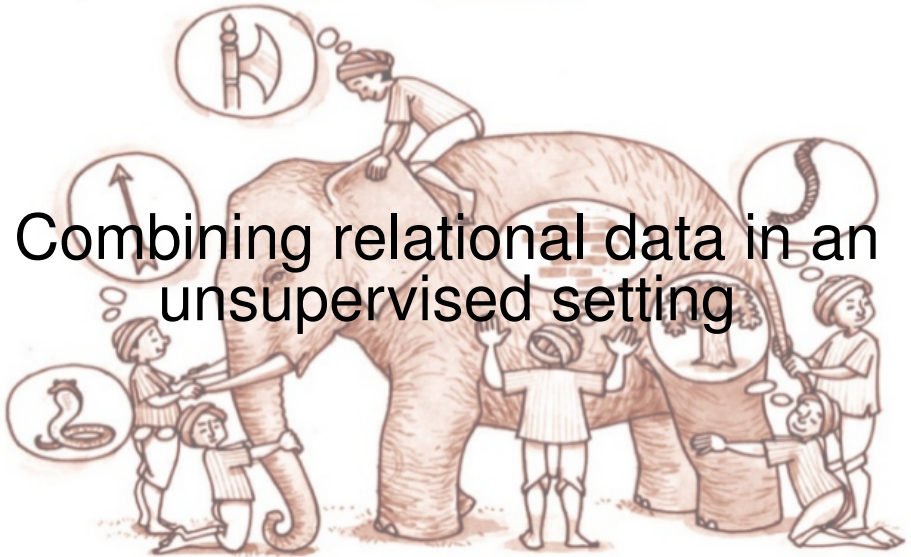


RSOM for typology of school-to-time transitions

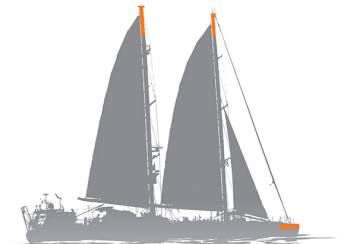
Edit distance between 12,000 categorical time series



Combining relational data in an unsupervised setting



TARA Oceans datasets



**TARA
OCEANS**

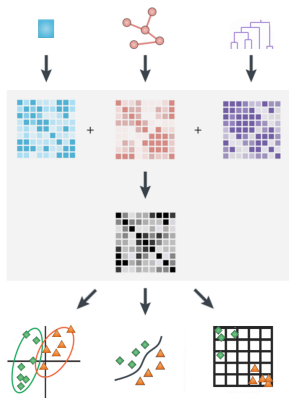


The 2009-2013 expedition

- Co-directed by Étienne Bourgois and Éric Karsenti
 - 7,012 datasets collected from 35,000 samples of plankton and water (11,535 Gb of data)
 - Study the plankton: bacteria, protists, metazoans and viruses (more than 90% of the biomass in the ocean)
- Metagenomic datasets similarity is well captured by unifracs distances



Multi-kernel/distances integration



How to “optimally” combine several relational datasets in an unsupervised setting?

for kernels $\mathbf{K}^1, \dots, \mathbf{K}^M$ obtained on the same n objects, search: $\mathbf{K}_\beta = \sum_{m=1}^M \beta_m \mathbf{K}^m$ with $\beta_m \geq 0$ and $\sum_m \beta_m = 1$

- [Mariette and Villa-Vialaneix, 2018]
- Package R **mixKernel**
<https://cran.r-project.org/package=mixKernel>



STATIS like framework

[L'Hermier des Plantes, 1976, Lavit et al., 1994]

Similarities between kernels:

$$C_{mm'} = \frac{\langle \mathbf{K}^m, \mathbf{K}^{m'} \rangle_F}{\|\mathbf{K}^m\|_F \|\mathbf{K}^{m'}\|_F} = \frac{\text{Trace}(\mathbf{K}^m \mathbf{K}^{m'})}{\sqrt{\text{Trace}((\mathbf{K}^m)^2) \text{Trace}((\mathbf{K}^{m'})^2)}}.$$

($C_{mm'}$ is an extension of the RV-coefficient [Robert and Escoufier, 1976] to the kernel framework)

STATIS like framework

[L'Hermier des Plantes, 1976, Lavit et al., 1994]

Similarities between kernels:

$$C_{mm'} = \frac{\langle \mathbf{K}^m, \mathbf{K}^{m'} \rangle_F}{\|\mathbf{K}^m\|_F \|\mathbf{K}^{m'}\|_F} = \frac{\text{Trace}(\mathbf{K}^m \mathbf{K}^{m'})}{\sqrt{\text{Trace}((\mathbf{K}^m)^2) \text{Trace}((\mathbf{K}^{m'})^2)}}.$$

($C_{mm'}$ is an extension of the RV-coefficient [Robert and Escoufier, 1976] to the kernel framework)

$$\begin{aligned} \text{maximize}_{\mathbf{v}} \quad & \sum_{m=1}^M \left\langle \mathbf{K}^*(\mathbf{v}), \frac{\mathbf{K}^m}{\|\mathbf{K}^m\|_F} \right\rangle_F = \mathbf{v}^\top \mathbf{C} \mathbf{v} \\ \text{for } \mathbf{K}^*(\mathbf{v}) = & \sum_{m=1}^M v_m \mathbf{K}^m \text{ and } \mathbf{v} \in \mathbb{R}^M \text{ such that } \|\mathbf{v}\|_2 = 1. \end{aligned}$$

STATIS like framework

[L'Hermier des Plantes, 1976, Lavit et al., 1994]

Similarities between kernels:

$$C_{mm'} = \frac{\langle \mathbf{K}^m, \mathbf{K}^{m'} \rangle_F}{\|\mathbf{K}^m\|_F \|\mathbf{K}^{m'}\|_F} = \frac{\text{Trace}(\mathbf{K}^m \mathbf{K}^{m'})}{\sqrt{\text{Trace}((\mathbf{K}^m)^2) \text{Trace}((\mathbf{K}^{m'})^2)}}.$$

($C_{mm'}$ is an extension of the RV-coefficient [Robert and Escoufier, 1976] to the kernel framework)

$$\begin{aligned} \text{maximize}_{\mathbf{v}} \quad & \sum_{m=1}^M \left\langle \mathbf{K}^*(\mathbf{v}), \frac{\mathbf{K}^m}{\|\mathbf{K}^m\|_F} \right\rangle_F = \mathbf{v}^\top \mathbf{C} \mathbf{v} \\ \text{for } \mathbf{K}^*(\mathbf{v}) = & \sum_{m=1}^M v_m \mathbf{K}^m \text{ and } \mathbf{v} \in \mathbb{R}^M \text{ such that } \|\mathbf{v}\|_2 = 1. \end{aligned}$$

Solution: first eigenvector of $\mathbf{C} \Rightarrow$ Set $\beta = \frac{\mathbf{v}}{\sum_{m=1}^M v_m}$ (consensual kernel).



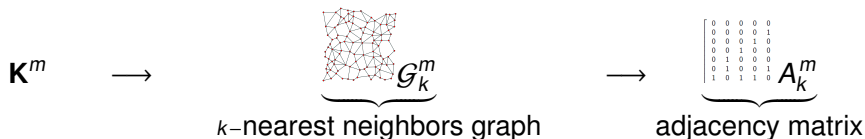
A kernel preserving the original topology of the data I

Similarly to [Lin et al., 2010], preserve the local geometry of the data in the feature space.

A kernel preserving the original topology of the data I

Similarly to [Lin et al., 2010], preserve the local geometry of the data in the feature space.

Proxy of the local geometry

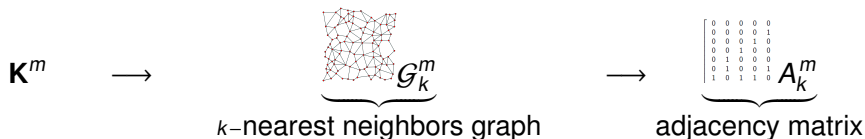


$$\Rightarrow W = \sum_m \mathbb{I}_{\{A_k^m > 0\}} \text{ or } W = \sum_m A_k^m$$

A kernel preserving the original topology of the data I

Similarly to [Lin et al., 2010], preserve the local geometry of the data in the feature space.

Proxy of the local geometry



$$\Rightarrow \mathbf{W} = \sum_m \mathbb{I}_{\{A_k^m > 0\}} \text{ or } \mathbf{W} = \sum_m \mathbf{A}_k^m$$

Feature space geometry measured by

$$\Delta_i(\beta) = \left\langle \phi_\beta^*(x_i), \begin{pmatrix} \phi_\beta^*(x_1) \\ \vdots \\ \phi_\beta^*(x_n) \end{pmatrix} \right\rangle = \begin{pmatrix} \mathbf{K}_\beta^*(x_i, x_1) \\ \vdots \\ \mathbf{K}_\beta^*(x_i, x_n) \end{pmatrix}$$

A kernel preserving the original topology of the data II

Sparse version

$$\begin{aligned} \text{minimize}_{\beta} \quad & \sum_{i,j=1}^N w_{ij} \|\Delta_i(\beta) - \Delta_j(\beta)\|^2 \\ \text{for } \mathbf{K}_{\beta}^* = & \sum_{m=1}^M \beta_m \mathbf{K}^m \text{ and } \beta \in \mathbb{R}^M \text{ st } \beta_m \geq 0 \text{ and } \sum_{m=1}^M \beta_m = 1. \end{aligned}$$

Non sparse version

$$\begin{aligned} \text{minimize}_{\mathbf{v}} \quad & \sum_{i,j=1}^N w_{ij} \|\Delta_i(\beta) - \Delta_j(\beta)\|^2 \\ \text{for } \mathbf{K}_{\mathbf{v}}^* = & \sum_{m=1}^M v_m \mathbf{K}^m \text{ and } \mathbf{v} \in \mathbb{R}^M \text{ st } v_m \geq 0 \text{ and } \|\mathbf{v}\|_2 = 1. \end{aligned}$$

A kernel preserving the original topology of the data II

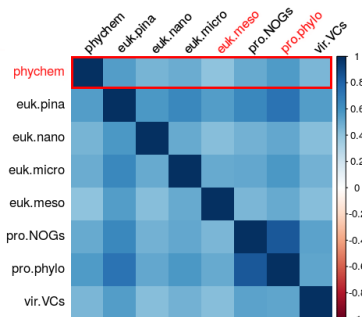
Sparse version

equivalent to a standard QP problem with linear constraints (ex: package **quadprog** in R)

Non sparse version

equivalent to a QPQC problem (harder to solve) solved with “Alternating Direction Method of Multipliers” (ADMM [[Boyd et al., 2011](#)])

Application to TARA oceans

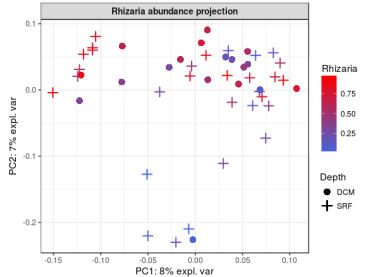
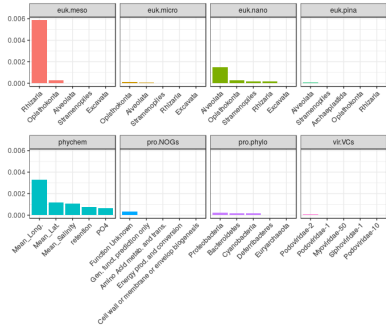


Similarity between datasets (STATIS)

- **phychem** and small size organisms are the most similar (confirmed by [de Vargas et al., 2015] et [Sunagawa et al., 2015]).



Application to *TARA* oceans



Important variables

- *Rhizaria* abundance strongly structure the differences between samples (analyses restricted to some organisms found differences mostly based on water depths)
- and waters from Arctic Oceans and Pacific Oceans differ in terms of *Rhizaria* abundance





SOMbrero

Madalina Olteanu,

Fabrice Rossi, Marie Cottrell,

Laura Bendhaïba and

Julien Boelaert



SOMbrero and mixKernel



Jérôme Mariette

adjclust

Pierre Neuvial, Guillem Rigail, Christophe Ambroise and

Shubham Chaturvedi



Google
Summer of Code



Toulouse
useR!
2019

Don't miss useR! 2019

user2019.r-project.org



Credits for pictures

- Slide 2: Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentsch and Richard Cyganiak. <http://lod-cloud.net/>
- Slide 3: Picture of Castelnau Montratier from https://commons.wikimedia.org/wiki/File:Place_Gambetta,_Castelnau-Montratier.JPG by Duch.seb CC BY-SA 3.0
- Slide 4: image based on ENCODE project, by Darryl Leja (NHGRI), Ian Dunham (EBI) and Michael Pazin (NHGRI)
- Slide 6: *Astraptes* picture is from <https://www.flickr.com/photos/39139121@N00/2045403823/> by Anne Toal (CC BY-SA 2.0), Hi-C experiment is taken from the article Matharu *et al.*, 2015 DOI:10.1371/journal.pgen.1005640 (CC BY-SA 4.0) and metagenomics illustration is taken from the article Sommer *et al.*, 2010 DOI:10.1038/msb.2010.16 (CC BY-NC-SA 3.0)
- Slide 12: TADS picture is from the article Fraser *et al.*, 2015 DOI:10.15252/msb.20156492 (CC BY-SA 4.0)
- Slide 27: Adjacency matrix image from: By S. Mohammad H. Oloomi, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=35313532>

References



Boulet, R., Jouve, B., Rossi, F., and Villa, N. (2008).
Batch kernel SOM and related Laplacian methods for social network analysis.
Neurocomputing, 71(7-9):1257–1273.



Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011).
Distributed optimization and statistical learning via the alternating direction method of multipliers.
Foundations and Trends in Machine Learning, 3(1):1–122.



Chen, Y., Garcia, E., Gupta, M., Rahimi, A., and Cazzanti, L. (2009).
Similarity-based classification: concepts and algorithm.
Journal of Machine Learning Research, 10:747–776.



de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, P., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Acinas, S., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and Karsenti, E. (2015).
Eukaryotic plankton diversity in the sunlit ocean.
Science, 348(6237).



Dehman, A. (2015).
Spatial Clustering of Linkage Disequilibrium blocks for Genome-Wide Association Studies.
PhD thesis, Université Paris Saclay.



Dixon, J., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J., and Ren, B. (2012).
Topological domains in mammalian genomes identified by analysis of chromatin interactions.
Nature, 485:376–380.



Goldfarb, L. (1984).
A unified approach to pattern recognition.
Pattern Recognition, 17(5):575–582.



Grimm, E. (1987).

CONISS: a fortran 77 program for stratigraphically constrained analysis by the method of incremental sum of squares.
Computers & Geosciences, 13(1):13–35.



Kohonen, T. (2001).

Self-Organizing Maps, 3rd Edition, volume 30.
Springer, Berlin, Heidelberg, New York.



Kondor, R. and Lafferty, J. (2002).

Diffusion kernels on graphs and other discrete structures.
In Sammut, C. and Hoffmann, A., editors, *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, Sydney, Australia. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA.



Lavit, C., Escoufier, Y., Sabatier, R., and Traissac, P. (1994).

The ACT (STATIS method).
Computational Statistics and Data Analysis, 18(1):97–119.



L'Hermier des Plantes, H. (1976).

Structuration des tableaux à trois indices de la statistique.
PhD thesis, Université de Montpellier.
Thèse de troisième cycle.



Lin, Y., Liu, T., and CS., F. (2010).

Multiple kernel learning for dimensionality reduction.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 33:1147–1160.



Mariette, J., Rossi, F., Olteanu, M., and Villa-Vialaneix, N. (2017).

Accelerating stochastic kernel som.
In Verleysen, M., editor, *XXVth European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2017)*, pages 269–274, Bruges, Belgium. i6doc.



Mariette, J. and Villa-Vialaneix, N. (2018).

Unsupervised multiple kernel learning for heterogeneous data integration.
Bioinformatics.



Forthcoming.



Needleman, S. and Wunsch, C. (1970).

A general method applicable to the search for similarities in the amino acid sequence of two proteins.
Journal of Molecular Biology, 48(3):443–453.



Olteanu, M. and Villa-Vialaneix, N. (2015a).

On-line relational and multiple relational SOM.
Neurocomputing, 147:15–30.



Olteanu, M. and Villa-Vialaneix, N. (2015b).

Using SOMbrero for clustering and visualizing graphs.
Journal de la Société Française de Statistique, 156(3):95–119.



Robert, P. and Escoufier, Y. (1976).

A unifying tool for linear multivariate statistical methods: the rv-coefficient.
Applied Statistics, 25(3):257–265.



Rossi, F. (2014).

How many dissimilarity/kernel self organizing map variants do we need?
In Villmann, T., Schleif, F., Kaden, M., and Lange, M., editors, *Advances in Self-Organizing Maps and Learning Vector Quantization (Proceedings of WSOM 2014)*, volume 295 of *Advances in Intelligent Systems and Computing*, pages 3–23, Mittweida, Germany. Springer Verlag, Berlin, Heidelberg.



Rossi, F., Villa-Vialaneix, N., and Hautefeuille, F. (2013).

Exploration of a large database of French notarial acts with social network methods.
Digital Medievalist, 9.



Sunagawa, S., Coelho, L., Chaffron, S., Kultima, J., Labadie, K., Salazar, F., Djahanschiri, B., Zeller, G., Mende, D., Alberti, A., Cornejo-Castillo, F., Costea, P., Cruaud, C., d'Oviedo, F., Engelen, S., Ferrera, I., Gasol, J., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B., Royo-Llonch, M., Sarmiento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S., and Bork, P. (2015).





Structure and function of the global ocean microbiome.

Science, 348(6237).

Villa-Vialaneix, N. (2017).

Stochastic self-organizing map variants with the R package SOMbrero.

In Lamirel, J., Cottrell, M., and Olteanu, M., editors, *12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (Proceedings of WSOM 2017)*, Nancy, France. IEEE.

Dendrogram corrections when reversals are detected

