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 Castelnau Montratier

AD 4648 J6 page 37, acte 26 (analyse détaillée id_acte=72, id_transaction=142)

références documentaires

1365, le mercredi avant la Pentecôte

Bail à fief par messire Amand de Roquefeuil et Dame Hélène de Castelnau à son époux en faveur de Bernard Carrazes, fille de feu Amand, de la paroisse de Saint Jean de Comps, d'une maison située à la Grauvière, paroisse de Cornus, tenant d'une part avec la terre de Jean Carrazes et de deux parts avec les rues publiques du dit lieu de la Grauvière.

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

sous la redevance de 6 d'avoine d'acquére à mutation de seigneur et de 3 (4 quartiers) mesures d'avoine et 1 poule à notre Dame de septembre.

Jean de Comberleau, seigneur et commissaire d'actes de monseigneur l'officiel de Cahors.
A medieval social network [Boulet et al., 2008, Rossi et al., 2013] corpus with more than 6,000 transactions, 3 centuries, all related to Castelnau Montratier

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

What can we learn from the French medieval society?
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

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

1. How to cluster career paths into homogeneous groups?

It is all about distance...

- $\chi^2$ 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)
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.

Similarly, \( K \) measures a "relation" between \( n \) entities in \( X \) (arbitrary space) if \n
\[
K(x, x') = \langle \phi(x), \phi(x') \rangle
\]
Formally, relational data are:

Euclidean distances or (non-Euclidean) dissimilarities between \( n \) entities: symmetric \((n \times n)\)-matrix \( D \) 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 \( X \) (arbitrary space)

\[
K(x, x') = \langle \phi(x), \phi(x') \rangle
\]
Formally, relational data are:

Euclidean distances or (non-Euclidean) dissimilarities between $n$ entities: symmetric $(n \times n)$-matrix $D$ 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)

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 $W$

$$K(x, x') = \langle \phi(x), \phi(x') \rangle$$
Formally, relational data are:

Euclidean distances or (non-Euclidean) dissimilarities between $n$ entities: symmetric $(n \times n)$-matrix $D$ 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 $X$ (arbitrary space).

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

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

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:

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

- from a dissimilarity, similarities can be computed:

\[
S(x, x) := a(x) \text{ (arbitrary)}, S(x, x') = \frac{1}{2} \left( a(x) + a(x') - D^2(x, x') \right)
\]

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

- From a dissimilarity, similarities can be computed:
  \[
  S(x, x) := a(x) \text{ (arbitrary)},
  S(x, x') = \frac{1}{2} \left( a(x) + a(x') - D^2(x, x') \right)
  \]

- 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 $K$)
 rewrite all quantities using:

- $K$ to compute distances and dot products
- linear or convex combinations of $\phi(x_i)$ 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

$\exists$ two Euclidean spaces $E^+$ and $E^-$ and two mappings $\phi^+$ and $\phi^-$ s.t:

$$D(x, x') = \| \phi^+(x) - \phi^+(x') \|_2^2 - \| \phi^-(x) - \phi^-(x') \|_2^2$$
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, ...
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

\[ D(x, x') = \|\phi_+(x) - \phi_+(x')\|_{E_+}^2 - \|\phi_-(x) - \phi_-(x')\|_{E_-}^2 \]

[Goldfarb, 1984]
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

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
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]

- \((x_i)_{i=1,...,n} \subset \mathbb{R}^d\) are affected to a unit \(f(x_i) \in \{1, \ldots, U\}\)
- 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, \(p_u(x)\) in \(\mathbb{R}^d\)

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

\[
p_{t+1}(u) \leftarrow p_t(u) + \mu(t) H_t(d(f(x_i)), u)(x_i - p_t(u))
\]
Iterative learning (assignment step): \( x_i \) is picked at random within \( (x_k)_k \) and affected to best matching unit:

\[
    f^t(x_i) = \arg \min_u \|x_i - p^t_u\|_2
\]
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: \((x_i)_{i=1,...,n} \in \mathbb{R}^d\)

1: Initialization:
   randomly set \(p_1^0, ..., p_U^0\) in \(\mathbb{R}^d\)

2: \(\text{for } t = 1 \rightarrow T \text{ do}\)

3: pick at random \(i \in \{1, \ldots, n\}\)

4: Assignment
   \[f^t(x_i) = \arg \min_{u=1,...,U} ||x_i - p_u^t||^2\]

5: \(\text{for all } u = 1 \rightarrow U \text{ do Representation}\)

6: \[p_u^{t+1} = p_u^t + \mu(t)H^t(d(f^t(x_i), u))(x_i - p_u^t)\]

7: \(\text{end for}\)

8: \(\text{end for}\)
Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

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

1: Initialization:
   randomly set \(p^0_1, ..., p^0_U\) in \(\mathbb{R}^d\)
2: \(\textbf{for } t = 1 \rightarrow T \textbf{ do} \)
3: pick at random \(i \in \{1, \ldots, n\}\)
4: Assignment
   \[ f^t(x_i) = \arg \min_{u=1,...,U} \|x_i - p^t_u\|^2 \]
5: \(\textbf{for all } u = 1 \rightarrow U \textbf{ do} \textbf{ Representation} \)
6:
   \[ p^{t+1}_u = p^t_u + \mu(t)H^t(d(f^t(x_i), u))(x_i - p^t_u) \]
7: \(\textbf{end for} \)
8: \(\textbf{end for} \)
Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

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

1: Initialization:
\[ p_u^0 = \sum_{i=1}^{n} \beta_{ui}^0 \phi(x_i) \] (convex combination)

2: for \(t = 1 \rightarrow T\) do

3: pick at random \(i \in \{1, \ldots, n\}\)

4: Assignment
\[ f^t(x_i) = \arg \min_{u=1,\ldots,U} \|x_i - p_u^t\|^2 \]

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: \( (x_i)_{i=1,...,n} \in X \)

1: Initialization:
   \[
   p_u^0 = \sum_{i=1}^n \beta_{ui}^0 \phi(x_i) \quad \text{(convex combination)}
   \]

2: for \( t = 1 \rightarrow T \) do

3: pick at random \( i \in \{1, \ldots, n\} \)

4: Assignment
   \[
   f^t(x_i) = \arg \min_{u=1,\ldots,U} \| \phi(x_i) - p_u^t \|_X^2
   \]

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

1: Initialization:
\[ p^0_u = \sum_{i=1}^{n} \beta_{ui}^0 \phi(x_i) \] (convex combination)

2: for \(t = 1 \rightarrow T\) do

3: pick at random \(i \in \{1, \ldots, n\}\)

4: Assignment
\[ f^t(x_i) = \arg \min_{u=1,\ldots,U} \|\phi(x_i) - p^t_u\|_X^2 \]

5: for all \(u = 1 \rightarrow U\) do Representation

6:
\[ p^{t+1}_u = p^t_u + \mu(t)H^t(d(f^t(x_i), u))(\phi(x_i) - p^t_u) \]

7: end for

8: end for
Extension of SOM to data described by a kernel or a dissimilarity

[Olteanu and Villa-Vialaneix, 2015a]

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

1: Initialization:
   \[ p^0_u = \sum_{i=1}^n \beta^0_{ui} \phi(x_i) \] (convex combination)

2: for \(t = 1 \rightarrow T\) do

3:   pick at random \(i \in \{1, \ldots, n\}\)

4:   Assignment
   \[ f^t(x_i) = \arg \min_{u=1,\ldots,U} (\beta^t_u)^\top K \beta^t_u - 2(\beta^t_u)^\top K(., x_i) \]

5:   for all \(u = 1 \rightarrow U\) do Representation

6:   \[ \beta^{t+1}_u = \beta^t_u + \mu(t) H^t(d(f^t(x_i), u))(1_i - \beta^t_u) \]

7:   end for

8: end for
**Extension of SOM to data described by a kernel or a dissimilarity**

[Olteanu and Villa-Vialaneix, 2015a]

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

1: **Initialization:**
   
   \(p_u^0 \sim \sum_{i=1}^{n} \beta_{ui}^0 x_i\) (convex combination)

2: **for** \(t = 1 \rightarrow T\) **do**

3: pick at random \(i \in \{1, \ldots, n\}\)

4: **Assignment**
   
   \(f^t(x_i) = \arg \min_{u=1,\ldots,U} D(p_u^t, x_i)\)

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

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 $i \in \{1, \ldots, n\}$

4:    Assignment
   $$f^t(x_i) = \arg \min_{u=1,...,U} (\beta_{ui}^t)^\top D(., x_i) - \frac{1}{2} (\beta_{ui}^t)^\top D \beta_{ui}^t$$

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

6:    $$\beta_{ui}^{t+1} = \beta_{ui}^t + \mu(t) H^t(d(f^t(x_i), u))(1_i - \beta_{ui}^t)$$

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](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!
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 $O(\gamma Udn)$ 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 $O(UU_n)$ (implemented in SOMbrero)

For the non Euclidean case, the learning algorithm can be very unstable (saddle points) [Chen et al., 2009]
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 $O(\gamma Udn)$ 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 \( O(\gamma Udn) \) 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 $O(\gamma Udn)$ 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

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

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 unifrac distances
Multi-kernel/distances integration

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

for kernels $K^1, \ldots, K^M$ obtained on the same $n$ objects, search: $K_\beta = \sum_{m=1}^{M} \beta_m K^m$

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:

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

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

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

Similarities between kernels:

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

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

maximize \(v\)

\[
\sum_{m=1}^{M} \left\langle K^*(v), \frac{K^m}{\|K^m\|_F} \right\rangle_F = v^\top Cv
\]

for \(K^*(v) = \sum_{m=1}^{M} v_m K^m\) and \(v \in \mathbb{R}^M\) such that \(\|v\|_2 = 1\).
STATIS like framework

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

Similarities between kernels:

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

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

maximize_\mathbf{v} \sum_{m=1}^{M} \left\langle K^*(\mathbf{v}), \frac{K^m}{\|K^m\|_F} \right\rangle_F = \mathbf{v}^\top \mathbf{Cv}

for \(K^*(\mathbf{v}) = \sum_{m=1}^{M} v_m K^m\) and \(\mathbf{v} \in \mathbb{R}^M\) such that ||\(\mathbf{v}||_2 = 1\).

Solution: first eigenvector of \(\mathbf{C}\) ⇒ 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

\[
K^m \longrightarrow \underbrace{G^m_k}_{\text{\( k \)-nearest neighbors graph}} \longrightarrow \underbrace{A^m_k}_{\text{adjacency matrix}}
\]

\[
\Rightarrow W = \sum_m I\{A^m_k > 0\} \text{ or } W = \sum_m A^m_k
\]
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

\[ K^m \rightarrow G^m_k \rightarrow A^m_k \]

\[ k \text{-nearest neighbors graph} \]

\[ \Rightarrow W = \sum_m I\{A^m_k > 0\} \text{ or } W = \sum_m A^m_k \]

Feature space geometry measured by

\[ \Delta_i(\beta) = \left\langle \phi^*_\beta(x_i), \begin{pmatrix} \phi^*_\beta(x_1) \\ \vdots \\ \phi^*_\beta(x_n) \end{pmatrix} \right\rangle = \begin{pmatrix} K^*_\beta(x_i, x_1) \\ \vdots \\ K^*_\beta(x_i, x_n) \end{pmatrix} \]
A kernel preserving the original topology of the data II

Sparse version

\[
\text{minimize}_{\beta} \sum_{i,j=1}^{N} W_{ij} \| \Delta_i(\beta) - \Delta_j(\beta) \|^2
\]

for \( \mathbf{K}_\beta^* = \sum_{m=1}^{M} \beta_m \mathbf{K}^m \) and \( \beta \in \mathbb{R}^M \) st \( \beta_m \geq 0 \) and \( \sum_{m=1}^{M} \beta_m = 1 \).

Non sparse version

\[
\text{minimize}_{\mathbf{v}} \sum_{i,j=1}^{N} W_{ij} \| \Delta_i(\beta) - \Delta_j(\beta) \|^2
\]

for \( \mathbf{K}_\mathbf{v}^* = \sum_{m=1}^{M} v_m \mathbf{K}^m \) and \( \mathbf{v} \in \mathbb{R}^M \) st \( v_m \geq 0 \) and \( \| \mathbf{v} \|_2 = 1 \).
A kernel preserving the original topology of the data II

Sparse version
equivalent to a standard QP problem with linear constrains (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
Don’t miss useR! 2019
user2019.r-project.org
Credits for pictures


- Slide 3: Picture of Castelnau Montratier from [https://commons.wikimedia.org/wiki/File:Place_Gambetta,_Castelnau-Montratier.JPG](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/](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)

References

Batch kernel SOM and related Laplacian methods for social network analysis.
*Neurocomputing*, 71(7-9):1257–1273.

Distributed optimization and statistical learning via the alternating direction method of multipliers.

Similarity-based classification: concepts and algorithm.

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., Ludicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S.,
Eukaryotic plankton diversity in the sunlit ocean.
*Science*, 348(6237).

*Spatial Clustering of Linkage Disequilibrium blocks for Genome-Wide Association Studies*.
PhD thesis, Université Paris Saclay.

Topological domains in mammalian genomes identified by analysis of chromatin interactions.

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


A general method applicable to the search for similarities in the amino acid sequence of two proteins.

On-line relational and multiple relational SOM.

Using SOMbrero for clustering and visualizing graphs.

A unifying tool for linear multivariate statistical methods: the rv-coefficient.

How many dissimilarity/kernel self organizing map variants do we need?

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

Structure and function of the global ocean microbiome.
*Science*, 348(6237).

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