SOMbrero: an R package for numeric and non-numeric self-organizing maps

Nathalie Villa-Vialaneix
with J. Boelaert, L. Bendhaïba, M. Olteanu

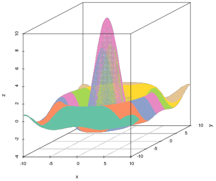
nathalie.villa@toulouse.inra.fr
http://www.nathalievilla.org

WSOM 2014 - Mittweida, Germany - July 4th

Outline

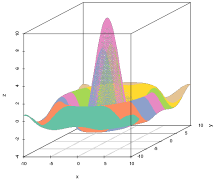
1. a short review of Self-Organizing Maps for non vectorial data

2. **SOMbrero**



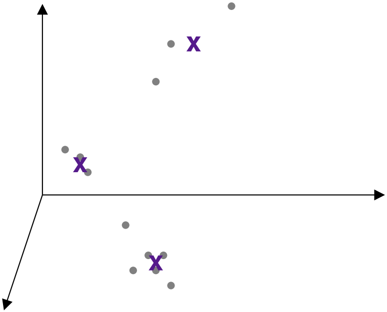
Outline

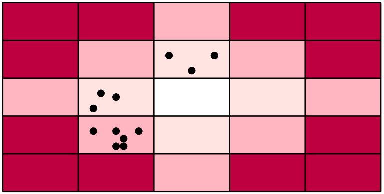
1. a short review of Self-Organizing Maps for non vectorial data
2. SOMbrero



Basics on stochastic SOM

[Kohonen, 2001]

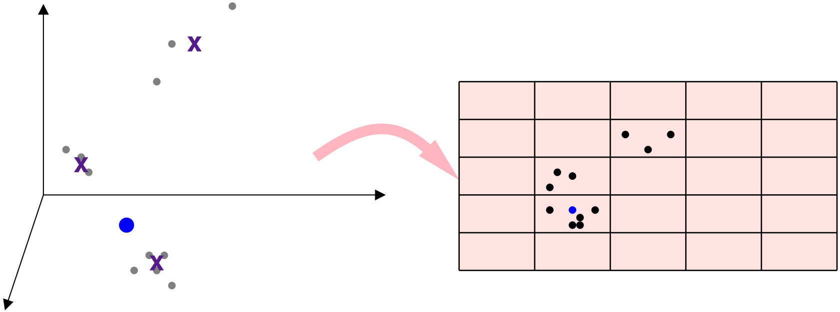




- $(x_i)_{i=1,\dots,n} \subset \mathbb{R}^d$ are affected to a unit $\mathcal{C}(x_i) \in \{1, \dots, U\}$
- the grid is equipped with a "distance" between units: $d(u, u')$ and observations affected to close units are close in $\mathbb{R}^d$
- every unit u corresponds to a prototype, p_u ($\mathbf{x}$)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

Basics on stochastic SOM

[Kohonen, 2001]

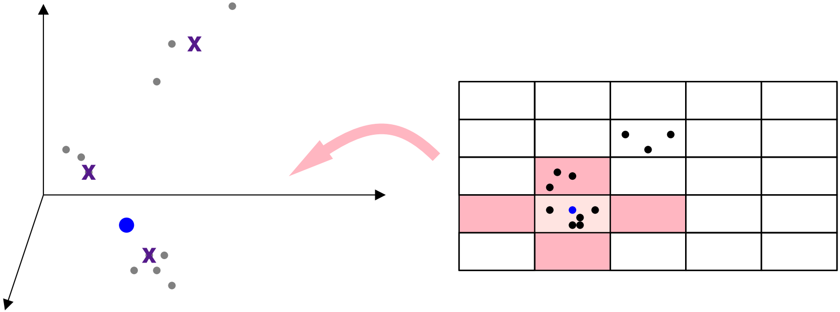


Iterative learning (affectation step): x_i is picked at random within $(x_k)_k$ and affected to *best matching unit*:

$$\mathcal{C}(x_i) = \arg\min_u \|x_i - p_u\|^2$$

Basics on stochastic SOM

[Kohonen, 2001]



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

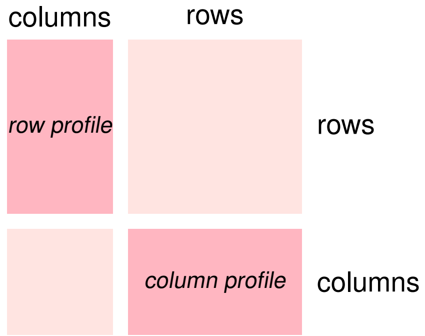
$$p_u^{t+1} \longleftarrow p_u^t + \mu(t) H^t(d(\mathcal{C}(x_i), u))(x_i - p_u^t)$$

Extensions to non vectorial data 1

KORRESP [Cottrell et al., 1993]

Data: contingency table $\mathbf{T} = (n_{ij})_{ij}$ with p rows and q columns transformed into a numeric dataset $\mathbf{X}$:

$\mathbf{X} =$



with

- $\forall\, i = 1, \ldots, p$ and $\forall\, j = 1, \ldots, q$, $\mathbf{x}_{ij} = \frac{n_{ij}}{n_{i.}} \times \sqrt{\frac{n}{n_{.j}}}$

Extensions to non vectorial data 1

KORRESP [Cottrell et al., 1993]

Data: contingency table $\mathbf{T} = (n_{ij})_{ij}$ with p rows and q columns transformed into a numeric dataset $\mathbf{X}$:

$$\mathbf{X} = \begin{array}{cc|c} \text{columns} & \text{rows} & \\ \hline & \textit{augmented row profile} & \text{rows} \\ \textit{augmented column profile} & & \text{columns} \end{array}$$

with

- $\forall\, i = 1, \ldots, p$ and $\forall\, j = q+1, \ldots, q+p$, $\mathbf{x}_{ij} = \mathbf{x}_{k(i)+p,j}$ with $k(i) = \arg\max_{k=1,\ldots,q} \mathbf{x}_{ik}$

Extensions to non vectorial data 1

KORRESP [Cottrell et al., 1993]

Data: contingency table $\mathbf{T} = (n_{ij})_{ij}$ with p rows and q columns transformed into a numeric dataset $\mathbf{X}$:

$$\mathbf{X} = \begin{array}{cc}
\text{columns} & \text{rows} \\
\left[\begin{array}{c|c}
\textit{row profile} & \textit{augmented row profile} \\
\hline
\textit{augmented column profile} & \textit{column profile}
\end{array}\right] & \begin{array}{l} \text{rows} \\ \text{columns} \end{array}
\end{array}$$

- affectation uses reduced profile
- representation uses augmented profile
- alternatively process row profiles and column profiles

Extensions to non vectorial data 2

Relational SOM

[Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ not necessarily vectorial)

Extensions to non vectorial data 2

Relational SOM

[Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ not necessarily vectorial)

Adaptations of the SOM algorithm:

- prototypes: expressed as (symbolic) convex combination of $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ and $\sum_i \gamma_{ui} = 1$

Extensions to non vectorial data 2

Relational SOM

[Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ not necessarily vectorial)

Adaptations of the SOM algorithm:

- prototypes: expressed as (symbolic) convex combination of $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ and $\sum_i \gamma_{ui} = 1$
- distance computation: $\|x_i - p_u\|^2$ replaced by

$$(\mathbf{D}\gamma_u)_i - \frac{1}{2}\gamma_u^T \mathbf{D}\gamma_u$$

in reference to a pseudo-Euclidean framework [Goldfarb, 1984]

Extensions to non vectorial data 2

Relational SOM

[Hammer and Hasenfuss, 2010, Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ not necessarily vectorial)

Adaptations of the SOM algorithm:

- prototypes: expressed as (symbolic) convex combination of $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ and $\sum_i \gamma_{ui} = 1$
- distance computation: $\|x_i - p_u\|^2$ replaced by

$$(\mathbf{D}\gamma_u)_i - \frac{1}{2}\gamma_u^T \mathbf{D}\gamma_u$$

  in reference to a pseudo-Euclidean framework [Goldfarb, 1984]
- representation: replaced by an update of $(\gamma_u)_u$:

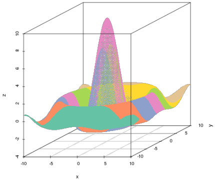
$$\gamma_u^{t+1} \leftarrow \gamma_u^t + \mu(t) H^t(d(\mathcal{C}(x_i), u)) \left(\mathbf{1}_i - \gamma_u^t\right)$$

  with $\mathbf{1}_{il} = 1$ if $l = i$ and 0 otherwise.

Outline

1 a short review of Self-Organizing Maps for non vectorial data

2 **SOMbrero**



What is it?

- **SOMbrero** is an R package implementing stochastic variants of SOM for non vectorial data (see **yasomi** for batch versions)
- first release: March 2013; latest release: November 2013 (version 0.4-1)
- depends on R (version $\geq$ 3.0) http://www.r-project.org  and on several packages available on CRAN:

  wordcloud, **igraph**, **RColorBrewer**, **scatterplot3d**, **knitr**, **shiny**
- available at http://sombrero.r-forge.r-project.org (licence GPL) and can be installed from inside R using

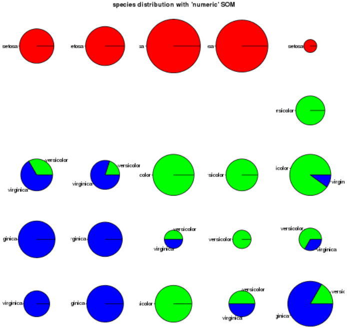
```
install.packages("SOMbrero",
  repos="http://R-Forge.R-project.org")
```

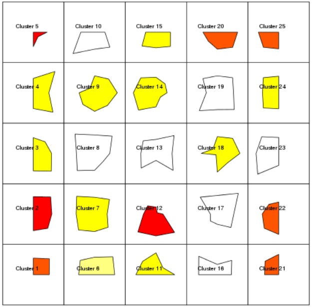
Features

1. 3 algorithms available through one function `trainSOM`
   - numeric SOM (input: $(n \times p)$-matrix with n observations of p variables)
   - KORRESP (input: $(p \times q)$-contingency table)
   - relational SOM (input: $(n \times n)$-dissimilarity matrix for n individuals)

Features

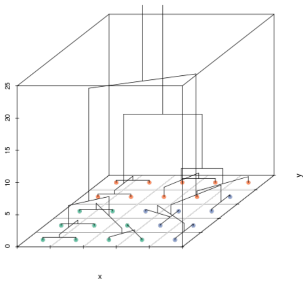
1. 3 algorithms available through one function `trainSOM`
2. many graphics available through one function `plot` with two main arguments `what` (prototypes, observations, additional variable) and `type` (`color`, `3d`, `barplot`, `poly.dist`, `words`, `pie`...)

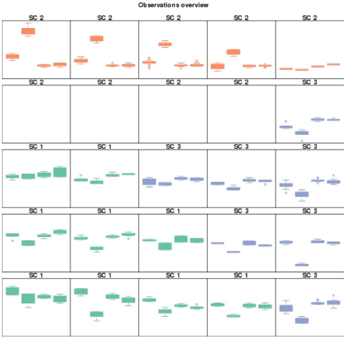




Features

1. 3 algorithms available through one function `trainSOM`
2. many graphics
3. super-clustering (HC on prototypes) with associated graphics through functions `superClass` and `plot`





Features

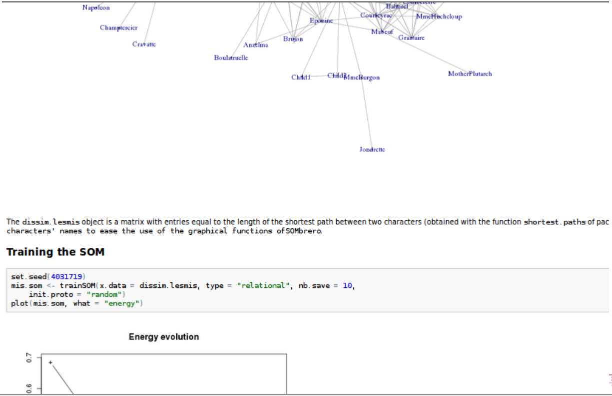
1. 3 algorithms available through one function `trainSOM`
2. many graphics
3. super-clustering (HC on prototypes) with associated graphics
4. quality measures (quantization error, topographic error) with `quality`

Start with SOMbrero

- 3 datasets corresponding to the three algorithms (`iris`, `presidentielles2002` and `lesmis`, a graph from « Les Misérables »)

Start with SOMbrero

- 3 datasets corresponding to the three algorithms (`iris`, `presidentielles2002` and `lesmis`, a graph from « Les Misérables »)
- comprehensive (HTML) vignettes included in the package and available on the website



The `dissim.lesmis` object is a matrix with entries equal to the length of the shortest path between two characters (obtained with the function `shortest.paths` of pac characters' names to ease the use of the graphical functions of SOMbrero.

Training the SOM

```
set.seed(4031719)
mis.som <- trainSOM(x.data = dissim.lesmis, type = "relational", nb.save = 10,
    init.proto = "random")
plot(mis.som, what = "energy")
```

Start with SOMbrero

- 3 datasets corresponding to the three algorithms (`iris`, `presidentielles2002` and `lesmis`, a graph from « Les Misérables »)
- comprehensive (HTML) vignettes included in the package and available on the website
- Web User Interface (made with **shiny**) for using the package even if you do not know R programming language (included in the package or available online at `http://shiny.nathalievilla.org/sombrero` but can be very slow)

The demo...

Let's go into **SOMbrero**...

disclaimer: as is standard during demos, something nasty might happen and nothing would work due to some weird technical issues... and the speaker would look like an idiot!

Conclusion

SOMbrero

- is easy to use (with a simple graphical interface)
- can be used with various data
- contains many tools for interpreting the results

INRA SCIENCE & IMPACT

Conclusion

SOMbrero

- is easy to use (with a simple graphical interface)
- can be used with various data
- contains many tools for interpreting the results
- ... has unfortunately been implemented by girls so default colors may not be suited for men (but they can easily change them)

Perspectives

- speed up the code
- more quality criteria
- more options (i.e., Gaussian neighborhood, weighted observations...)

Thank you for your attention...

... questions?

Cottrell, M., Letrémy, P., and Roy, E. (1993).
Analyzing a contingency table with Kohonen maps: a factorial correspondence analysis.
In Cabestany, J., Mary, J., and Prieto, A. E., editors, *Proceedings of International Workshop on Artificial Neural Networks (IWANN 93)*, Lecture Notes in Computer Science, pages 305–311. Springer Verlag.

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

Hammer, B. and Hasenfuss, A. (2010).
Topographic mapping of large dissimilarity data sets.
Neural Computation, 22(9):2229–2284.

Kohonen, T. (2001).
Self-Organizing Maps, 3rd Edition, volume 30.
Springer, Berlin, Heidelberg, New York.

Olteanu, M. and Villa-Vialaneix, N. (2014).
On-line relational and multiple relational SOM.
Neurocomputing.
Forthcoming.