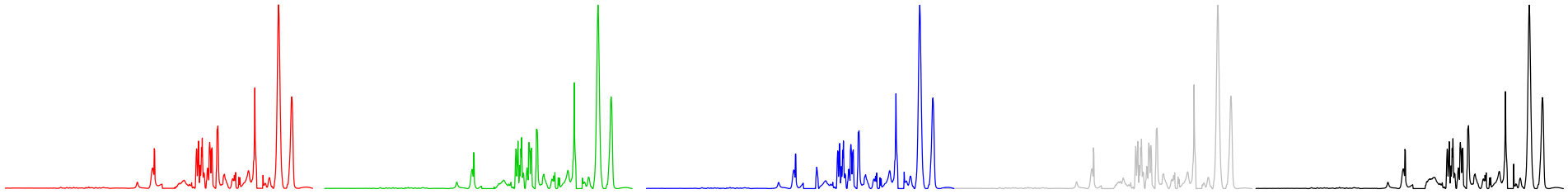
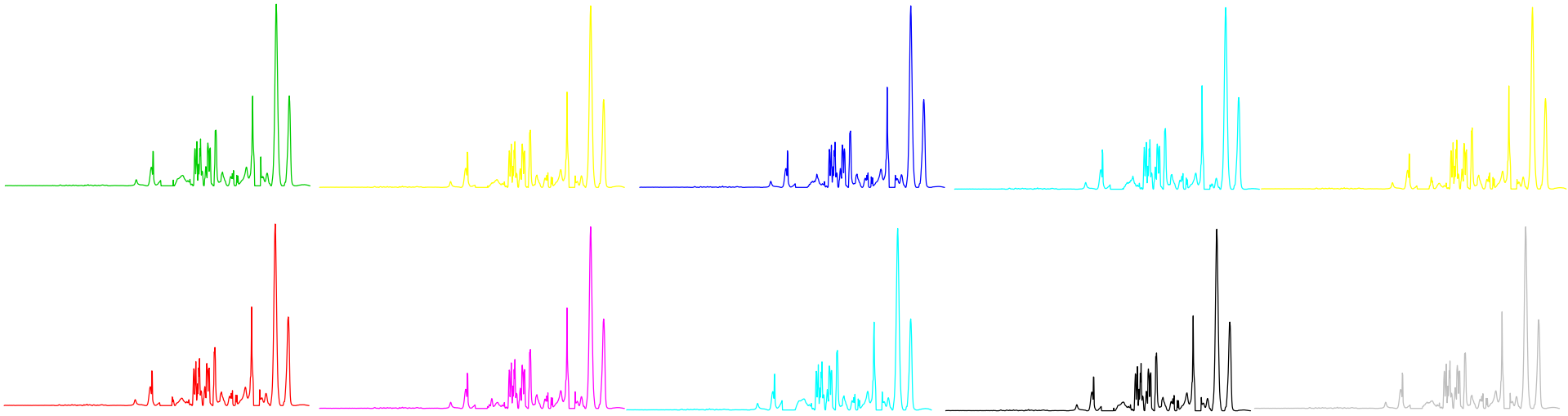


BAGIDIS

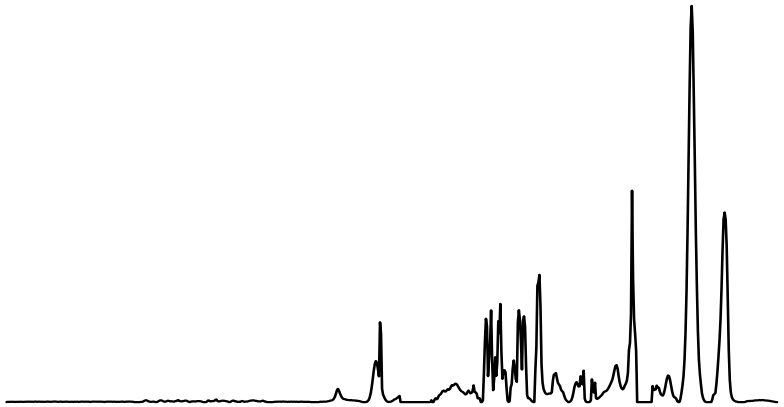
A new way of measuring distances between curves with sharp peaks.

Catherine Timmermans



Our dataset

- 24 serum spectra, collected using H-NMR, and pre-processed.



Ref: *De Tullio P, Frédérich M, Université de Liège; Rousseau R, Université catholique de Louvain.*

3 factors:

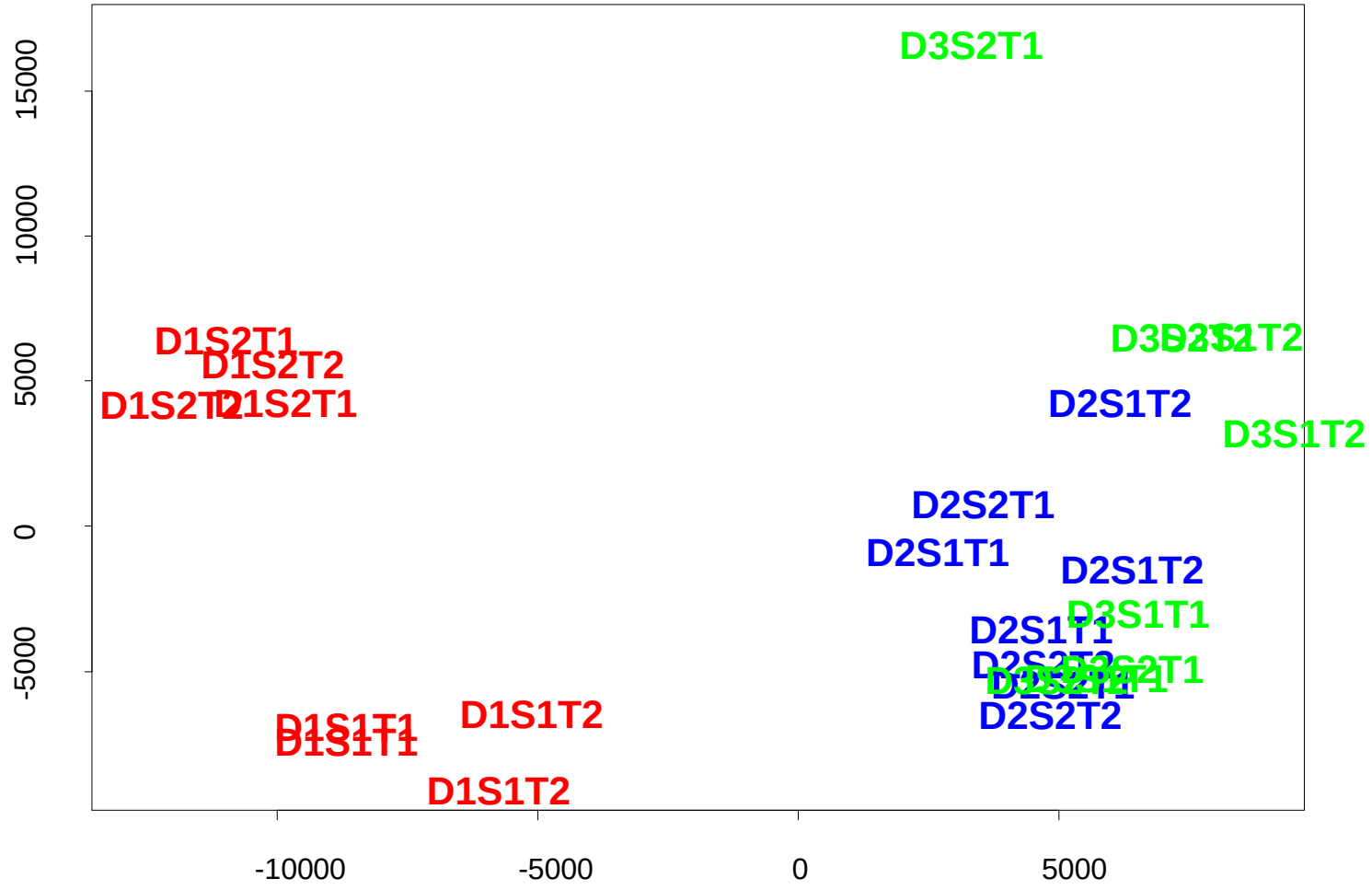
- Sample: S1,S2
- Day: D1,D2,D3
- Time: T1,T2

+ 1 repetition of the trials.

→ Can we identify an effect of those factors ?

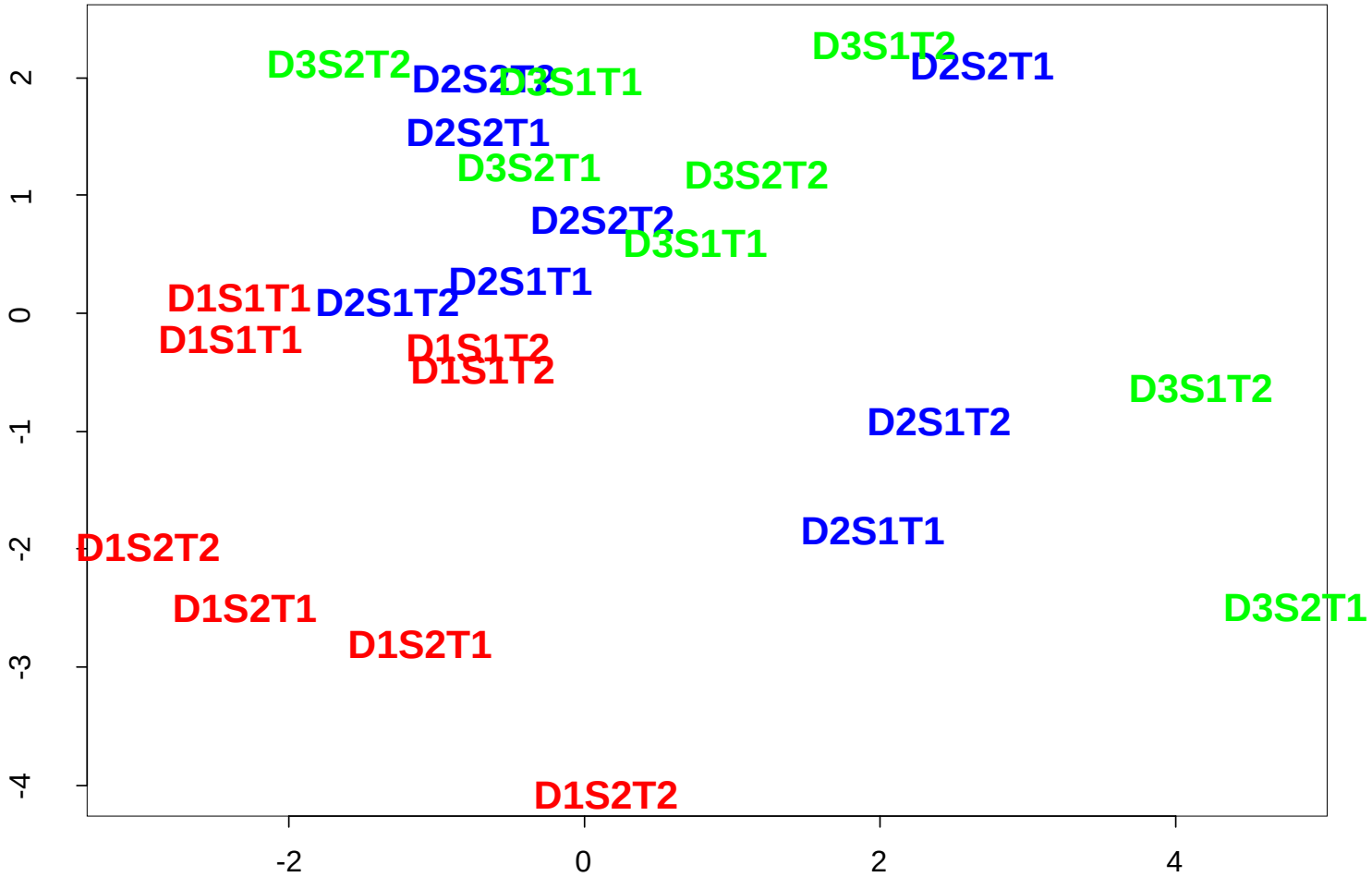
Visual analysis using BAGIDIS

- (Semi-)distances evaluation using **the BAGIDIS method**.
- *Multidimensional scaling* representation.



Visual analysis using “classical” methods

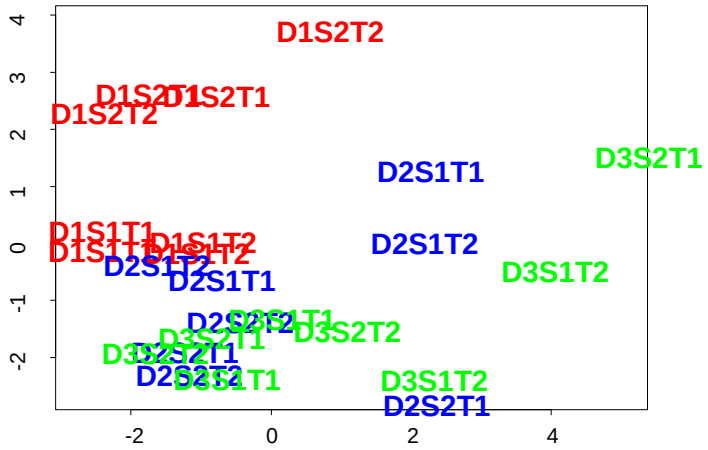
- Projection on the first factorial plane of principal component analysis = *Multidimensional scaling* using the L2-distance .



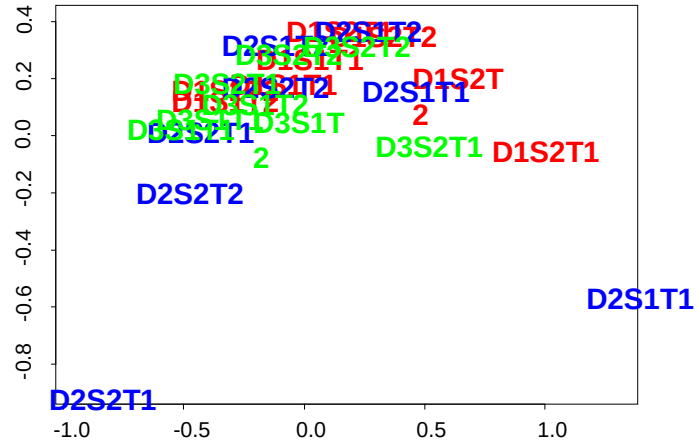
Visual analysis using “classical” methods

- *Multidimensional scaling* using some famous dissimilarities and distances….

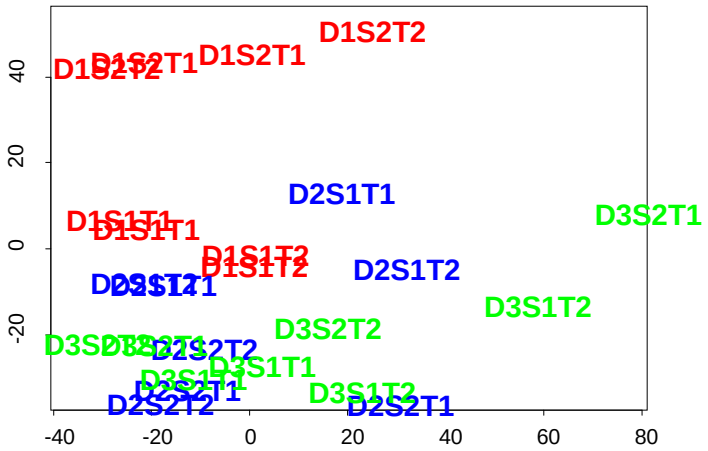
- **L1- Distance**



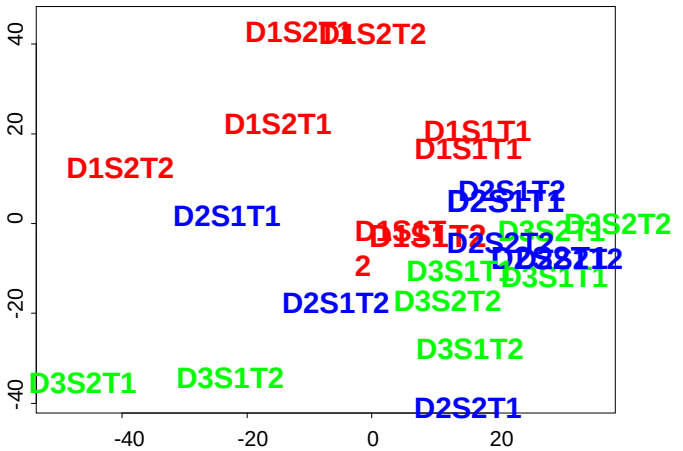
- **L2-Distance between derivatives** *[Ferraty,Vieu, 2006]*



- **Dynamic time warping**



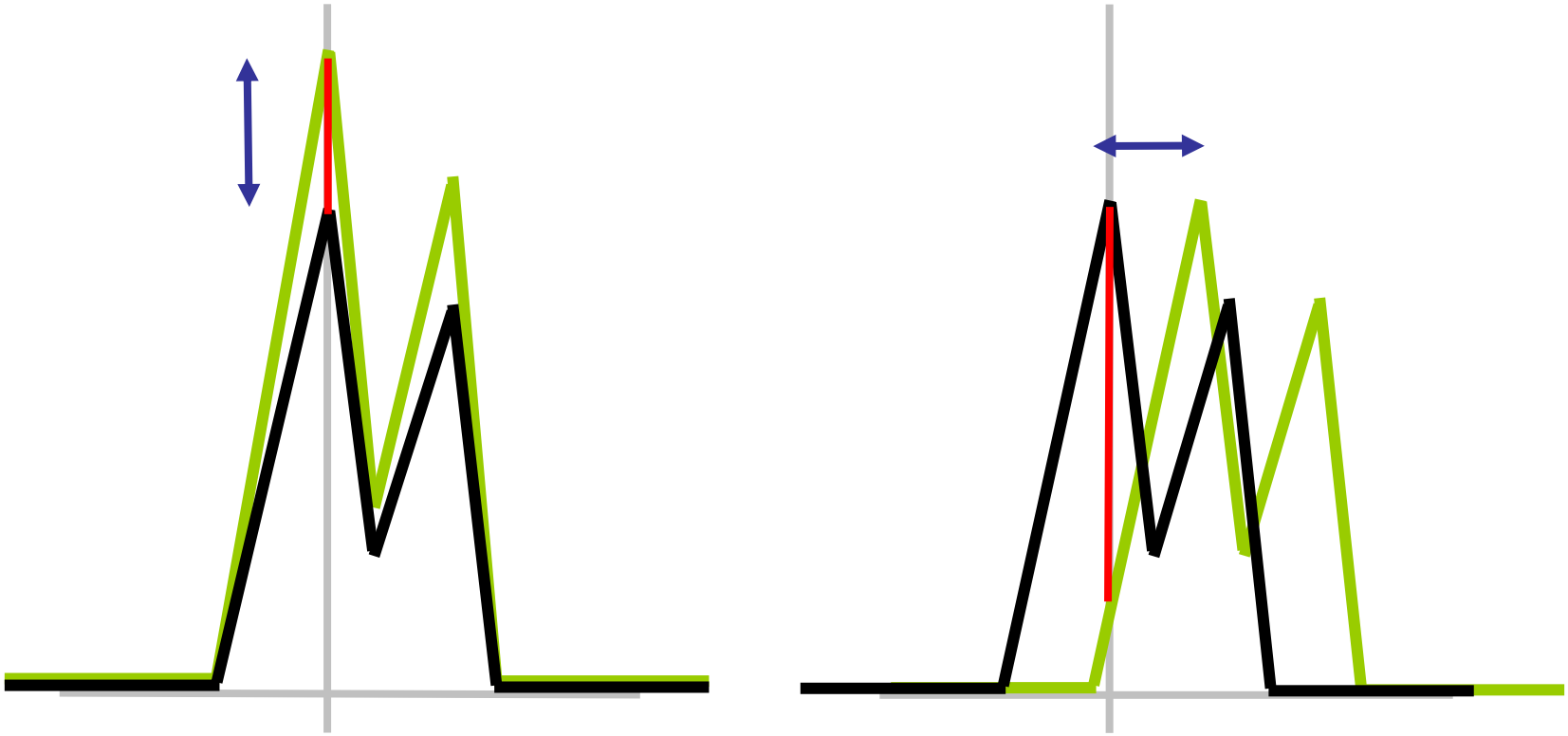
- **L2-Distance between functional principal components** *[Ferraty,Vieu,2006]*



- **Motivation**
- **The BAGIDIS methodology**
- **Discussions, variations and extensions**

What happens with « classical » methods

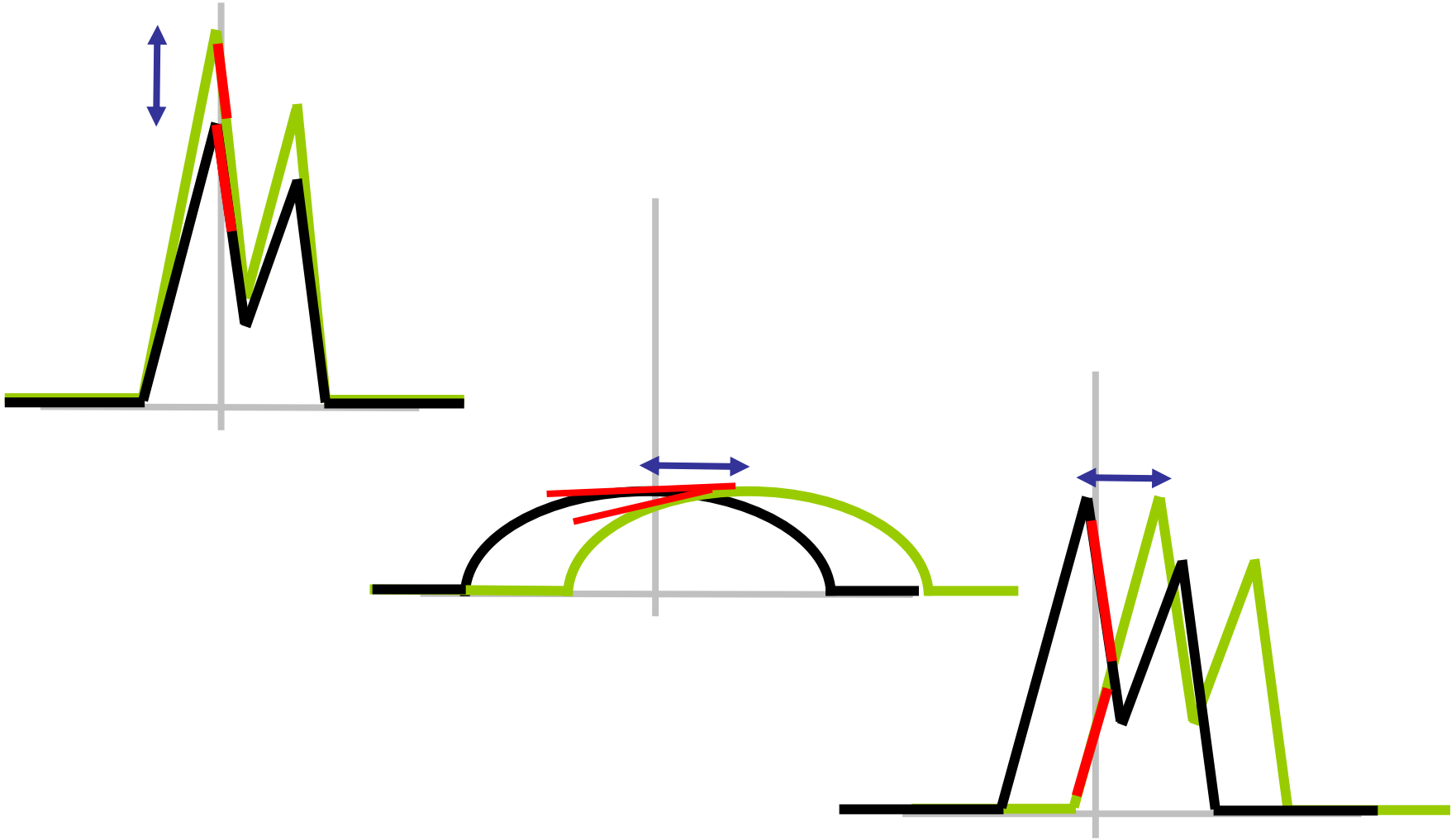
- Numerous « classical » methods *(Distance L2, Distance L1, PCA…)* compare curves at each given abscissa…



- Curves similar but a vertical variation are measured close.
- Curves similar but an horizontal variation are measured distant.

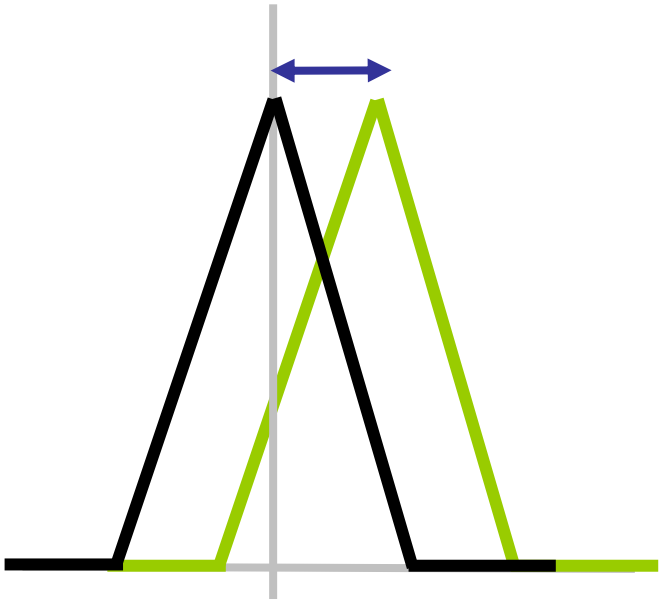
What happens with « classical » methods

- *Functional* methods help to overcome that difficulty….
  … **as long as** the peaks are not too sharp.

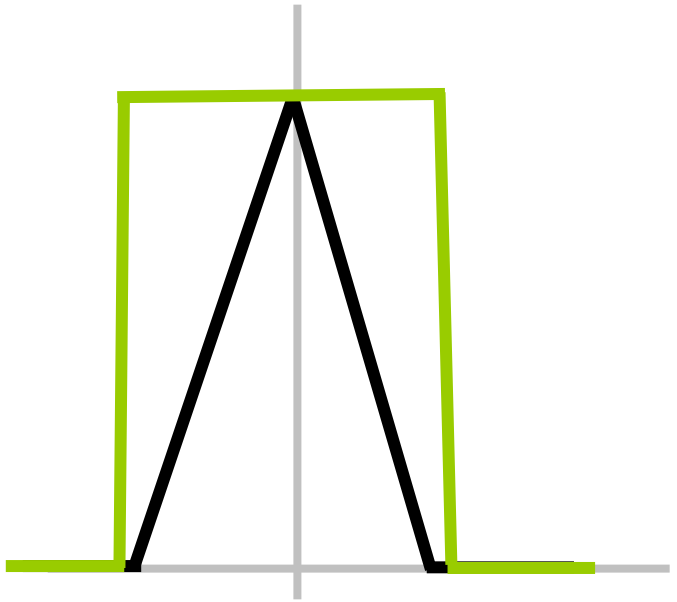


What happens with « classical » methods

- *Dynamic time warping* allows for speed variations along the X-axis so that the distance between Y-values is minimum.



$Dist_{dtw} = 0$

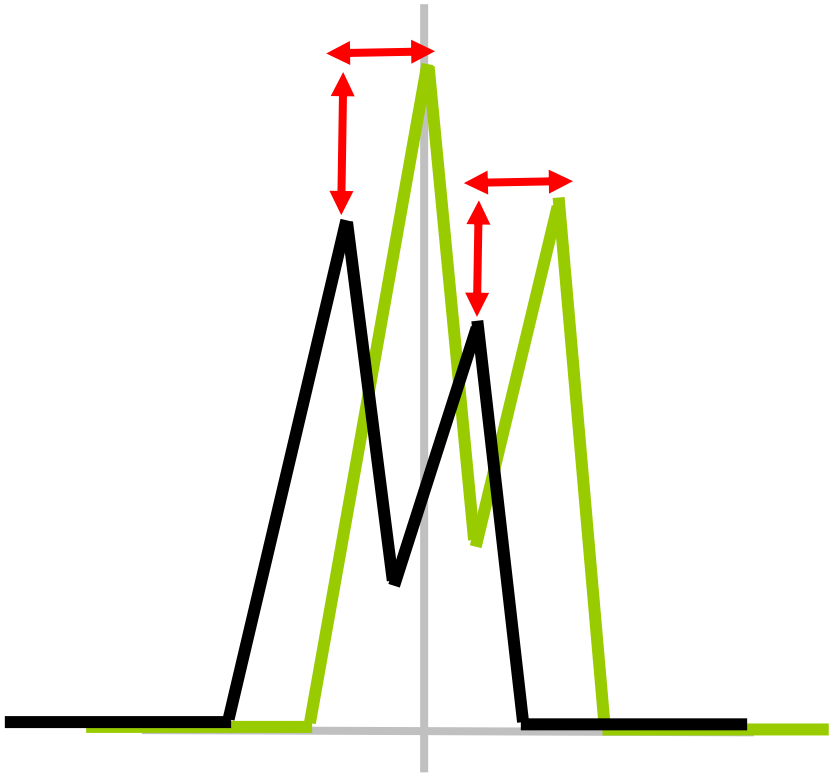


$Dist_{dtw} = 0$

- Curves similar but an horizontal shift are measured identical.
- Curves whose shapes are different and amplitudes are similar are measured identical.

What the BAGIDIS method does propose

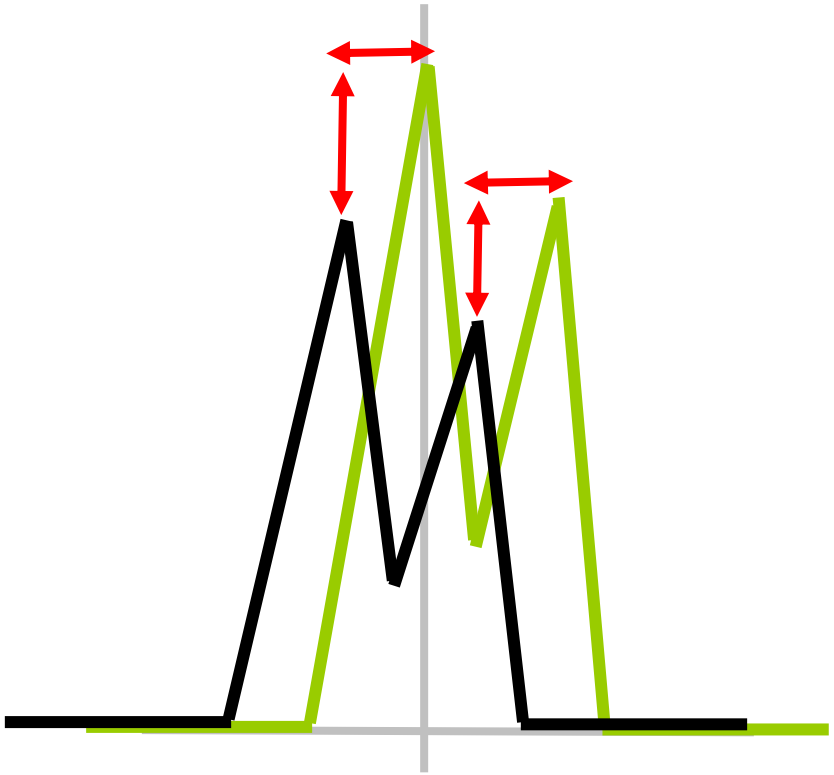
- Compare **shapes** of the curves by quantifying both their vertical and horizontal differences.



- **Motivation**
- **The BAGIDIS methodology**
- **Discussions, variations and extensions**

What the BAGIDIS method does propose

- Compare **shapes** of the curves by quantifying both their vertical and horizontal differences.



- Shapes can be described as level changes.

What the BAGIDIS method does propose

- **Find an “optimal” basis for each curve**

  The firsts basis vectors encode the most significant level changes while subsequent vectors refer to less important ones.

- **Take advantage of the ‘hierarchical property’ of those bases**

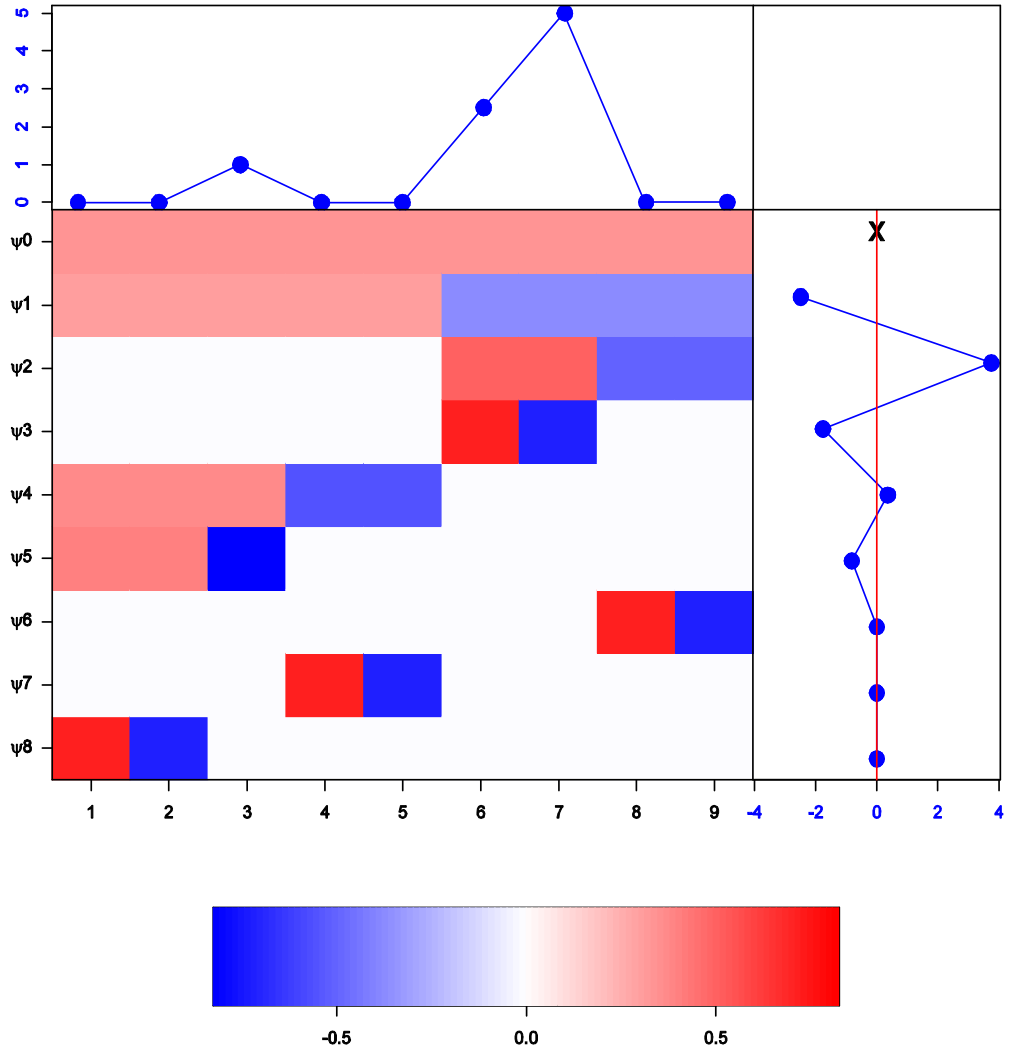
  Compare basis vectors of **similar importance** and their **associated coefficients**.

  Associate **large weight** to comparisons of basis vectors encoding **significant level changes**.

BAGIDIS = **Ba**ses **Gi**ving **Dis**tances.

Unbalanced Haar wavelets expansion

[Girardi & Sweldens, 1997]



$$x^{(i)} = \sum_{k=0}^{N-1} d_k^{(i)} \psi_k^{(i)}$$

Unbalanced Haar wavelets expansion

[Girardi & Sweldens, 1997]

- We can compute the unbalanced Haar wavelets basis that is best suited to a given series.

[Fryzlewicz, 2007]

- Let's note $\{ \psi_k^{(i)} \}$ the unbalanced Haar wavelets basis that is best suited to the series $x^{(i)}$:

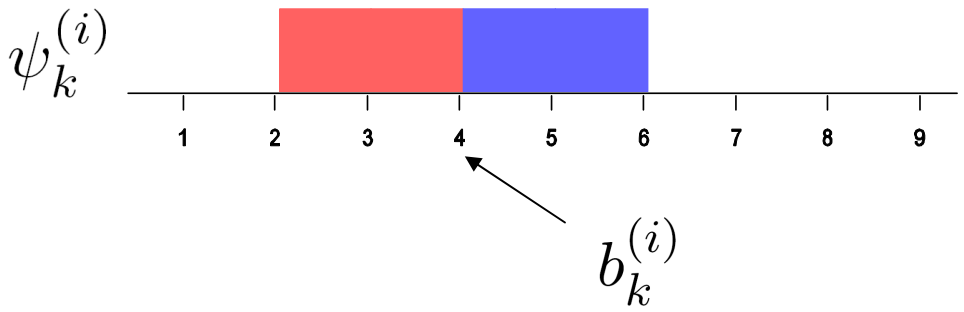
$$x^{(i)} = \sum_{k=0}^{N-1} d_k^{(i)} \psi_k^{(i)}$$

Unbalanced Haar wavelets expansion

[Girardi & Sweldens, 1997]

- Unbalanced Haar wavelets bases are orthonormal. The ordered set of the breakpoints $b_k^{(i)}$ determines the basis $\{ \psi_k^{(i)} \}$ uniquely.

[Fryzlewicz, 2007]



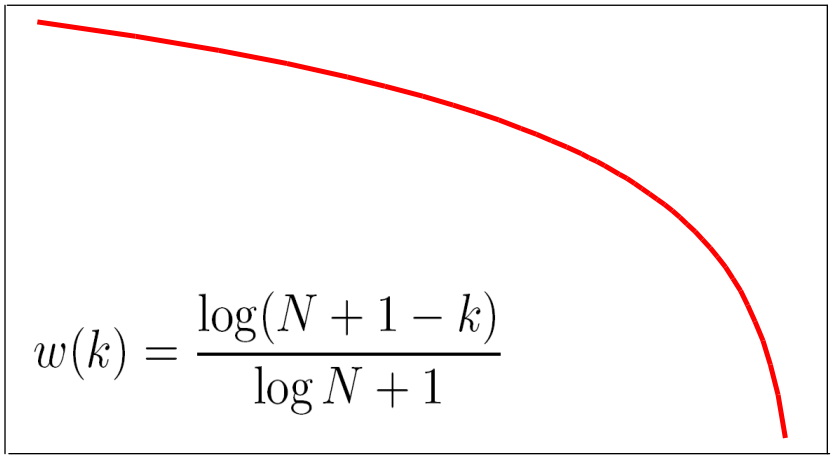
Definition of the distance

BAGIDIS = **Ba**ses **gi**ving **dis**tances.

- (Semi-)distance :

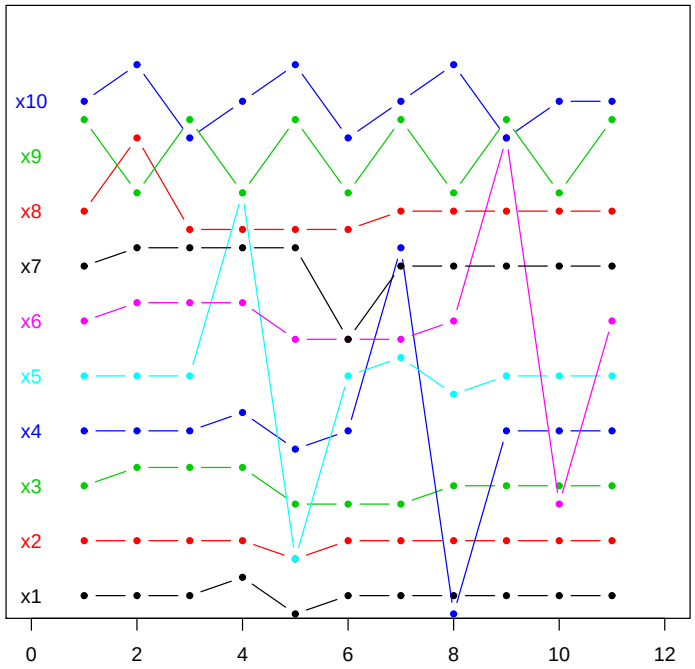
$$d(x^{(i)}, x^{(j)}) = \sum_{k=1}^{N-1} w_k \{|b_k^{(i)} - b_k^{(j)}| + |d_k^{(i)} - d_k^{(j)}|\}$$

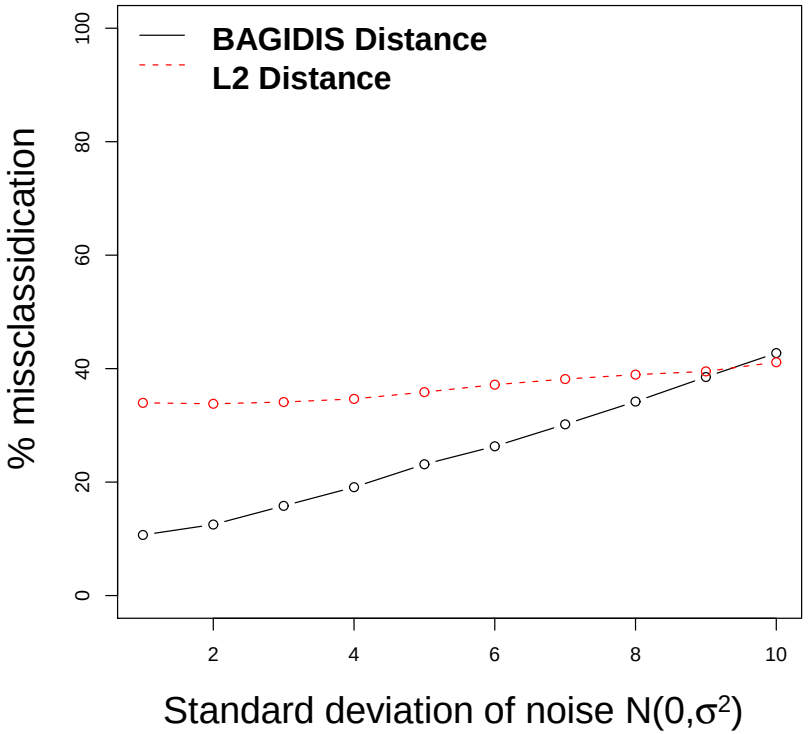
with w_k , decreasing weight function.



A first example

- Simulated example of supervised classification (*nearest neighbour* classification)





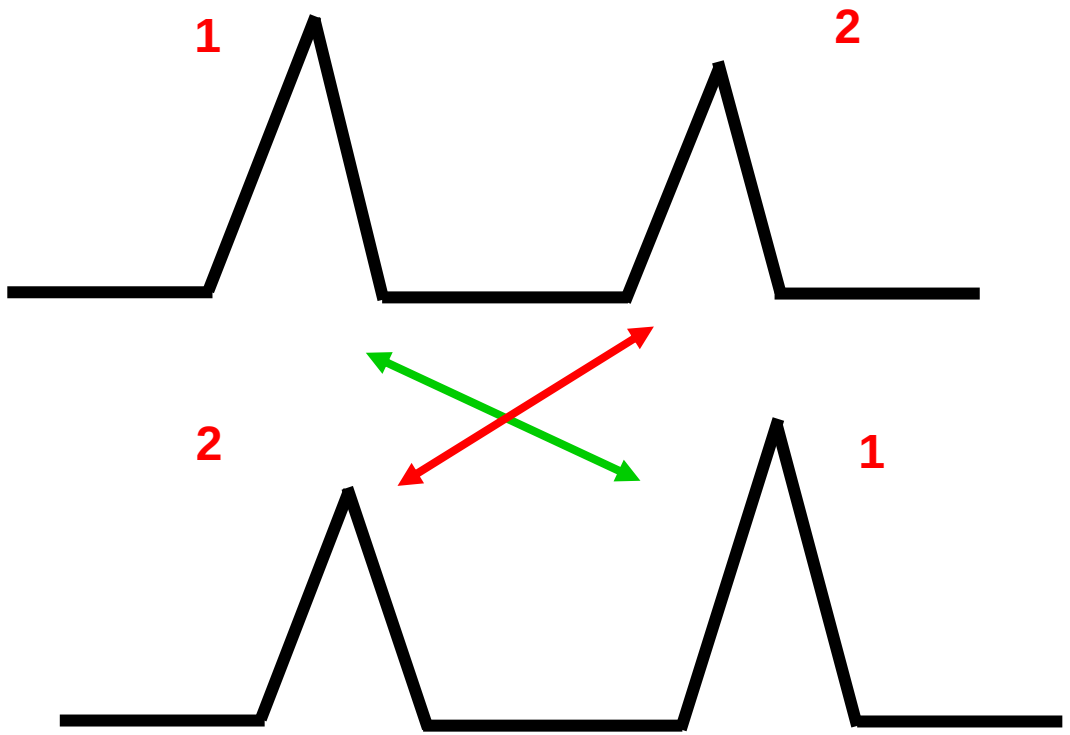
For each familly: 5 x 20 simulated series:

- noisy N(0,σ^2),
- shifted to the right and noisy N(0,σ^2),
- shifted to the left and noisy N(0,σ^2),
- multiplied by 1.25 and noisy N(0,σ^2),
- multiplied by 0.75 and noisy N(0,σ^2).

- **Motivation**
- **The BAGIDIS methodology**
- **Discussions, variations and extensions**

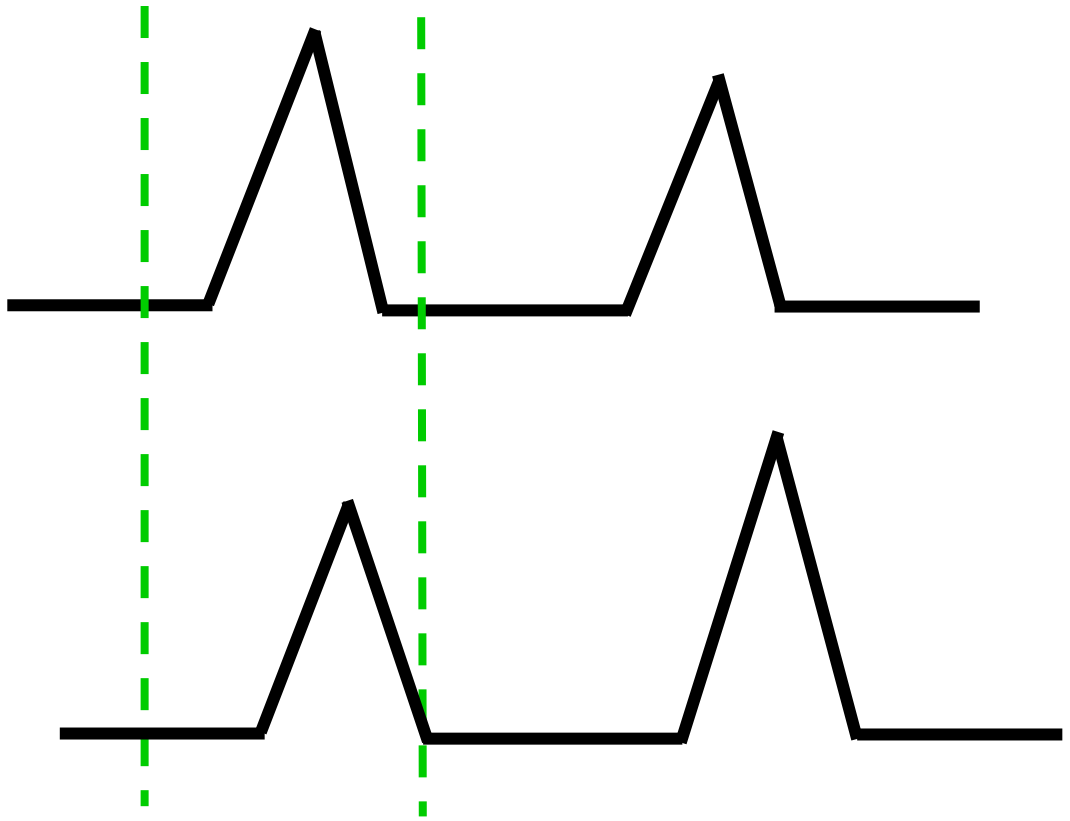
Avoiding features confusion in long series

- There are numerous patterns in long series.



Avoiding features confusion in long series

- Defining a *sliding distance*
- Global distance is the sum of the distances in every window.



- The choice of the length Δ of the window is problem-dependant.
- Experience shows little sensitivity to small variations of that parameter.

Being flexible w.r.t. scaling effects

- Finding the best balance between position and amplitude differences

- λ - Distance:

$$d(x^{(i)}, x^{(j)}) = \sum_{k=1}^{N-1} w_k \{\lambda . |b_k^{(i)} - b_k^{(j)}| + (1-\lambda) . |d_k^{(i)} - d_k^{(j)}|\}$$

with $\lambda \in [0; 1]$.

Being flexible w.r.t. scaling effects

- Finding the best balance between position and amplitude differences

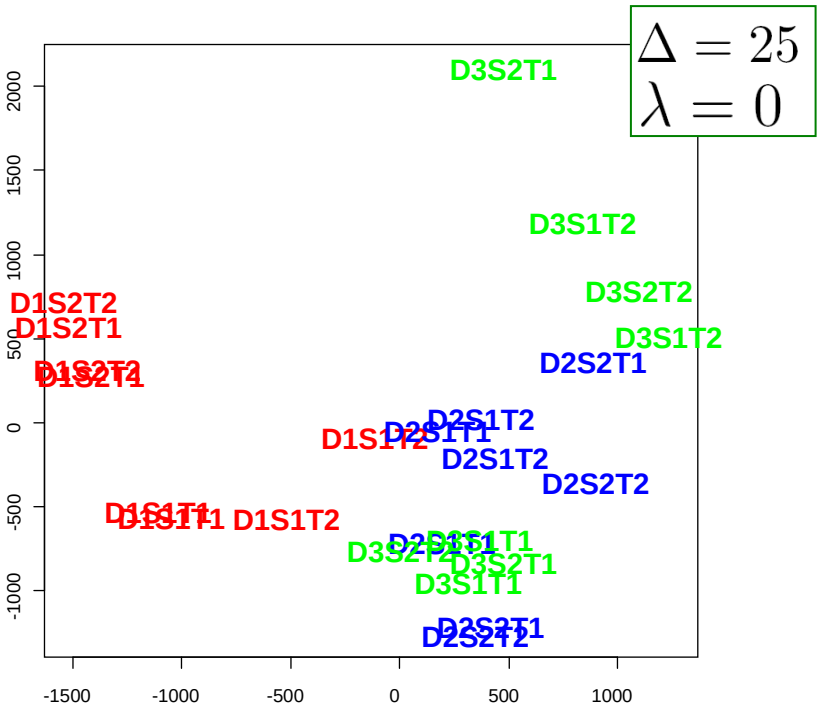
- Find best value of λ :

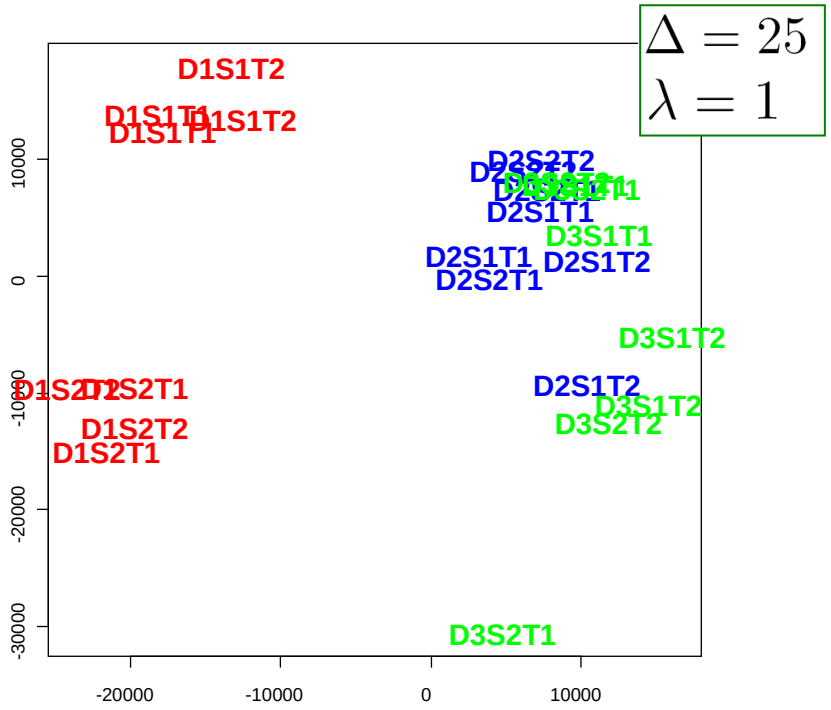
  Supervised problem : cross-validation

  Unsupervised problem: work in progress…

Using flexibility as a diagnostic tool

- Exploring the balance between position and amplitude
- *Multidimensional scaling* of the spectra using BAGIDIS method





$$d(x^{(i)}, x^{(j)}) = \sum_{k=1}^{N-1} w_k \{|d_k^{(i)} - d_k^{(j)}|\}$$

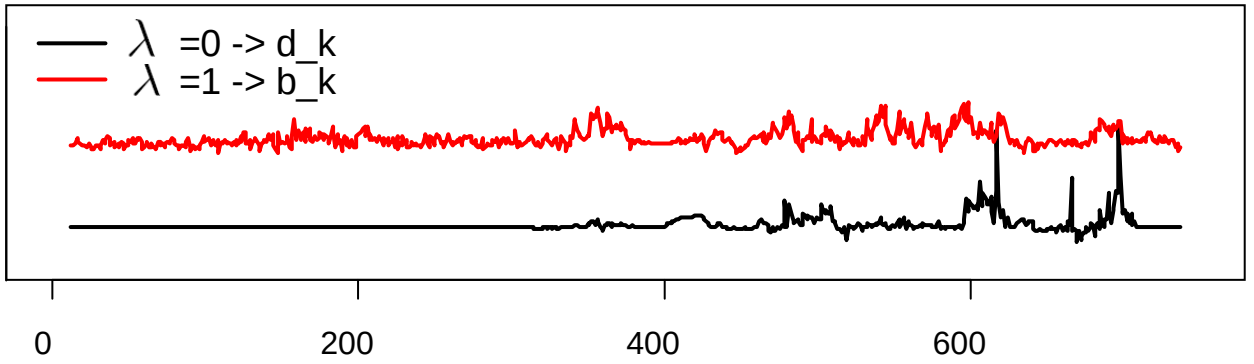
$$d(x^{(i)}, x^{(j)}) = \sum_{k=1}^{N-1} w_k \{|b_k^{(i)} - b_k^{(j)}|\}$$

Using flexibility as a diagnostic tool

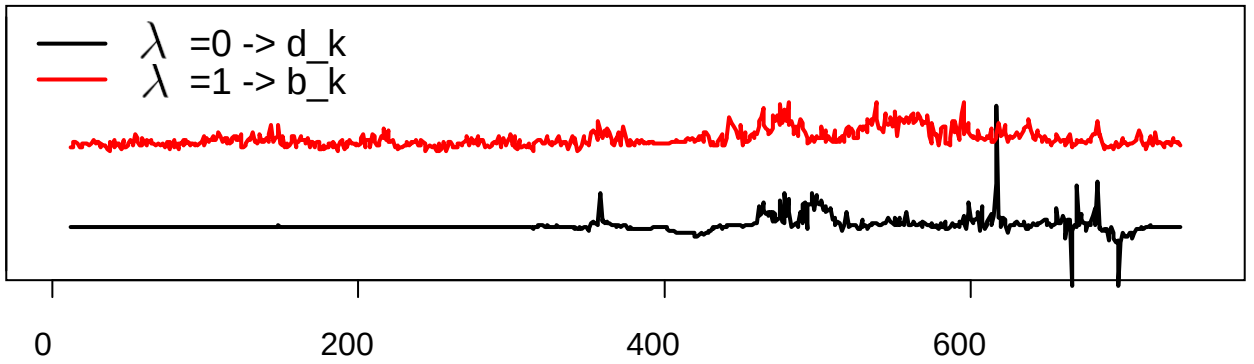
- Exploring the balance between position and amplitude
- Analysing differences between spectra measured on D1 and D3.

Discussions, variations and extensions

- Mean distance D1 x D3 – mean distance D1 x D1



- Mean distance D1 x D3 – mean distance D3 x D3



Using flexibility as a diagnostic tool

- Exploring the balance between position and amplitude
- Analysing differences between spectra measured on D1 and D3.

Discussions, variations and extensions

- Significance of a t-test on the means (correction of p-values: Bonferroni)

$\Delta = 25$
$\lambda = 0$

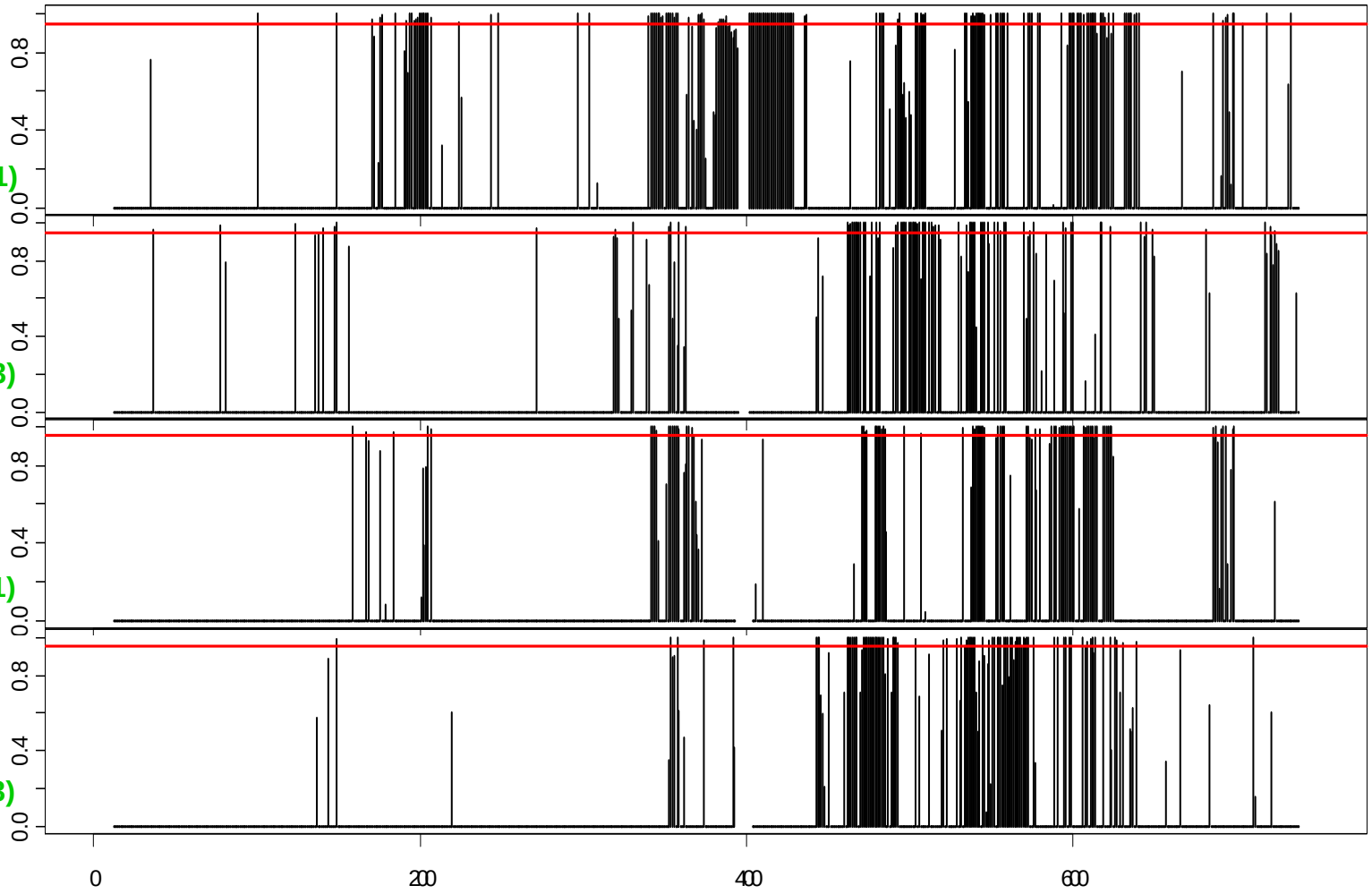
d(D1,D3) > d(D1,D1)

d(D1,D3) > d(D3,D3)

$\Delta = 25$
$\lambda = 1$

d(D1,D3) > d(D1,D1)

d(D1,D3) > d(D3,D3)



- **Motivation**
- **The BAGIDIS methodology**
- **Discussions, variations and extensions**

BAGIDIS

Bases **Gi**ving **Dis**tances

Conclusion

- **Finding an "optimal basis" for each curve:**

  Using unbalanced Haar wavelet expansion.

- **Taking advantage of the ' hierarchical property' of those bases:**

  Comparing **basis vectors** of similar importance and their **associated coefficients,** with a weighting decreasing with the rank .

$$d(x^{(i)}, x^{(j)}) = \sum_{k=1}^{N-1} w_k \{|b_k^{(i)} - b_k^{(j)}| + |d_k^{(i)} - d_k^{(j)}|\}$$

- **Using a sliding distance** when comparing 'long' series.

- **Explore the balance between position and amplitude** as a diagnostic tool and to be flexible w.r.t scaling effects.

References:

- FERRATY F., VIEU P., *Nonparametric Functional Data Analysis : Theory and Practice,* Springer Series in Statistics, 2006.
- FRYZLEWICZ P., *Unbalanced Haar technique for non parametric function estimation*, J. Am. Stat. Assoc., 2007.
- GIRARDI M., SWELDENS W., *A new class of unbalanced Haar wavelets that form an unconditional basis for Lp on general measure spaces*, J. Fourier Anal. Appl., 1997.

Dataset :

- DE TULLIO P., FREDERICH M., Centre Intrafacultaire de Recherche du Médicament, Laboratoire de Pharmacognosie et de Chimie Pharmaceutique, Université de Liège, 2009.
- ROUSSEAU R., Institut de Statistique, Université Catholique de Louvain, 2009.

Software:

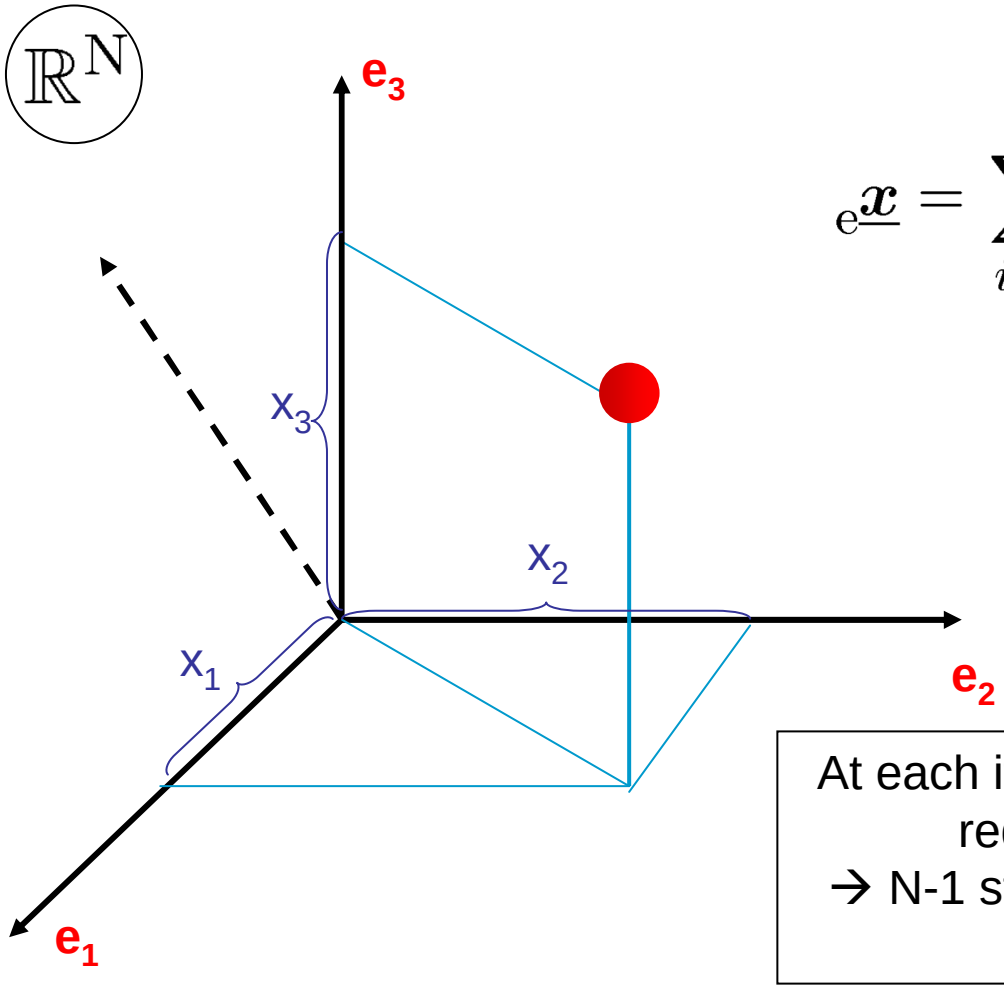
- R DEVELOPMENT CORE TEAM, *R : A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008.

Appendix

Bottom-up unbalanced Haar wavelets basis.

A data reduction algorithm as a first step to basis construction.

→ **Step 0: Series in the canonical basis $\{e_i\}$ (i=1..N) of R^N**



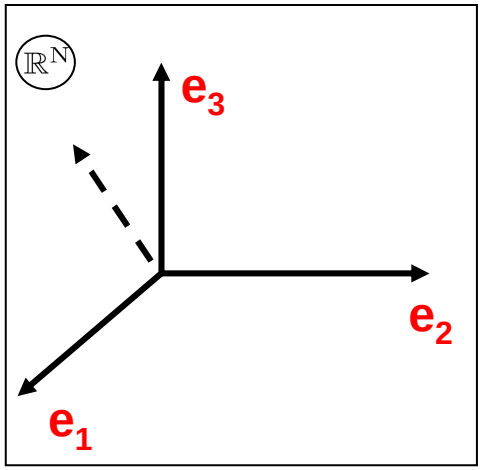
$$\underline{x} = \sum_{i=1}^{N} x_i \underline{e_i} = \sum_{i=1}^{N} \langle \underline{x} \mid \underline{e_i} \rangle \cdot \underline{e_i}$$

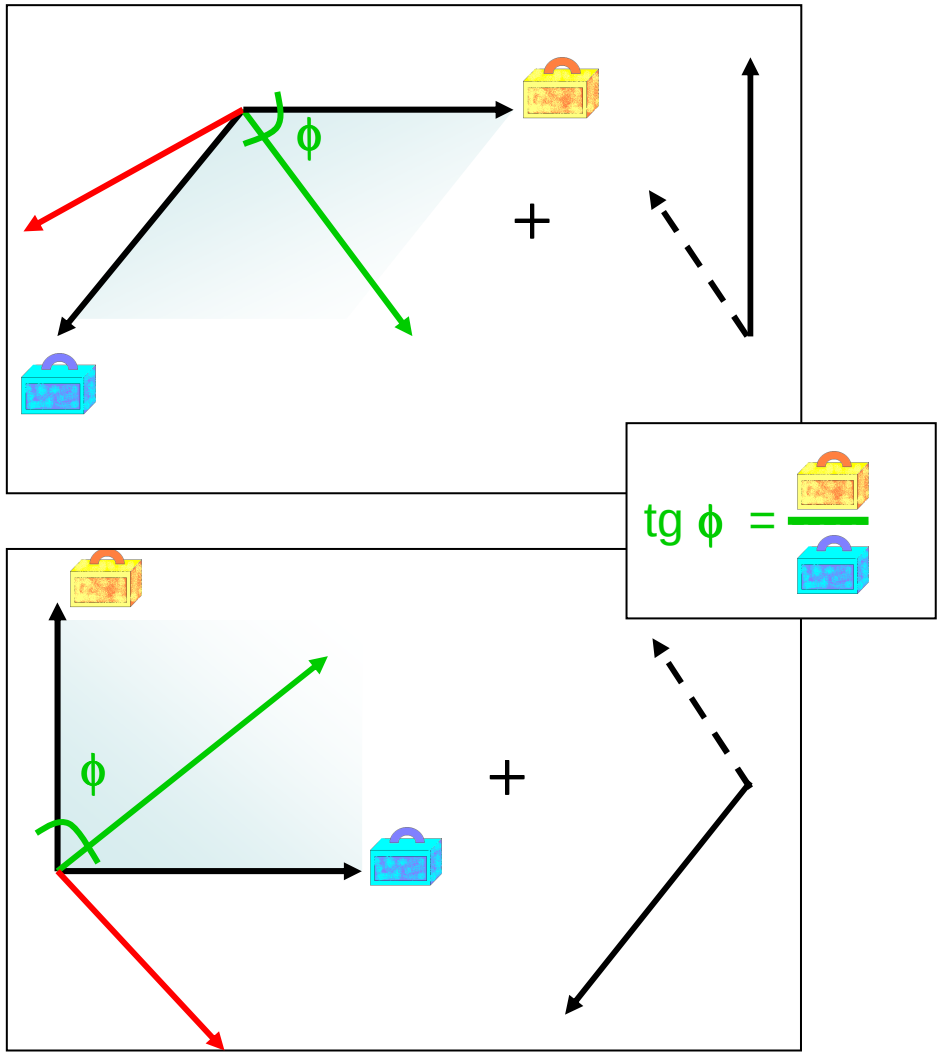
At each iteration of the algorithm, we will reduce the dimension by 1.
→ N-1 steps for a reduction to a unique scalar value.

A data reduction algorithm as a first step to basis construction.

We want to summarize 1 “well chosen” plane by 1 “well chosen” axis.

→ **Step 1: Define optimal summary-axis in each plane $\{e_i, e_{i+1}\}$.**





.

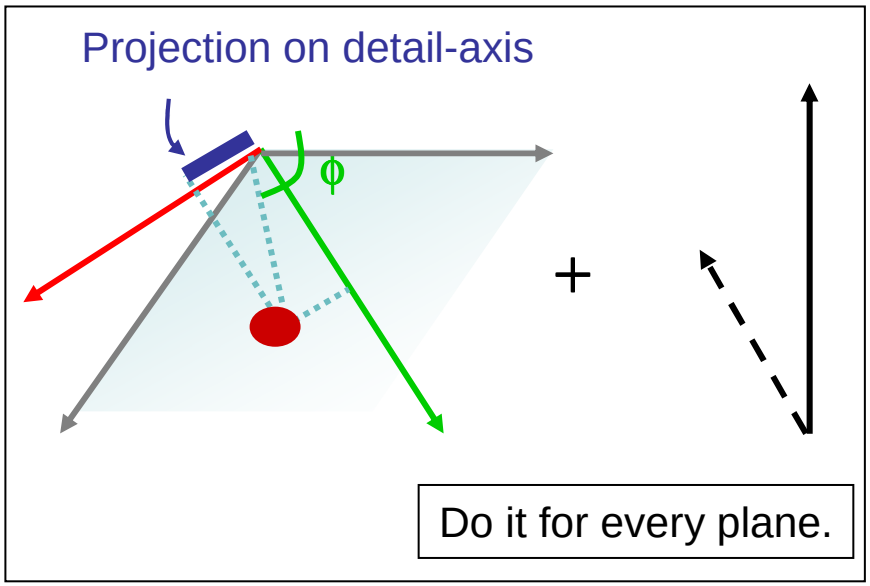
In each plan, there is 1 optimal summary-axis: (↗),
and 1 optimal detail-axis: (↗)

It depends on the importance - weights given “a priori” - (🧰 ‘s) of the information carried by the original axes.
First, this importance is = 1 for all axes.

A data reduction algorithm as a first step to basis construction.

We want to summarize 1 “well chosen” plane by 1 “well chosen” axis.

→ Step 2: Select the plane you will actually summarized.



You should lose as less information as possible when summarizing 1 plan by 1 axis.

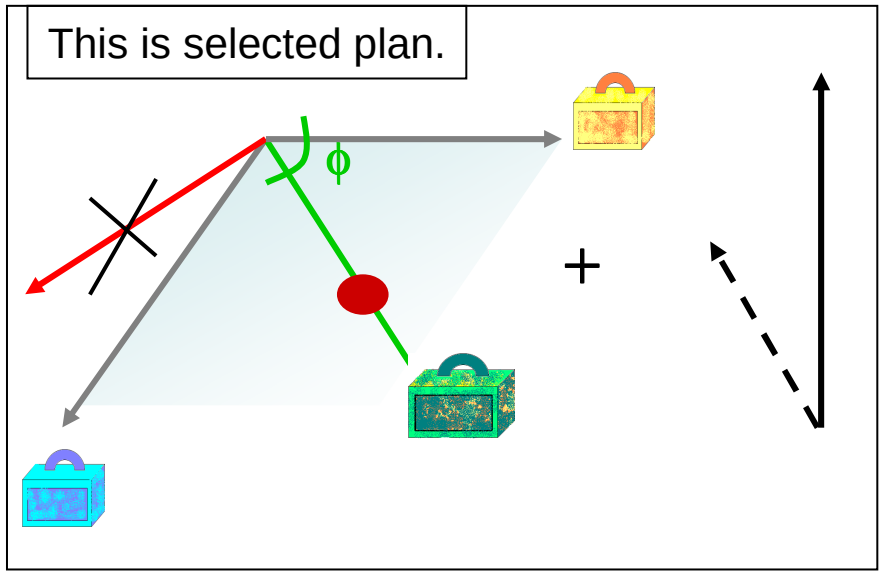
→ Compare projections of the data on all the possible detail-axis ( ) and select the plane were this projection is minimum.

C. Timmermans

A data reduction algorithm as a first step to basis construction.

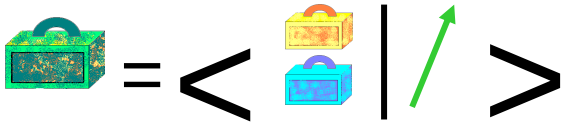
We want to summarize 1 “well chosen” plan by 1 “well chosen” axis.

→ Step 3: Reduce data, adapt weights.



- In the selected plane, suppress the detail- axis. Keep only summary-axis –and the projection of the data on it.

• Compute weight for the summary-axis.



• Others axes remain unchanged.

→Step 4: Back to Step 1 - until reduced series is a scalar (or until you’re satisfied of your data reduction).

From data reduction to basis definition.

Our new basis vectors are the “by-products” of our data reduction.

Unbalanced Haar wavelets

Series in the canonical basis $\{e_i\}$ (i=1..N) of R^N

→ **Step 1: Define optimal summary-axis in each plan $\{e_i, e_i+1\}$.**

→ **Step 2: Select the plan you will actually summarized.**

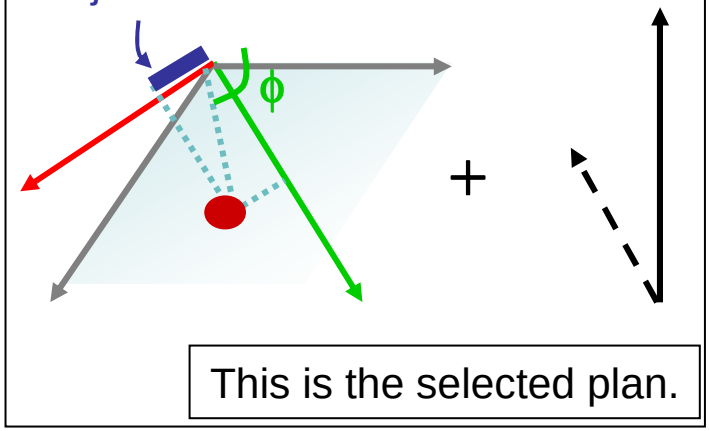
→ **Step 3: Reduce data, adapt weights.**

→ **Step 4: Back to Step 1. Until reduced series is a scalar.**

→ **Final step: a scalar.**

Iterate and go from R^N to R.

Store selected detail-vector and the projection of the data on it. This shall be a new basis vector and its associate coefficient.

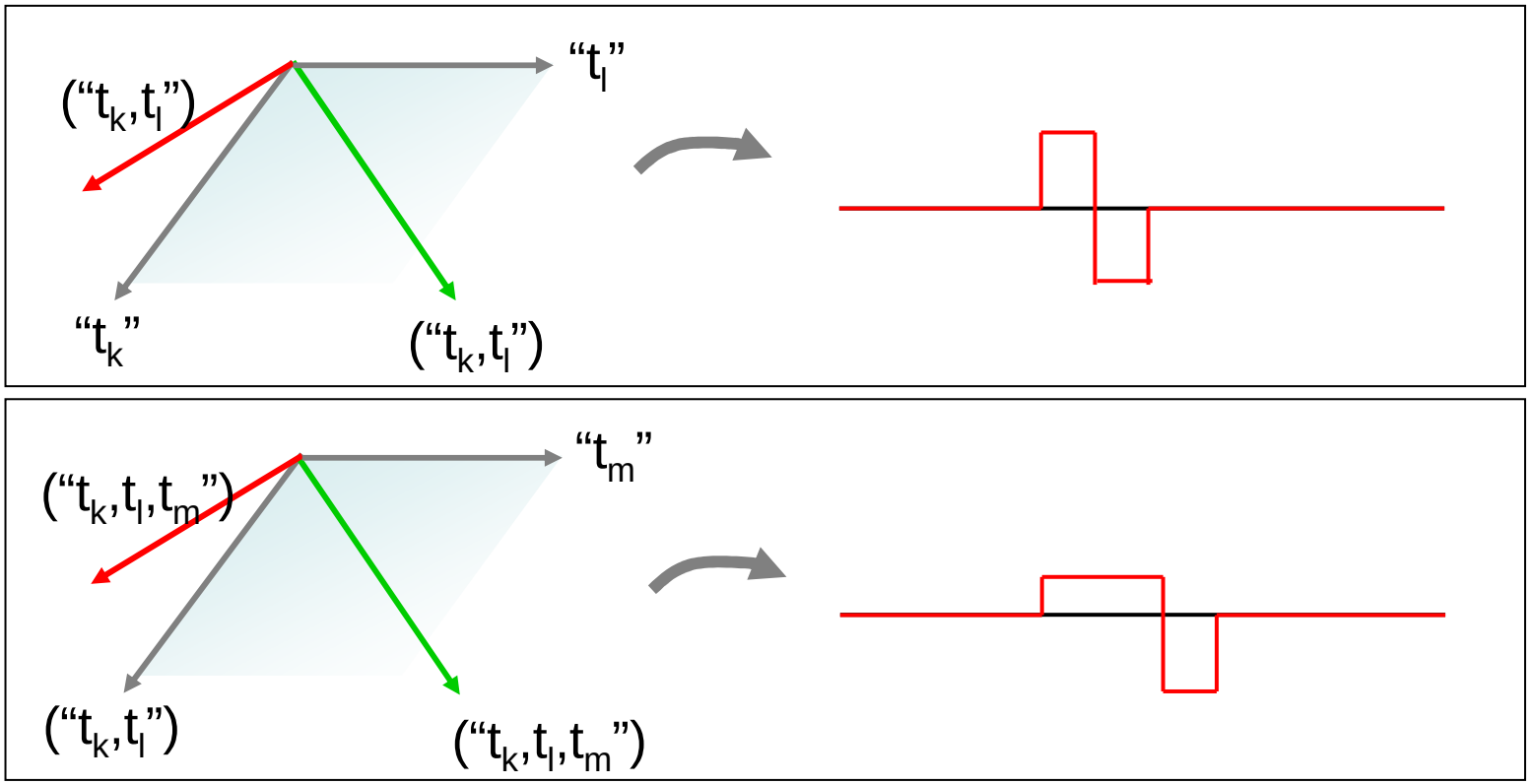


First basis vector is a constant and its associate coefficient is the remaining scalar.

Iterate and define basis vectors, starting from the last one.

Series in its unbalanced Haar wavelet basis of R^N.

Unbalanced Haar wavelet basis.
Interpretation, example and the wavelets.

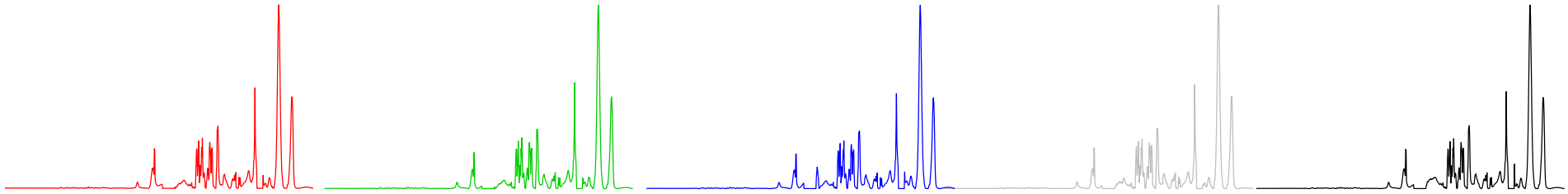


→Looks like Haar wavelets. One vector encode a difference between 2 adjacent values (or groups of values) .

But non-zero part is not symmetrical. This is the **“unbalanced” property**.

It allows us to overcome the “dyadic restriction” of traditional wavelets.

Our basis is truly adaptive – and remains orthonormal.



BAGIDIS

A new way of measuring distances between curves with sharp peaks.

Catherine Timmermans