PCA and Classification in Symbolic Context

SDA 2011, Beijing, China

Richard Emilion

MAPMO Lab., University of Orléans, France

October 27, 2011

- I - Symbolic PCA
- II - Unsupervised Classification of multihistograms
- III - Conclusion

I - Principal Component Analysis in symbolic context

PCA in $\mathbb{R}^p$

PCA in Hilbert space

• $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{R}^p, \alpha_i, i = 1, \ldots, n$: cloud of weighted vectors.

Affine line in $\mathbb{R}^p$ minimizing the distance between the cloud and its orthogonal projection on this line ?

Affine plane containing the above line minimizing the the distance between the cloud and its orthogonal projection on this plane ? etc

Unique solution: line $= (g, v_1)$, plane $= (g, v_1, v_2)$, etc ...

g : barycenter of the cloud, $v_1, v_2, \ldots, v_p$ orthogonal eigenvectors of the positive definite symmetric correlation matrix C of the p column vectors $y_j = (x_{i,j}) - g_j \in \mathbb{R}^n$ associated to the eigenvalues $\lambda_1 \geq \lambda_2, \ldots \geq \lambda_p \geq 0$ of C, resp.

• Linear method (linear algebra, linear space)
• Dimension reduction method (from p to 1 or 2)
• Graphical representation: trends, clusters, correlations, outliers

x_i : a class of data

• Examples of classes of data: a district of individuals, an image of pixels ...

• Description of the classes: mean, variance, quantiles, interval, histogram, ...

• How to perform PCA on units which are classes of data: symbolic PCA

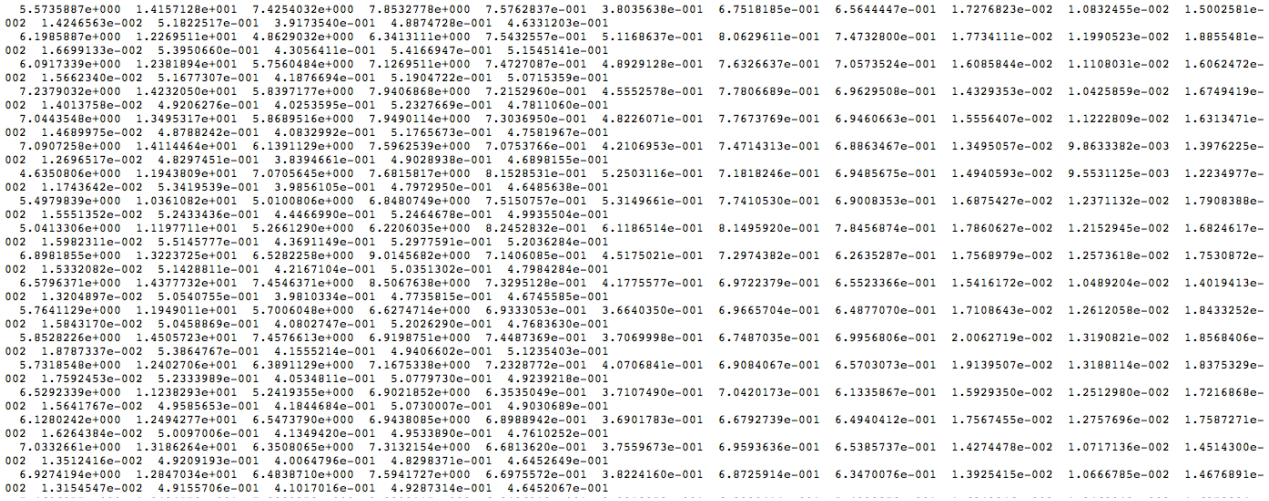

Figure: Each image: 400 pixels, Each pixel: 16 measurements

One image x_i = One matrix $k = 400 \times p = 16$

$x_{i,1}$ image i measurement 1 : column vector $k \times 1 \in \mathbb{R}^k$

...

$x_{i,p}$ image i measurement p : column vector $k \times 1 \in \mathbb{R}^k$

• E. Diday et. al.:
- Columns of raw data are summarized by intervals, histograms, ...
- extended PCA
→ a lot of raw informations are lost: multivariate dist., correlations
→ basic vector space is changed and not precised
→ consitency problems (results depend on the bins)
• R. Emilion:
- Hilbert space PCA considering both **raw data** (if available) **and classes**
- Summarize just at the end the results of PCA for visualization purposes of classes

More generally, our approach of the symbolic case: higher level math. representation using as far as possible both huge raw datasets nowadays available, classes, power of computers.
Sometimes raw data are not available (private, cost, confidentiality, security ...)

•$\mathbb{H}$: Hilbet space, i.e. vector space with a inner product $<>$, complete.
Examples : matrices, square integrable functions, product of Hilbert spaces, ...
• $x_i = (x_{i,1}, \ldots, x_{i,p}) \in \mathbb{H}^p, \alpha_i, i = 1, \ldots, n$: cloud of weighted vectors.

$$g = \Sigma_{i=1}^{n} \alpha_i x_i \in \mathbb{H}^p$$

: barycenter of the cloud
$v_1, v_2, \ldots, v_p$ orthogonal eigenvectors of the positive definite symmetric correlation, $p \times p$ real matrix C of the p column vectors $(x_{i,j}) - g_j \in \mathbb{H}^n$ associated to the eigenvalues
$\lambda_1 \geq \lambda_2, \ldots \geq \lambda_p \geq 0$ of C, resp.
•Change of coordinates from the canonical basis of $\mathbb{R}^p$ to the (v_i)
•Apply the same linear changes to the $x_i's$ to get new coordinates $y_i's$: principal components.
• Dimension reduction (from $\mathbb{H}^p$ to $\mathbb{H}^1$ or $\mathbb{H}^2$

- First and second component: n vectors in $\mathbb{R}^k$
- If k is large: statistics on these vectors.

Example: if use the means of the components (and not the components of the means): each image is represented by a point in the plane.

II- Classification of multihistograms
Example
Dirichlet mixtures
Consistency theorem

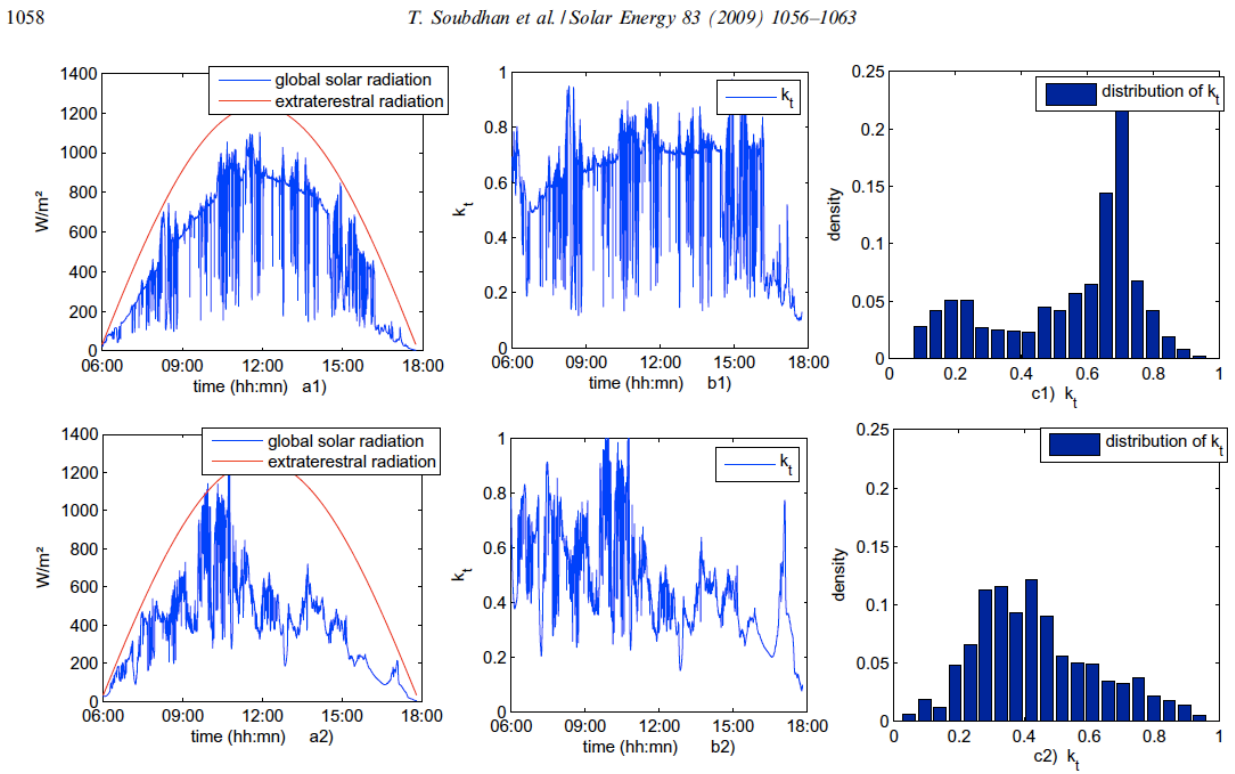

Figure: Each solar day stochastic process: 12h x 3600 measurements, histogram

- Each histogram is a vector of probability
- histograms are not those of any parametric distributions
- Considered as outputs of a random variable: random distribution
- Estimating this random distribution as a mixture of Dirichlet distributions (Solar Energy 2009)
- The Classification by estimating this mixture is consitent if the true distribution is an outcome of a mixture of Dirichlet processes: proof uses the Martingale convergence theorem (submitted to SAM)

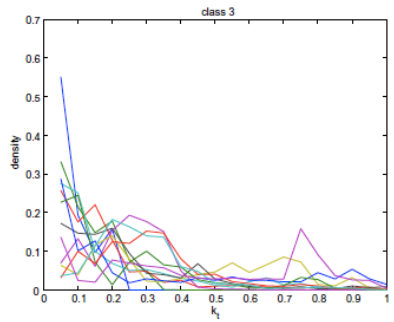

Fig. 6. Daily distributions of k_t in class 3.

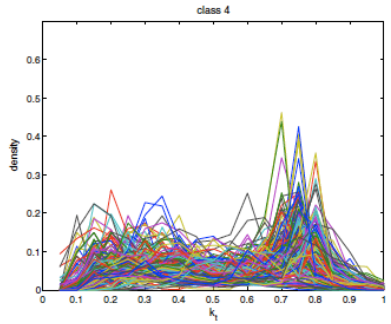

Fig. 8. Daily distributions of k_t in class 4.

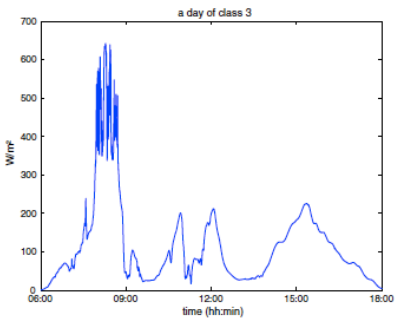

Fig. 7. A day of class 3.

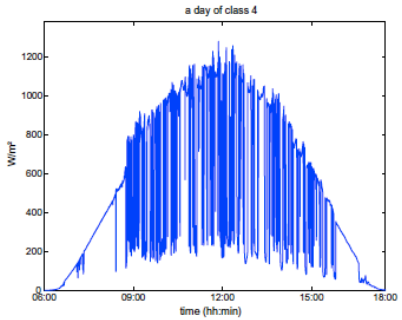

Fig. 9. A day of solar radiation in class 4.

Figure: 4 Classes of histograms were identified

- One day: Solar radiation, humidity, Temperature, ... stochastic processes
- Marginal histogram for each process: do not capture correlations
- Consider one multihistogram for each day
- Outputs of a Dirichlet mixture with Dirichlet's defined on a product space
- Same story

• Link between the two parts: Symbolic PCA and then classification of 2-dim histograms of the 2 best principal components

• SDA requires more complex mathematics

• What is a symbolic object:

Symbolic Galois concepts (Diday-Emilion 1995)

Concepts: representation of classes, First order logic (Diday).

Both can be derived from Galois connections.

• Historically point. J. F. C. Kingmann (1975). J. Royal Stat. B, 37. Random discrete distributions.

Spoke of many previous works where people were interested with 'Random objects which are themselves probability distributions'.