Nonparametric Functional Data Analysis

Yoan Miche

CIS, HUT

March 27, 2007

Outline

1. Functional Stuff
2. Some examples

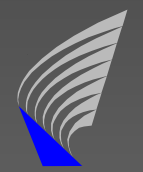
Functional Variable

Consider a random family $\{X(t_j)\}_{j=1,\dots,J}$, discrete.

If you increase *a lot* the number of samples

Can get close to a continuous case

$$\chi = \{X(t); t \in (t_{\min}, t_{\max})\}$$



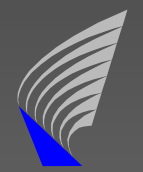
Functional Variable

Consider a random family $\{X(t_j)\}_{j=1,\dots,J}$, discrete.

If you increase *a lot* the number of samples

Can get close to a continuous case

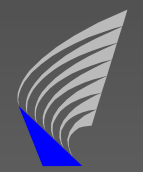
$$\chi = \{X(t); t \in (t_{\min}, t_{\max})\}$$



Functional Variable continued...

Not always time, can be different wavelengths, case of the spectrometric curves for meat balls

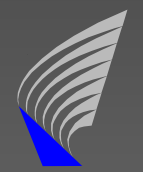
Functional refers to the form of the data and parametric/nonparametric to the form of the constraints (on the model)



Functional Variable/Dataset

Definition A functional variable $\boldsymbol{\chi}$ takes values in a infinite dimensional space. An observation χ of $\boldsymbol{\chi}$ is called a functional data

Definition A functional dataset $\chi_1, ..., \chi_n$ is the observation of n functional variables $\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n$ identically distributed as $\boldsymbol{\chi}$

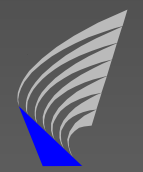


(Non)Parametric Models

Definition Let $\mathbf{X}$ be a random vector valued in $\mathbb{R}^p$ and let Φ be a function defined on $\mathbb{R}^p$ and depending on the distribution of $\mathbf{X}$. A model for the estimation of Φ consists in introducing some constraint of the form

$$\Phi \in \mathbf{C}.$$

The model is called parametric for the estimation of Φ if $\mathbf{C}$ is indexed by a finite number of elements of $\mathbb{R}$. Otherwise, the model is called nonparametric.



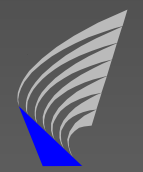
Functional (non)parametric Models

This definition can be extended to

Definition Let $\mathbf{Z}$ be a random variable valued in some infinite dimensional space F and let Φ be a mapping defined on F and depending on the distribution of $\mathbf{Z}$. A model for the estimation of Φ consists in introducing some constraint of the form

$$\Phi \in \mathbf{C}.$$

The model is called functional parametric for the estimation of Φ if $\mathbf{C}$ is indexed by a finite number of elements of F. Otherwise, the model is called functional nonparametric.



Some Examples

Meat Balls again...

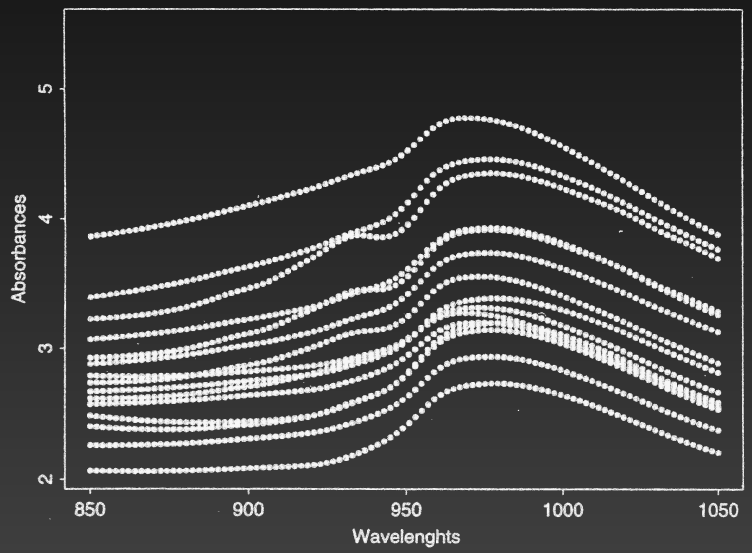


Fig. 2.1. Original Chemometric Data Concerning 15 Subjects

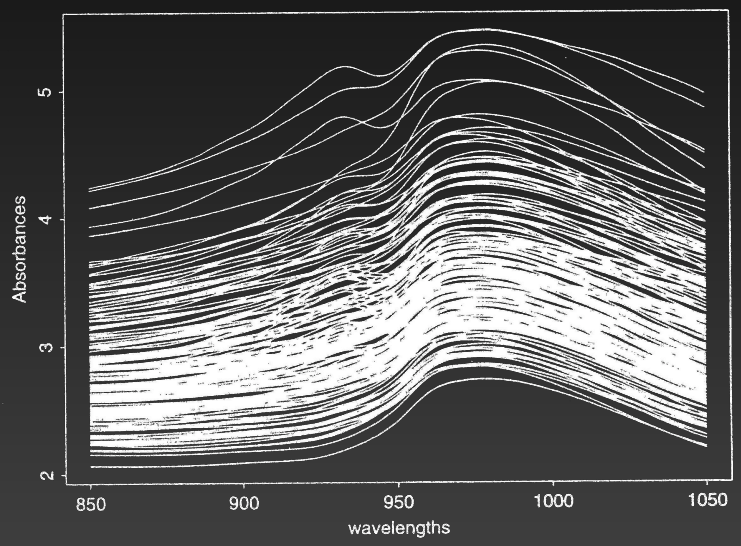
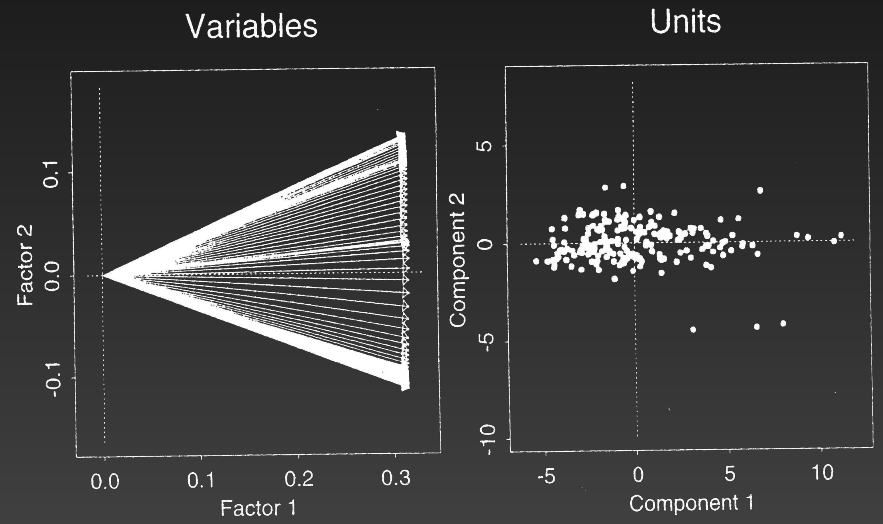


Fig. 2.2. The Spectrometric Curves

The original discrete data (sub-sample) and the whole data set in curve form.

Meat Balls continued...

Is there a shift in the curves ? Use of a PCA to look for a relationship

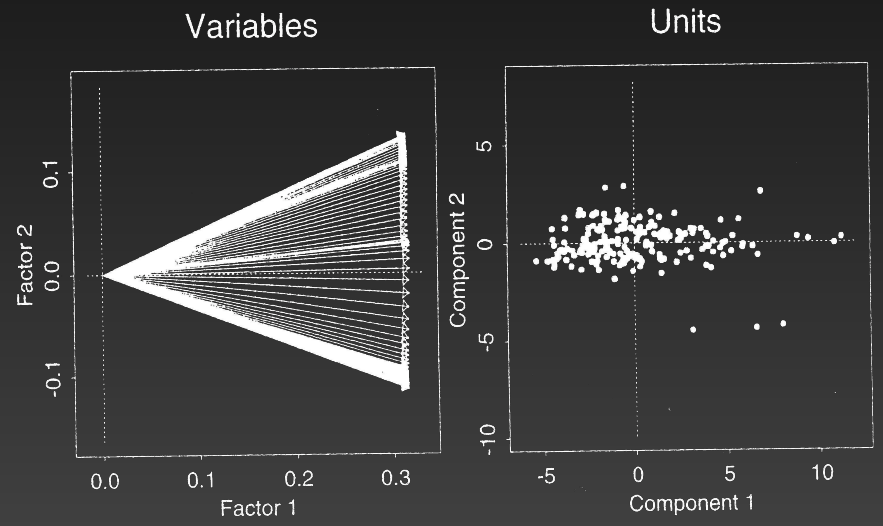


Short Analysis

- Scale factor for the variables: Linked to the shift ?
- Units seem to be OK, no special structure to see...

Meat Balls continued...

Is there a shift in the curves ? Use of a PCA to look for a relationship



Short Analysis

- Scale factor for the variables: Linked to the shift ?
- Units seem to be OK, no special structure to see...

Speech Recognition Data

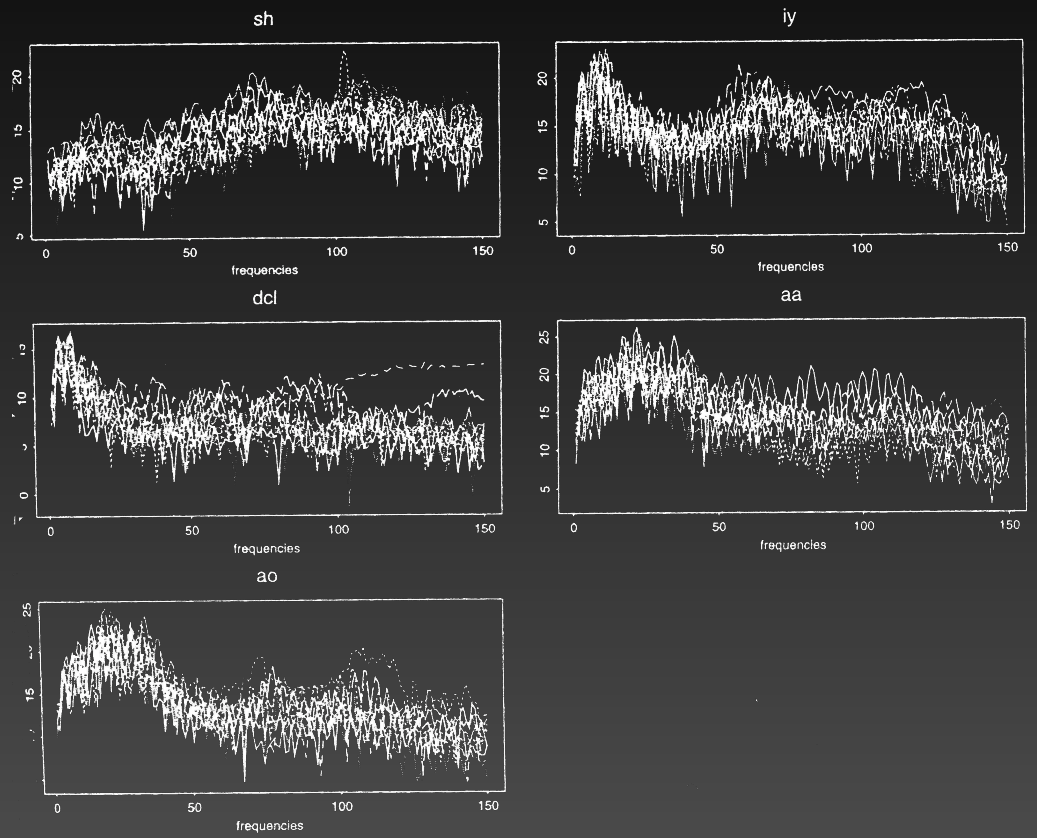


Fig. 2.4. A Sample of 10 Log-Periodograms (Curves Data) for each of the Five Phoneme Classes

10 Log-periodograms for each of the five Phoneme Classes.

Speech Recognition Data continued...

PCA again: Enables to exhibit one group (3), the sound "dcl"

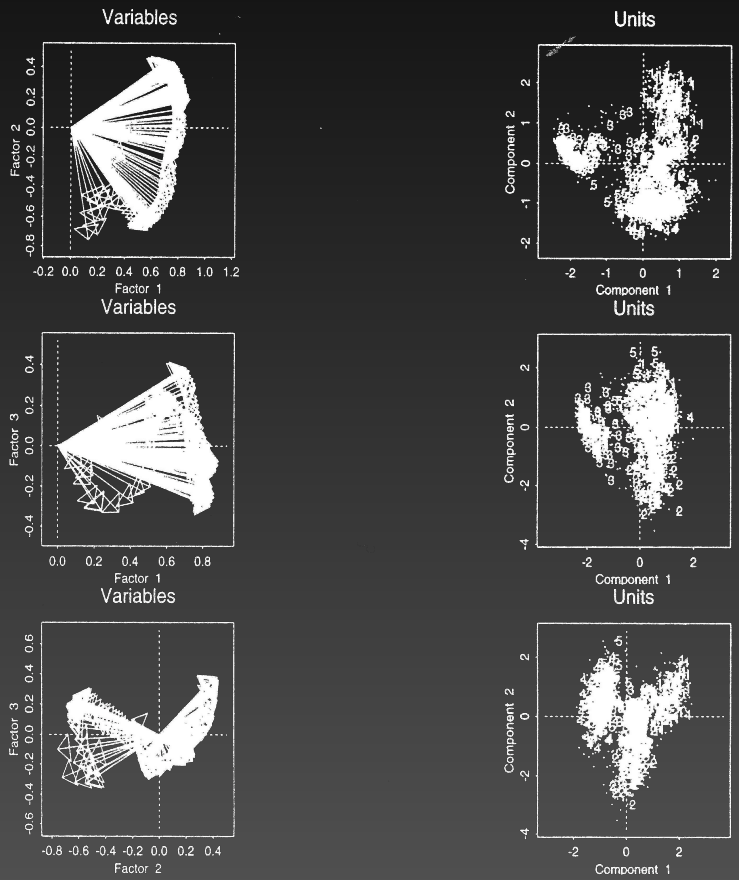


Fig. 2.5. Standard PCA: Phoneme Data

Knowledge beforehand of the classes: Supervised curve classification

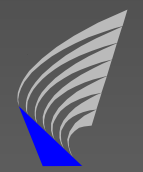
Saving you the last one. . .

Regression on the Electricity consumption dataset.

Use of differentiated log data, separated in years: Issue in this case in low number of samples compared to the number of variables.

Now for the real fun

Adapted spaces for functional data, and local weighting of functional variables.



Saving you the last one. . .

Regression on the Electricity consumption dataset.

Use of differentiated log data, separated in years: Issue in this case in low number of samples compared to the number of variables.

Now for the real fun

Adapted spaces for functional data, and local weighting of functional variables.