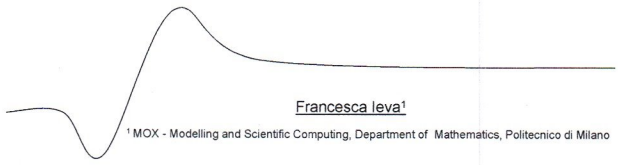
Nonparametric Inference for vital signs

Nonparametric Statistics



Francesca Ieva[1]

[1] MOX - Modelling and Scientific Computing, Department of Mathematics, Politecnico di Milano

Outline

(1) Depth measures for (multivariate) Functional Data

- ❑ Depth measures for (multivariate) functional data
- ❑ Epigraph and Hypograph indexes
- ❑ Spearman correlation index & matrix

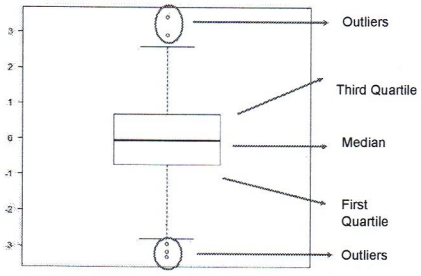
(2) Graphical tools (more than the multivariate case)

- ❑ (Multivariate) Outliergram
- ❑ Functional Boxplot
- ❑ Outlier Detection
- ❑ roahd package

(3) Case Study: ECG signals

(4) References

The very beginning



Rank data according to their depth

Depths measure

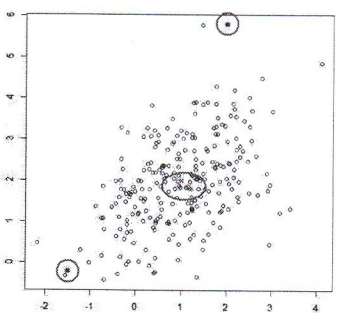
Outlyingness

Centrality

Outlyingness

From real random variables to random vectors

Data generated according to a multivariate distribution with law P_X



- Green data are deeper than **red** ones
- More difficult to rank the two green points or the two red ones
- No total order in R^d when d>1

Aim: rank data

⇩

Need: depth measure in R^d

Depth measures - multivariate case

Denote $\wp$ the class of probability distributions on the Borel sets of $\mathbb{R}^d$ and $\mathbf{P_X}$ the law of a given random vector **X**

Let $D(\cdot,\cdot): \mathbb{R}^d \times \wp \to \mathbb{R}$ be a bounded nonegative map such that the following properties hold:

1. **Affine invariance:** for any pxp non singular matrix A and any vector $\mathbf{b} \in \mathbb{R}^d$
$$D(Ax + b, P_{AX+b}) = D(\boldsymbol{x}, \boldsymbol{P_X}) \quad \forall\, \boldsymbol{X} \in \mathbb{R}^d$$
2. **Maximality at center**
$$D(\theta, P) = \sup_{x \in \mathbb{R}^d} D(\boldsymbol{x}, P) \quad \forall\, P \in \wp \text{ centered at } \theta$$
3. **Monotonicity wrt the deepest point:** for any P having deepest point at θ
$$\mathrm{D}(\mathbf{x}, P) < \mathrm{D}(\theta + \alpha(\boldsymbol{x} - \theta), P) \quad \forall\, \alpha \in [0; 1]$$
4. **Vanishing to infinity**
$$D(\boldsymbol{x}, P) \to 0 \quad \text{as} \quad ||\boldsymbol{x}|| \to \infty \quad \forall\, P \in \wp$$

Then D(·, $\wp$) is called a ***Statistical Depth Function***

The ***sample version*** of D(x;P), denoted by D(x) := D(x,$\mathbf{P_n}$) is defined by replacing **P** by the empirical version $\mathbf{P_n}$. It can be proved to be consistent.

Depth measures - multivariate case

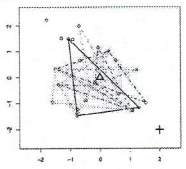
Simplicial Depth (SD)

[Liu, 1990]

The SD of a point x in $\mathbb{R}^d$ with respect to a probability measure P on $\mathbb{R}^d$ is defined to be

$$SD(\mathbf{x}, P) = \mathbb{P}(\mathbf{x} \in S[X_1, \ldots, X_{p+1}])$$

being S[X1, ..., Xp+1] the closed simplex formed by p+1 random vectors from P

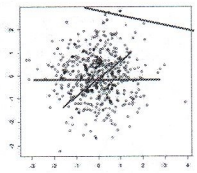


Halfspace (Tukey) Depth (HD)

[Tukey, 1975]

The HD of a point x in $\mathbb{R}^d$ with respect to a probability measure P on $\mathbb{R}^d$ is defined as the minimum probability mass carried by any closed half space containing x.

$$HD(\mathbf{x}, P) = \inf_H P(H) \quad \mathbf{x} \in \mathbb{R}^d$$



Depth measures - multivariate case

Spatial Depth (SpaD)

[Serfling, 2000]

Invariant for affine transformations with matrix A proportional to an orthogonal matrix be

$$SpaD(x; \mathbf{P}) = 1 - \left\| \mathbb{E}\left\{ \frac{\mathbf{X} - x}{\|\mathbf{X} - x\|} \right\} \right\|.$$

$SpaD(x, P) = 1$ (i.e. the maximum value) when

$$\mathbb{E}\left\{ \frac{X - x}{\|X - x\|} \right\} = P(X > x) - P(X < x) = 0$$

For d = 1 the point x with maximum SpaD(x;P) is the usual ***median***

Mahalanobis Depth (MHD)

[Liu, 1992]

Useful in case of nondegenerate d-variate normal distribution, but as d increases parametric assumptions becomes difficult to check.

$$MHD(x; \mathbf{P}) = \frac{1}{\left(1 + (x - \mathbb{E}[\mathbf{X}])^T \Sigma_{\mathbf{X}}^{-1} (x - \mathbb{E}[\mathbf{X}])\right)}$$

↓

Mahalanobis distance between x and E[X]

MHD fails in identifying the underlying distribution, since only the first two moments are used.

Depth measures - functional case

Univariate Functional Data

Let X be a stochastic process in the space of continuous functions from a compact interval I to R, C(I;R)

Generalization of the *simplicial depth* to univariate functional data

[López-Pintado and Romo 2009,2011]

Multivariate Functional Data

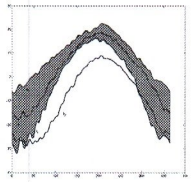
Let $\boldsymbol{X}$ be a stochastic process in the space of continuous functions from a compact interval I to R, C(I; $\mathbb{R}^d$)

Depth measures - functional case

Univariate Functional Data [López-Pintado and Romo 2009]

- Generalisation of the simplicial depth to univariate functional data
- Let X be a stochastic process in the space of continuous functions from a compact interval I to R, C(I;R)
- For J > 1 let us denote the random band depth of order J the quantity:

$$BD_{P_X}^J(f) = \sum_{j=2}^{J} P_X\{G(f) \subset B(X_1, X_2, ..., X_j)\}$$



Band given by two curves: one function belongs completely to the band, whereas the other does not

where G(f) indicates the graph of f, and $B(X_1,...X_j)$ the random band in R^2 delimited by the i.i.d. copies $X_1,...X_j$ of X:

$$B(X_1, ..., X_j) = \{(t, y(t)) : t \in I, \min_{r=1,..,J} X_r(t) \leq y(t) \leq \max_{r=1,..,J} X_r(t)\}, \quad j = 2, ..., J.$$

Depth measures - functional case

Sample version

Can be set to J=2, see [Tarabelloni et al. 2015]

BD_n expresses the proportion of bands determined by j different curves containing the whole graph of X

$$BD_{n,J}(x) = \sum_{j=2}^{J} BD_n^{(j)}(x) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots i_j \leq n} \mathbb{I}\{C(x) \subset B(x_{i_1}, x_{i_2}, ..., x_{i_j})\}$$

Sample Modified Band Depth (avoid ties)

Normalised Lebesgue measure on I

$$MBD_{n,J}(x) = \sum_{j=2}^{J} MBD_n^{(j)}(x) = \sum_{j=2}^{J} \binom{n}{j}^{-1} \sum_{1 \leq i_1 < i_2 < \cdots i_j \leq n} \hat{\lambda}\{E(x, x_{i_1}, x_{i_2}, ..., x_{i_j})\}$$

where $E(x, x_{i_1}, x_{i_2}, ..., x_{i_j}) = \left\{t \in I, \min_{r=i_1,...,i_j} x_r(t) \leq x(t) \leq \max_{r=i_1,...,i_j} x_r(t)\right\}$

The modified band depth measures the proportion of time interval I where the graph of x belongs to the envelopes of the j-tuples $(x_{i_1}, ..., x_{i_j})$, j = 1, . . . , J

Depth measures - functional case

Univariate Functional Data

Let *X* be a stochastic process in the space of continuous functions from a compact interval I to R, C(I;R)

Generalization of the *simplicial depth* to univariate functional data
[López-Pintado and Romo 2007,2009]

Multivariate Functional Data

Let ***X*** be a stochastic process in the space of continuous functions from a compact interval I to R, $C(I; R^d)$

Generalization of the *simplicial depth* to multivariate functional data
[Ieva and Paganoni 2013
López-Pintado, Sun, Lin and Genton 2014]

Depth measures - functional case

Univariate Functional Data

Let *X* be a stochastic process in the space of continuous functions from a compact interval I to R, C(I;R)

Generalization of the *simplicial depth* to univariate functional data
[López-Pintado and Romo 2007,2009]

Particular case (d=1)
[Claeskens, Hubert, Slaets and Vakili 2014]

Multivariate Functional Data

Let ***X*** be a stochastic process in the space of continuous functions from a compact interval I to R, $C(I; R^d)$

Generalization of the *simplicial depth* to multivariate functional data
[Ieva and Paganoni 2013
López-Pintado, Sun, Lin and Genton 2014]

Generalization of the *halfspace depth* to multivariate functional data
[Claeskens, Hubert, Slaets and Vakili 2014]

Depth measures - the functional case

Multivariate Simplicial Band Depth [López-Pintado, Sun, Lin and Genton 2014]

$$MSBD(\mathbf{f}, P_X) = \int_I SD(\mathbf{f}(t), P_{X(t)})dt$$

Multivariate Simplicial Depth

Multivariate Band Depth [Ieva and Paganoni 2013, Tarabelloni et al. 2015]

$$MBD^J_{P_X}(\mathbf{f}) = \sum_{k=1}^{d} p_k MBD^J_{P_{X_k}}(f_k), \qquad p_k > 0, \sum_{k=1}^{d} p_k = 1$$

Multivariate Functional Halfspace Depth [Claeskens, Hubert, Slaets and Vakili 2014]

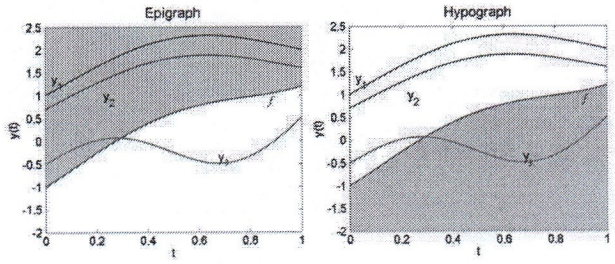
$$MFHD(\mathbf{f}, P_X) = \int_I w(t) HD(\mathbf{f}(t), P_{X(t)})dt$$

Function weighting the point wise variability

Multivariate Halfspace Depth

Epigraph/Hypograph Index

- Functional data ordering that is not based on a center-outward order, but on a down-upward order. Hence, we need an appropriate index that expresses this ordering.
- Hypograph and Epigraph indexes compute, in their modified version, the amount of time spent by the curves of the sample below (above) a given curve



(Modified) Epigraph/Hypograph Index

The Epigraph Index and Modified Epigraph Index of a function f w.r.t a sample $f_1,\dots,f_N$ are:

$$EI(f) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\{f_i(t) \geq f(t), \forall t \in I\}) \qquad MEI(f) = \frac{1}{N}\sum_{i=1}^{N} \bar{\lambda}(\{t \in I : f_i(t) \geq f(t)\})$$

Analogously the Hypograph Index and Modified Hypograph Index are:

$$HI(f) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\{f_i(t) \leq f(t), \forall t \in I\}) \qquad MHI(f) = \frac{1}{N}\sum_{i=1}^{N} \bar{\lambda}(\{t \in I : f_i(t) \leq f(t)\})$$

- MEI and MHI provide top-down and bottom-up ordering for functional datasets
- MEI are related to MBD by the following relation:

$a_0 = a_2 = 2/(N(N-1))$

$$MBD(f) \leq a_0 + a_1 MEI(f) + a_2 N^2 MEI^2(f)$$

$a_1 = 2(n+1)/(n-1)$

It will be used to build the Outliergram

Spearman's correlation index

Grades for a stochastic process

Let X_t be a stochastic process with sample paths in $C(I)$, with $I \subset \mathbb{R}$ compact. The Inferior and Superior Length grade of X_t are defined as

$$IL\text{-}grade(X_t) = \frac{1}{\lambda(I)}\mathbb{E}_{Z_t}[\lambda\{t \in I : X_t \geq Z_t\}],$$

$$SL\text{-}grade(X_t) = \frac{1}{\lambda(I)}\mathbb{E}_{Z_t}[\lambda\{t \in I : X_t \leq Z_t\}]$$

where Z_t is a stochastic process with the same distribution of X_t and λ is the Lebesgue measure on $\mathbb{R}$.

Spearman index for two stochastic processes

Let (X_t, Y_t) be a stochastic process with sample paths in $C(I;\mathbb{R}^2)$, with $I \subset \mathbb{R}$ compact. The Spearman index for (X_t, Y_t) is defined as

$$\rho_s(X_t, Y_t) = \rho_p(IL\text{-}grade(X_t), IL\text{-}grade(Y_t)),$$

where ρ_p denotes the Pearson correlation coefficient. An alternative definition can be provided considering also the Superior Length grade.

Spearman's correlation index

Sample grades for a function

Consider a functional dataset $x_1(t), ..., x_n(t)$, with $t \in I$. Fixing $x = x(t)$, the sample versions of both *IL-grade* and *SL-grade* are defined as

$$IL_n\text{-}grade(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \geq x_i(t)\},$$

$$SL_n\text{-}grade(x) = \frac{1}{n\lambda(I)} \sum_{i=1}^{n} \lambda\{t \in I : x(t) \leq x_i(t)\}.$$

Remark

$IL_n\text{-}grade(x)$ and $SL_n\text{-}grade(x)$ quantify the relative position of x with respect to the other curves of the sample.

Spearman's correlation index

Spearman index for bivariate functional data

Consider the bivariate functional dataset,

$$[\mathbf{x} \quad \mathbf{y}] = \begin{bmatrix} x_1(t) & y_1(t) \\ x_2(t) & y_2(t) \\ \vdots & \vdots \\ x_n(t) & y_n(t) \end{bmatrix}_{t \in I}$$

composed by n realizations of the stochastic process (X_t, Y_t). Then, the sample Spearman index is defined as

$$\hat{\rho}_s(\mathbf{x}, \mathbf{y}) = \hat{\rho}_p(IL_n\text{-}grade(\mathbf{x}), IL_n\text{-}grade(\mathbf{y})),$$

where $\hat{\rho}_p$ is the sample Pearson correlation coefficient and

$$IL_n\text{-}grade(\mathbf{x}) = (IL_n\text{-}grade(x_1), IL_n\text{-}grade(x_2), ..., IL_n\text{-}grade(x_n)),$$
$$IL_n\text{-}grade(\mathbf{y}) = (IL_n\text{-}grade(y_1), IL_n\text{-}grade(y_2), ..., IL_n\text{-}grade(y_n)).$$

Spearman's correlation index

In the multivariate framework:

- Given two random variables X,Y, the Spearman index captures with a value in [-1,1] the possible dependence among them. In particular, its absolute value increases in magnitude as X and Y become closer to be a perfect monotone function one of each other.
- The Spearman index is 0 when two random variables are independent.

In the functional framework:

- The Spearman index quantifies with a value in [-1,1] the tendency of X_t and Y_t to be perfect monotone functions one of each other
- The Spearman index is 0 when two processes are stochastically independent

Spearman's Matrix

Spearman Matrix for h-variate functional data (h> 2)

Let

$$\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_h] = \begin{bmatrix} x_{1,1}(t) & x_{1,2}(t) & \dots & x_{1,h}(t) \\ x_{2,1}(t) & x_{2,2}(t) & \dots & x_{2,h}(t) \\ \dots & \dots & \dots & \dots \\ x_{n,1}(t) & x_{n,2}(t) & \dots & x_{n,h}(t) \end{bmatrix}_{t \in I}$$

be a multivariate functional dataset, composed by n realizations of the stochastic process $\mathbf{X}_t = (X_t^1, X_t^2, ..., X_t^h)$, with sample paths in $C(I; \mathbb{R}^h)$ and $I \subset \mathbb{R}$ compact. The sample Spearman Matrix $\widehat{SM}(\mathbf{X})$ is given by

$$\widehat{SM}(\mathbf{X}) = \begin{bmatrix} \hat{\rho}_s(\mathbf{x}_1, \mathbf{x}_1) & \hat{\rho}_s(\mathbf{x}_1, \mathbf{x}_2) & \dots & \hat{\rho}_s(\mathbf{x}_1, \mathbf{x}_h) \\ \hat{\rho}_s(\mathbf{x}_2, \mathbf{x}_1) & \hat{\rho}_s(\mathbf{x}_2, \mathbf{x}_2) & \dots & \hat{\rho}_s(\mathbf{x}_2, \mathbf{x}_h) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_s(\mathbf{x}_h, \mathbf{x}_1) & \hat{\rho}_s(\mathbf{x}_h, \mathbf{x}_2) & \dots & \hat{\rho}_s(\mathbf{x}_h, \mathbf{x}_h) \end{bmatrix},$$

where $\hat{\rho}_s(\mathbf{x}_i, \mathbf{x}_j)$ is the sample Spearman index computed on the bivariate functional dataset $[\mathbf{x}_i \quad \mathbf{y}_j]$.

$\widehat{SM}(\mathbf{X})$ provides an effective insight of the pattern of dependence among components of a multivariate functional dataset.

Indipendence Test

Independence test

Let (X_t, Y_t) be a stochastic process with sample paths in $C(I;\mathbb{R}^2)$ and $I \subset \mathbb{R}$ compact. Suppose to have the functional dataset

$$[\mathbf{x} \quad \mathbf{y}] = \begin{bmatrix} x_1(t) & y_1(t) \\ x_2(t) & y_2(t) \\ \vdots & \vdots \\ x_n(t) & y_n(t) \end{bmatrix}_{t\in T}$$

composed by n realizations of (X_t, Y_t). We wish to check the following hypotheses:

$$H_0 : \rho_s(X_t, Y_t) = 0 \quad \text{vs} \quad H_1 : \rho_s(X_t, Y_t) \neq 0$$

Our test procedure

Reject H_0, with respect to a significance level equal to α, if the confidence interval for ρ_s of level 1-α does not contain zero.

Outline

(1) Depth measures for (multivariate) Functional Data

- ❑ Depth measures for (multivariate) functional data
- ❑ Epigraph and Hypograph indexes
- ❑ Spearman correlation index & matrix

(2) Graphical tools

- ❑ (Multivariate) Outliergram
- ❑ Functional Boxplot
- ❑ Outlier Detection
- ❑ roahd package

(3) Case Study: ECG signals

(4) References

Outliergram – univariate functional data

[Arribas-Gil and Romo 2014]

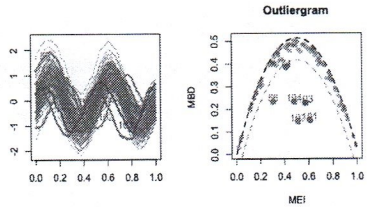
- The outliergram is a recently proposed graphical tool to identify shape outliers
- It is based on a scatterplot of sample MBD vs. MEI, that are related by:

$$MBD(f) \leq a_0 + a_1 MEI(f) + a_2 N^2 MEI^2(f)$$

- Points lying far away from the upper parabolic boundary are likely to be flagged as outliers
- We compute $d_i = a_0 + a_1 MEI(f) + a_2 N^2 MEI^2(f) - MBD(f)$
- We flag as outliers those points s.t. $d_i \geq Q_3(d) + F \cdot IQR(d)$

Optionally

- A secondary procedure to purge identified outliers of possible amplitude outliers lying at the bottom of the parabola can be employed
- the value *F = 1.5* can be adjusted, so that in case of gaussian distribution of data we have a desired probability δ



Outliergram – multivariate functional data

[Ieva and Paganoni 2017]

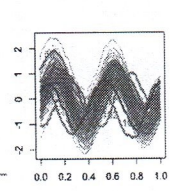
- Based on the generalized version of MEI to the multivariate functional case

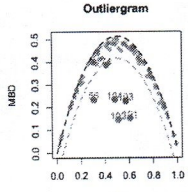
$$MEI_{\{f_1,\dots,f_n\}}(\mathbf{f}) = \sum_{k=1}^{h} p_k MEI_{\{f_{1k},\dots,f_{nk}\}}(f_k) \qquad p_k > 0 \;\forall\; k = 1,\dots,h, \quad \sum_{k=1}^{h} p_k = 1$$

- It is based on a scatterplot of sample MBD vs. MEI, that are related by:

$$MBD(f) \leq a_0 + a_1 MEI(f) + a_2 N^2 MEI^2(f)$$

- Points lying far away from the upper parabolic boundary are likely to be flagged as outliers
- We compute $d_i = a_0 + a_1 MEI(f) + a_2 N^2 MEI^2(f) - MBD(f)$
- We flag as outliers those points s.t. $d_i \geq Q_3(d) + F \cdot IQR(d)$





Functional boxplot

[Sun and Genton 2011]

- The Functional Boxplot is a visualization/detection tool based on functional depths

Procedure

(i) Compute data depths (e.g., MBD)

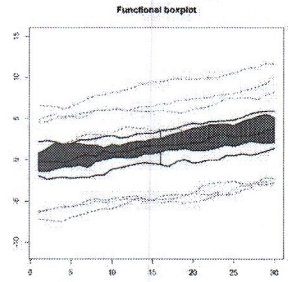
(ii) Rank data and consider the 50% central region

$$C_\alpha = \left\{(t, z(t)) : \min_{i=1,\dots,\lceil \alpha N \rceil} X_i(t) \le z(t) \le \max_{i=1,\dots,\lceil \alpha N \rceil} X_i(t)\right\}$$

(iii) inflate it by a factor F > 1

(iv) obtain the fences given by the envelope of functions contained inside the inflated region

(v) flag the observations crossing the fences as amplitude outliers



Adjusted functional boxplot

[Sun and Genton 2012]

- The standard choice F = 1.5 is tuned for the gaussian case
- For univariate (scalar) data: $\delta = P(Z > Q_3 + F \cdot IQR) = 2\Phi(4Z_{0.25}) =\approx 7e-3$
- An automatic procedure can be used to adjust for the required value 7e-3 even for univariate functional data

Procedure

1. Robustly estimate location and covariance matrix from data
2. For i = 1,…,N_{iter}
   - Simulate a Gaussian population of M elements with same location/covariance
   - Compute depths, build central regions and fences
   - Compute the optimal value F_i
3. Take $F^{opt} = \sum_{i=1}^{N_{trials}} F_i / N_{trials}$
4. Perform the functional boxplot on original data with F^{opt}

Flaws

- The originally proposed robust estimator for covariance comes from multivariate analysis
- Its *spd*-ness must be enforced (e.g. through shrinkage)
- Its spectrum is not clearly related to that of C_X

Robust alternatives to C_X

Spherical Covariance Operator [Gervini 2008]

- Let X be a random function with values in $L^2(I)$. The spherical covariance estimator $\mathcal{C}_S$ is:

$$\mathcal{C}_S = \mathbb{E}\left[\frac{(X - \tilde{\mu}_X) \otimes (X - \tilde{\mu}_X)}{\|X - \tilde{\mu}_X\|^2}\right] \qquad \widehat{\mathcal{C}}_S = \frac{1}{N}\sum_{i=1}^{N} \frac{(X_i^{(L)} - \tilde{\mu}_X^{(L)}) \otimes (X_i^{(L)} - \tilde{\mu}_X^{(L)})}{\|X_i^{(L)} - \tilde{\mu}_X^{(L)}\|^2}$$

- where $\tilde{\mu}_X$ is a robust measure of location of X, e.g. the median and $X^{(L)}$ denotes the orthogonal projection onto $V_L \subset V, dim(V_L) = L$
- under some rather general hypotheses, given

$$\mathcal{C}_X = \sum_{i \ge 1} \lambda_i \, \psi_i \otimes \psi_i \qquad \text{(estimation needed, e.g. with } Q_N\text{)}$$

$$\mathcal{C}_S^{(L)} = \sum_{i=1}^{L} \sigma_i \, \psi_i \otimes \psi_i, \qquad \text{with} \qquad \sigma_i \neq \lambda_i$$

- $\mathcal{C}_S$ is the covariance of data X projected onto the unit sphere in V
- eigenvalues' breakdown is decreasing along the spectrum (if decreasingly sorted by value)
- it also depends on the spacing (decay rate) of eigenvalues

Robust alternatives to C_X

Median Covariation [Kraus and Panaretos 2012, Cardot and Godichon 2015]

- Solves and analogous L_1 minimisation to the spatial median's one

$$Q_X(0) = \operatorname{argmin}_{z \in V} \left[\mathbb{E}\|X - z\| - \|X\|\right]$$

- Given the space H(V,V) of Hilbert-Schmidt operators on V, it solves

$$\mathcal{C}_M = \operatorname{argmin}_{\mathcal{R} \in \mathcal{H}_+} \left\{\mathbb{E}\left\|\mathcal{R} - (X - \tilde{\mu}_X) \otimes (X - \tilde{\mu}_X)\right\| - \left\|(X - \tilde{\mu}_X) \otimes (X - \tilde{\mu}_X)\right\|\right\}$$

- its modal sample version originally solved by quasi-Newton BFGS algorithm (quite burdensome)
- an averaged stochastic-gradient algorithm can be used to compute a nodal approx. $\widehat{\mathcal{C}}_M^h$ and adapted to provide a modal approx. $\widehat{\mathcal{C}}_M^{(L)}$
- similarly to $\mathcal{C}_S$, $\mathcal{C}_M$ has the same principal directions of $\mathcal{C}_X = \sum_{i \ge 1} \lambda_i \, \psi_i \otimes \psi_i$

$$\mathcal{C}_M = \sum_{i=1}^{L} \rho_i \, \psi_i \otimes \psi_i \qquad \text{with} \qquad \rho_i \neq \lambda_i \qquad \text{(estimation needed, e.g. with } Q_N\text{)}$$

Adjustment of functional boxplot

[Tarabelloni and Ieva 2017]

- To adjust the F factor in the functional boxplot, we have to simulate a gaussian population with same location and covariance of X
- To be general, we exploit a truncated Karhunen-Loève expansion of X with gaussian scores:

$$Y_i^{*(L)} = \mu_X^{(L)} + \sum_{j=1}^{L} \sqrt{\lambda_j}\, \xi_{i,j} \psi_j^{(L)}$$

- We estimate $\psi_j^{(L)}$ by $\widehat{\psi}_j^{(L)}$, j-th eigenfunction of either $\widehat{C}_S$ or $\widehat{C}_M$
- We could estimate the values $\sqrt{\lambda_i}$ with robust estimators like $Q_N(\xi_{1,j}, \dots, \xi_{N,j})$ (or MAD or S_N):

$$Q_N(Z_1, \dots, Z_N) = d\,\{|Z_i - Z_j|, 1 \le i \ne j \le N\}_{(K)}, \quad K \approx \binom{N}{2}$$

- Unfortunately, these estimators rely on some distribution-dependent constants that have to be to chosen in order to enforce (Fisher's) consistency, (e.g. $d = \left(\sqrt{2}\,\Phi^{-1}(5/8)\right)^{-1}$ for gaussian data
- We simulate the family $\widehat{Y}_i^{*(L)} = \sum_{j=1}^{L} \sqrt{\rho_j}\, \xi_{i,j}\, \psi_j^{(L)}, \quad \rho_j = \frac{\lambda_i}{\lambda_1}$, using ratios $Q_N(\xi_{\cdot,i})/Q_N(\xi_{\cdot,1})$
- We exploit affine invariance of depths:

$$D(\sqrt{\lambda_1}\widehat{Y}^{*(L)} + \mu_X | P_{Y^{*(L)}}) = D(\widehat{Y}^{*(L)} | P_{\widehat{Y}^{*(L)}})$$

Simulation of gaussian data

Back

- In the adjustment procedure, a synthetic Gaussian population Y* associated to the original population X must be simulated
- Same covariance C_X and location μ_X as the original data
- A general Karhunen-Loève expansion can be used to this aim
- Spherical covariance and median covariation have same eigenfunctions
- We have to estimate eigenvalues!

Estimation Strategy

- compute C_S or C_M and a robust location parameter μ_X (e.g. spatial median)
- compute the first K eigenfunctions, $\varphi_1, \dots, \varphi_K$
- project data on these eigenfunctions and estimate $\lambda^{1/2}$ using the robust estimator (50% breakdown and 82% efficient at gaussian pop.) :

$$Q_N(Z_1, \dots, Z_N) = d\,\{|Z_i - Z_j|, 1 \le i \ne j \le N\}_{(K)}, \quad K \approx \binom{N}{2}$$

- unfortunately, the constant d is distribution-dependent and guarantees consistency (it is generally chosen to address gaussian data)

Outlier Detection in FDA

- During statistical analysis, **outliers are often considered as an error or noise**
- Instead, they may carry important information on the phenomenon under study.

They may lead to model misspecification,
biased parameter estimation and incorrect results.

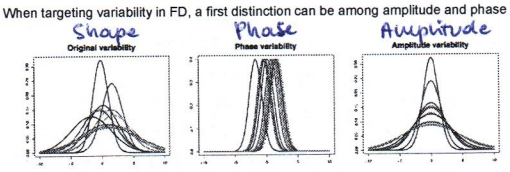
- It is important to identify them prior to model data and to carry out the analysis
- There is **no general definition of outliers**, since their presence often depends on assumptions regarding the hidden structure of data and the applied detection method.

"[An outlier is] an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism".
[D. M. Hawkins. Identification of outliers. Springer, 1980]

- Outlier detection is particularly important in those contexts where contaminations may heavily bias estimation and inference (ex: **Functional Data Analysis**)

Outlyingness patterns

- When targeting variability in FD, a first distinction can be among amplitude and phase variability:



- This distinction drives a generally accepted analogous distinction among outlyingness patterns:
  - amplitude outliers (related to amplitude variability)
  - shape outliers (related to phase variability)
- Amplitude outlyingness is directly related to *standard* outlyingness of multivariate data
- Shape outlyingness has no counterpart in multivariate analysis (it directly stems from the nature of functional data)
- This distinction motivates the development of separate methods to target functional outliers

---

Magnitude, Shape and Covariance Outliers

- Amplitude and phase variability inspired the main distinction currently accepted between outlyingness patterns => magnitude, shape outliers and covariance outliers.
    - **Magnitude outliers**: related to amplitude variability (direct analogue of the multivariate outlyingness concept)
    - **Shape outliers**: related to phase variability (new concept => no counterpart in classic statistics).
    - **Covariance outliers**: generated by a model that is different from the model of the central bulk of data in term of the variance-covariance operator that tunes the second order moments of data
- The different nature of these kinds of outliers motivates the need for different tools to detect and handle them.
    - **Registration -> using proper warping functions to map the timings of each observation to a common time. This synchronisation procedure may identify those data with degenerate warping.**
    - **Outliergram -> identify and remove abnormal observations using detection methods tailored to shape outliers.**

Outlier Detection in FDA

There are two ways, in general, to deal with outliers in a data sample:

1. to **apply outlier-detection** tools and **remove outlying observations** from the dataset;

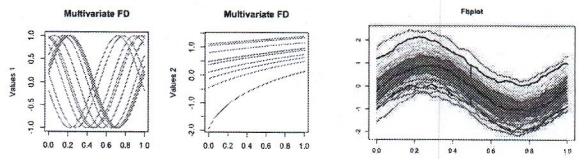
   Outliergram *(shape outliers)*: joint use of depths and epigraphic index to detect and discard both shape and magnitude outliers in order to robustify the reference sample data, composed by G different known groups

2. to robustify the estimators adopted for carrying out the inference.

   Adjusted functional boxplot *(amplitude outliers)*: based on the notion of statistical depth and adapted to the dataset at hand, which targets amplitude outliers. There is the possibility to combine the strengths of both approaches in order to build a robust version of the adjusted functional box plot.

Package roahd

- **roahd** (RObust Analysis of High dimensional Data) merges methodological contributions of MOX and UC3M research groups on depths/robust methods for functional data.
- It relies on S3 implementations of univariate (fData) and multivariate functional data (mfData)
- So far, (m)fData are represented as point wise measurements over 1D grids
- It supports to model-based simulation of synthetic functional data (method benchmarking)
- Uni/multivariate functional depths, graphical and correlation indices are implemented
- Visualisation and outlier detection methods (fbplot and outliergram)
- Currently hosted on CRAN



Outline

(1) Depth measures for (multivariate) Functional Data

- ❑ Depth measures for (multivariate) functional data
- ❑ Epigraph and Hypograph indexes
- ❑ Spearman correlation index & matrix

(2) Graphical tools

- ❑ (Multivariate) Outliergram
- ❑ Functional Boxplot
- ❑ Outlier Detection
- ❑ roahd package

(3) Case Study: ECG signals

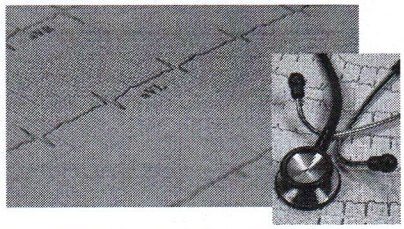
(4) References

The PROMETEO project

PROMETEO

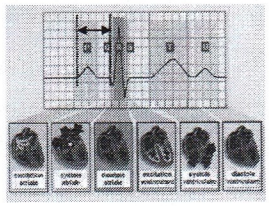
(PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'ExtraOspedaliero).

Anticipating diagnostic time, reducing infarction complications and optimizing the number of hospital admissions are three main goals of this project. Thanks to the partnerships of Azienda Regionale Emergenza Urgenza (AREU) and Abbott Vascular, ECG machinery with GSM transmission have been installed on all Basic Rescue Units of Milan urban area. PROMETEO project, planned and realized by 118 Dispatch Center of Milan since the end of **2008**, made possible to send quickly the ECG from territory to 118 Dispatch Center itself, and then to the hospital where patient would have been admitted to, even when a Basic Rescue Unit (Unit managed by volunteers only, without physicians on board) is sent to the patient.



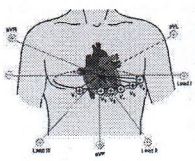
ECG signals

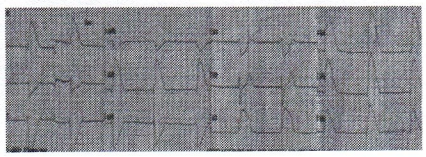
Electrocardiography (ECG) is a transthoracic recording of the heart electrical activity over time. Each phase of the cardiac cycle can be associated to a different segment of the ECG.



Nowadays, the most commonly used clinical ECG-system is the 12-lead ECG system, which enhance pattern recognition starting from signals registered by each of the 12 leads.

Since the voltages measured by some leads are proportional to the projections of the electric heart vector on the sides of some others, we can consider only the **8 leads I, II, V1, V2, V3, V4, V5, V6.**





Bundle Brunch Block

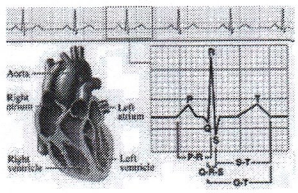
Bundle Brunch Block

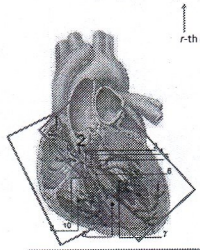
Bundle branch or fascicle injuries result in **altered pathways** for ventricular depolarization.

For i = 1,...,n, define a ***multivariate function*** which describes curves in ***8*** dimensions.

$$\mathbf{f}_i : t \mapsto (f_{i1}(t), \ldots, f_{i8}(t)) \in \mathbb{R}^8, \quad t \in T \subset \mathbb{R}$$

↑ *r*-th *component* of $\mathbf{f}_i(t)$





1. Sinoatrial node
2. Atrioventricular node
3. Bundle of His
4. **Left bundle branch**
5. Left posterior fascicle
6. Left-anterior fascicle
7. Left ventricle
8. Ventricular septum
9. Right ventricle
10. **Right bundle branch.**

Dataset

6758 signals

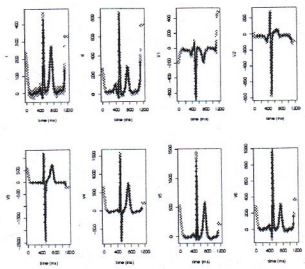
Physiological, LBBB, RBBB, Atrial Fibrillation, ...

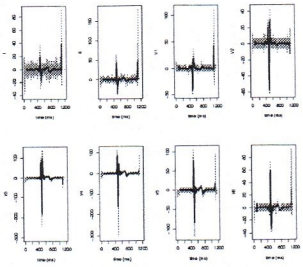
Pre processing

Wavelets Smoothing ⟶ All the components are jointly taken into account (they describe the same dynamic)

Estimation of functional form of the ECG signal and its first derivative, for each patient *i* = 1,...,149.

$$\mathbf{f}_i(t) = (I_i(t), II_i(t), V1_i(t), V2_i(t), V3_i(t), V4_i(t), V5_i(t), V6_i(t))$$





Raw data of the eight leads (gray) and wavelet functional estimates (blue) for a normal subject

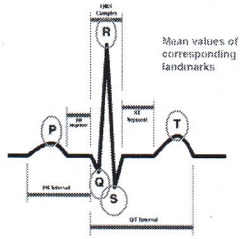
First central finite difference of the eight leads and wavelet estimates of the first derivatives for a normal subject

Pre processing

Landmark Registration

Functional observations usually show both phase and amplitude variation, i.e., each curve has its own biological time so that the same feature can appear at different times among the patient. It is well known that a correct separation between these two kinds of variability is necessary for a successful analysis.

For each patient i, we look for a non linear warping function (cubic splines) h_i such that:



Mean values of corresponding landmarks

$h_i(P^i_{onset}) = P^0_{onset}$ $\quad h_i(P^i_{offset}) = P^0_{offset}$

$h_i(QRS^i_{onset}) = QRS^0_{onset}$ $\quad h_i(I^i peak) = I^0 peak$

$h_i(QRS^i_{offset}) = QRS^0_{offset}$ $\quad h_i(T^i_{offset}) = T^0_{offset}$

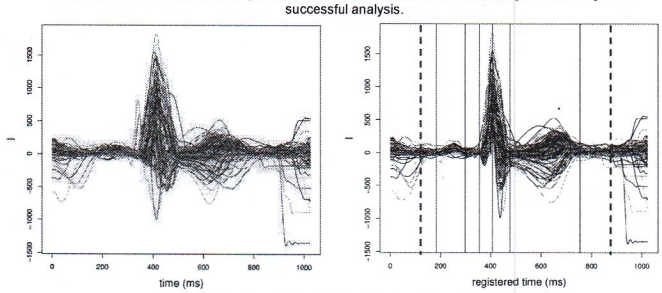
The registered vectorial function is then:

$$\mathbf{F}_i(t) = \mathbf{f}_i(h_i(t)) \qquad i = 1,\ldots,198$$

Pre processing

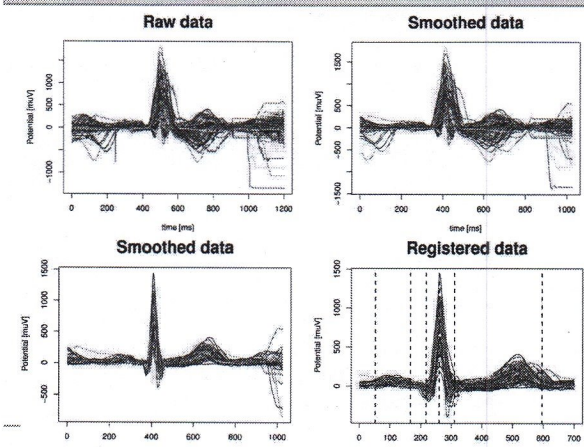
Landmark Registration

Functional observations usually show both phase and amplitude variation, i.e., each curve has its own biological time so that the same feature can appear at different times among the patient. It is well known that a correct separation between these two kinds of variability is necessary for a successful analysis.

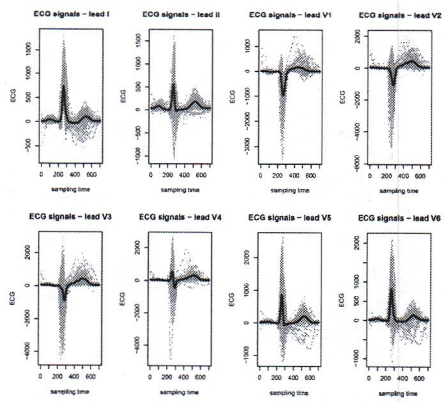


Original I leads for the 149 patients (left) and registered ones (right).
Vertical lines indicate position of mean landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $I^0 peak$, QRS^0_{offset}, T^0_{offset}.

Pre processing



Pre processing

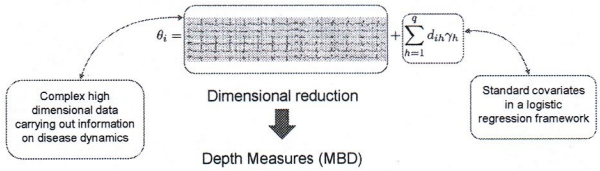


Depths for Dimensional Reduction

- Let Y_i be the indicator of the presence of LBBB for the i-th patient.
- Consider a logistic regression model, where the response variable is

$$Y_i \sim Be(p_i) \longleftrightarrow \theta_i = \log\left(p_i/(1-p_i)\right)$$

We model the logit of the mean as linear transformation of the covariates related to i-th statistical unit.



[Tarabelloni et al. 2015]

[Ieva & Paganoni, 2015]

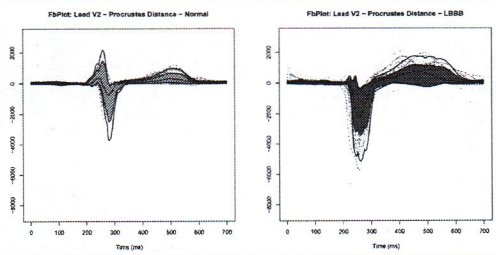
Depth for Dimensional Reduction

Analyses carried out with all the (pseudo) distances
*Results are presented for **Procrustes** pseudo-distances*
(robustness wrt the choice of the (pseudo) distance)

Weights

Lead	V2	V3	V1	V4	V5	V6	I	II
Weights	0.1722	0.1607	0.1385	0.1357	0.1132	0.1104	0.0872	0.0821

Functional Boxplots



Depth for Dimensional Reduction

Parameter estimation of the generalized model

- Fit a logistic regression model for i = 1, . . . ,51 (physiological) and i = 1, . . . ,48 (LBBB)
- MBDs of the derivatives dropped out because not significant (due to the high correlation with the corresponding MBDs of signals).

$$Y_i \sim Be(p_i)$$
$$\theta_i = \log\left(p_i/(1-p_i)\right)$$
$$\theta_i = \beta_0 + \beta_1 MBD_i$$

Parameter	Estimate	Std. Error	p-value
β^0 (Intercept)	11.484	2.483	$3.75*10^{-06}$
β^1 (MBD)	-46.268	9.619	$1.51*10^{-06}$

Prediction of the presence of disease

- Confusion matrix (threshold = 0.5) (representative case over 20)

	Normal	LBBB
Classified as Normal	47	8
Classified as LBBB	4	40

Indexes of performance

- **sensitivity 84.48 (±2.29)%**
- **specificity 89.80 (±1.87)%**
- **correct classification rate 87.22 (±1.58)%**

Rank Test for Functional Data

Population 1

$f_1,\ldots,f_N$ sample of (multivariate) curves generated according to a distribution P_X.

Population 2

$g_1,\ldots,g_M$ sample of (multivariate) curves generated according to a distribution P_Y.

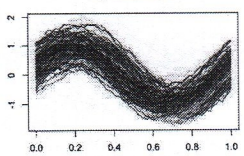
Reference Population

$h_1,\ldots,h_W$ sample of (multivariate) curves generated according to one among P_X or P_Y (say P_X).

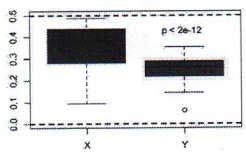
H_0: there is no difference between the distributions generating data

Compute the depth of each (multivariate) curve from Population 1 and 2 with respect to the Reference population and apply the usual Wilcoxon sum rank test.

Synthetic data



MBDs and W-test



Rank Test for Functional Data

Population 1

$f_1,\dots,f_N$ sample of (multivariate) curves generated according to a distribution P_X.

Population 2

$g_1,\dots,g_M$ sample of (multivariate) curves generated according to a distribution P_Y.

Reference Population

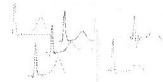
$h_1,\dots,h_W$ sample of (multivariate) curves generated according to one among P_X or P_Y (say P_X).

H_0: there is no difference between the distributions generating data

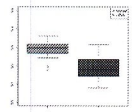
Compute the depth of each (multivariate) curve from Population 1 and 2 with respect to the Reference population and apply the usual Wilcoxon sum rank test.

Healthy

Left Bundle Branch Block

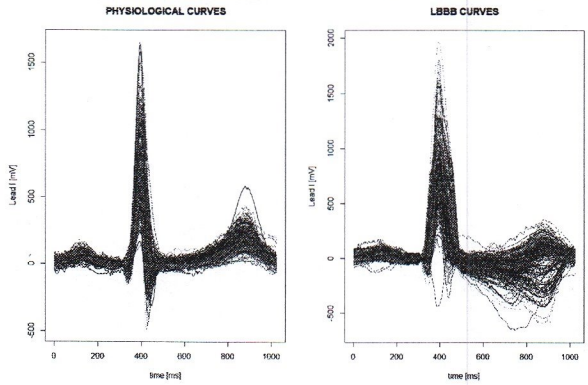


Healthy



p-value $\leq 3.02 \times 10^{-12}$

Spearman stuffs for testing differences in patterns of dependence among ECGs



Spearman stuffs for testing differences in patterns of dependence among ECGs

$$\widehat{SM}(\mathbf{X}) = \begin{array}{c} \\ \text{I} \\ \text{II} \\ \text{V1} \\ \text{V2} \\ \text{V3} \\ \text{V4} \\ \text{V5} \\ \text{V6} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{I} & \text{II} & \text{V1} & \text{V2} & \text{V3} & \text{V4} & \text{V5} & \text{V6} \end{array} \\ \begin{bmatrix} 1 & 0.357 & -0.039 & 0.193 & 0.166 & 0.185 & 0.226 & 0.264 \\ & 1 & 0.045 & 0.212 & 0.457 & 0.554 & 0.589 & 0.630 \\ & & 1 & 0.713 & 0.451 & 0.304 & 0.255 & 0.173 \\ & & & 1 & 0.709 & 0.571 & 0.501 & 0.361 \\ & & & & 1 & 0.879 & 0.761 & 0.575 \\ & & & & & 1 & 0.905 & 0.710 \\ & & & & & & 1 & 0.843 \\ & & & & & & & 1 \end{bmatrix} \end{array}$$

- High and significant entries on the upper diagnonal => the dynamics of the heart on a lead is strictly related to the dynamics on the following one.
- Physiological ECGs have a cordinated pattern in which most of the components depend on the others => Evidence for a regular heart dynamics.
- The significant entries are positive, indicating that the leads tend to be monotone increasing functions of each other.

Spearman stuffs for testing differences in patterns of dependence among ECGs

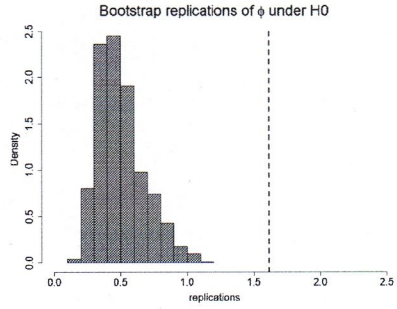
$$\widehat{SM}(\mathbf{Y}) = \begin{array}{c} \\ \text{I} \\ \text{II} \\ \text{V1} \\ \text{V2} \\ \text{V3} \\ \text{V4} \\ \text{V5} \\ \text{V6} \end{array} \begin{array}{c} \begin{array}{cccccccc} \text{I} & \text{II} & \text{V1} & \text{V2} & \text{V3} & \text{V4} & \text{V5} & \text{V6} \end{array} \\ \begin{bmatrix} 1 & 0.451 & -0.378 & -0.043 & -0.037 & 0.337 & 0.612 & 0.659 \\ & 1 & -0.077 & 0.053 & 0.182 & 0.459 & 0.589 & 0.569 \\ & & 1 & 0.756 & 0.559 & 0.136 & -0.210 & -0.366 \\ & & & 1 & 0.723 & 0.368 & 0.018 & -0.144 \\ & & & & 1 & 0.682 & 0.220 & -0.058 \\ & & & & & 1 & 0.715 & 0.438 \\ & & & & & & 1 & 0.844 \\ & & & & & & & 1 \end{bmatrix} \end{array}$$

- Again, high and significant entries on the upper diagonal.
- Pathological ECGs have a sparse pattern due to the presence of several pairs of independent leads => evidence for a chaotic heart dyamics.
- Some significant entries are negative

Spearman stuffs for testing differences in patterns of dependence among ECGs

Conjecture: the pattern of dependence among leads of the ECGs is different in the two populations of patients due to the presence of heart disease.

Validation: hypothesis test to check the dissimilarity of Spearman Matrices



Robustification of ECGs

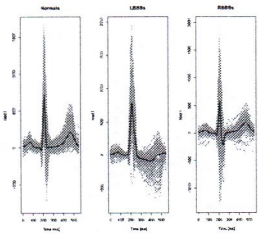
- We apply the graphical tools to identify and remove outliers in a sample of 8-variate ECG data
- By removing atypical observations, we aim at improving inference
- We come up with N = 2020 subjects (1564 Normals, 205 LBBB, 251 by RBBB).

Classification Procedure

1. Randomly sample a subset of curves (15% of the dataset)
2. Classify them computing the L^2 distances between each curve of the test set and the median of the three groups.
3. Assign a curve to the group that minimises this distance.
4. Repeat this analysis k = 40 times to avoid selection bias.

CCR = 0.9108 (±0.0163) on the robustified data
CCR = 0.9014 (±0.0148) on the original data

Performances in robustified case are stochastically greater than the others (p-value 0.0075)



[Ieva and Paganoni 2017]

Outline

(1) Depth measures for (multivariate) Functional Data

- ❑ Depth measures for (multivariate) functional data
- ❑ Epigraph and Hypograph indexes
- ❑ Spearman correlation index & matrix

(2) Graphical tools

- ❑ (Multivariate) Outliergram
- ❑ Functional Boxplot
- ❑ Outlier Detection
- ❑ roahd package

(3) Case Study: ECG signals

(4) References

Main references

- Ieva, Palma, Romo (2017) Bootstrap based inference for dependence in multivariate functional data. *Submitted*
- Ieva, F., Paganoni, A.M. (2020, online first 2016) Component-wise outlier detection methods for robustifying multivariate functional samples . Statistical Papers, 61: 595–614
- Tarabelloni, Ieva (2016) A robust Functional Boxplot for Outlier Detection in Functional Data Analysis. *Submitted*. [MOX report]
- Tarabelloni, Ieva, Paganoni, Biasi (2015) Use of depth measure for multivariate functional data in disease prediction: an application to electrocardiograph signals. International Journal of Biostatistics, 11(2), 189–201.
- Ieva, Paganoni (2013) Depth Measures for Multivariate Functional Data. Communications in Statistics–Theory and Methods. 42 (7), 1265–1276.
- Tarabelloni, Schenone, Collin, Ieva, Paganoni and Gerbeau, (2016) Statistical assessment and calibration of ECG models. *JP Journal of Biostatistics*, 15(2): 151 - 173

- Valencia, Lillo, Romo (2016) Dependence for functions: Spearman coefficient. *Submitted*
- Martin-Barragan, Lillo, Romo (2016) Functional boxplots based on epigraphs and hypographs. JAS, 43(6).
- Lopez-Pintado, Sun, Lin, Genton (2014) Simplicial band depth for multivariate functional data. ADAC, 8, 321-338.
- Claeskens, Hubert, Slaets, Vakili (2014) Multivariate Functional Halfspace Depth, JASA, 109(505), 411-423.
- Arribas-Gil, Romo (2014) Shape outlier detection and visualization for functional data: the outliergram. Biostatistics, 15 (4), 603-619.
- Lopez-Pintado, Romo (2009) On the Concept of Death for Functional Data, JASA, 104 (486), 718-734.
- Lopez-Pintado, Romo (2007) Depth-based inference for functional data, CSDA, 51(10), 4957-4968.

Main references

- Cardot, Godichon (2015) Robust principal components analysis based on the median covariation matrix. arXiv:1504.02852
- Hubert, Rousseeuw, Segaert (2015) Multivariate functional outlier detection. SMA, 24.
- Chakraborty, Chaudhuri (2014) On data depth in infinite dimensional spaces. Ann. Inst. Stat. Math. 66:303–324
- Chakraborty Chaudhuri (2014) The spatial distribution in infinite dimensional spaces and related quantiles and depths. Ann. Stat. 42 (3): 1203–1231
- Sun, Genton (2012) Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics 23 (1): 53–64.
- Kraus, Panaretos (2012) Dispersion operators and resistant second-order functional data analysis. 99 (4): 813–832.
- Sun, Genton (2011) Functional boxplots, JCGS 20: 316–334.
- Gervini (2008) Robust functional estimation using the median and spherical principal components. Biometrika 95 (3): 587–600.
- Chakraborty (2003) On multivariate quantile regression. J. Stat. Planning and Inference. 110 :109-132
- Ma, Genton (2001) Highly robust estimation of dispersion matrices. JMA 78 (1): pp. 11–36.
- Serfling, Zuo, (2000) General notions of statistical depth function, Annals of Statistics, 28, 461-482.

Packages

- **roahd** – Tarabelloni, Arribas-Gil, Ieva, Paganoni, Romo (2016), https://bitbucket.org/ntarabelloni/roahd
- **depth** - Genest, Masse, Plante, (2012). depth: Depth functions tools for multivariate analysis
- **aplpack** – Wolf, Bielefeld (2014) aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions
- **fda** - Ramsay, Wickham, Graves, Hooker, (2013) fda: Functional Data Analysis

Acknowledgments

This is a team work: thanks to

- prof. Anna Paganoni,
- PhD. Nicholas Tarabelloni,
- dr. Francesco Palma
- dr. Rachele Biasi

@ MOX

- prof. Juan Romo
- dr. Ana Arribas-Gil

@ UC3M