# FUNCTIONAL DATA ANALYSIS // ¿HOW TO DETECT OUTLIERS in UNIVARIATE FUNCTIONAL DATA (CURVES)?

RUBEN DARIO GUEVARA GONZALEZ, PhD

UNIVERSIDAD NACIONAL DE COLOMBIA - SEDE BOGOTÁ
DEPARTAMENTO DE ESTADÍSTICA

May 2, 2024

# ¿HOW TO DETECT OUTLIERS in UNIVARIATE FUNCTIONAL DATA (CURVES)?

## CONTENTS

- Functional DEPTH
- Functional OUTLYINGNESS

# ¿HOW TO DETECT OUTLIERS in UNIVARIATE FUNCTIONAL DATA (CURVES)?

**CONTENTS**

- Functional DEPTH
- Functional OUTLYINGNESS

UNIVERSIDAD NACIONAL DE COLOMBIA

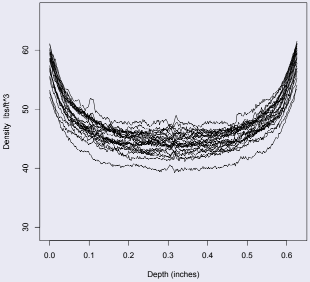
# ¿HOW TO DETECT OUTLIERS in UNIVARIATE FUNCTIONAL DATA (CURVES)?

**FUNCTIONAL DEPTH: CONTENTS**

- Functional Depth: Concept and properties
- Statistics based on Depth
  - Functional Trimmed Mean
  - Median Function
  - Functional Quartiles
- Types of Functional depth
- Functional Boxplot
- Adjusted Functional Boxplot

# PROBLEM

## How to order these curves?



# Definition

## DEPTH in $\mathbb{R}^d$

Given a distribution of probability $F$ in $\mathbb{R}^d$, a statistical depth function $D$ is a bounded function that assigns a real non-negative value $D(\mathbf{x}, F)$ to each point $\mathbf{x}$. This function satisfies the following properties:

# Definition

## DEPTH in $\mathbb{R}^d$: Properties Zuo, Y., & Serfling, R. (2000).

- Affine invariance: The depth of a point $\mathbf{x} \in \mathbb{R}^d$ should not depend on the underlying coordinate system or the scales of the underlying measurements.
- Maximality at the centre. For a distribution having a uniquely defined centre, the depth function should attain maximum value at this centre.
- Monotonicity relative to the deepest point. As a point $\mathbf{x} \in \mathbb{R}^d$ moves away from the centre, the depth at $\mathbf{x}$ should decrease monotonically.
- Vanishing at infinity. The depth of a point, $\mathbf{x}$, should approach zero as $||\mathbf{x}|| \longrightarrow \infty$, where $||\mathbf{x}||$ is the Euclidean norm of $\mathbf{x}$.

# DEPTH in $\mathbb{R}^d$

## Spatial Depth

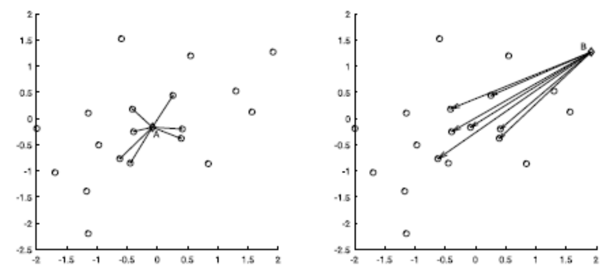



Figure: 20 observations from $N_2(\mathbf{0}, \mathbf{I}_2)$. LEFT: Vectors from a few data points towards a given point, marked as point A, in the central region of the data cloud. The average direction vector has a length close to zero.

# DEPTH in $\mathbb{R}^d$

## Spatial Depth

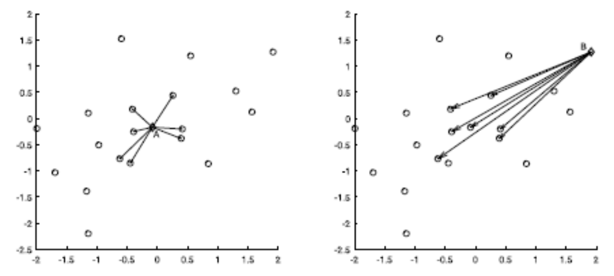



Figure: 20 observations from $N_2(\mathbf{0}, \mathbf{I}_2)$. RIGHT: Vectors from a few data points towards a given point, marked as point B, in the periphery of the data cloud. The average direction vector has a length close to one.

# DEPTH in $\mathbb{R}^d$

## Multivariate spatial sign function

Let $\mathbf{Y}$ be a $d$-dimensional random vector having cumulative distribution function $F$. Then, multivariate spatial sign function (the spatial distribution) is given by

$$S(\mathbf{x}) = \begin{cases} \frac{\mathbf{x}}{\|\mathbf{x}\|}, & \mathbf{x} \neq \mathbf{0}, \\ \mathbf{0}, & \mathbf{x} = \mathbf{0}. \end{cases}$$

# DEPTH in $\mathbb{R}^d$

## Multivariate spatial sign function

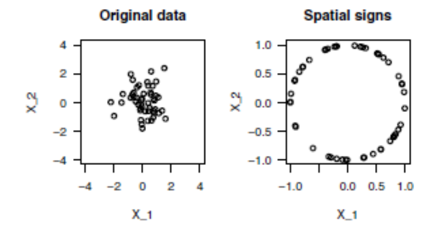



Figure: The scatterplots for a random sample of size 50 from $N_2(\mathbf{0}, \mathbf{I}_2)$ with scatterplots for corresponding observed spatial signs.

# DEPTH in $\mathbb{R}^d$

**Multivariate spatial depth**

Let $\mathbf{Y}$ be a $d$-dimensional random vector having cumulative distribution function $F$. Then, multivariate spatial depth is given by

$$SD(\mathbf{x}) = 1 - \| E[S(\mathbf{x} - \mathbb{Y})] \|$$

This definition ensures that a point near the center of the point cloud has higher depth than a point at the peripheral regions of the point cloud, and also that the depth lies between 0 and 1.

# Definition

## DEPTH in any infinite-dimensional Hilbert space $\mathbb{H}$.

- The notion of DEPTH for functional data provides a center-outward ordering of the set of curves .
- The curve with maximum depth may be defined as an estimate of the center of the functional distribution.
- The depth function in any infinite-dimensional Hilbert space $\mathbb{H}$ must satisfy the properties: Affine invariance, Maximality at the centre, Monotonicity relative to the deepest point and Vanishing at infinity.

UNIVERSIDAD NACIONAL DE COLOMBIA

# FUNCTIONAL DEPTH: Proposals

## 1. Fraiman and Muniz depth (FMD, Fraiman and Muniz 2001)

Let $\{Y(s), s \in I\}$ be a stochastic process in $C(I)$, the space of continuous functions on the interval $I$. The Fraiman and Muniz depth of $x(s) \in C(I)$ relative to $Y(s)$ is given by

$$FMD(x, Y) = \int_I D(x(s), Y(s))dt,$$

where $D(\cdot, \cdot)$ is a univariate depth.

# FUNCTIONAL DEPTH: Proposals

## 1. Fraiman and Muniz depth (FMD, Fraiman and Muniz 2001)

In particular, if the univariate simplicial depth is used as $D(\cdot, \cdot)$, we have:

$$D(x(s), Y(s)) = F_s(x(s))[1 - F_s(x(s))]$$

for any $s \in I$, where $F_s(\cdot)$ is the cumulative distribution function of $Y(s)$ at any fixed $s \in I$.

Other option is define the univariate depth as:

$$D(x(s), Y(s)) = 1 - \left|\frac{1}{2} - F(s)\right|$$

# FUNCTIONAL DEPTH: Proposals

## 1. Fraiman and Muniz depth (FMD, Fraiman and Muniz 2001)

In particular, if the univariate simplicial depth is used as $D(\cdot,\cdot)$, we have:

$$D(x(s), Y(s)) = F_s(x(s))[1 - F_s(x(s))]$$

for any $s \in I$, where $F_s(\cdot)$ is the cumulative distribution function of $Y(s)$ at any fixed $s \in I$.

Other option is define the univariate depth as:

$$D(x(s), Y(s)) = 1 - \left|\frac{1}{2} - F(s)\right|$$

# FUNCTIONAL DEPTH: Proposals

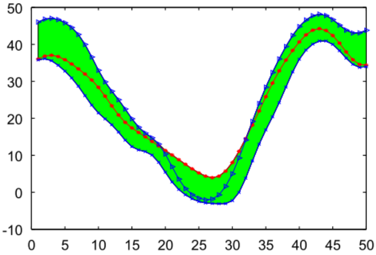
## 1. Fraiman and Muniz depth (FMD, Fraiman and Muniz 2001)

The idea consists in considering the integral of the univariate depths of $x(s)$ at each single point $s \in I$.
The sampling version $FMD_n(x)$ is defined by replacing $F_s$ by the empirical distribution $F_{n,s}$, so we have

$$FMD_n(x) = \int_I D_n(x(s), Y(s))ds = \int_I F_{n,s}(x(s))[1 - F_{n,s}(x(s))]ds,$$

# FUNCTIONAL DEPTH: Proposals

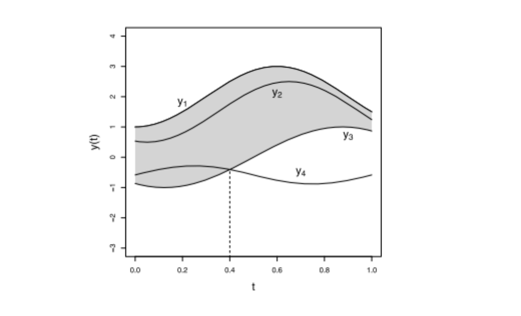
## Band in $\mathbb{R}^2$: López-Pintado and Romo(2009)




Band defined by three curves.

# FUNCTIONAL DEPTH: Proposals

## Band in $\mathbb{R}^2$: López-Pintado and Romo(2009)




Gray area: band determined by two curves, y1 and y3.

# FUNCTIONAL DEPTH: Proposals

The space of reference is restricted to the space $C_{\mathfrak{T}}$ of real continuous functions on the compact interval $\mathfrak{T}$.

**Band Depth: López-Pintado and Romo(2009)**

If $X_1, X_2, \ldots, X_j$ are independent copies of the stochastic process $X$, the population version of Band Depth is given by:

$$BD_J(x, P) = \sum_{j=2}^{J} P\left\{G(x) \subset B(X_1, \ldots, X_j)\right\},$$

where

$G(x) = \{(t, x(t) : t \in \mathfrak{T}\}$ is the subset of the plane that represents the function $x$, and

$B(X_{i1}, \ldots, X_{ik}) =$

$$\left\{(t, y) : t \in \mathfrak{T}, \min_{r=1,\ldots,k} x_{ir}(t) \leq y \leq \max_{r=1,\ldots,k} x_{ir}(t)\right\}$$

# FUNCTIONAL DEPTH: Proposals

## Band Depth: López-Pintado and Romo(2009)

$$BD_m^{(j)}(y) = \binom{m}{j}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_j \le m} I\left\{G(y) \subseteq B(y_{i1}, y_{i2}, \ldots, y_{ij})\right\}, \tag{1}$$

where $I\{\cdot\}$ represents the indicator function.

$BD_m^{(j)}(y)$ expresses the proportion of bands $B(y_{i1}, y_{i2}, \ldots, y_{ij})$ determined by $j$ different curves $y_{i1}, y_{i2}, \ldots, y_{ij}$ containing the whole graph of $y$.

# FUNCTIONAL DEPTH: Proposals

## Band Depth: López-Pintado and Romo(2009)

Let $J$ be a fixed value with $2 \leq J \leq m$. For the functions $y_1, \ldots, y_m$ the band depth of any of these curves $y$ is

$$BD_{m,J}(y) = \sum_{j=2}^{J} BD_m^{(j)}(y) \tag{2}$$

Where $J$ is an exogenous parameter that indicates the maximum number of curves used to construct a band.

# FUNCTIONAL DEPTH: Proposals

## Median

A sample median function $\widehat{y}_{0.5}$ is a curve from the sample with highest depth value,

$$y_{0.5} = \arg \max_{y \in \{x_1, \ldots, x_n\}} BD_n(x).$$

Population median is a function $y$ in $C_{\mathfrak{T}}$ maximizing $BD(\cdot)$.

If they are not unique, the median will be the average of the curves maximizing depth.

# FUNCTIONAL DEPTH: Proposals

## Modified Band Depth: López-Pintado and Romo(2009)

MBD Instead of consider the indicator function uses the proportion of times that a curve $y(t)$ is in the band.

$$MBD_m^{(j)}(y) = \binom{m}{j}^{-1} \sum_{1 \leq i_1 < \ldots < i_j \leq m} \lambda_r \left( A(y_{i1}, y_{i2}, \ldots, y_{i_j}) \right), \quad (3)$$

$2 \leq j \leq m$.

# FUNCTIONAL DEPTH: Proposals

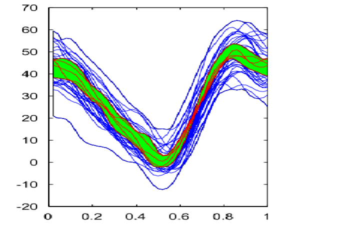
## Central Region

The $p$-central region is the band defined by the fraction $p$ of deepest sample curves:

$$C_{m,p} = B(y_{(1)}, \dots, y_{(\lceil mp \rceil)}), \tag{4}$$

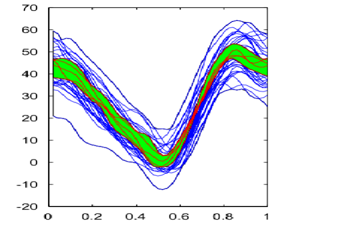
where $\lceil mp \rceil$ is the nearest integer greater than or equal to $mp$.

# FUNCTIONAL DEPTH: Proposals




The equation $p = |2q - 1|$ presents the relation between the probability used in $C_{m,p}$ and the quantiles $q$ that envelope this region.

# FUNCTIONAL DEPTH: Proposals




The equation $p = |2q - 1|$ presents the relation between the probability used in $C_{m,p}$ and the quantiles $q$ that envelope this region.

# FUNCTIONAL DEPTH: Proposals

## 3. Functional spatial depth

Let $x$ be an functional observation defined in any infinite-dimensional Hilbert space $\mathbb{H}$. The spatial sign function for $x$ is given by

$$FS(x) = \begin{cases} \frac{x}{\|x\|}, & x \neq 0, \\ 0, & x = 0. \end{cases}$$

(Chakraborty and Chaudhuri 2014)

UNIVERSIDAD NACIONAL DE COLOMBIA

# FUNCTIONAL DEPTH: Proposals

## 3. Functional spatial depth

The spatial depth function for $x$ is given by

$$FSD(x, P) = 1 - ||E\left[FS(x - Y)\right]||\,,$$

where $Y$ is a functional random variable with probability distribution $P$ on $\mathbb{H}$ (Chakraborty and Chaudhuri 2014).

# FUNCTIONAL DEPTH: Proposals

## 3. Functional spatial depth

When a sample of curves is observed, $y_1, \ldots, y_n$, $FSD(x, P)$ must be replaced with its corresponding sample version, i.e.,

$$FSD_n(x) = 1 - \frac{1}{n} \left\| \sum_{i=1}^{n} FS(x - y_i) \right\|,$$

# FUNCTIONAL DEPTH: Proposals

## 4. TOTAL VARIATION DEPTH. Huang, H., & Sun, Y. (2019)

Let $X$ be a real-valued stochastic process on $T$ with distribution $F_X$. Let $f$ a given function. For a given $t \in T$. Let

$$R_f(t) = \mathbb{I}\left[X(t) \leq f(t)\right]$$

where $\mathbb{I}$ is the indicator function. Then

$$E\left(R_f(t)\right) = p_f(t) = P\left[X(t) \leq f(t)\right].$$

If $f(t)$ is the true median, then $p_f(t) = 1/2$.

## The POINTWISE TOTAL VARIATION DEPTH of $f(t)$

$$D_f(t) = VAR\left(R_f(t)\right) = p_f(t)\left(1 - p_f(t)\right)$$

# FUNCTIONAL DEPTH: Proposals

## 4. TOTAL VARIATION DEPTH. Huang, H., & Sun, Y. (2019)

The total variation depth (TVD) for the given function $f$ on $T$ is defined as

$$TVD(f) = \int_T w(t) D_f(t) dt,$$

where $w(t)$ is a weight function defined on $T$ .

## Ways to choose the weight function

- $w(t) = \frac{1}{|T|}$
- $w(t)$ is proportional to the vertical variability at different time points, Claeskens et al. (2014).

# FUNCTIONAL DEPTH

How is the trimmed mean?

# FUNCTIONAL DEPTH

## Other Functional Depths

- Extremal Depth (Narisetty, N. N., & Nair, V. N. (2016)).
- Total Variation Depth (Huang, H., & Sun, Y. (2019)).
- Pareto Depth (Helander, S., Van Bever, G., Rantala, S., & Ilmonen, P. (2020)).
- Curve Depth (de Micheaux, P. L., Mozharovskyi, P., & Vimond, M. (2020))

# FUNCTIONAL DEPTH

## Packages in R

- fda
- fda.usc
- fdaoutlier
- curveDepth

# FUNCTIONAL OUTLIERS

**What is a functional Outlier?**

According to Febrero et al. (2007, 2008), a functional outlier is a curve generated by a stochastic process with a different distribution than the distribution of the normal curves.

# FUNCTIONAL OUTLIERS

## Types of functional Outliers

- Magnitude Outliers
- Shape Outliers
- Partial Outliers: curves having atypical behaviors only in some segments of the domain.

Shape and partial outliers are typically harder to detect than magnitude outliers (in the case of high magnitude, outliers can even be recognized by simply looking at a graph), and therefore entail more difficult outlier detection problems

# FUNCTIONAL OUTLIERS

## Types of functional Outliers

- Magnitude Outliers
- Shape Outliers
- Partial Outliers: curves having atypical behaviors only in some segments of the domain.

Shape and partial outliers are typically harder to detect than magnitude outliers (in the case of high magnitude, outliers can even be recognized by simply looking at a graph), and therefore entail more difficult outlier detection problems

# FUNCTIONAL OUTLIERS

## Types of functional Outliers

- Magnitude Outliers
- Shape Outliers
- Partial Outliers: curves having atypical behaviors only in some segments of the domain.

Shape and partial outliers are typically harder to detect than magnitude outliers (in the case of high magnitude, outliers can even be recognized by simply looking at a graph), and therefore entail more difficult outlier detection problems

# FUNCTIONAL OUTLIERS

## Types of functional Outliers

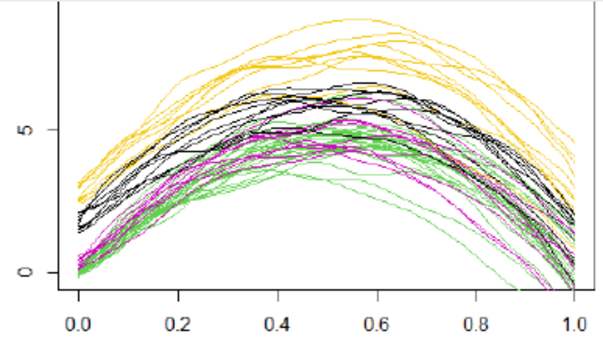
- Magnitude Outliers
- Shape Outliers
- Partial Outliers: curves having atypical behaviors only in some segments of the domain.

Shape and partial outliers are typically harder to detect than magnitude outliers (in the case of high magnitude, outliers can even be recognized by simply looking at a graph), and therefore entail more difficult outlier detection problems

UNIVERSIDAD NACIONAL DE COLOMBIA

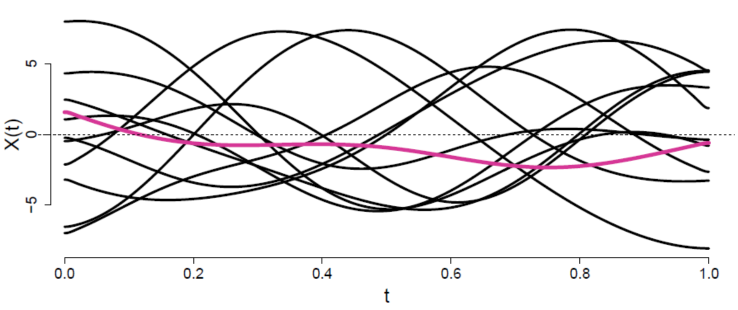
# OUTLIERS

## Magnitude Outliers



# OUTLIERS

## Shape Outliers




A sample of Gaussian processes with a shape outlier highlighted in pink.

# FUNCTIONAL OUTLIERS

## Outlier Detectionfor Functional Data

The outlier detection problem can be described as follows: let $Y_n = \{y_1, \ldots, y_n\}$ be a sample that has been generated from a mixture of two functional random variables in an infinite-dimensional Hilbert space, one for normal curves and one for outliers, say $Y_{nor}$ and $Y_{out}$, respectively. Let $Y_{mix}$ be this mixture, which is given by:

$$Y_{mix} = \begin{cases} Y_{nor}, & \text{with probability } 1-\alpha \\ Y_{out}, & \text{with probability } \alpha \end{cases}$$

where $\alpha \in [0,1]$ is the contamination probability (usually, a value rather close to 0). See Sguera, Galeano and Lillo (2016).Q

# FUNCTIONAL OUTLIERS

## Outlier Detectionfor Functional Data

The outlier detection problem can be described as follows: let $Y_n = \{y_1, \ldots, y_n\}$ be a sample that has been generated from a mixture of two functional random variables in an infinite-dimensional Hilbert space, one for normal curves and one for outliers, say $Y_{nor}$ and $Y_{out}$, respectively. Let $Y_{mix}$ be this mixture, which is given by:

$$Y_{mix} = \begin{cases} Y_{nor}, & \text{with probability } 1-\alpha \\ Y_{out}, & \text{with probability } \alpha \end{cases}$$

where $\alpha \in [0, 1]$ is the contamination probability (usually, a value rather close to 0). See Sguera, Galeano and Lillo (2016).Q

# FUNCTIONAL OUTLIERS

## Outlier Detectionfor Functional Data

The outlier detection problem can be described as follows: let $Y_n = \{y_1, \ldots, y_n\}$ be a sample that has been generated from a mixture of two functional random variables in an infinite-dimensional Hilbert space, one for normal curves and one for outliers, say $Y_{nor}$ and $Y_{out}$, respectively. Let $Y_{mix}$ be this mixture, which is given by:

$$Y_{mix} = \begin{cases} Y_{nor}, & \text{with probability } 1-\alpha \\ Y_{out}, & \text{with probability } \alpha \end{cases}$$

where $\alpha \in [0, 1]$ is the contamination probability (usually, a value rather close to 0). See Sguera, Galeano and Lillo (2016).Q

# FUNCTIONAL OUTLIERS

## Outlier Detection for Functional Data

The curves composing $Y_n$ are all unlabeled, and the goal of the analysis is to decide whether each curve is a normal curve or an outlier.

Since any functional depth measures the degree of centrality (or extremality) of a given curve relative to a distribution or a sample, outliers are expected to have low depth values.

# FUNCTIONAL OUTLIERS

## Outlier Detection for Functional Data

The curves composing $Y_n$ are all unlabeled, and the goal of the analysis is to decide whether each curve is a normal curve or an outlier.

Since any functional depth measures the degree of centrality (or extremality) of a given curve relative to a distribution or a sample, outliers are expected to have low depth values.

# Functional Boxplot

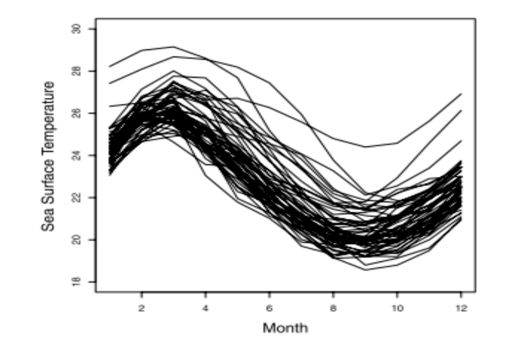
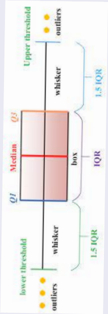
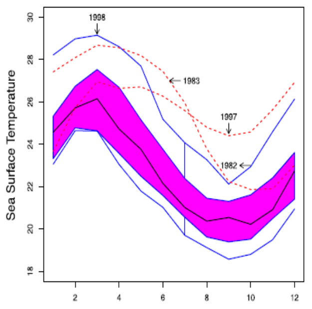



Figure: Data of monthly sea surface temperatures measured in degrees Celsius over the east-central tropical Pacific Ocean from 1951 to 2007.

# OUTLIERS

## Functional Boxplot Sun & Genton (2011)

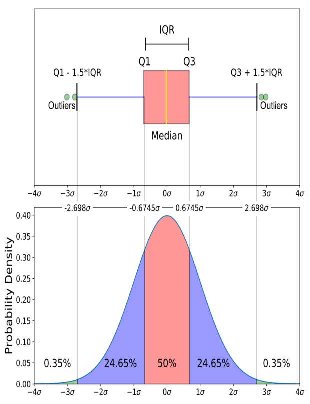



Data of monthly sea surface temperatures measured in degrees Celsius over the east-central tropical Pacific Ocean from 1951 to 2007.

# OUTLIERS

## Boxplot



# ADJUSTED FUNCTIONAL BOXPLOT

## F=1.5?

In a functional boxplot, potential outliers can be detected by the 1.5 times the 50% central region empirical rule, analogous to the rule for classical boxplots. The factor $F = 1.5$ in a classical boxplot can be justified by a standard normal distribution, because it leads to a probability of 99.3% for correctly detecting no outliers.

# ADJUSTED FUNCTIONAL BOXPLOT

## F=1.5?

For functional data, there will be necessarily dependence in time for each curve. And for spatio-temporal data, curves from different locations will be spatially correlated as well. The outlier detection performance may be affected by the dependence in time and space. Sun & Genton (2012) study the relationship between the dependence and the constant factor, and then propose a method based on simulation to adjust the factor in a functional boxplot by controlling the probability of detecting not outliers to be 99.3% when actually no outliers are present.

# ADJUSTED FUNCTIONAL BOXPLOT

## F=1.5?

Sun & Genton (2012) study the relationship between the dependence and the constant factor, and then propose a method based on simulation to adjust the factor in a functional boxplot by controlling the probability of detecting not outliers to be 99.3% when actually no outliers are present.

# OUTLIERS

## Adjusted Functional Boxplot Sun & Genton (2012)

- The 1.5 constant factor in the functional boxplot can be replaced by a different quantity with the aim of controlling the probability of correctly detecting no outliers.
- The method consists in simulating observations without outliers on the basis of a robust estimator of the covariance function of the data.
- The factor is then selected as the one for which, when used with the functional boxplot, the probability of detecting no outliers is the closest to 0.993. The selected factor is applied to the functional boxplot of the original data.

# ADJUSTED FUNCTIONAL BOXPLOT

## F=1.5?

Sun & Genton (2012) proposed:

- To estimate the covariance matrix of the data
- To Generate observations without any outliers. They generated a small number of curves, $n = 100$, without any outliers at $p$ time points from the model $Z(\mathbf{s}, t) = g(\mathbf{s}, t) + e(s, t)$, with $E(g(\mathbf{s}, t)) = 0, (\mathbf{s}, t) \in R^2 \times R^2$. Here $e(\mathbf{s}, t)$ is a Gaussian random field with mean zero and covariance function estimated from the standarized original data.

UNIVERSIDAD NACIONAL DE COLOMBIA

# ADJUSTED FUNCTIONAL BOXPLOT

## F=1.5?

Sun & Genton (2012) proposed:

- To calculate the coverage probabilities for different values of the constant factor.
- To select the constant factor, $F$, closes to 99.3%.
- To run the functional boxplot with the adjusted constant factor.

UNIVERSIDAD NACIONAL DE COLOMBIA

# OUTLIERS

## Robust estimator of a dispersion matrix Ma & Genton (2001)

$$Cov(X,Y) = \frac{\sigma_X \sigma_Y}{4}\left[Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) - Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)\right],$$

where $\sigma_X = Var(X)$ and $\sigma_Y = Var(Y)$.

## Highly robust estimator of scale Rousseeuw and Croux (1992, 1993)

Let $z_1, \ldots, z_n$ be a sample of a random variable $Z$.

$$\hat{\sigma}_Z = Q_n(Z) = d\,[|z_i - z_j|; i < j, i, j = 1, 2, \ldots, n]_{(k)}$$

where $d = 2.2191$ is a consistency factor and
$k = \lfloor \left(\binom{n}{2} + 2\right)/4 \rfloor + 1$, approximately the first quartile for large $n$.

# Functional Boxplot

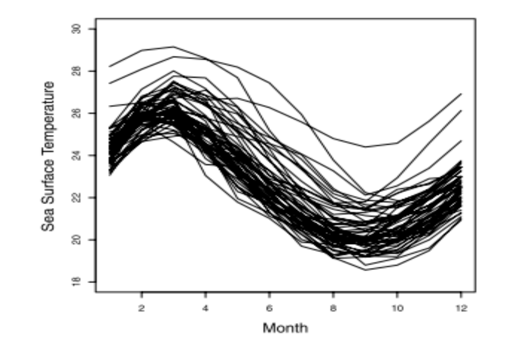



Figure: Data of monthly sea surface temperatures measured in degrees Celsius over the east-central tropical Pacific Ocean from 1951 to 2007.

# OUTLIERS

## Adjusted Functional Boxplot Sun & Genton (2012)

- We assume that these annual temperature curves are independent copies of each other and estimated the $12 \times 12$ covariance matrix in time. Robust techniques are needed when considering the potential presence of outliers in the original data.

UNIVERSIDAD NACIONAL DE COLOMBIA

# FUNCTIONAL OUTLIERS: DETECTION based on DEPTH

## Adjusted Functional Boxplot Sun & Genton (2012)

- In simulations, by generating $n = 100$ curves at $p = 12$ time points from a Gaussian process with zero mean and estimated covariance function the coverage probabilities for different values of the constant factor are listed in this Table with 1.000 replications.

| Factor $F$ | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | **1.8** | 1.9 | 2.0 |
|---|---|---|---|---|---|---|---|---|---|
| Coverage | 0.768 | 0.859 | 0.922 | 0.956 | 0.979 | 0.988 | **0.995** | 1.000 | 1.000 |

Note: The selected factor is in bold font.

Data of monthly sea surface temperatures

- We select the constant factor with the coverage probability close to 99.3%.

# FUNCTIONAL OUTLIERS: DETECTION based on DEPTH

**The shape outlyingness plot + functional boxplot, Huang, H., & Sun, Y. (2019)**

Let $X$ be a real-valued stochastic process on $T$ with distribution $F_X$. Let $f$ a given function. For a given $t \in T$. Let

$$R_f(t) = \mathbb{I}\left[X(t) \leq f(t)\right]$$

The POINTWISE TOTAL VARIATION DEPTH of $f(t)$

$$D_f(t) = VAR\left(R_f(t)\right) = p_f(t)\left(1 - p_f(t)\right)$$

# FUNCTIONAL OUTLIERS: DETECTION based on DEPTH

**The shape outlyingness plot + functional boxplot,** Huang, H., & Sun, Y. (2019)

**Decomposition of pointwise TVD**

from the law of the total variance

$$\begin{aligned} D_f(t) &= VAR\,(R_f(t)) \\ &= VAR[E\{Rf(t)|Rf(s)\}] + E[VAR\{Rf(t)|Rf(s)\}] \\ &= \text{The shape componente} + \text{The magnitude componente.} \end{aligned}$$

where $s, t \in T$ with $s = t - \Delta$.

# FUNCTIONAL OUTLIERS: DETECTION based on DEPTH

## The shape outlyingness plot + functional boxplot, Huang, H., & Sun, Y. (2019)

### The modified shape similarity (MSS) of $f$

$$MSS(f, \Delta) = \int_T \nu(t, \Delta) S_{\tilde{f}}(t, \Delta) dt$$

where

$$\tilde{f}(s, \Delta) = \begin{cases} \text{median}\,(X(s)) & \text{if } s = t, \\ f(s) - f(t) + \text{median}\,(X(t)) & \text{if } s = t - \Delta. \end{cases}$$

with $S_{\tilde{f}}(t, \Delta) = \frac{VAR[E\{Rf(t)|Rf(s)\}]}{D_{\tilde{f}}(t)}$.
The population estimation is done on the discrete time points, and at each $t_i$, $i > 1$, we let $\Delta = t_i - t_{i-1}$.

# FUNCTIONAL OUTLIERS: DETECTION based on DEPTH

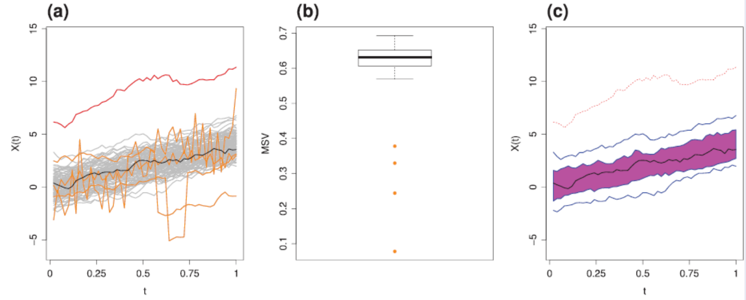
**The shape outlyingness plot + functional boxplot,** Huang, H., & Sun, Y. (2019)

Suppose we observe $n$ sample curves, then the outlier detection procedure is:

1. Estimate the total variation depth and modified shape similarity (MSS) for each curve.
2. Draw a classical boxplot for the $n$ values of the modified shape similarity (MSS) and detect outliers. This plot detects all shape outliers.
3. Remove detected shape outliers and draw a functional boxplot using the total variation depth. This plot detects all themagnitude outliers.

# FUNCTIONAL OUTLIERS: DETECTION based on DEPTH

## The shape outlyingness plot + functional boxplot, Huang, H., & Sun, Y. (2019)




(a) Observations with magnitude outliers (red) and shape outliers (orange). (b) MSS plot: orange points for shape outliers. (c) Functional boxplot using TVD after removing detected shape outliers.

# FUNCTIONAL DEPTH

## References


- Biasi, R., Ieva, F., Paganoni, A. M., & Tarabelloni, N. (2014). Multivariate functional data depth measure based on variance-covariance operators. Contributions in Infinite-Dimensional Statistics and Related Topics, Società Editrice Esculapio, Bologna, 49-54.
- Claeskens, G., Hubert, M., Slaets, L., & Vakili, K. (2014). Multivariate functional halfspace depth. Journal of the American Statistical Association, 109(505), 411-423.
- Febrero, M., Galeano, P., & González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. Environmetrics, 19(4), 331-345.

# FUNCTIONAL DEPTH

## References


- Fraiman, R., & Muniz, G. (2001). Trimmed means for functional data. Test, 10(2), 419-440.
- Guevara, R. D., & Vargas, J. A. (2016). Evaluation of process capability in multivariate nonlinear profiles. Journal of Statistical Computation and Simulation, 86(12), 2411-2428.
- Huang, H., & Sun, Y. (2019). A decomposition of total variation depth for understanding functional outliers. Technometrics.
- Ieva, F., & Paganoni, A. M. (2013). Depth measures for multivariate functional data. Communications in Statistics-Theory and Methods, 42(7), 1265-1276.

# FUNCTIONAL DEPTH

## References


- López-Pintado, S., & Romo, J. (2009). On the concept of depth for functional data. Journal of the American Statistical Association, 104(486), 718-734.
- Sguera, C., Galeano, P., & Lillo, R. E. (2016). Functional outlier detection by a local depth with application to NO x levels. Stochastic Environmental Research and Risk Assessment, 30(4), 1115-1130.
- Sun, Y. & Genton, M. (2011): Functional Boxplots, Journal of Computational and Graphical Statistics, 20:2, 316-334
- Sun, Y., & Genton, M. G. (2012). Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. Environmetrics, 23(1), 54-64.