Machine Learning with Functional Data

Exploratory analysis and alignment

The Concept of "Functional Expectation": Strong and Weak Integrals

Let X be a random element defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ taking values in a separable Banach space $(\mathcal{X}, \|\cdot\|)$. The expectation

$$\mathbb{E}(X) = \int_{\Omega} X \, d\mathbb{P}$$

can be defined in the following ways:

- If X is a simple function (a linear combination of indicators), $\mathbb{E}(X)$ is defined naturally (linearity).
- When X is a general random element, it is expressed as the limit of simple functions, and $\mathbb{E}(X)$ is defined as the corresponding limit of the integral.

This is the Bochner integral or *strong integral*, which exists if and only if $\mathbb{E}\|X\| < \infty$. Bosq (2012).

The Concept of "Functional Expectation": Strong and Weak Integrals

We usually work with random functions $X = X(t, \omega)$, $t \in [0, 1]$, $\omega \in \Omega$. In this context, there is another natural way to define $\mathbb{E}(X) = \mathbb{E}(X)(t)$ by calculating the ordinary expectation for each t, $\int_\Omega X(t, \omega)\, d\mathbb{P}(\omega)$.

The general version of this idea is called the Pettis integral or *weak integral*. A random variable taking values in a Banach space $\mathcal{X}$ is Pettis integrable if there exists an element $EX \in \mathcal{X}$ such that $\mathbb{E}(x^*(X)) = x^*(EX)$ for all $x^* \in \mathcal{X}^*$ (the continuous dual of $\mathcal{X}$).

If the Bochner integral exists and is finite, then it coincides with the Pettis integral.

As in the case of classical inference models the Gaussian distribution plays an outstanding role in FDA. Let us recall that a stochastic process $X = X(t)$ is Gaussian if and only if all the finite-dimensional projections $(X(t_1), \ldots, X(t_k))$, for $t_1, \ldots, t_k \in [0, 1]$ and $k \in \mathbb{N}$ have Gaussian distributions.

Thus, the distribution of a Gaussian process is uniquely determined by the mean function $m(t) = \mathbb{E}(X(t))$ and its covariance function $\gamma(s, t) = \mathrm{Cov}(X(s), X(t))$. Two examples of special interest are (standard) *Brownian motion* for which $m = 0$ and $\gamma(s, t) = \min(s, t)$ and the (standard) *Brownian bridge*, for which $m = 0$ and $\gamma(s, t) = \min(s, t)(1 - \max(s, t))$.

La expansión de Karhunen-Loève

Desarrollo en serie de funciones ortonormales para procesos estocásticos.

KARHUNEN-LOÈVE THEOREM. Let $X = X(t)$, $t \in [0,1]$, be a stochastic process with $\mathbb{E}(X(t)) = 0$ and $\mathbb{E}(X^2(t)) < \infty$ for all $t \in [0,1]$. Suppose that the covariance function $\gamma(s,t)$ is continuous.
Then $X(t)$ can be expressed in the form

$$X(t) = \sum_{k=1}^{\infty} Z_k e_k(t), \tag{1}$$

where the convergence is in L^2, uniform in t, $\{e_k\}_{k\in\mathbb{N}}$ is an orthonormal basis of $L^2[0,1]$ given by the eigenfunctions of the covariance operator Γ, associated with $\gamma(s,t)$, whose corresponding eigenvalues are λ_k (that is $\lambda_k e_k(t) = \int_0^1 \gamma(s,t)e_k(s)ds$) and $Z_k = \int_0^1 X(t)e_k(t)dt$ is a sequence of orthogonal uncorrelated random variables with $\mathbb{E}(Z_k) = 0$, $\mathbb{E}(Z_i^2) = \lambda_i$.

Remark: The Z_k's in (1) are always uncorrelated, but independence is only guaranteed in the Guassian case. This a further reason why Gaussianity is a so common assumption in FDA.

Functional Law of Large Numbers

Functional Strong Law of Large numbers [Mourier, 1953]. Let $\{X_n\}$ be a sequence of iid random elements with values in a separable Banach space $\mathcal{X}$. If $\mathbb{E}\|X_1\| < \infty$ then

$$\frac{\sum_{i=1}^{n} X_i}{n} \xrightarrow{a.s.} \mathbb{E}(X_1),$$

where $\mathbb{E}(X_1)$ denotes the strong expectation of X_1.

The functional Central Limit Theorem

Teorema Central del Límite funcional [Varadhan, 1962].
Let $\{X_n\}$ be a sequence of iid random elements with values in a separable Hilbert space $\mathcal{X}$. If $\mathbb{E}\|X_1\|^2 < \infty$ then

$$\sqrt{n}\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mathbb{E}(X_1)\right) \xrightarrow{w} \mathcal{G}(0, \Gamma_{X_1}),$$

where $\mathcal{G}(0, \Gamma_{X_1})$ denotes the Gaussian probability measure on $\mathcal{X}$ with expectation 0 and covariance operator $\Gamma_{X_1}(x^*, y^*) = Cov(x^*(X_1), y^*(X_1))$, for $x^*, y^* \in \mathcal{X}^*$ (the continuous dual of $\mathcal{X}$)

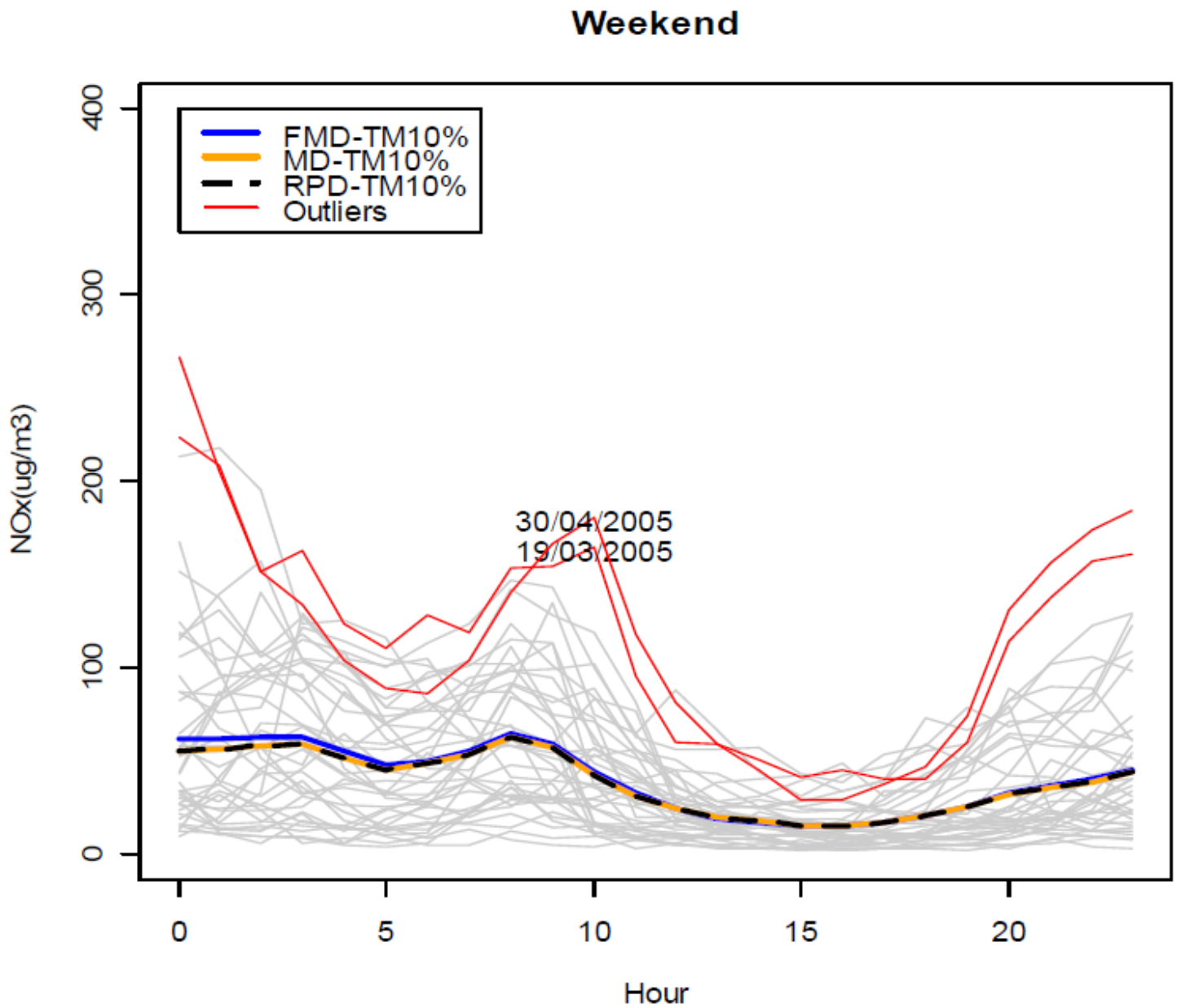
Random projections

This result from Cuesta, Fraiman, and Ransford (2007) generalizes the Cramér-Wold theorem, which characterizes a probability measure in terms of one-dimensional projections.

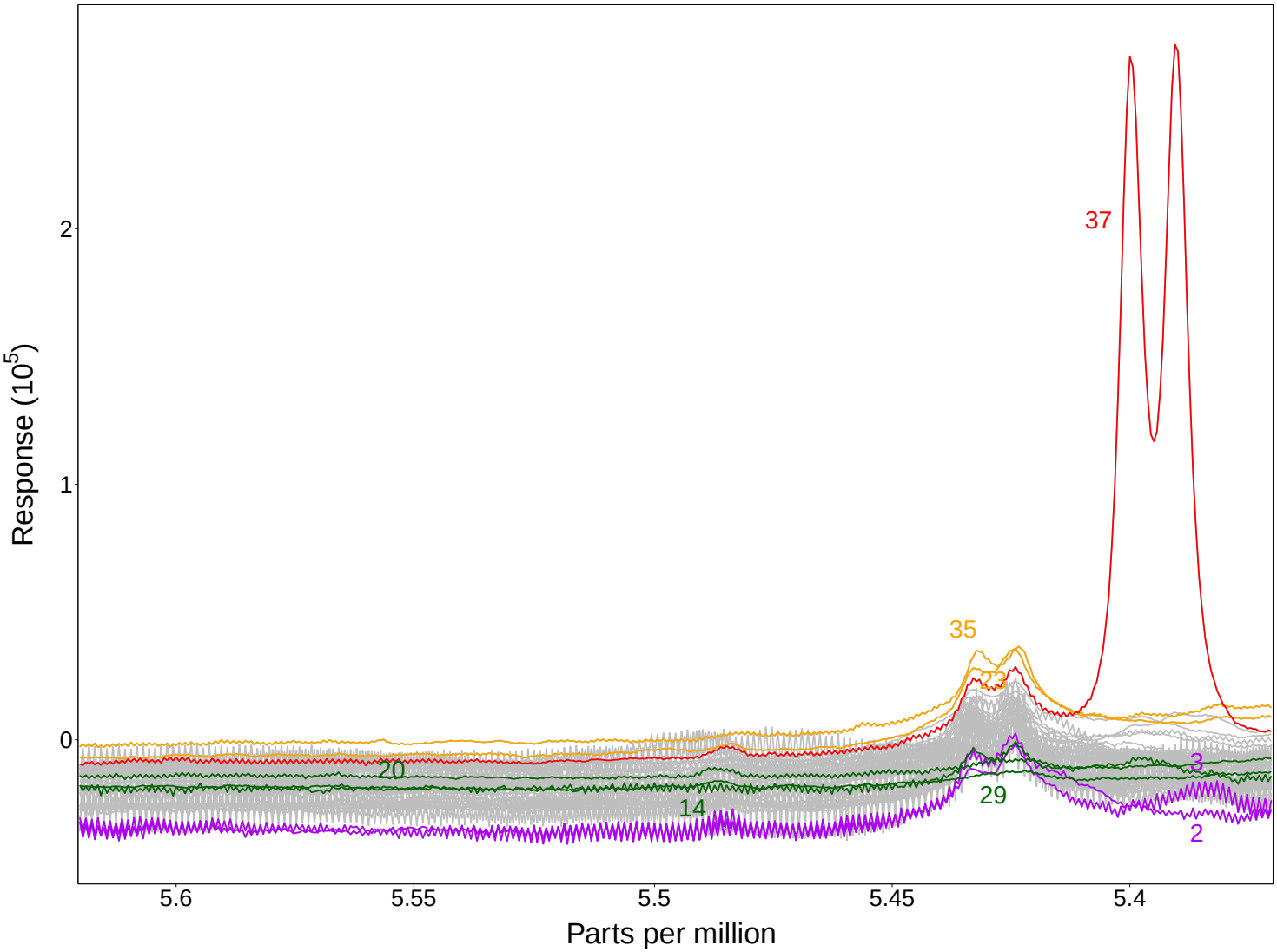
THEOREM. *Let X, Y be random elements taking values in a separable Hilbert space $(\mathbb{H}, \langle\cdot\rangle)$ with distribution P and Q, respectively. Let μ be a non-degenerate Gaussian measure on $\mathbb{H}$. Denote by $\|\cdot\|$ the norm in $\mathbb{H}$ associated with the inner product $\langle\cdot\rangle$. Assume that*
(a) The absolute moments $m_n = \mathbb{E}\|X\|^n = \int \|x\|^n dP(x)$ satisfy the Carleman condition $\sum_n m_n^{-1/n} = \infty$.
(b) $\mu\{v \in \mathbb{H} : \langle v, X\rangle \stackrel{d}{=} \langle v, Y\rangle\} > 0$, where the notation $\stackrel{d}{=}$ stands for equality in distribution.
Then, $P = Q$.

The idea is that under the conditions of the theorem, the probability distribution in a Hilbert space is determined by the one-dimensional projections on any set of positive measure. In other words, given a Gaussian reference measure μ, if two probability distributions are distinct, the μ-probability of finding two equally distributed one-dimensional linear projections is zero.

The centrality notion



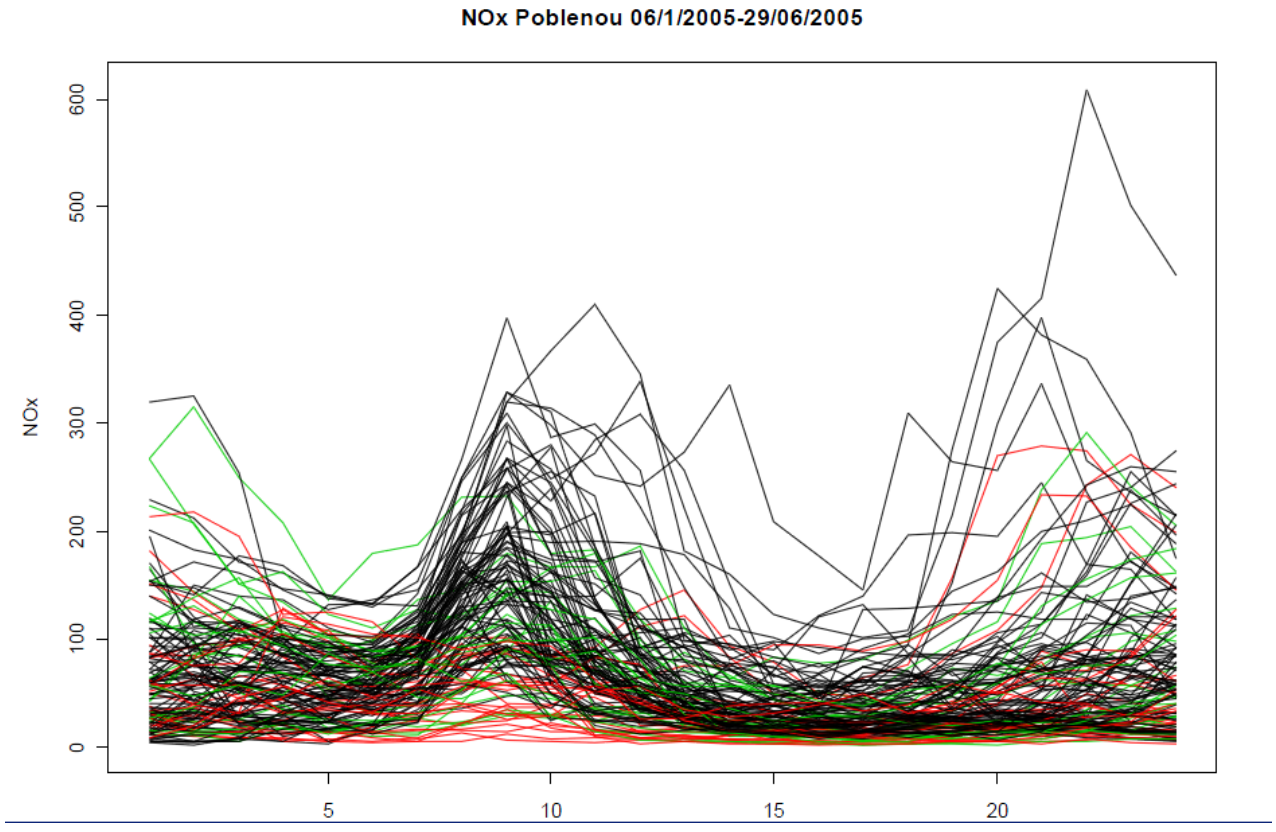
The centrality notion



Hubert et al. (2015)

NOx example

Hourly levels of NOx (nitrogen oxides) measured in an environmental control station in Poblenou (Barcelona) in $\mu g/m^3$. 127 daily records from January 6, 2005 to June 26, 2005. Febrero et al. (2008)



In $L^2[0,1]$ (or assimilable spaces) the sample mean is computed as,

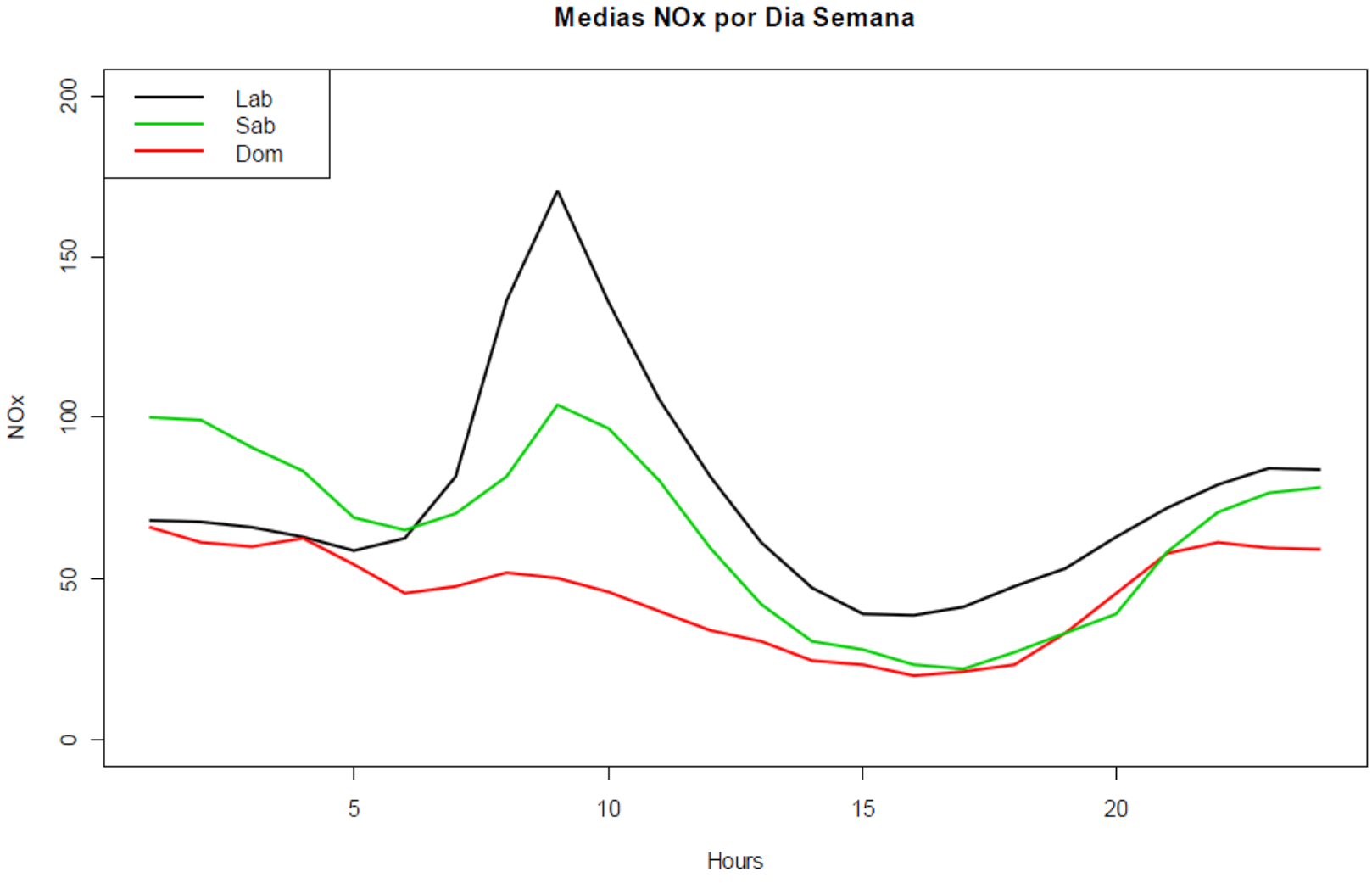
$$\bar{X}(t) = \frac{1}{N} \sum X_i(t).$$

Otherwise, we can refer to the definition:

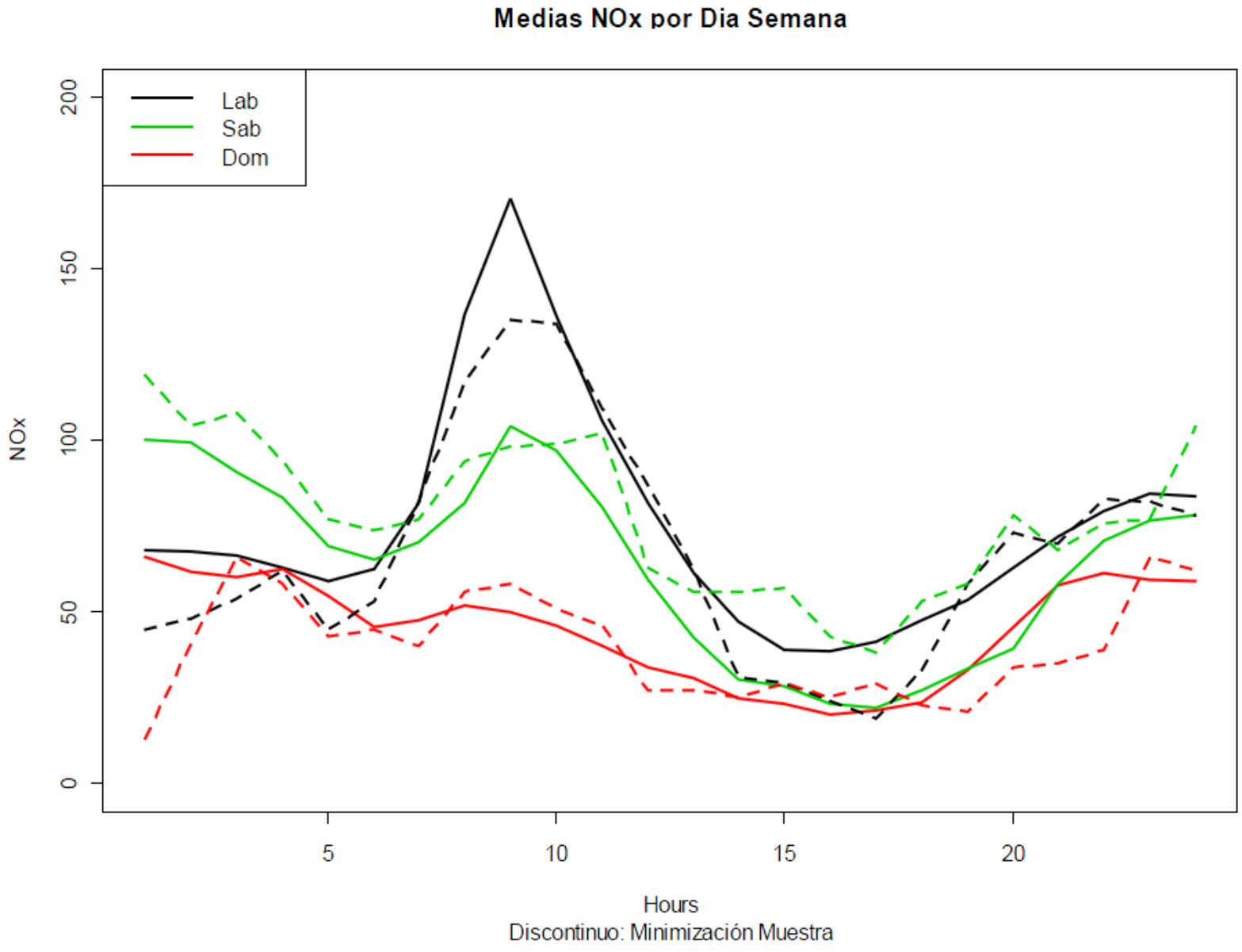
$$\bar{X}(t) = argmin_{f \in \mathcal{F}} \mathbb{E} \parallel X - f \parallel^2$$

It can be difficult to get the minimum in the whole space (solution: minimize in the sample).

Mean


Medias NOx por Dia Semana
Lab
Sab
Dom
NOx
0
50
100
150
200
5
10
15
20
Hours

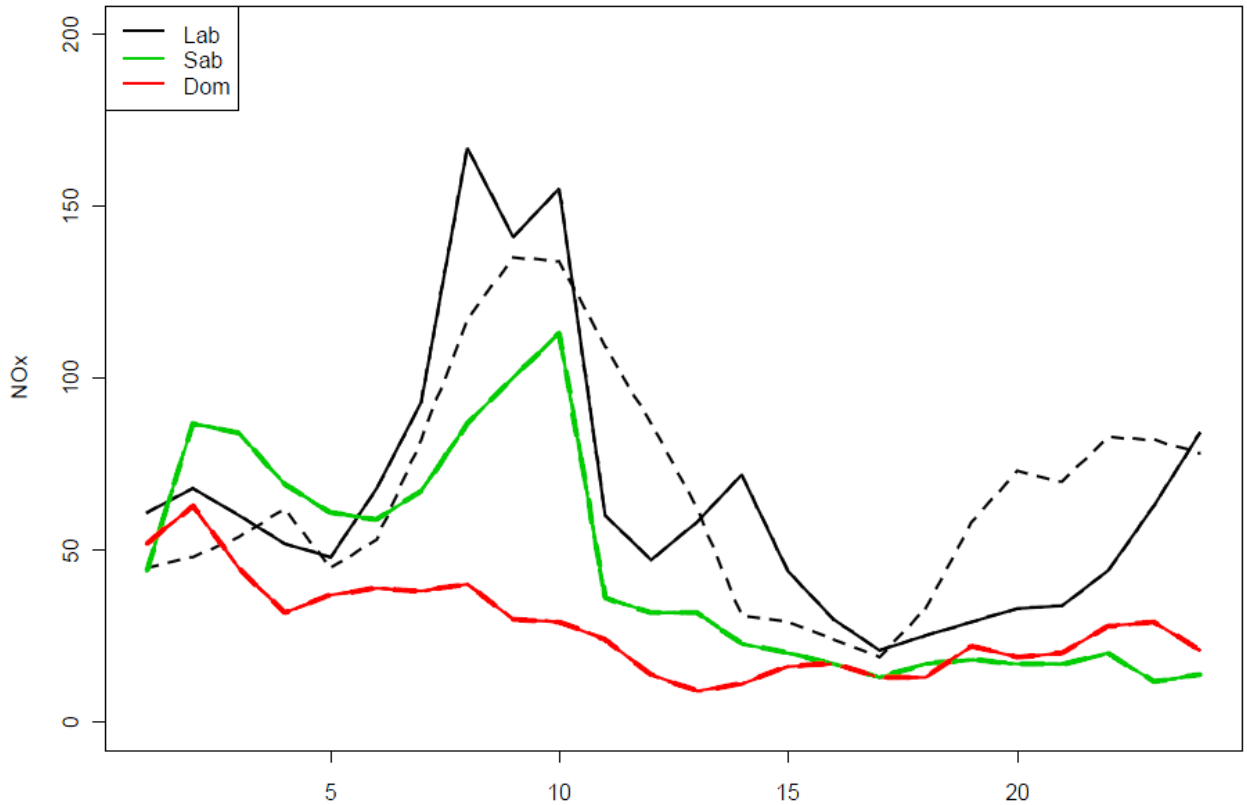
Mean


Medias NOx por Dia Semana
Lab
Sab
Dom
NOx
0
50
100
150
200
5
10
15
20
Hours
Discontinuo: Minimización Muestra

Median

The *median* $M = M(X)$ can be defined as the argument that minimizes $\Psi(a) = \mathbb{E}(\|X - a\| - \|X\|)$.
Again, it can be replaced by a minimization along the sample functions,

$$\hat{M} = \text{argmin}_a \sum_{i=1}^{n} \|X_i - a\|.$$



Functional mode

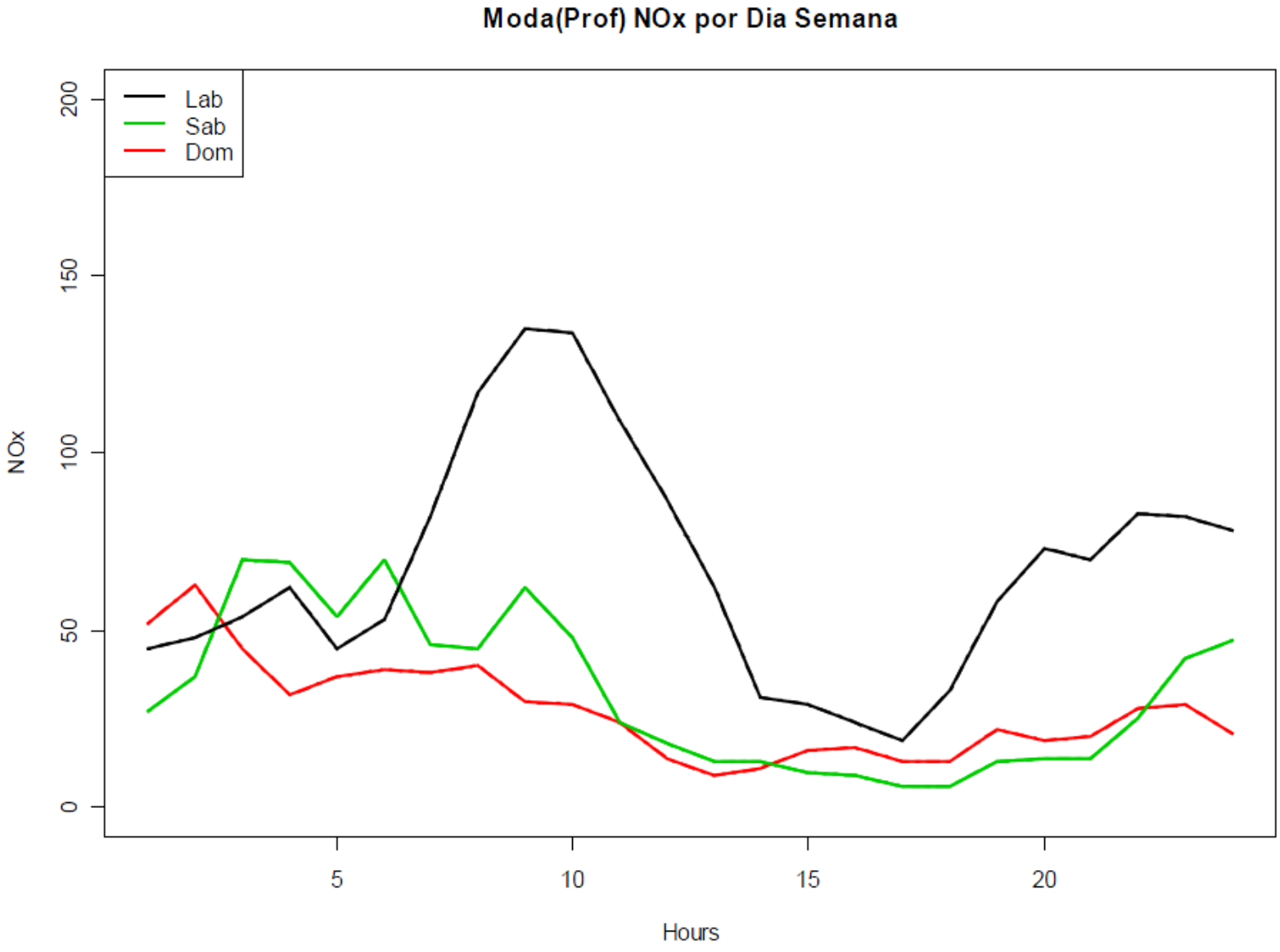
Problem: The lack of a single and natural notion of density of functional spaces.

h-mode Cuevas, Febrero and Fraiman (2007): given a proper kernel function K (ej., the standard Gaussian one or the cuadratic kernel $K(t) = \frac{3}{2}(1 - t^2)\mathbb{I}_{[0,1]}(t)$) the functional h-mode can be defined as follows,

$$M_0 = \text{argmax}_a \mathbb{E} K\left(\frac{\|a - X\|}{h}\right).$$

In some sense the mode is the "most surrounded" datum.

Moda funcional



Depth measures

The study of *depth measures* has become a recurrent topic in FDA.

Given a probability measure P on the sample space $\mathcal{X}$, a *depth function* $D(P, x)$ is a non-negative function defined on $\mathcal{X}$ that indicates how deep an observation x is in the distribution P. Let $\mathbb{P}_n$ be the empirical distribution of the sample $X_1, \ldots, X_n$, then $D(\mathbb{P}_n, x)$ indicates the depth of x in that sample.

For univariate data, typical depth functions include the mean, median, and mode. For multivariate data, computing depth measures becomes increasingly difficult as the dimension increases.

Desirable properties of a depth measure include: affine invariance, maximality at the center, monotonicity, zero at infinity, and invariance under dimensionality reduction.

Given a depth measure, one can define the median as the deepest point in the sample. Similarly, percentiles and truncated means can be defined.

Fraiman-Muniz Depth is an integral depth measure based on the naturalness of order statistics in one dimension.

Let $F_{n,t}$ be the empirical distribution function of $x_1(t), \ldots, x_n(t)$. The univariate depth of each data point $x_i(t)$ is denoted by $D_i(t) = 1 - |\frac{1}{2} - F_{n,t}(x_i(t))|$. The depth index for each i is defined as follows:

$$I_i = \int_0^1 D_i(t)dt.$$

Integrated Depth: A general approach to get an infinite-dimensional depth function from the one-dimensional depths of the projections is proposed Cuevas and Fraiman (2009): the basic idea is just to define a new depth function by integrating out the one-dimensional depths. This leads to a definition of type

$$D(P, x) = \int D_i(P_f, f(x)) dQ(f),$$

where D_i is a one-dimensional depth function, f denotes an element in the continuous dual space, P_f is the (one-dimensional) distribution of $f(X)$ (the projection of X via the real linear continuous f) and Q is a probability measure on the continuous dual space of f's.

the methods of class IntegratedDepth can be used to compute integrated depth measures. The default choice is the measure propose by Fraiman and Muniz.

The practical aspects of this idea, as well as some comparisons with other methods, have been considered in Cuevas, Febrero and Fraiman (2007).
We shall not discuss here depth functions in further detail. Let us just remark the fact that every depth function has an associated notion of median (the deepest point), quantiles (defined in the obvious way after sorting the points according to their depths) and α-trimmed mean

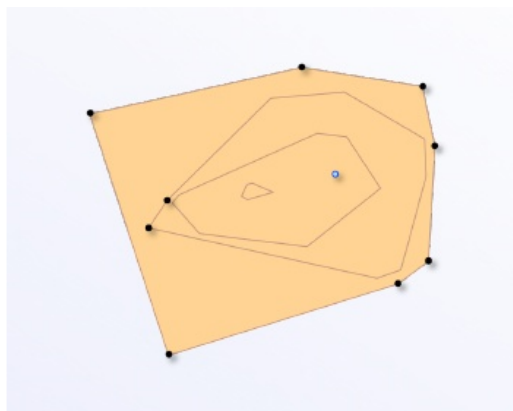
Tukey depth

In the case $\mathcal{X} = \mathbb{R}$, the usual depth functions (which are equivalent in various cases) are given by

$$D_0(P, x) = P(-\infty, x]P[x, \infty) \text{ and } D_1(P, x) = \min(P[x, \infty), P(-\infty, x]).$$

Tukey depth: In both cases, we obtain the usual definition of median and quantiles. However, the multivariate case $\mathcal{X} = \mathbb{R}^d$ is more interesting. The generalization of D_1 is called the Tukey depth (or halfspace depth), which is defined as the infimum of the probabilities of all halfspaces containing x. (see Zuo and Serfling, 2000)





Tukey's Depth II

Cuesta-Albertos and Nieto-Reyes (2008) propose a random version of D_T to minimize computational cost. Let us take k projection directions $v_1, \ldots, v_k$ iid.

$$D_{RT}(P, x) = \inf_{1 \leq i \leq k} D_1(P_{v_i}, \langle v_i, x \rangle),$$

where P_{v_i} is the distribution of $\langle v_i, X \rangle$ (the one-dimensional projection of X onto the direction v_i).

This definition can be extended (maintaining almost the same definition) to the case where $\mathcal{X}$ is a separable Hilbert space (for example $L^2[0, 1]$).

This definition can be extended (with minor modifications) to the case where $\mathcal{X}$ is a separable Hilbert space, such as $L^2[0,1]$. In this case, the random vectors v_i are replaced with elements from the continuous dual space $\mathcal{X}^*$. By the Riesz Representation Theorem, these dual elements v are associated with kernel functions a that define a linear continuous operator $x \mapsto \int_0^1 a(s)x(s)ds$, which replaces the finite-dimensional operator $x \mapsto \langle v, x \rangle$. Therefore, choosing the dual elements v corresponds to randomly selecting the corresponding kernel functions a using an appropriate L^2 process.

Spatial depth:

$$SD(x) = 1 - \mathbb{E}\| \frac{x - X}{\|x - X\|} \|$$

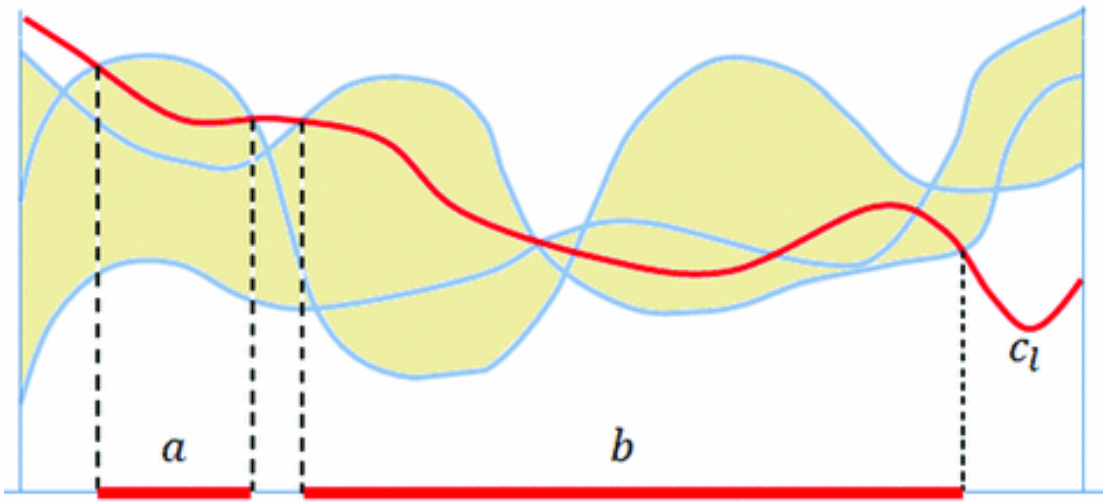
The idea behind the theorem of random projections is that under certain conditions, the probability distribution in a separable Hilbert space is determined by the one-dimensional linear projections onto any set of positive measure. That is, given a Gaussian reference measure μ, if two probability distributions are different, the μ-probability of finding two identically distributed one-dimensional linear projections is zero.

Depth Given a set $X_1, \ldots, X_n$, a random and independent direction is chosen, and the data is projected onto it. Then, the depth of each X_i is defined with some simple one-dimensional depth measure. In the case of functional data, we can assume that we are in $L^2[0,1]$ and use its inner product for the projection. Once this depth estimator is obtained, other directions can be taken and the results averaged.

Depth measures in FDA

An alternative definition is the band depth (BD), introduced by López-Pintado and Romo (2009). To compute this functional depth, one needs to identify the bands that are delimited by all possible pairs of functional observations in the sample. The BD value is the fraction of bands that completely encompass the curve. In *scikit-fda*, this quantity can be computed using methods of the class BandDepth.

A related, less restrictive measure, is the modified band depth (MBD). This measure takes into account not only the number of bands that contain the function, but also the time that $x(t)$ lies within each band. The MBD has better statistical properties than the original BD, in part because it is an integrated depth measure (Nagy et al. 2016). In *scikit-fda*, MBD is implemented in the class ModifiedBandDepth.

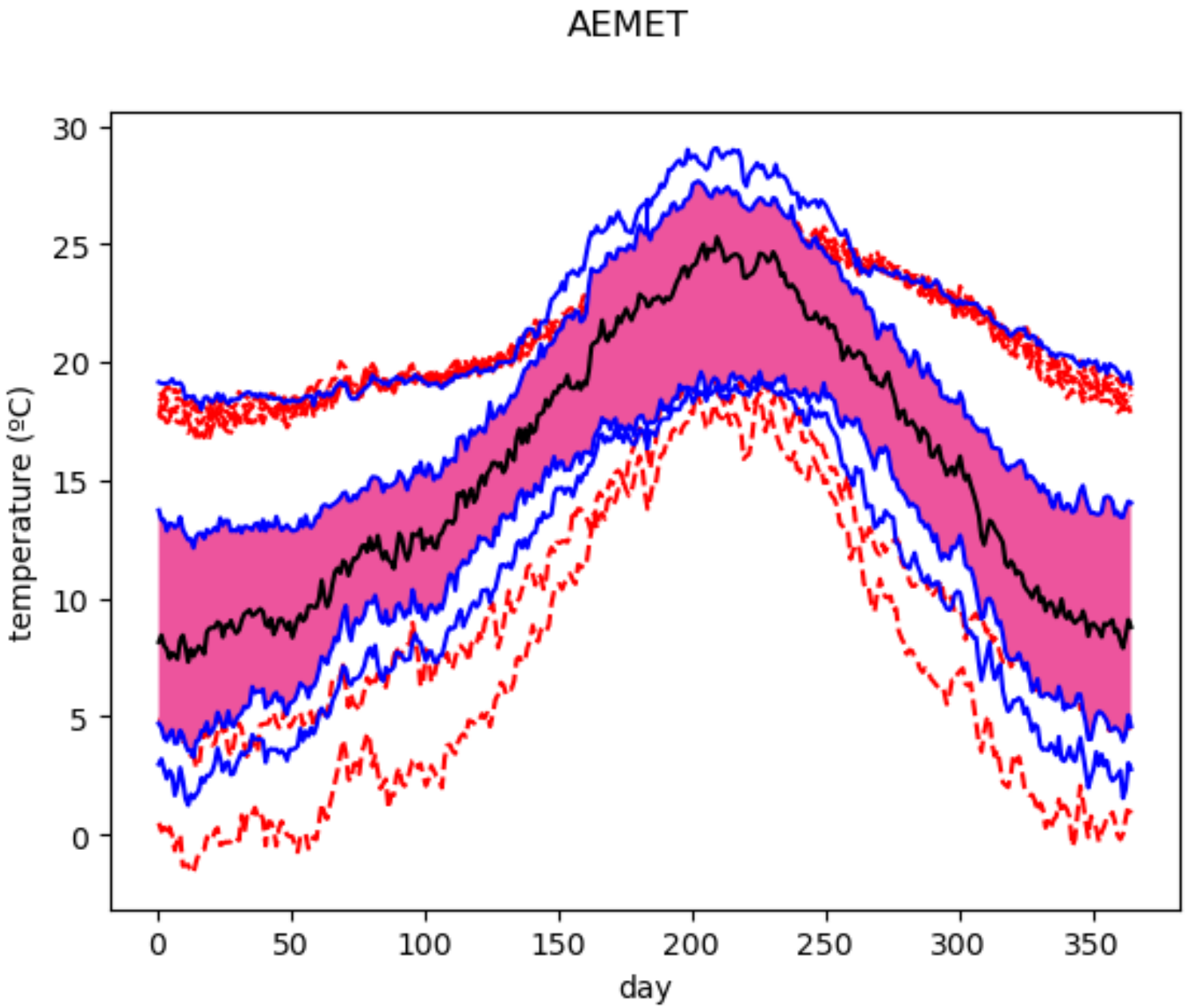


Boxplot funcional

The functional boxplot (Sun and Genton 2011) is a generalization of the univariate boxplot for functional data. I consists of a graph of the functional median surrounded by a central envelope, which encompasses the deepest 50% of the observations, and a maximum non-outlying envelope. The width of this outer envelope is determined by scaling the central one by a constant factor. Its default value is 1.5. but it can be selected by the user. The class Boxplot can be used to generate and customize functional boxplots.

In this plot, a trajectory is marked as an outlier if it lies beyond the maximum non-outlying envelope for some interval. The class BoxplotOutlierDetector can be used for outlier detection based on this criterion. Some customizable elements of Boxplot objects are the depth measure, and the fraction of the deepest observations that define the inner bands.

Boxplot funcional



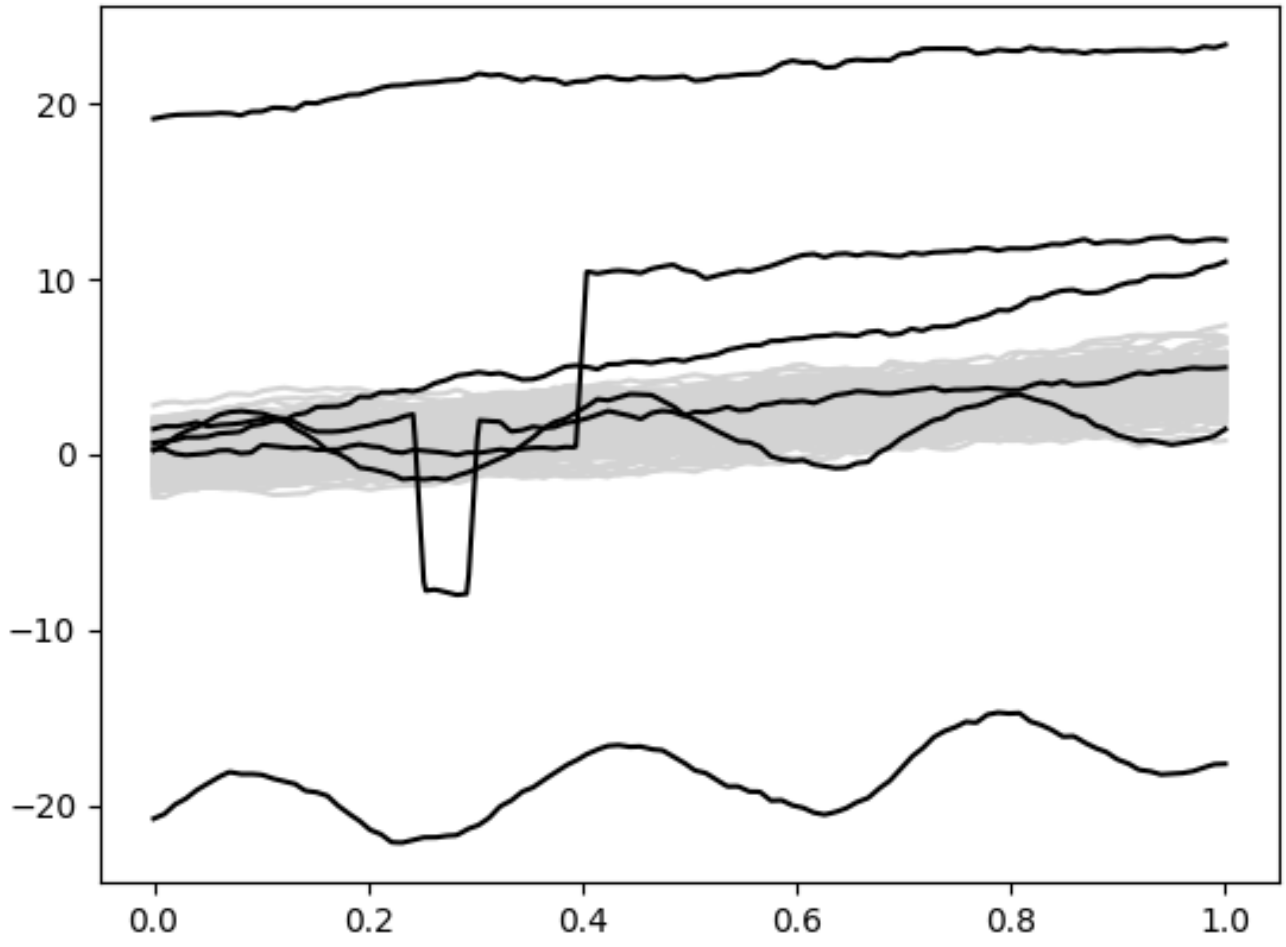
Magnitud-Shape plot

The magnitude-shape (MS-plot) (Dai and Genton 2018, 2019) characterizes the degree of outlyingness of a functional observation is characterized in terms of two quantities: the magnitude outlyingness (MO) and the shape outlyingness (VO).

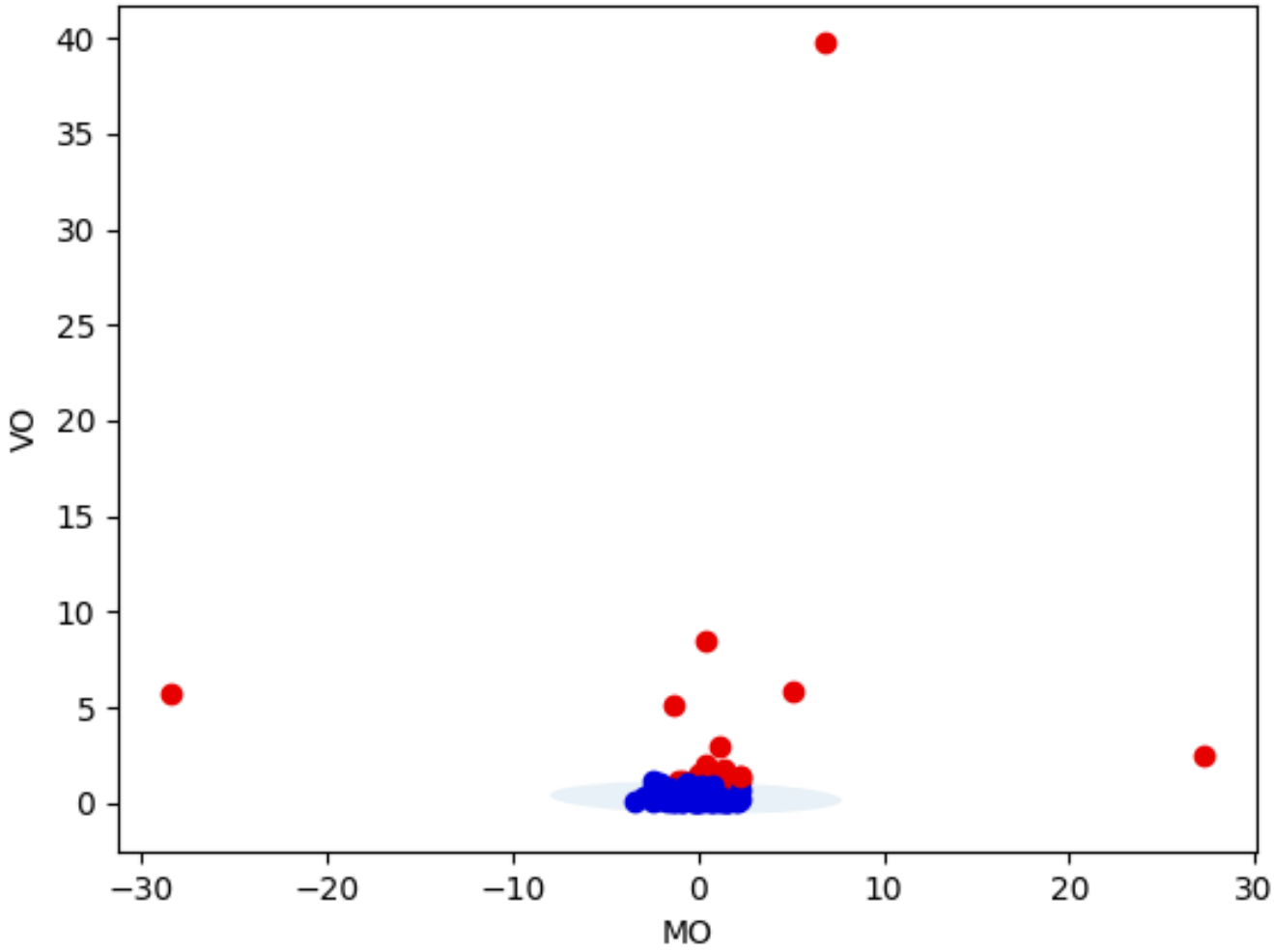
The MS-plot is the scatter plot of the values MO and VO for each functional observation. This two-dimensional representation of the data can be used, for instance, to identify clusters of functions, or detect potential outliers, either in shape or in magnitude.

The class MagnitudeShapePlot generates the MS-plot and uses internally the methods of the class MSPlotOutlierDetector for outlier detection.

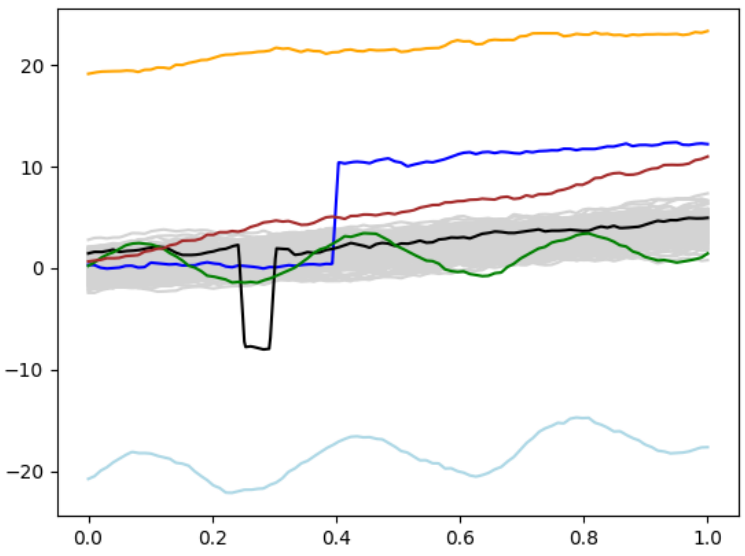
Magnitude-Shape plot

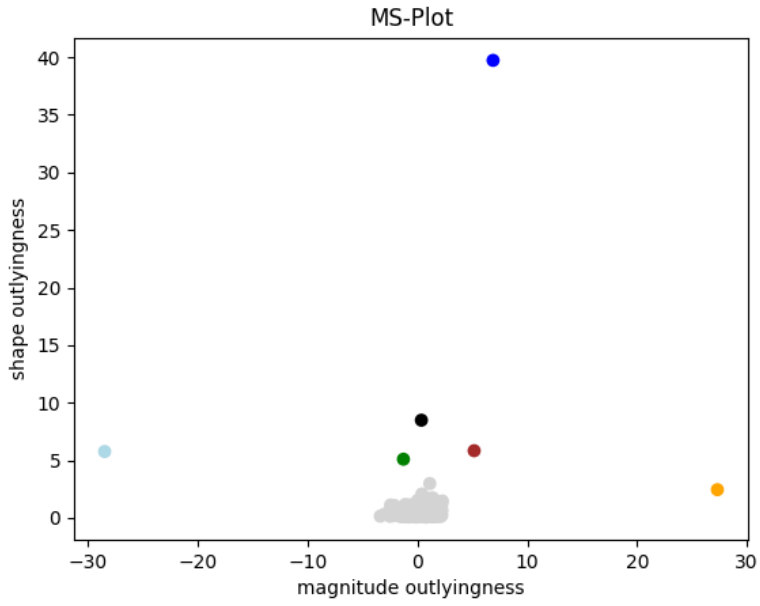


Magnitude-Shape plot



Magnitude-Shape plot





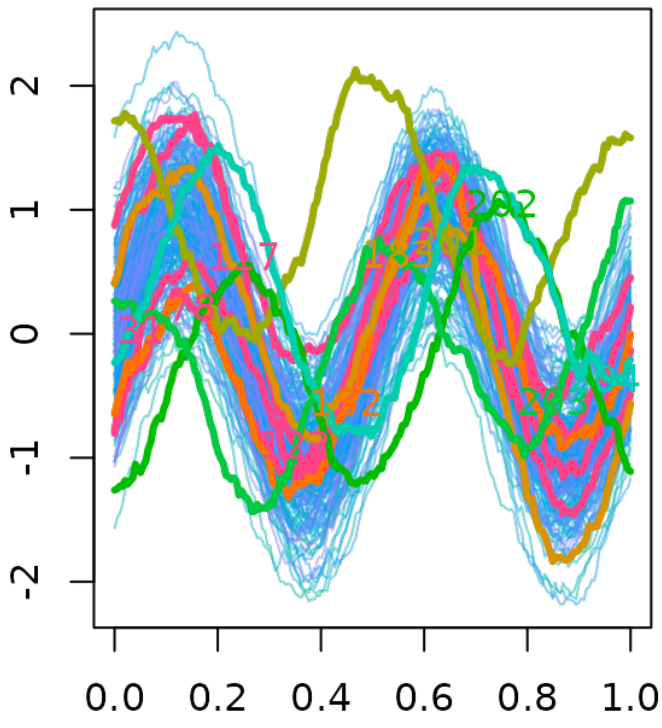
Outliergram

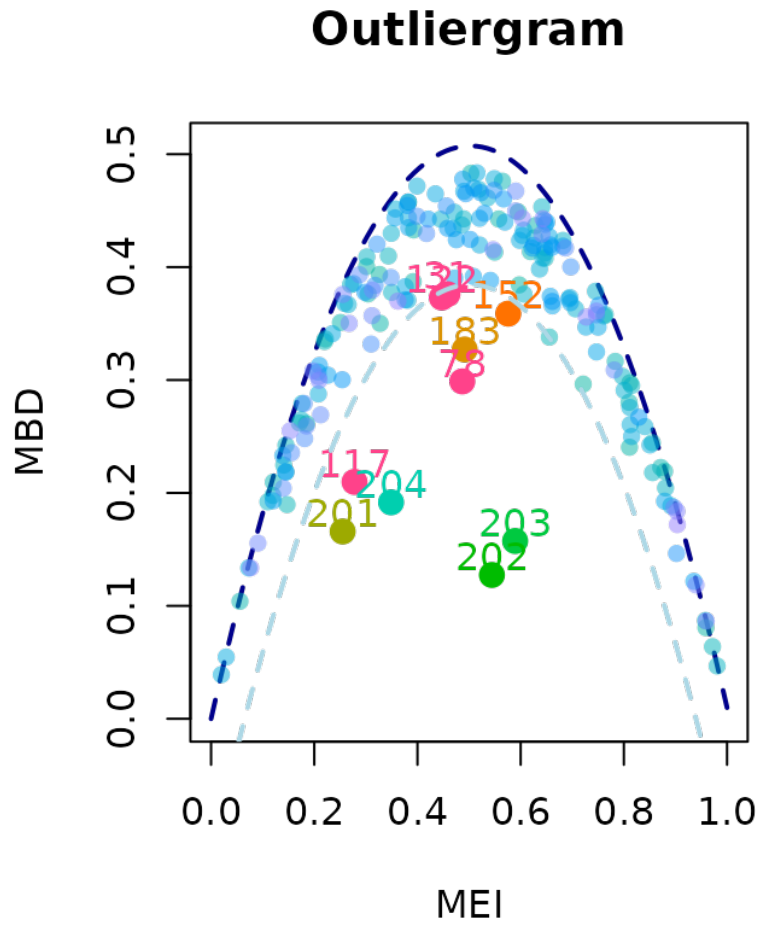
The class Outliergram provides an additional method for data visualization and detection of shape outliers (Arribas-Gil and Romo 2014). The graph is defined in terms of two related quantities: the modified epigraph index (MEI) and the MBD.

The MEI of a trajectory is the average over time of the fraction of curves in the sample that lie above it. Each curve is a point (MEI, MBD) in the scatter plot.

The outliergram takes advantage of the fact that points corresponding to typical functional observations lie on a parabola, whose analytical form is known. This parabola is used as a reference for the identification of shape outliers. Specifically, the degree of outlyingness of a curve is quantified in terms of its vertical distance to the parabola.

Outliergram





Registration

Registration consists in applying transformations to the raw data so that the functional observations are properly aligned.

A number of strategies can be used for registration. For instance, maxima, minima, zeros, and other landmarks can be used as reference. Alternatively, some measure of dispersion between the observations can be minimized. It is also possible to register a set of functional observations to a reference function.

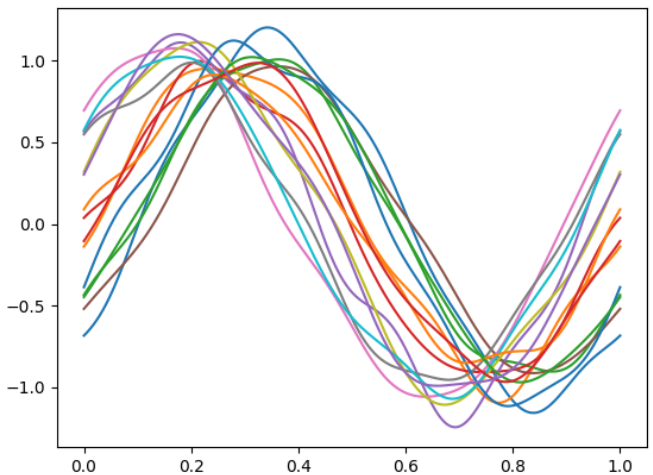
After registration, it may be necessary to evaluate the functional observations at points in the domain that are different from the ones in the original grid. This can be made utilizing the interpolation and extrapolation techniques.

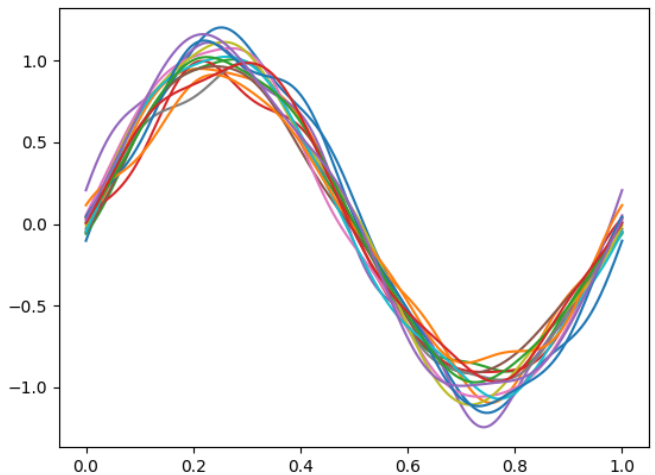
Shift registration

Shift registration consists in aligning the functional observations by a translation

$$\hat{x}_i(t) = x_i(t + \delta_i), \quad i = 1, \ldots, n,$$

The function landmark_shift_registration() can be used to carry out this transformation. The values of the δ_i can be retrieved using the landmark_shift_deltas() function.



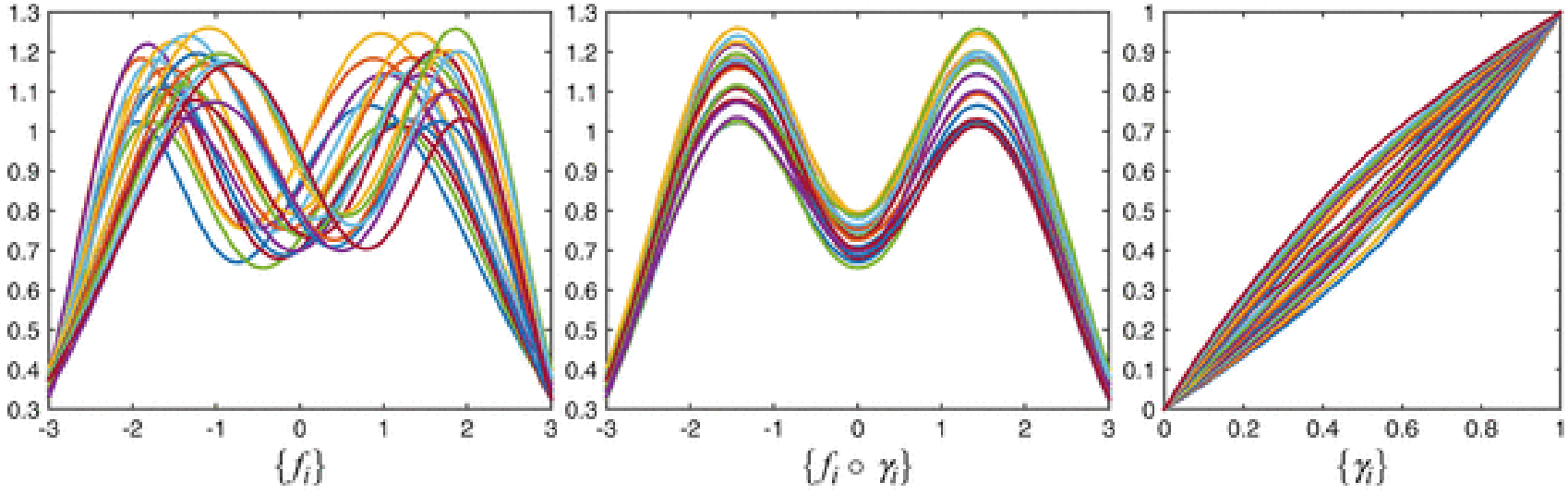


Elastic registrationIn elastic

In elastic registration, one attempts to align the data by applying a warping transformation

$$\hat{x}_i(t) = x_i(\gamma_i(t)), \quad i = 1, \ldots, n,$$

In scikit-fda we have landmark_elastic_registration() for elastic landmark registration (the warpings can be retrieved with landmark_elastic_registration_warping()), and FisherRaoElasticRegistration for elastic registration to a common template.



Clustering

Unsupervised classification: the k-means procedure

Given a probability measure P on the space $\mathcal{X}$, an $\mathcal{X}$-valued random element X with distribution P and $k \in \mathbb{N}$, let us define the k-mean parameter, associated with P as any set $\{h_1^P, \ldots, h_k^P\}$ of k *cluster centers*, $h_i^P \in \mathcal{X}$ minimizing (on all possible sets $\{h_1, \ldots, h_k\}$, with $h_i \in \mathcal{X}$) the following expression

$$I_k(P; h_1, \ldots, h_k) := \mathbb{E}\left(\min_{i=1,\ldots,k} \|X - h_i\|^2\right) \tag{2}$$

The intuitive idea is very simple: we are just looking for the cluster centers h_i such that the expected distance from a random observation X to its nearest center in $\{h_1, \ldots, h_k\}$ is minimal.

Unsupervised classification: the k-means procedure

The sample version of (2), based on data $X_1 \ldots, X_n$, leads to minimize on $\{h_1, \ldots, h_k\}$ the natural empirical approximation of $I_k(P; h_1, \ldots, h_k)$, that is,

$$I_k(\mathbb{P}_n; h_1, \ldots, h_k) := \frac{1}{n} \sum_{i=1}^{n} \|X_i - h_{c(i)}\|^2, \tag{3}$$

where $h_{c(i)}$ denotes the cluster center, in $\{h_1, \ldots, h_k\}$, closest to X_i.
Now, given the set of cluster centers $\{h_1, \ldots, h_k\}$, those observations having the same closest center are in the same cluster.
Of course, the exact computation of the optimal cluster centers, even in the empirical version (3), is a formidable task. Different approximate (often randomized) algorithms have been proposed.

A simple algorithm for the k-means procedure

(i) We divide randomly the n data into k groups and calculate the value, denoted by I_k^0 of I_k for such partition. Here the centers $h_{c(i)}$ would be just the averages of the corresponding groups.

(ii) Then, we first re-assign (if needed) X_1 from its original cluster to another one, in order to minimize I_k^0 . Denote by I_k^1 the resulting value of I_k, after re-assigning X_1 (the centers $h_{c(i)}$ can of course change after that).

(iii) Now, we proceed sequentially by re-assigning X_2, X_3, etc.

(iv) The algorithm stops when no observation needs to be re-assigned. The resulting centers are estimators of the true centers h_i.

The consistency of the k-means procedure

It is not trivial to prove that the minimizers $\{\hat{h}_1, \ldots, \hat{h}_k\}$ of $I_k(\mathbb{P}_n; h_1, \ldots, h_k)$ converge to the minimizers of $I_k(P; h_1, \ldots, h_k)$. Note that the order is not relevant here, so that we just need that the sequence of sets $\{\hat{h}_1, \ldots, \hat{h}_k\}$ "converges" to $\{h_1, \ldots, h_k\}$. We can use the Hausdorff metric, defined for compact non-empty sets,

$$d_H(A, C) = \inf\{\epsilon > 0 : A \subset B(C, \epsilon),\ C \subset B(A, \epsilon)\}.$$

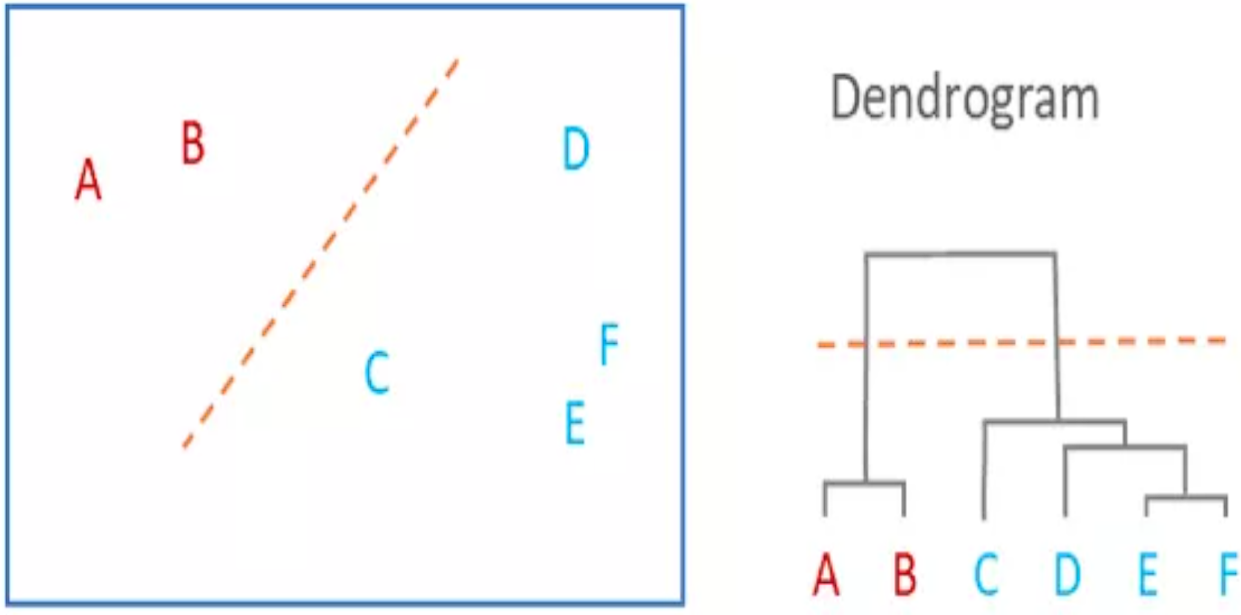
where $B(A, \epsilon) := \bigcup_{a \in A} B(a, \epsilon)$ denotes the ϵ-parallel set.
See Pollard (1981,1982), Cuesta and Matrán (1988).

A different approach to clustering: agglomerative (hierarchical) algorithms

- Choose a criterion of proximity between two groups of observations (clusters). For example: the distance between the two closest observations of both groups (single linkage), or the distance between the two farthest observations (complete linkage), or the distance between the centroids (averages) of both groups.
- start with a extreme situation in which there are as many clusters as observations. So, each cluster is made of just one observation.
- Then, join the two closest clusters.
- Recalculate the distances between clusters and iterate the previous step until the procedure leads us two clusters which are "too far from each other".

A different approach to clustering: agglomerative (hierarchical) algorithms

When we don't have too many observations this algorithm is usually summarized in a **dendogram** as this one



Equivalent or mutually singular distributions

Let P_0 and P_1 be two probability distributions defined on the same space.

- They are equivalent ($P_0 \sim P_1$) if, for all $A \in \mathcal{F}$,

$$P_0(A) = 0 \iff P_1(A) = 0.$$

- They are mutually singular ($P_0 \perp P_1$) if there exists $A \in \mathcal{F}$ with $P_0(A) = 0$ and $P_1(A) = 1$.

If P_0 and P_1 are Gaussian, then they are equivalent or mutually singular.

If $P_0 \sim P_1$, there exist Radon-Nikodym derivatives satisfying

$$P_1(A) = \int_A \frac{dP_1}{dP_0} dP_0 \quad \text{and} \quad P_0(A) = \int_A \frac{dP_0}{dP_1} dP_1.$$

When measures are equivalent, the optimal classification rule is a function of the R-N derivative.

Theorem 1 (Baíllo et al., Scand. J. Stat. 2011)

$$g^*(X) = 1 \Leftrightarrow \frac{dP_1}{dP_0}(X) > 1$$

The interesting point is that the Radon-Nikodym derivative is explicitly known (and not too complicate) in several important examples when P_0 and P_1 are Gaussian. See e.g. Parzen (Annals Math. Stat. 1961) Varberg (Pacific J. Math. 1961, Trans. Amer. Math. Soc 1964), Shepp (Annals Math. Stat. 1966), Kailath (IEEE Trans. Inf. Theory 1971) Lipster and Shiryaev (2013, 1st ed. 1977) or applications of Cameron-Martin Theorem in Mörters and Peres (2010).

The model

$$\begin{cases} P_0: & X(t) = & m_0(t) + \xi(t), & t \in [0,1] \\ P_1: & X(t) = & m_1(t) + \xi(t), & t \in [0,1] \end{cases}$$

- $\xi(t)$ Gaussian with $\mathbb{E}(\xi(t)) = 0$.
- $K(s,t) = \mathbb{E}(\xi(s)\xi(t))$
- Prior probabilities: $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$.
- $m(t) = m_1(t) - m_0(t)$.

The Brownian case: Cameron-Martin Theorem

Cameron-Martin Theorem Mörters and Peres (2010, p.24) Let $F \in \mathcal{C}[0,1]$ such that $F(0) = 0$. Let P_0 and P_F be the distribution of the standard Brownian motion, $B(t)$ and $B_F(t) = B(t) + F(t)$, respectively. Denote by $\mathcal{D}[0,1]$ de Dirichlet space $\mathcal{D}[0,1] = \{F : [0,1] \to \mathbb{R} : F(t) = \int_0^t f(s)ds, \text{for some } f \in L^2[0,1]\}$. Then,

a) If $F \notin \mathcal{D}[0,1]$, then $P_F \perp P_0$.

b) If $F \in \mathcal{D}[0,1]$, then $P_F \sim P_0$. Moreover,

$$\frac{dP_F}{dP_0}(B) = \exp\left(-\frac{1}{2}\int_0^1 F'(s)^2 ds + \int_0^1 F' dB\right),$$

for P_0-almost all $B \in \mathcal{C}[\prime, \infty]$.

"It turns out, in my opinion, that reproducing kernel Hilbert spaces are the natural setting in which to solve problems of statistical inference on time processes".

Emanuel Parzen (1962)

RKHS approach

Why natural? RKHS provides an intrinsic inner product depending on the covariance structure.

- Overlooked in FDA
  - "Inadequate" time series label.
  - *"Curiously, despite a huge research activity in this area, few attempts have been made to connect the rich theory of stochastic processes with functional data analysis."* Biau et al. 2015
- Explicit expressions of the Bayes rule (equivalent distributions).
- Approximate optimal rule under mutually singular distributions.
- Insight into the near "perfect classification phenomenon" (Delaigle and Hall 2012)
- Natural setting to formalize variable selection problems (RK-VS and associated classifier).

Definition: If $X = \{X_t, t \in [0, T]\}$ is a L^2-process with covariance function $K(s, t)$, define $(\mathcal{H}_0(K), \langle \cdot, \cdot \rangle)$ by

$$\mathcal{H}_0(K) := \{f : \ f(s) = \sum_i a_i K(s, t_i), \ a_i \in \mathbb{R}, \ t_i \in [0, 1], \ n \in \mathbb{N}\}$$

$$\langle f, g \rangle_K = \sum_{i,j} \alpha_i \beta_j K(s_j, t_i),$$

where $f(x) = \sum_i \alpha_i K(x, t_i)$ and $g(x) = \sum_j \beta_j K(x, s_j)$.

The RKHS associated wit h K, $\mathcal{H}(K)$, is defined as the completion of $\mathcal{H}_0(K)$. More precisely, $\mathcal{H}(K)$ is the set of functions $f : [0, T] \to \mathbb{R}$ obtained as t pointwise limit of a Cauchy sequence $\{f_n\}$ in $\mathcal{H}_0(K)$.

Some background (II)

Reproducing property: $f(t) = \langle f, K(\cdot, t)\rangle_K$, for all $f \in \mathcal{H}(K)$.

Natural congruence: If $\bar{\mathcal{L}}(X)$ is the L^2-completion of the linear span of X, $\Psi(\sum_i a_i X_{t_i}) = \sum_i a_i K(\cdot, t_i)$ defines a congruence between $\bar{\mathcal{L}}(X)$ and $\mathcal{H}(K)$.

$\mathcal{H}(K)$ coincides with the space of functions of the form $h(t) = \mathbb{E}(X_t U)$, for some $U \in \bar{\mathcal{L}}(X)$. Thus, in a very precise way, $\mathcal{H}(K)$ can be seen as the "natural Hilbert space" associated with a process $\{X(t), t \in [0, T]\}$.

Theorem 7A (Parzen, Ann. Math. Stat. 1961)

Under this model, if K is continuous

$$P_0 \sim P_1 \Leftrightarrow m(t) \in \mathcal{H}_K,$$

and if $P_0 \sim P_1$

$$\frac{dP_M}{dP_0}(X) = \exp\left\{\langle m, X\rangle_K - \frac{1}{2}\langle m, m\rangle_K\right\}$$

a) $\langle K(\cdot, t), X\rangle_K = X(t)$.

b) $\mathbb{E}_M\langle h, X\rangle_K = \langle h, m\rangle_K$.

c) $Cov[\langle h, X\rangle_K, \langle g, X\rangle_K] = \langle h, g\rangle_K$.

Equivalent measures: the new optimal rule

Bayes Rule (Berrendero et al. 2016, Theorem 2)

Under the given model, if $m(t) \in \mathcal{H}(K)$ then

$$g^*(X) = 1 \Leftrightarrow \eta^*(X) = \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0.$$

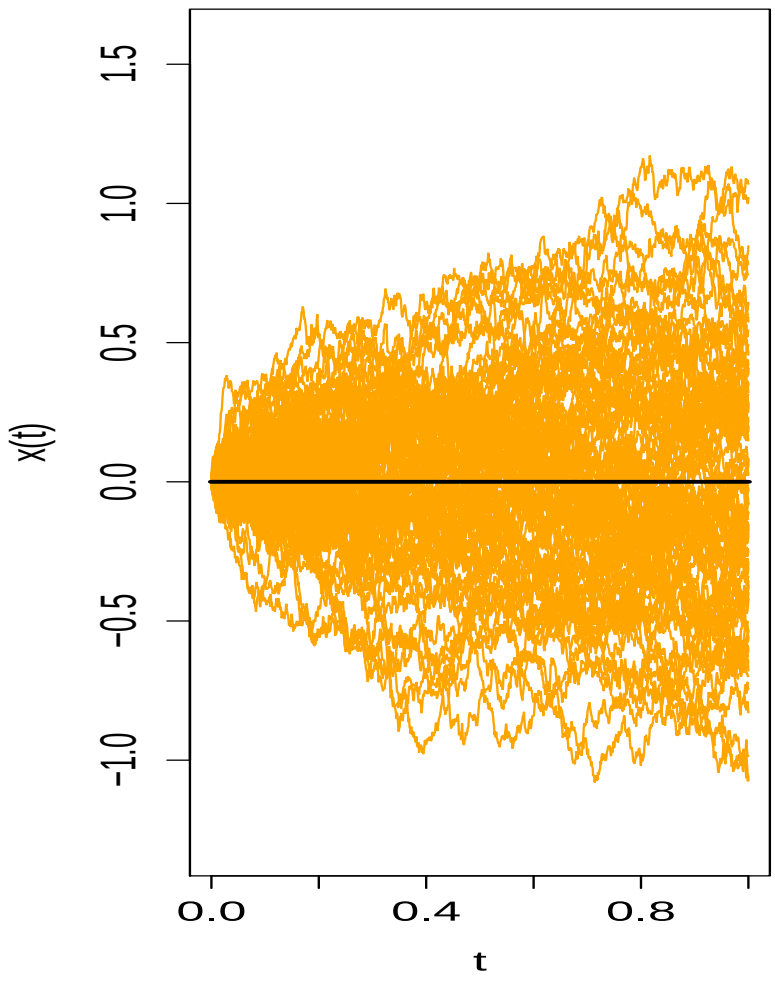
Bayes error

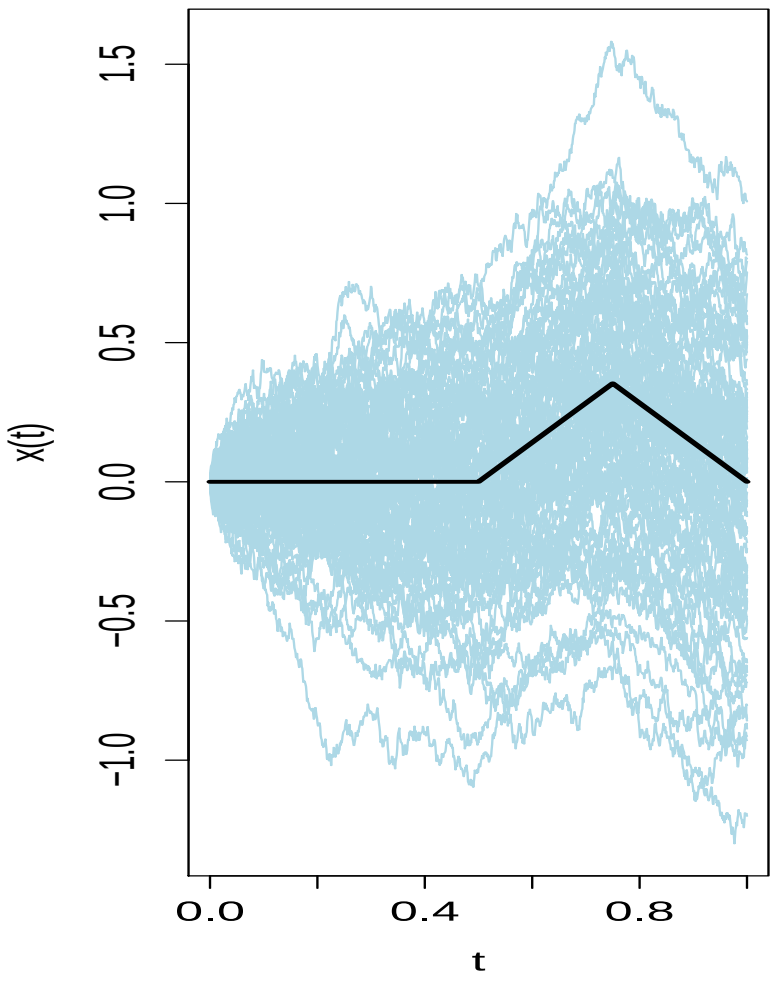
(1) $\eta^*(X)|Y = 0 \sim N\left(-\frac{1}{2} \parallel m \parallel_K^2, \parallel m \parallel_K\right).$

(2) $\eta^*(X)|Y = 1 \sim N\left(\frac{1}{2} \parallel m \parallel_K^2, \parallel m \parallel_K\right).$

$$L^* = \mathbb{P}\{g^*(X) \neq Y\} = 1 - \Phi\left(\frac{\parallel m \parallel_K}{2}\right)$$

Brownian example





Brownian example

$$\begin{cases} P_0: & X(t) = & B(t), & t \in [0,1] \\ P_1: & X(t) = m(t) + & B(t), & t \in [0,1] \end{cases}$$

$$m(t) = \begin{cases} 2\sqrt{2}t, & t \in [1/2, 3/4] \\ 2\sqrt{2}(1-t), & t \in [3/4, 1] \end{cases}$$

$K(s,t) = \min\{s,t\}$.

$\mathcal{H}_K = \{m \in \mathcal{C}[0,1] : \exists \Updownarrow \in L^2[0,1] \text{ with } m(t) = \int_0^t \Updownarrow(u)du, \ \forall u \in [0,1]\}$,
with $\langle h, g \rangle_K = \langle h', g' \rangle_{L^2}$.

Brownian example: Bayes rule

Then, since $m'(t) = \begin{cases} 2\sqrt{2}, & t \in [1/2, 3/4] \\ -2\sqrt{2}, & t \in [3/4, 1] \end{cases}$,

$$\eta(X) = 2\sqrt{2}\Big[2X(3/4) - X(1/2) - X(1)\Big] - 2.$$

Classify x in P_1 if

$$2X(3/4) - X(1/2) - X(1) > \frac{1}{\sqrt{2}} \approx 0.707.$$

$$L^* = 1 - \Phi(1) \approx 0.1587$$

Trajectories are only observed at a grid of points

As $N \to \infty$,

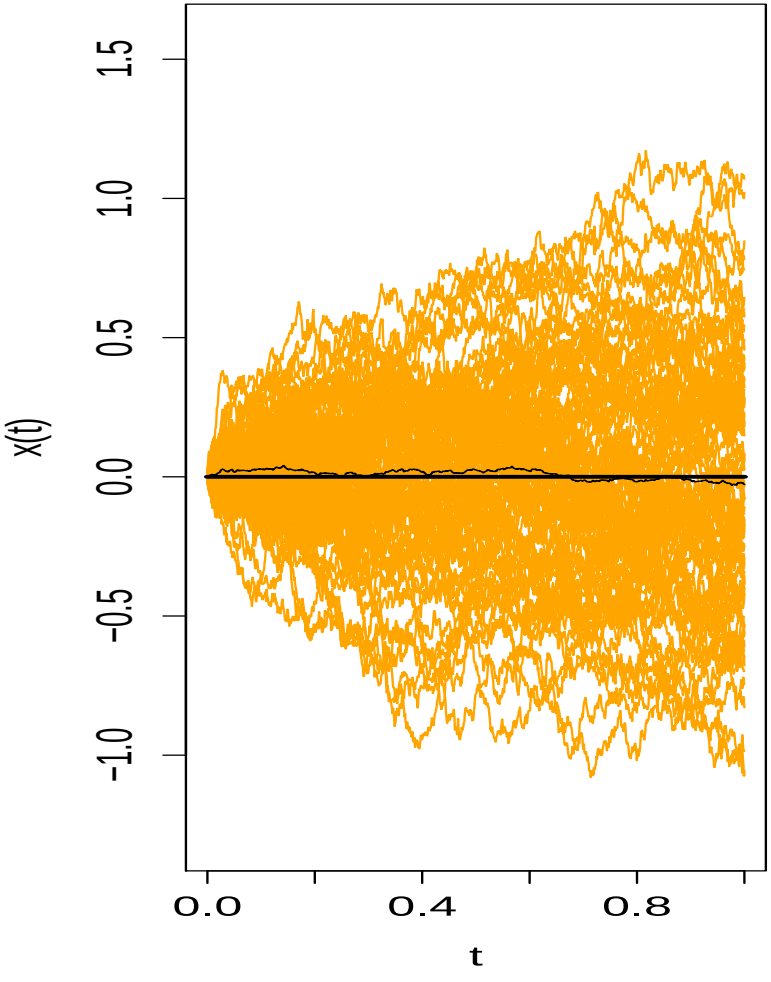
$$2^N \sum_{j=1}^{2^N} (\Delta_j m)(\Delta_j X) - 2^{N-1} \sum_{j=1}^{2^N} (\Delta_j m)^2 \to \eta(X)$$

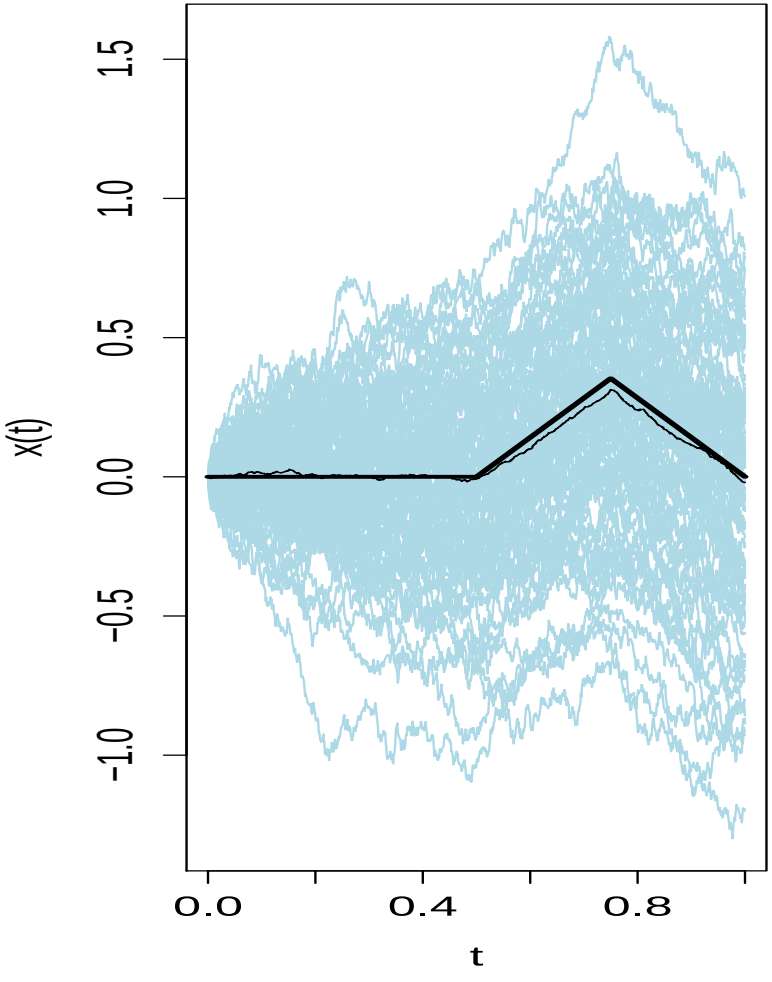
m is unknown

Naive choice: we can replace it with

$$\hat{m}(t) = \frac{1}{n} \sum_{i=1}^{n} x_i(t).$$

Example: estimated Bayes rule

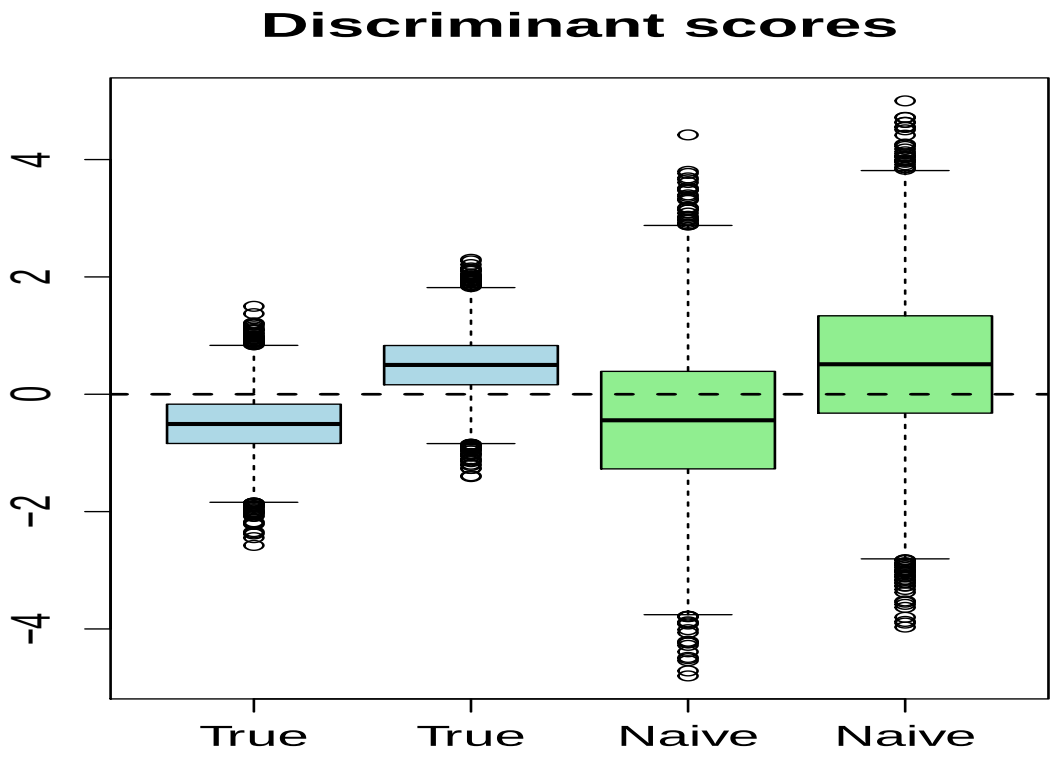




Discriminant scores and classification errors

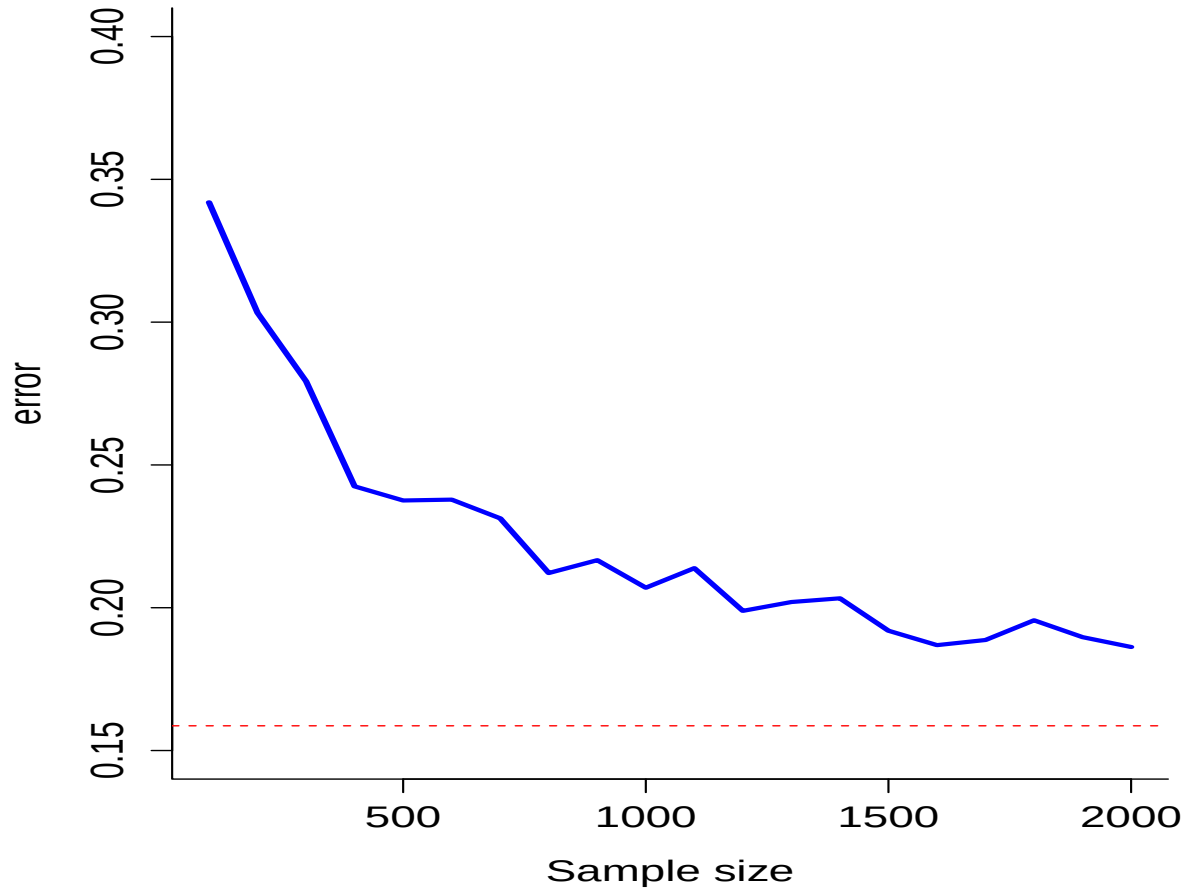
Based on 10000 test trajectories of each model ($n_0 = n_1 = 100$)

Rule	Bayes	True	Naive
Error	15.87%	15.98%	34.88%



Classification error as n increases

$$n_0 = n_1 = 100, 200, \ldots, 2000$$



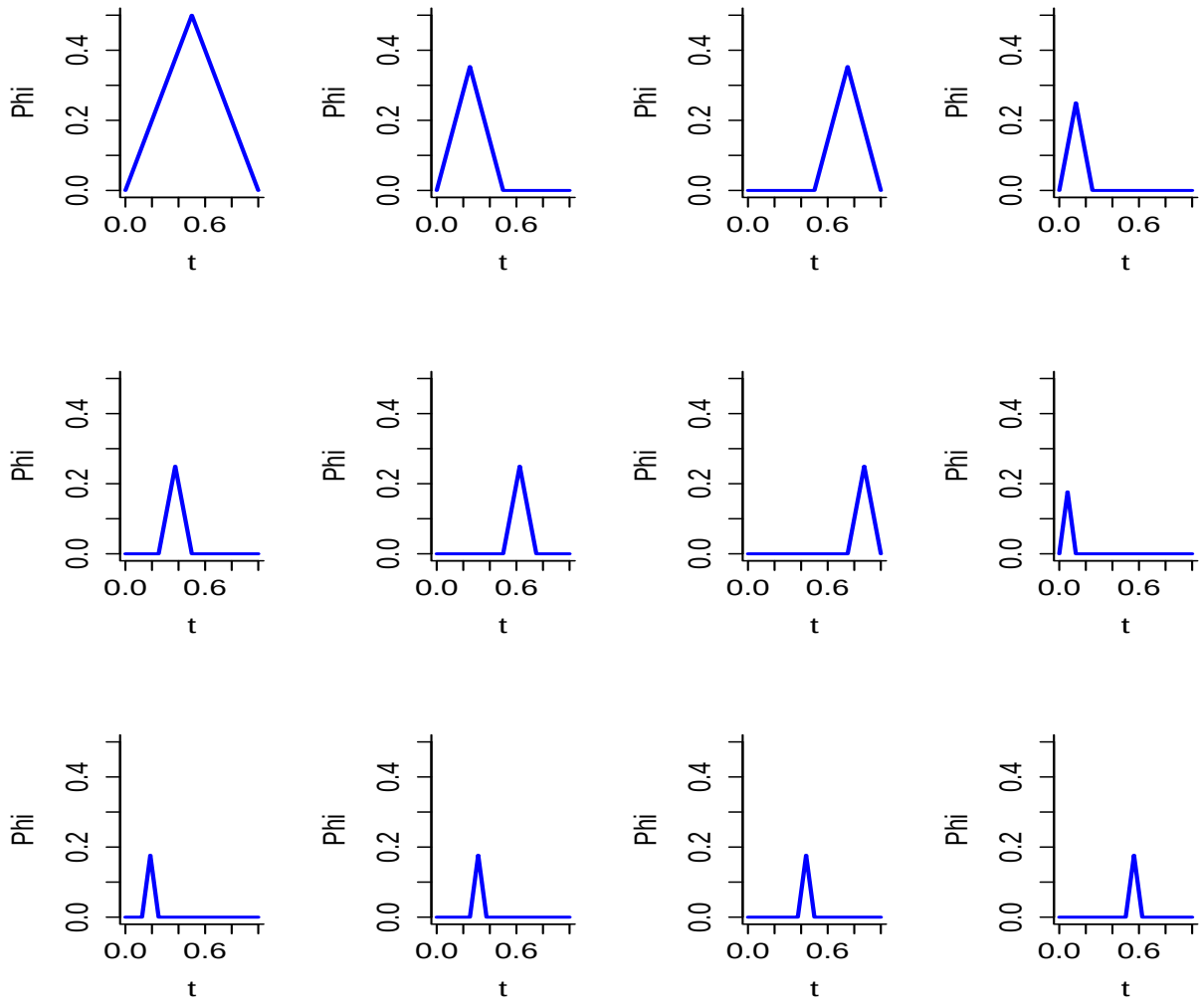
Alternatives

The following alternative rules can be viewed as two different methods of smoothing

- Expand the averages of the raw trajectories in terms of an appropriate basis of $\mathcal{H}_K$ and select the main terms of the expansion

- Smooth the trajectories before computing the means

Haar basis



Haar coefficients

$$c_{0,0} = m(1),$$
$$c_{k,j} = \sqrt{2^{k-1}} \left[2m\left(\frac{2j-1}{2^k}\right) - m\left(\frac{2j}{2^k}\right) - m\left(\frac{2j-2}{2^k}\right) \right].$$

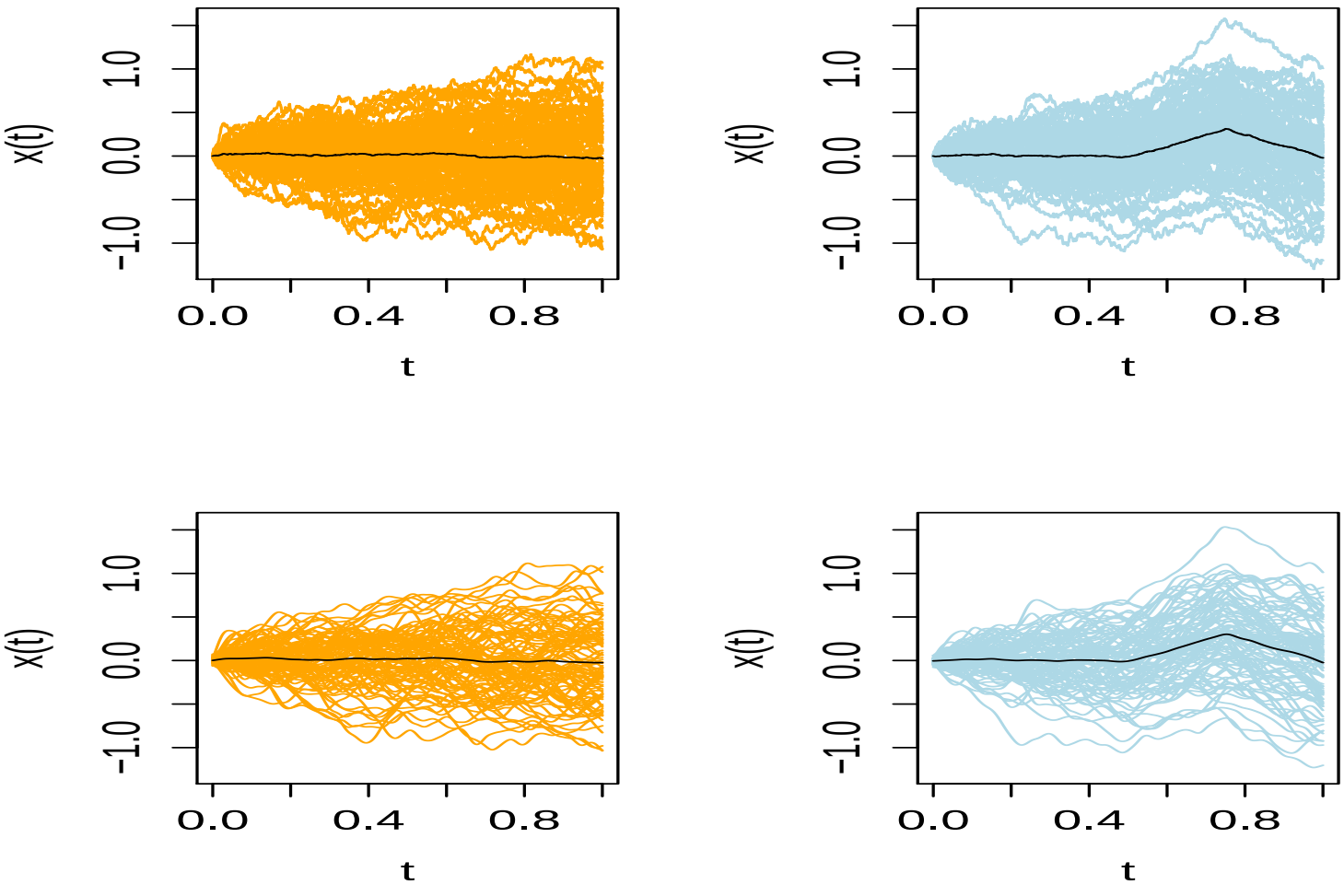
Natural estimators are obtained replacing $m(\cdot)$ with $\bar{X}(\cdot)$.

For an appropriate K,

$$\hat{m}(t) = \hat{c}_{0,0}\, t + \sum_{k=1}^{K} \sum_{j=1}^{2^{k-1}} \hat{c}_{k,j} \Phi_{k,j}(t)$$

Smoothed trajectories

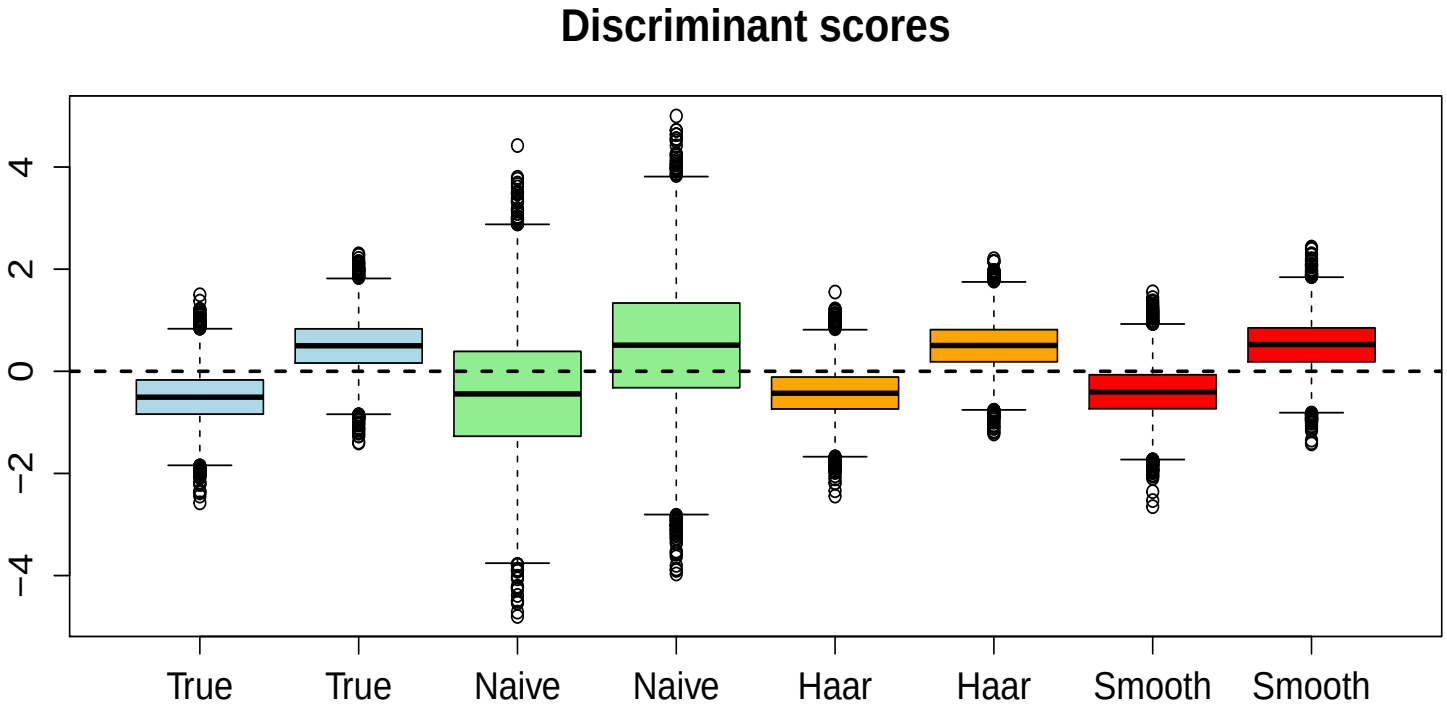
Using splines (`fda.usc` package in `R`)



Discriminant scores and classification errors

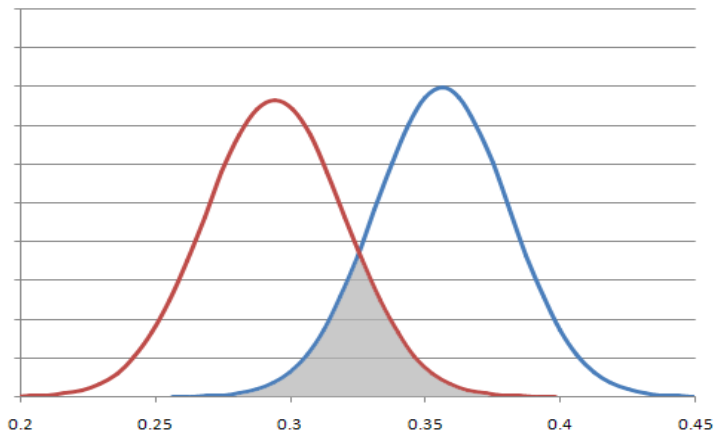
Based on 10000 test trajectories of each model ($n_0 = n_1 = 100$)

Rule	Bayes	True	Naive	Haar	Smooth
Error	15.87%	15.98%	34.88%	16.13%	17.95%



The singular case

"We argue that those [functional classification] problems have unusual, and fascinating, properties that set them apart from their finite dimensional counterparts. In particular we show that, in many quite standard settings, the performance of simple [linear] classifiers constructed from training samples becomes perfect as the sizes of those samples diverge [...]. That property never holds for finite dimensional data, except in pathological cases." Delaigle and Hall, J. R. Statist. Soc. B 2012



$$\begin{cases} P_0: & X(t) = & \xi(t), & t \in [0,1] \\ P_1: & X(t) = & m(t) + \xi(t), & t \in [0,1] \end{cases}$$

$\xi(t)$ gaussian with $\mathbb{E}(\xi(t)) = 0$.

$$K(s,t) = \mathbb{E}(\xi(s)\xi(t)) = \sum_{j=1}^{\infty} \theta_j \phi_j(s)\phi_j(t).$$

Where $\theta_1 \geq \theta_2 \geq \ldots$ and K is strictly positive definite and uniformly bounded.

$$m(t) = \sum_{j=1}^{\infty} \mu_j \phi_j.$$

Prior probabilities: $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$.

Centroid classifier

$$T(X) = 1 \Leftrightarrow D^2(X, \bar{X}_1) - D^2(X, \bar{X}_0) < 0.$$

$$D(X, \bar{X}_k) = |\langle X, \psi \rangle_{L^2} - \langle \bar{X}_k, \psi \rangle_{L^2}|,$$

where $\langle X, \psi \rangle_{L^2} = \int_{[0,1]} X(t)\psi(t)dt$.

and $\psi^{(r)} = \sum_{j=1}^{r} \theta_j^{-1} \mu_j \phi_j(t)$

Asymptotic centroid

$$T(X) \overset{n\to\infty}{\longrightarrow} T^0(X) = (\langle X, \psi \rangle_{L^2} - \langle m, \psi \rangle_{L^2})^2 - \langle X, \psi \rangle_{L^2}^2.$$

Theorem 1 (Delaigle and Hall, J. R. Statist. Soc. B 2012)

(a) When $\sum_{j\geq 1} \theta_j^{-1}\mu_j^2 < \infty$ the Bayes (minimal) error is $err_0 = 1 - \Phi\left(\frac{1}{2}(\sum_{j\geq 1} \theta_j^{-1}\mu_j^2)^{1/2}\right) > 0$ and the optimal classifier (that achieves this error) is the rule

$$T^0(X) = 1, \text{ if and only if } (\langle X,\psi\rangle_{L^2} - \langle m_1,\psi\rangle_{L^2})^2 - \langle X,\psi\rangle_{L^2}^2 < 0,$$

with $\psi(t) = \sum_{j=1}^{\infty} \theta_j^{-1}\mu_j\phi_j(t)$.

(b) If $\sum_{j\geq 1} \theta_j^{-1}\mu_j^2 = \infty$ then the minimal misclassification probability is $err_0 = 0$ and it is achieved, in the limit, by a sequence of classifiers constructed from T^0 by replacing the function ψ with $\psi^{(r)} = \sum_{j=1}^{r} \theta_j^{-1}\mu_j\phi_j(t)$, with $r = r_n \uparrow \infty$.

Near perfect classification

If $\sum_{j \geq 1} \theta_j^{-1} \mu_j^2 = \infty$ then the minimal misclassification probability is $err_0 = 0$.

Near perfect classification

If $\sum_{j \geq 1} \theta_j^{-1} \mu_j^2 = \infty$ then the minimal misclassification probability is $err_0 = 0$.

An unanswered question:

Why?

"The theoretical foundation for these findings is an intriguing dichotomy of properties and is as interesting as the findings themselves." Delaigle and Hall, 2012

Near perfect classification

If $\sum_{j \geq 1} \theta_j^{-1} \mu_j^2 = \infty$ then the minimal misclassification probability is $err_0 = 0$.

An unanswered question:

Why?

"The theoretical foundation for these findings is an intriguing dichotomy of properties and is as interesting as the findings themselves." Delaigle and Hall, 2012

Because of the singularity

Our view of the "near perfect classification"

Theorem 4 (Berrendero et al., 2016)

(a) $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 < \infty$ if and only if $P_1 \sim P_0$. In that case, the Bayes rule g^* is

$$g^*(X) = 1 \text{ if and only if } \langle X, m\rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0.$$

This is a coordinate-free, equivalent expression of the optimal rule given by D. & H. The corresponding optimal (Bayes) classification error is $L^* = 1 - \Phi\left(\parallel m \parallel_{\mathcal{H}_K} /2\right)$.

(b) $\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 = \infty$ if and only if $P_1 \perp P_0$. In this case the Bayes error is $L^* = 0$. Moreover, for any $\epsilon > 0$ we can construct a classification rule whose misclassification probability is smaller than ϵ (Berrendero et al., Theorem 5).

Bayes Rule

$$g^*(X) = 1 \Leftrightarrow \eta^*(X) = \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0.$$

Bayes Error

1. $\eta^*(X)|Y = 0 \sim N\left(-\frac{1}{2} \parallel m \parallel_K^2, \parallel m \parallel_K\right).$
2. $\eta^*(X)|Y = 1 \sim N\left(\frac{1}{2} \parallel m \parallel_K^2, \parallel m \parallel_K\right).$

$$L^* = \mathbb{P}\{g^*(X) \neq Y\} = 1 - \Phi\left(\frac{\parallel m \parallel_K}{2}\right)$$

$$m \in \mathcal{H}_K \Rightarrow m(\cdot) = \sum_{j=1}^{d} \alpha_j K(\cdot, t_j).$$

Theorem 4.12. Cucker and Zhou, 2007

Let $K : [0,1] \times [0,1] \longrightarrow \mathbb{R}$ a Mercer's kernel and let θ_j be the eigenvalues of the integral operator

$$Kf(t) = \int_0^1 K(t,u) f(u) du,$$

And let ϕ_j be the corresponding orthogonal eigenfunctions. Then, $\{\sqrt{\theta_j}\phi_j : \theta_j > 0\}$ forms an orthonormal basis of $\mathcal{H}_K$.

Equivalence or singularity

$$\sum_{j\geq 1} \theta_j^{-1} \mu_j^2 < \infty \Longleftrightarrow P_0 \sim P_1 (m \in \mathcal{H}_K)$$

$$\left(\sum_{j\geq 1} \theta_j^{-1} \mu_j^2 = \infty \Longleftrightarrow P_0 \perp P_1 (m \notin \mathcal{H}_K) \right)$$

Equivalence or singularity

$$\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 < \infty \Longleftrightarrow P_0 \sim P_1(m \in \mathcal{H}_K)$$

$$\left(\sum_{j\geq 1}\theta_j^{-1}\mu_j^2 = \infty \Longleftrightarrow P_0 \perp P_1(m \notin \mathcal{H}_K)\right)$$

$\{\sqrt{\theta_j}\phi_j(t)\}$ is an orthonormal basis of $\mathcal{H}_K$. Therefore, every element in $\mathcal{H}_K$ can be represented uniquely as a linear combination of elements of the basis.

$$m(t) = \sum_{j=1}^{\infty}\frac{\mu_j}{\sqrt{\theta_j}}\sqrt{\theta_j}\phi_j(t) \Longrightarrow \| m \|^2_{\mathcal{H}_K} = \sum_{j=1}^{\infty}\frac{\mu_j^2}{\theta_j} \text{ and}$$

$$m(t) \in \mathcal{H}_K \Longleftrightarrow \sum_{j=1}^{\infty}\frac{\mu_j^2}{\theta_j} < \infty \Longleftrightarrow P_0 \sim P_M$$

The optimal rule

If $P_0 \sim P_1$

$$(\langle X, \psi \rangle_{L^2} - \langle \mu, \psi \rangle_{L^2})^2 - \langle X, \psi \rangle_{L^2}^2 < 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0$$

The optimal rule

If $P_0 \sim P_1$

$$\langle X, \psi \rangle_{L^2}^2 - 2\langle m, \psi \rangle_{L^2} \langle X, \psi \rangle_{L^2} < 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0$$

The optimal rule

If $P_0 \sim P_1$

$$\langle X, \psi \rangle_{L^2}^2 - 2\langle m, \psi \rangle_{L^2} \langle X, \psi \rangle_{L^2} < 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0$$

$\psi = \sum_{j=1}^{\infty} \theta_j^{-1} \mu_j \phi_j$

$$m(t) = \sum_{j=1}^{\infty} \mu_j \phi_j(t) = \sum_{j=1}^{\infty} \frac{\mu_j}{\sqrt{\theta_j}} \sqrt{\theta_j} \phi_j(t)$$

$$\langle m, \psi \rangle_{L^2} = \sum_{j=1}^{\infty} \frac{\mu_j^2}{\theta_j} = \parallel m \parallel_{\mathcal{H}_K}^2$$

The optimal rule

If $P_0 \sim P_1$

$$\| m \|^4_{\mathcal{H}_K} - 2 \| m \|^2_{\mathcal{H}_K} \langle X, \psi \rangle_{L^2} < 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \| m \|^2_K > 0$$

The optimal rule

If $P_0 \sim P_1$

$$\langle X, \psi \rangle_{L^2} - \frac{1}{2} \parallel m \parallel^2_{\mathcal{H}_K} > 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel^2_K > 0$$

The optimal rule

If $P_0 \sim P_1$

$$\langle X, \psi \rangle_{L^2} - \frac{1}{2} \parallel m \parallel^2_{\mathcal{H}_K} > 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel^2_K > 0$$

$$\begin{aligned}
\langle X, m \rangle_K &= \langle X, \sum_{j=1}^{\infty} \mu_j \phi_j \rangle_K \\
&\overset{\theta_j > 0}{=} \langle X, \sum_{j=1}^{\infty} \frac{\mu_j}{\theta_j} \theta_j \phi_j \rangle_K \\
&\overset{\text{bilinearity of } \langle \cdot, \cdot \rangle_K}{=} \sum_{j=1}^{\infty} \frac{\mu_j}{\theta_j} \langle X, K\phi_j \rangle_K \\
&\overset{\phi_j \text{ eigenfunction of } K}{=} \sum_{j=1}^{\infty} \frac{\mu_j}{\theta_j} \langle X, \int_0^1 K(\cdot, u)\phi_j(u)du \rangle_K
\end{aligned}$$

The optimal rule

If $P_0 \sim P_1$

$$\langle X, \psi \rangle_{L^2} - \frac{1}{2} \parallel m \parallel^2_{\mathcal{H}_K} > 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel^2_K > 0$$

$$\begin{aligned}
\langle X, m \rangle_K &= \sum_{j=1}^{\infty} \frac{\mu_j}{\theta_j} \int_0^1 \langle X, K(\cdot, u) \rangle_K \phi_j(u) du \\
&= \sum_{j=1}^{\infty} \frac{\mu_j}{\theta_j} \int_0^1 X(u) \phi_j(u) du \\
&= \sum_{j=1}^{\infty} \frac{\mu_j}{\theta_j} \langle X, \phi_j \rangle_{L^2} \\
&= \langle X, \sum_{j=1}^{\infty} \frac{\mu_j}{\theta_j} \phi_j \rangle_{L^2} = \langle X, \psi \rangle_{L^2}
\end{aligned}$$

The optimal rule

If $P_0 \sim P_1$

$$\langle X, \psi \rangle_{L^2} - \frac{1}{2} \parallel m \parallel^2_{\mathcal{H}_K} > 0 \Leftrightarrow \langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel^2_K > 0$$

Remark

Note that the expression given for the Bayes error is also equivalent.

$$L^* = 1 - \Phi\left(\frac{1}{2} \parallel m \parallel_{\mathcal{H}_K}\right) = 1 - \Phi\left(\frac{1}{2}\left(\sum_{j \geq 1} \frac{\mu_j^2}{\theta}\right)^{1/2}\right) = err_0.$$

Some conclusions

- RKHS approach is a useful tool for classification of Gaussian processes.
- Radon-Nikodym derivatives have a straightforward application to FDA.
- Unlike the finite dimensional case, in the functional setting the classification problem of mutually singular distributions is meaningful.
- Equivalent and singular case could be identified empirically.

Bayes rule under heteroscedasticity

Theorem (Shepp 1966, Thm. 1)

Let P_0, P_1 be the distributions corresponding to the standard Brownian Motion $\{B(t),\ t \in [0,T]\}$ and to a Gaussian process $\{X(t),\ t \in [0,T]\}$ with mean function m_1 in the Dirichlet space $\mathcal{D}[0,T]$ and covariance function K. Then $P_1 \sim P_0$ if and only if there exists a function $K_1 \in L^2([0,T] \times [0,T])$ such that

$$K(s,t) = \min\{s,t\} - \int_0^s \int_0^t K_1(u,v)dudv,$$

with $1 \notin \sigma(K_1)$, the spectrum of K_1. In this case, the function K_1 is given by $K_1(s,t) = -\frac{\partial^2}{\partial s \partial t} K(s,t)$.

We will also need Lemmas 1 and 2 in Shepp (1966), p. 334-335 which give the expression of the Radon-Nikodym derivative dP_1/dP_0 in the case $P_1 \ll P_0$ under the conditions of Theorem 1

Theorem (Bayes rule under heteroscedasticity)

Let us consider the classification general Gaussian model under heteroscedasticity. Let us denote by $g(x) = \mathbb{I}_{\{\eta^*(x)>0\}}$ *the Bayes rule.*

(a) *If* $m_0 \equiv 0$, $m_1 \in \mathcal{D}[0,T]$, ϵ_0 *is the standard Brownian motion on* $[0,T]$, *with* $T < 1$, *and* ϵ_1 *is the standard Brownian bridge on* $[0,T]$, *then*

$$\eta^*(X) = -\frac{1}{2}\log(1-T) - \frac{TX(T)^2 + m_1(T)^2 - 2m_1(T)X(T)}{2T(1-T)} - \log\left(\frac{1-p}{p}\right).$$

Notice that if $m_1 \equiv 0$ (that is, no trend in the Brownian bridge) and $p = 1/2$, the rule $\mathbb{I}_{\{\eta^*(X)>0\}}$ in (a) reduces to just the indicator of

$$X(T)^2 < T(T-1)log(1-T).$$

Theorem (Bayes rule under heteroscedasticity)

Let us consider the classification general Gaussian model under heteroscedasticity. Let us denote by $g(x) = \mathbb{I}_{\{\eta^*(x)>0\}}$ *the Bayes rule.*

(b) *If the noise processes* ϵ_0, ϵ_1 *are both standard Brownian bridges on* $[0, T]$ *with* $T < 1$, *and both* m_0 *and* $m_1 \in \mathcal{D}[0, T]$, *then*

$$\eta^*(X) = \frac{(X(T) - m_0(T))^2 - (X(T) - m_1(T))^2}{2T(1-T)} - \log\left(\frac{1-p}{p}\right).$$

Notice that when $p = 1/2$, the rule $\mathbb{I}_{\{\eta^*(X)>0\}}$ for (b) reduces to the indicator of

$$|X(T) - m_0(T)| - |X(T) - m_1(T)| > 0.$$