Machine Learning with Functional Data

Classification, regression and clustering

The goals of FDA (Ramsay and Silverman, 2005)

- To represent the data in ways that aid further analysis.
- To display the data so as to highlight various characteristics.
- To study important sources of pattern and variation among the data.
- To explain variation in an outcome or dependent variable by using input or independent variable information.
- To compare two or more sets of data with respect to certain types of variation, where two sets of data can contain different sets of replicates of the same functions, or different functions for a common set of replicates.

The functional linear regression problem (I)

The **general model** would be

$$Y = g(X) + \epsilon,$$

where X is the explanatory (functional) variable, Y is the output (response) variable, g is a (usually unknown) function and ϵ is the error which is often assumed to fulfill $\mathbb{E}(\epsilon|X) = 0$.
The aim is to estimate g from a random sample (X_i, Y_i), $i = 1, \ldots, n$.
In the **linear case**,

$$Y_i = \alpha + \int_0^1 \beta(t) X_i(t) dt + \epsilon_i, \ \ i = 1, \ldots, n, \qquad (1)$$

where $\beta \in L^2[0,1]$, $X_i = X_i(\omega, .) \in L^2[0,1]$ are iid with $\mathbb{E}\|X_i\|^2 < \infty$, ϵ_i are iid with $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(X_i(t)\epsilon_i) = 0$, almost everywhere. The main problem here is the estimation of β ($\mathbb{E}(y|X) = \langle X, \beta \rangle_{L^2}$).

The functional linear regression problem (II)

Assume that $\gamma(s,t) = Cov(X_i(s), X_i(t))$ is continuous. The covariance operator of the underlying L^2-process $X(t)$ is given by

$$\Gamma u(t) = \int_0^1 \gamma(s,t)u(s)ds, \ \forall u \in L^2[0,1].$$

In a equivalent way it can be expressed by

$$\Gamma u = \mathbb{E}\left(\langle X_i - \mathbb{E}X_i, u\rangle (X_i - \mathbb{E}(X_i))\right)$$

The operator Γ is non-negative, symmetric, Hilbert-Schmidt ($\sum_i \|\Gamma e_i\|^2 < \infty$) and hence compact. The eigenvalues λ_j of Γ are all positive and zero is their only accumulation point. If we denote by e_j the sequence of orthogonal eigenfunctions associated with the eigenvalues λ_j (we assume they are sorted in decreasing order), we may perform the above mentioned Karhunen-Loève decomposition,

$$X_i(t) - \mathbb{E}(X_i(t)) = \sum_{j=1}^{\infty} Z_j e_j(t),$$

where the Z_j are centered uncorrelated with $V(Z_j) = \lambda_j$.

The functional linear regression problem (III)

If we center the variables in (1) it can be seen that our function β must fulfill

$$\Delta = \Gamma\beta \tag{2}$$

in the sense that $\Gamma\beta$ coincides with $\mathbb{E}\left((X_1 - \mathbb{E}(X_1))(s)(Y_1 - \mathbb{E}(Y_1))\right)$, which is the kernel function which defines the operator

$$\Psi \mapsto \Delta\Psi = \mathbb{E}\left(\langle\Psi, (X_1 - \mathbb{E}(X_1))\rangle(Y_1 - \mathbb{E}(Y_1))\right).$$

Let us note the analogy with the ordinary multivariate regression model $Y = X\beta + \epsilon$, with $\beta \in \mathbb{R}^p$, for which β appears as the solution of the *normal equations* $X'X\beta = X'Y$. However, in the functional case Γ is not an invertible operator unless its range is a finite-dimensional space

The functional linear regression problem (IV)

Then, we look for solutions in the closure of $\mathrm{Im}(\Gamma) = \{\Gamma x : \ x \in L^2[0,1]\}$; alternatively, we assume (w.l.o.g.) that $\ker(\Gamma) = \{0\}$. Expanding β in the orthonormal basis $\{e_i\}$ of eigenfunctions of Γ , we have $\beta = \sum_{j=1}^{\infty} \langle \beta, e_j \rangle e_j$. Using the normal equations (2),

$$\Delta e_j = \lambda_j \langle \beta, e_j \rangle, \ j = 1, 2, \ldots \tag{3}$$

Equation (3) implies

$$\beta = \sum_{j=1}^{\infty} \frac{\Delta e_j}{\lambda_j} e_j = \sum_{j=1}^{\infty} \frac{\mathbb{E}\left(Z_j (Y_1 - \mathbb{E}(Y_1)\right)}{\lambda_j} e_j, \tag{4}$$

where Z_j is the j-coordinate of $X_1 - \mathbb{E}(X_1)$.

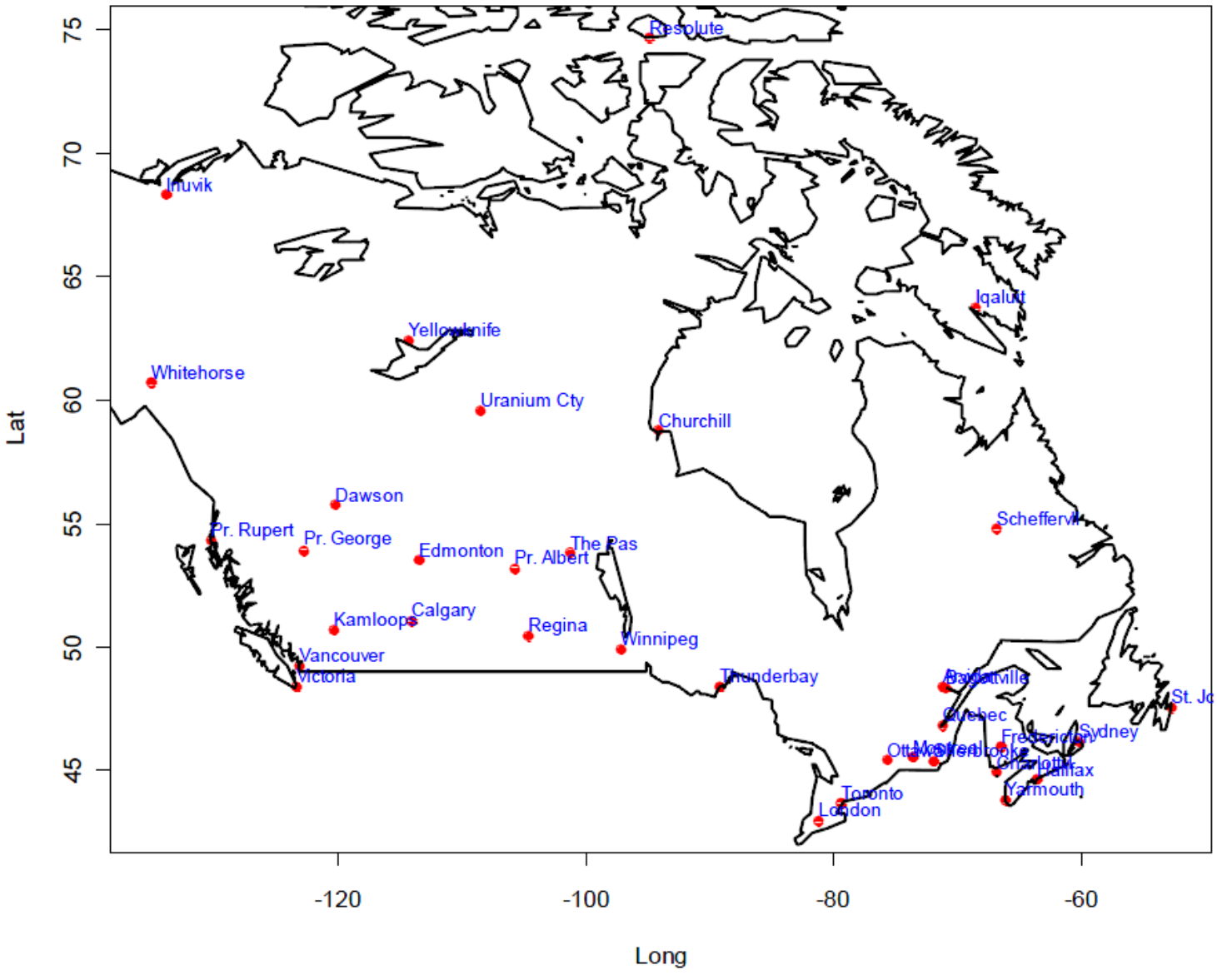
The functional linear regression problem (V)

Thus, a *Functional Principal Component Regression Estimator* for β could be defined by

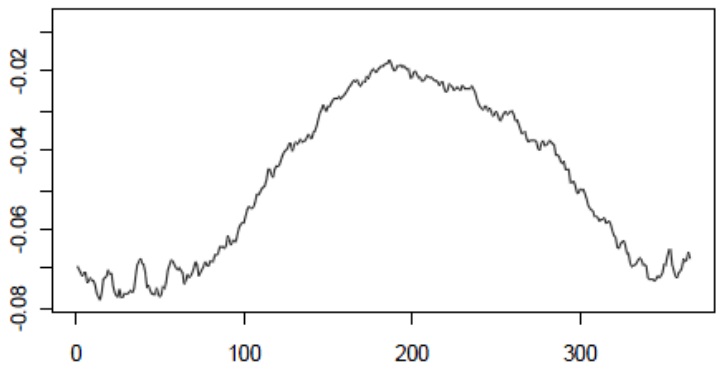
$$\hat{\beta} = \sum_{j=1}^{K} \frac{\Delta_n \hat{e}_j}{\hat{\lambda}_j} \hat{e}_j$$

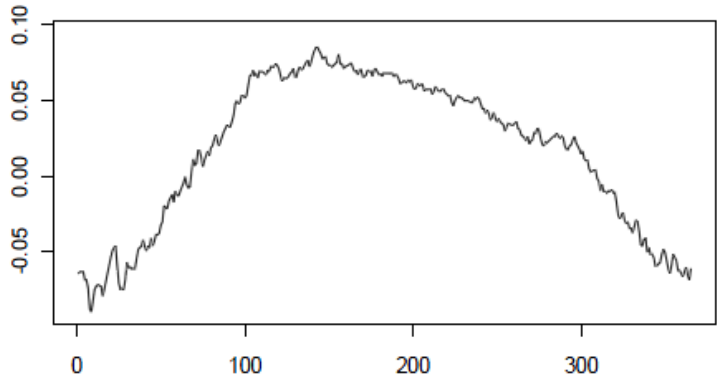
where Δ_n is the empirical version of the cross-covariance operator Δ defined above, $\hat{\lambda}_j$ and $\hat{e}_j$ are the (K-largest) eigenvalues and eigenvectors of the empirical covariance operator Γ_n which estimates Γ.

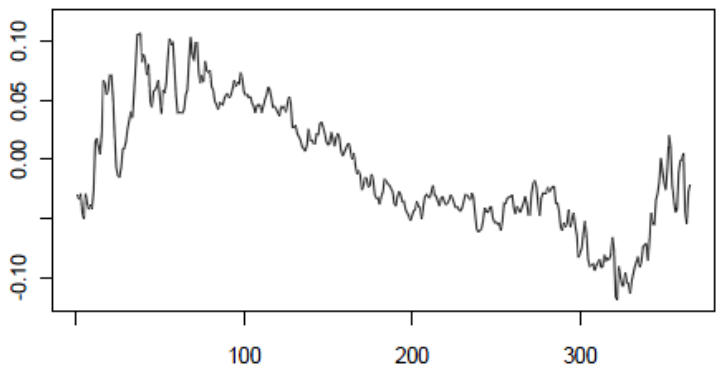
FPCA regression

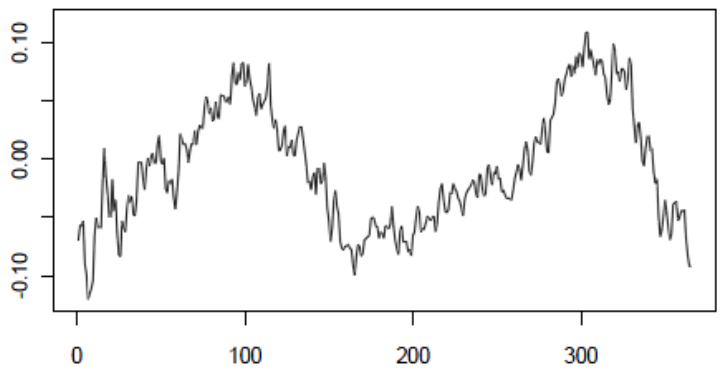


CWD. Primeras 4 componentes principales









FLR: estimation by penalized least squares (I)

An alternative approach is as follows:

We look for the estimator $\tilde{\beta}$ minimizing

$$\sum_{i=1}^{n} \left(Y_i - \int_0^1 \tilde{\beta}(t) X_i(t) dt \right)^2 + \lambda \int_0^1 (\tilde{\beta}^{(m)}(t))^2 dt$$

and the function $\tilde{\beta}$ is of type

$$g(t) = \sum_{k=1}^{K} \hat{b}_k \phi_k(t).$$

for some orthonormal basis $\{\phi_k\}$

FLR: estimation by penalized least squares (II)

This leads to minimize

$$\sum_{i=1}^{n}\left(Y_i - \sum_{k=1}^{K}\langle X_i, \phi_k\rangle b_k\right)^2 + \lambda \sum_{k=1}^{K} b_k^2 \int_0^1 (\phi^{(m)}(t))^2 dt$$

FLR: the case of functional response

Here the basic model is

$$Y_i(t) = \int_0^1 \beta(s,t) X_i(s) ds + \epsilon_i(t), \ \ i = 1, \ldots, n, \tag{5}$$

where $\beta(s,t)$ defines a Hilbert-Schmidt (hence, compact) integral operator, $x \mapsto \int_0^1 \beta(s,t)x(s)ds$ on $L^2[0,1]$, that is $\int_0^1 \int_0^1 \beta(s,t)^2 dsdt < \infty$, $X_i = X_i(t) \in L^2[0,1]$ and typically $\epsilon_i(t)$ are iid L^2-processes with $\mathbb{E}(\epsilon_i) = 0$, and ϵ_i independent from (X_i, Y_i).

Functional nonparametric regression (I)

Given a scalar random output Y and a (possibly functional) explanatory X it is natural to look for the "best approximation" of Y in terms of X. It is well-known that the minimizer (on the space of real measurable functions g such that $\mathbb{E}(g^2(X)) < \infty$) of the L^2-mean error $\mathbb{E}\left(Y - g(X)\right)^2$ is given the *regression function*

$$\eta(x) = \mathbb{E}(Y|X = x).$$

Functional nonparametric regression (II)

Two natural non-parametric estimators

$$\hat{\eta}(x) = \sum_{i \in k_n(x)} \frac{Y_i}{k},$$

where $k_n(x) = \{i : X_i \text{ is one of the } k \text{ closest neighbors of } x\}$ Biau et al. (2010)

$$\tilde{\eta}(x) = \frac{\sum_{i=1}^{n} Y_i K(D(x, X_i)/h)}{\sum_{i=1}^{n} K(D(x, X_i)/h)}$$

Ferraty and Vieu (2006, ch. 6)

Loss of the universal consistency (Forzani et al., 2012)

The problem is to check

$$m_1(t) = \ldots = m_k(t)$$

in a model

$$X_{ij}(t) = m_i(t) + \epsilon_i(t),\ j = 1, \ldots, n_i.$$

A possible test statistic (Cuevas, Febrero and Fraiman 2004) is

$$F_n = \frac{\sum_{i=1}^{k} n_i ||\bar{X}_{i.} - \bar{X}_{..}||^2/(k-1)}{\sum_{i,j} ||X_{ij} - \bar{X}_{i.}||^2/(n-k)},$$

where

$$\bar{X}_{i.} = \bar{X}_{i.}(t) = \sum_{j=1}^{n_i} \frac{X_{ij}(t)}{n_i},\ \ \bar{X}_{..} = \bar{X}_{..}(t) = \sum_{i=1}^{k} \frac{n_i \bar{X}_{i.}(t)}{n},\ \ n = \sum_{i=1}^{k} n_i,$$

and $\|.\|$ stands for the usual L^2 norm.

Other regression topics

- **Logistic regression** models with a functional explanatory variable have been considered, among others, by Lindquist and McKeague (2009), Berrendero, Bueno & Cuevas (2023).
- **Autoregressive linear models** for functional time series are studied in Bosq (2000) and Horváth and Kokoszka (2012).
- Some measures to detect **influential observations** in the functional linear model are analyzed in Febrero-Bande, M., Galeano, P. and González-Manteiga, W. (2010).
- The problem of **detecting a change**, during the observation period, in the operator which defines a functional linear model has been addressed by Horváth and Reeder (2012).

The classification problem

Classify:

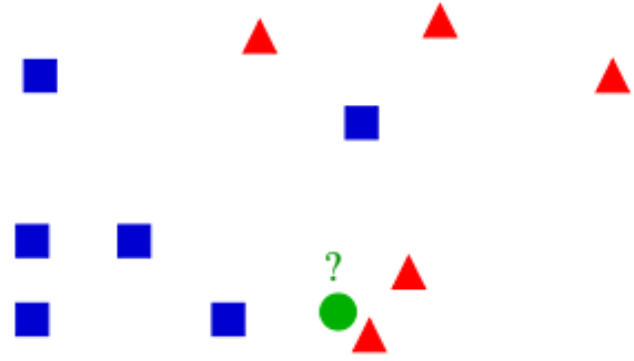
1.- Arrange (a group) in classes according to shared characteristics.

2.- Assign to a particular class or category.

Oxford English Dictionary

The classification problem

Supervised



Unsupervised

Unsupervised classification: the k-means procedure (I)

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

$$I_k(P; h_1, \ldots, h_k) := \mathbb{E}\left(\min_{i=1,\ldots,k} \|X - h_i\|^2\right) \tag{6}$$

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 (II)

The sample version of (6), 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,

$$I_k(\mathbb{P}_n; h_1, \ldots, h_k) := \frac{1}{n} \sum_{i=1}^{n} \|X_i - h_{c(i)}\|^2, \tag{7}$$

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 (7), is a formidable task. Different approximate (often randomized) algorithms have been proposed.

A simple algorithm for the k-means procedure

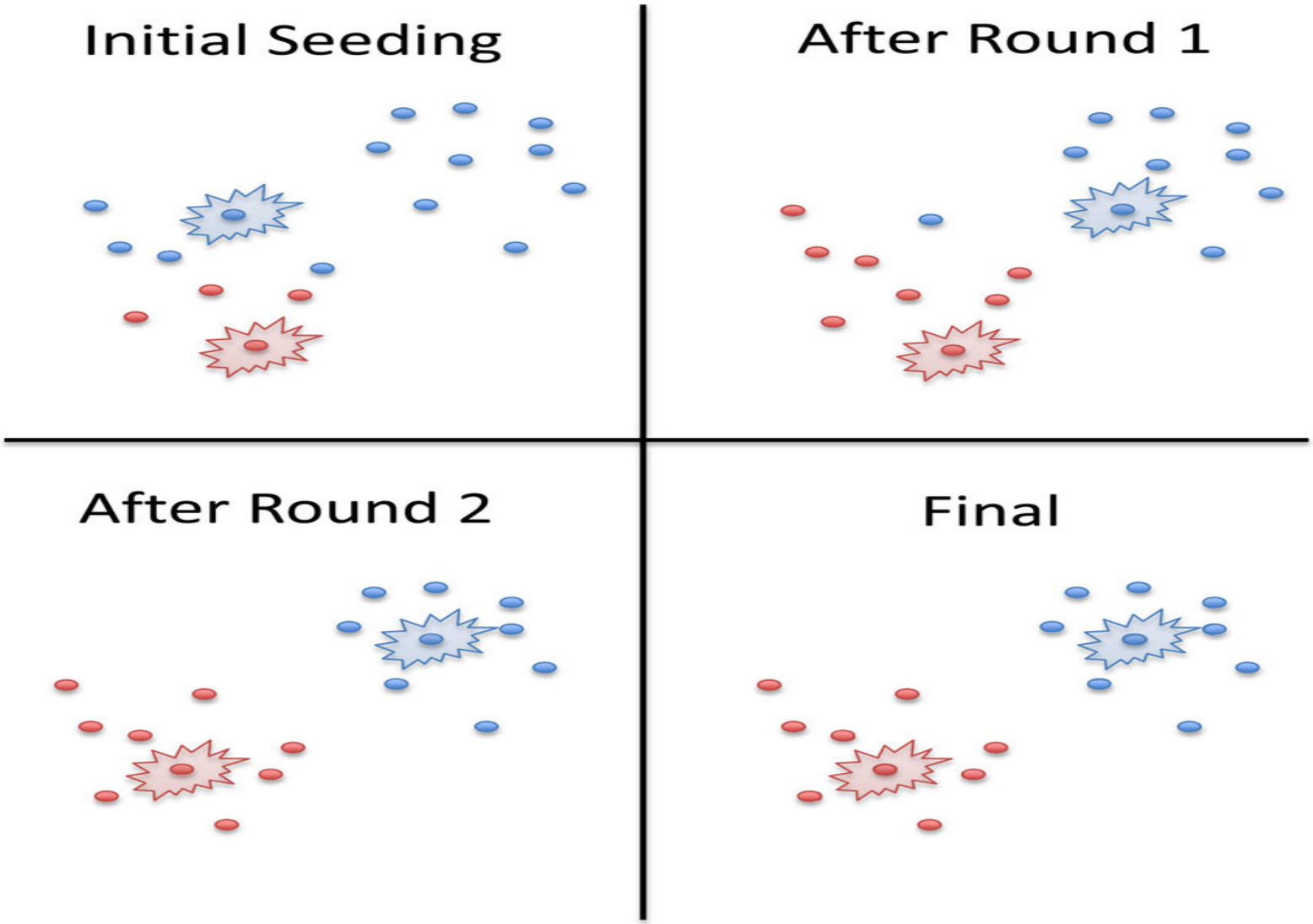
(i) We divide randomly the n data into k groups and calculate the value, denoted by I_k^0 of I_k for such partition. Here the centers $h_{c(i)}$ would be just the averages of the corresponding groups.

(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 k-means algorithm



https://devopedia.org/k-means-clustering

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,

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

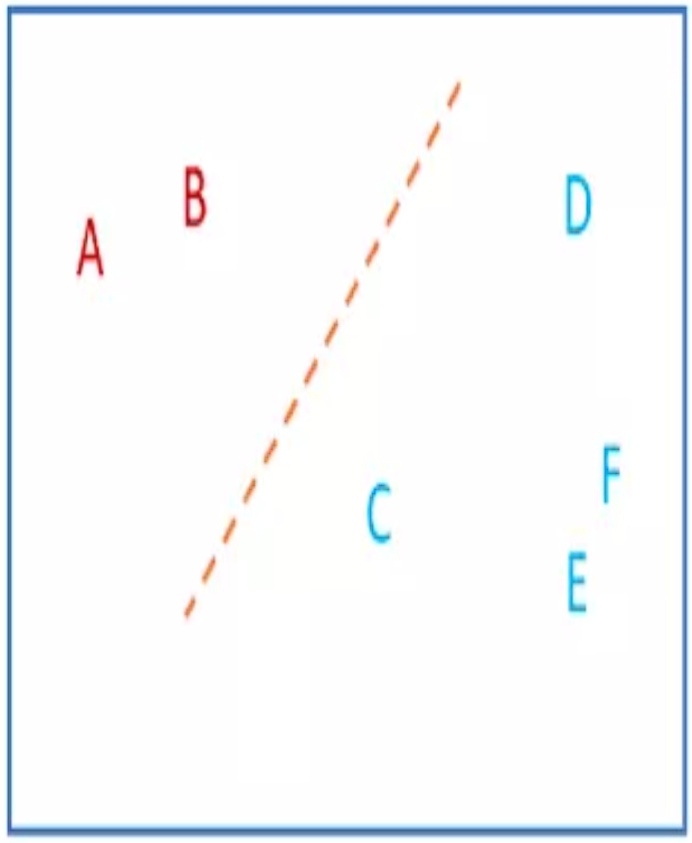
where $B(A, \epsilon) := \bigcup_{a \in A} B(a, \epsilon)$ denotes the ϵ-parallel set.
See Pollard (1981,1982), Cuesta and Matrán (1988).

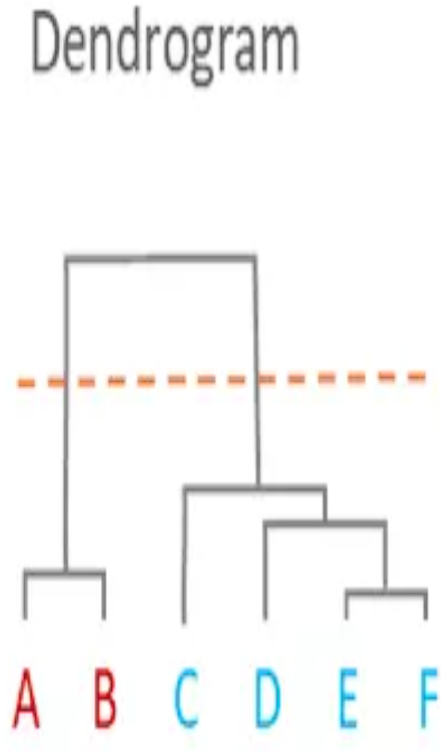
A different approach to clustering: agglomerative (hierarchical) algorithms (I)

- 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".

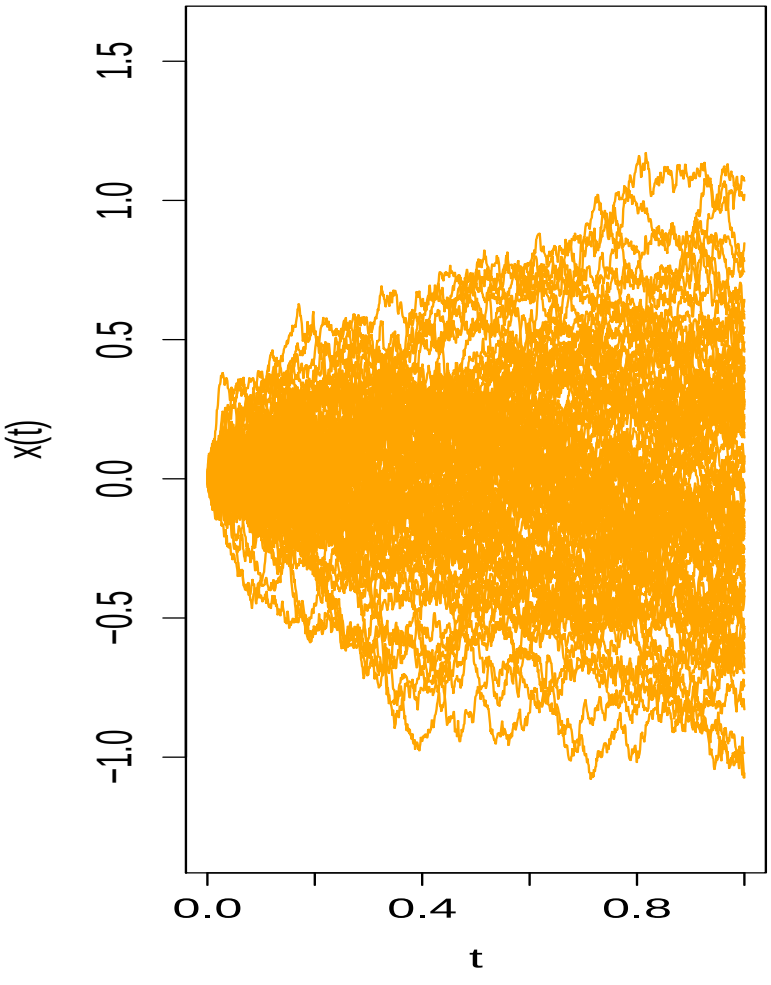
A different approach to clustering: agglomerative (hierarchical) algorithms (II)

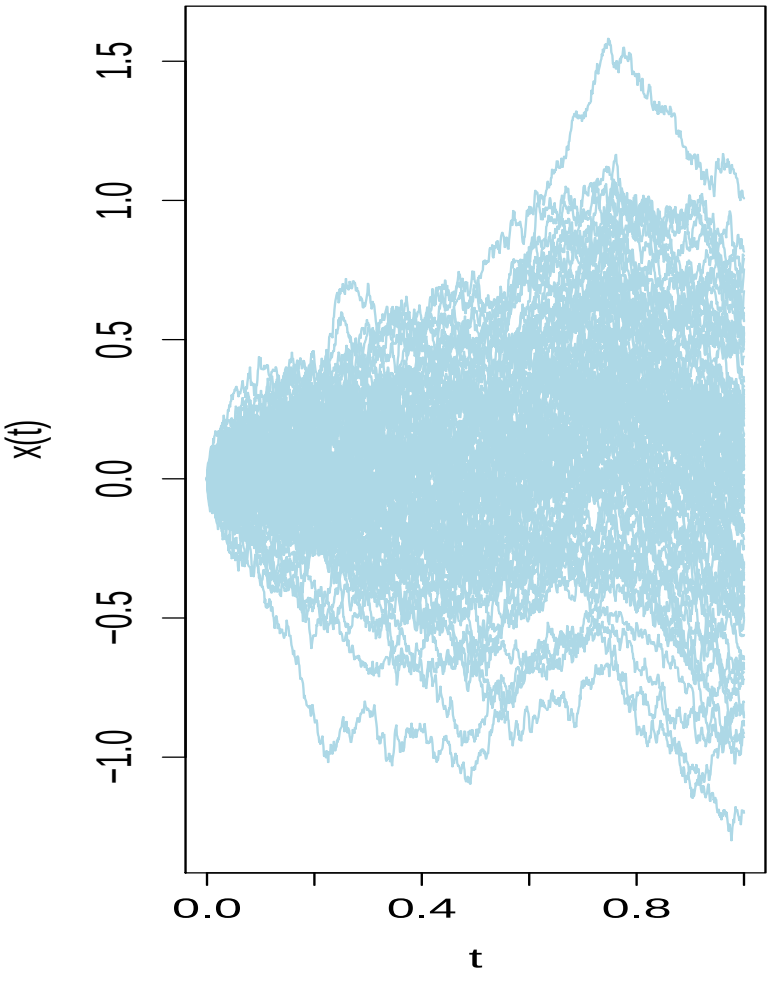
When we don't have too many observations this algorithm is usually summarized in a **dendogram** as this one





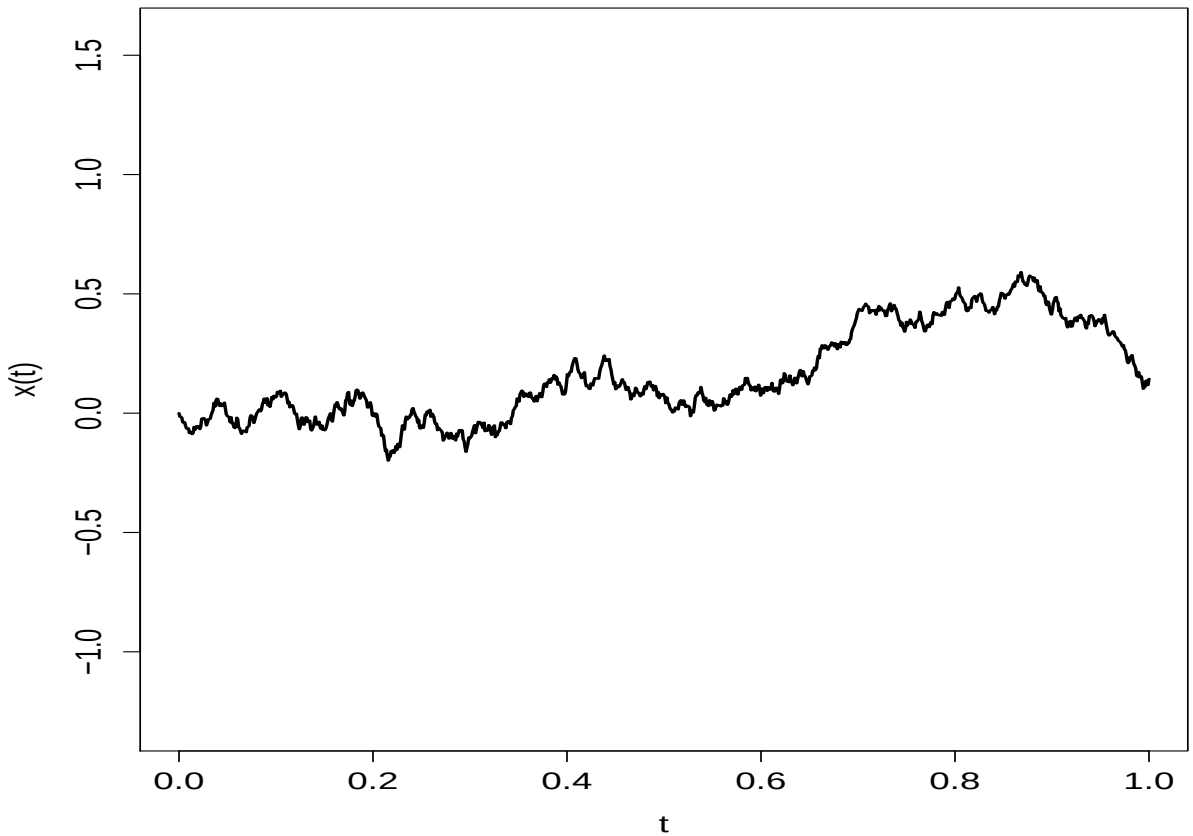
Functional classification problem





Functional classification problem (II)

Which is the class of this trajectory?



The problem of functional discrimination: statement

Let (X_i, Y_i), $i = 1, \ldots, n$, i.i.d., where the X_i toman valores en $\mathcal{X}$ and the Y_i take values on $\{0, 1\}$

$\{(X_i, Y_i), i = 1, \ldots, n\} =$ "training sample".

The aim is to predict the value of Y corresponding to a new incoming individual $Z = (X, Y)$ for which only X is observed.

Other usual names: "supervised classification", "statistical learning", "discriminant analysis", "pattern recognition", etc.

A few general references:

- Devroye, L., Györfi, L. and Lugosi, G. (2013). A Probabilistic Theory of Pattern Recognition. Springer-Verlag
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning, 2ª ed.. Springer.

Mathematical aspects: The optimal classifier

Classifier

A classifier (or classification rule) is a measurable function $g : \mathbb{R}^d \longrightarrow 0, 1$, that assigns a class to an observation x.

The optimal classifier

The "Bayes Rule" is a special classifier defined as

$$g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}},$$

where $\eta(x) = \mathbb{P}(Y = 1|X = x) = \mathbb{E}(Y|X = x)$.

Theorem.- For any classifier $g : \mathbb{R}^d \longrightarrow 0, 1$, we have that

$$\mathbb{P}(g^*(X) \neq Y) \leq \mathbb{P}(g(X) \neq Y). \tag{8}$$

Mathematical aspects: The optimal classifier

Proof:

$$\begin{aligned}
&\mathbb{P}(g(X) \neq Y | X = x) = 1 - \mathbb{P}(g(X) = Y | X = x) \\
&= 1 - (\mathbb{P}(Y = 1, g(X) = Y | X = x) + \mathbb{P}(Y = 0, g(X) = Y | X = x)) \\
&= 1 - \left(\mathbb{I}_{\{g(x)=1\}} \mathbb{P}(Y = 1 | X = x) + \mathbb{I}_{\{g(x)=0\}} \mathbb{P}(Y = 0 | X = x)\right) \\
&= 1 - \left(\mathbb{I}_{\{g(x)=1\}} \eta(x) + \mathbb{I}_{\{g(x)=0\}} (1 - \eta(x))\right)
\end{aligned}$$

Then, for all x,

$$\begin{aligned}
&\mathbb{P}(g(X) \neq Y | X = x) - \mathbb{P}(g^*(X) \neq Y | X = x) \\
&\eta(x) \left(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}\right) + (1 - \eta(x)) \left(\mathbb{I}_{\{g^*(x)=0\}} - \mathbb{I}_{\{g(x)=0\}}\right) \\
&= (2\eta(x) - 1) \left(\mathbb{I}_{\{g^*(x)=1\}} - \mathbb{I}_{\{g(x)=1\}}\right) \geq 0,
\end{aligned}$$

by definition of g^*.

Mathematical aspects: Classifiers and errors

An empirical classifier is a sequence of functions $g_n(x) := g_n(x; (X_1, Y_1), \ldots, (X_n, Y_n))$ defined on $\mathbb{R}^d$, taking values in $0, 1$. Given a new random observation X, the prediction for the corresponding Y is given by $g_n(X)$.
The conditional error of the classifier g_n is given by:

$$L_n = \mathbb{P}\{g_n(X) \neq Y | \mathcal{D}_n\},$$

where $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$. It can be estimated in different ways.
Let us define the Bayes error (or optimal error) as

$$L^* = \mathbb{P}(g^* \neq Y).$$

Consistency

A classifier is said to be consistent if $L_n \longrightarrow L^*$, as $n \to \infty$.

- Weak consistency: if convergence is in probability.
- Strong consistency: if convergence is almost sure.

Mathematical aspects: Construction of classifiers

The statistical theory of classification focuses on the construction of "good" classifiers using two methodologies:

Empirical risk minimization

The optimal classifier is sought by minimizing the empirical error on the training set.

Empirical error

$$\hat{L}_n = \frac{1}{n}\sum_{i=1}^{n} I_{\{g_n(X_i)\neq Y_i\}}$$

Cross validation error

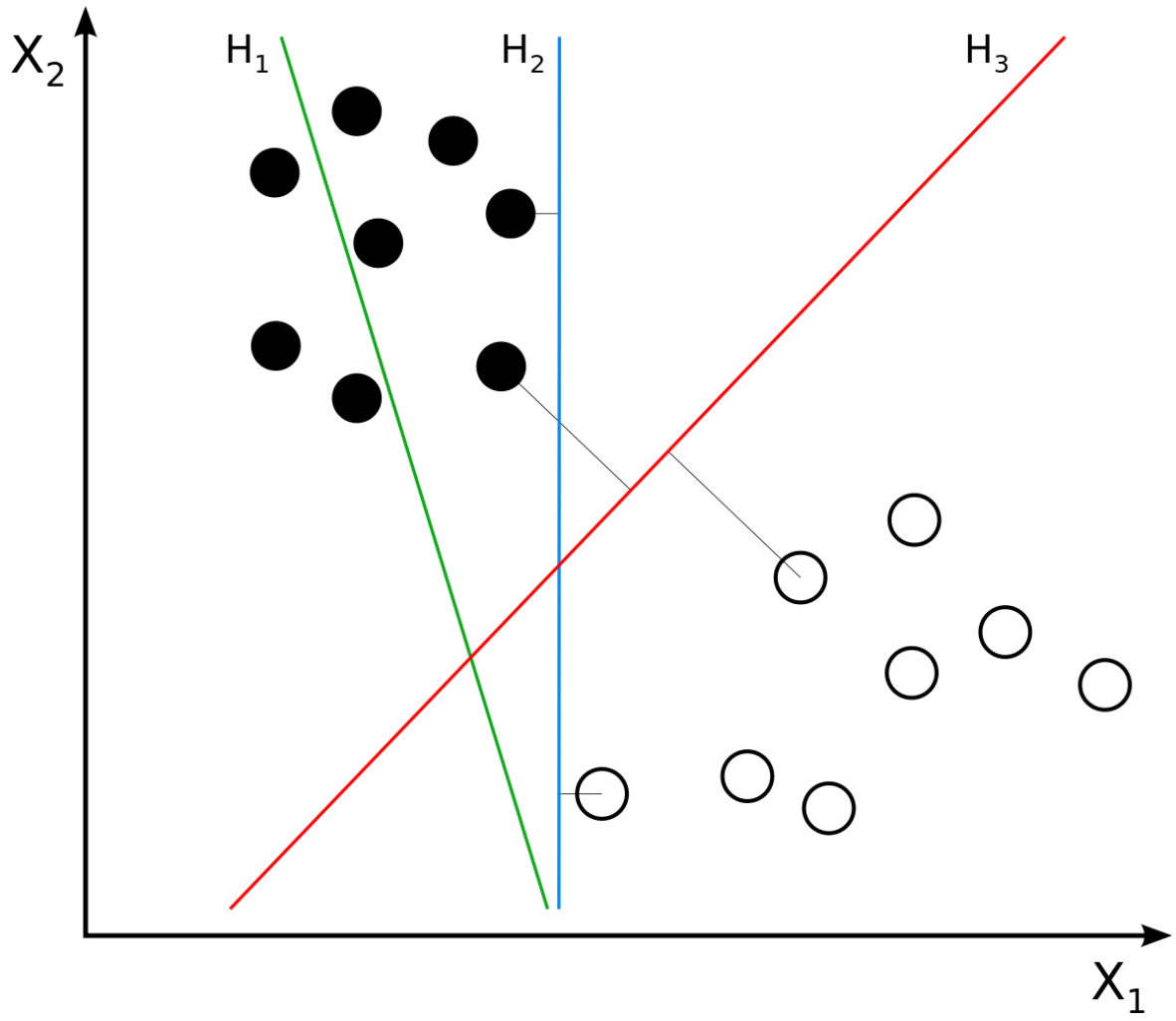
$$\tilde{L}_n = \frac{1}{n}\sum_{i=1}^{n} I_{\{g_{ni}(X_i)\neq Y_i\}}$$

where g_{ni} denotes the classifier trained without the i-th observation.

Plug-in methods

Estimating the function $\eta(x)$ to provide an expression for the optimal rule. Each nonparametric estimation of $\eta(x)$ yields a different classifier.

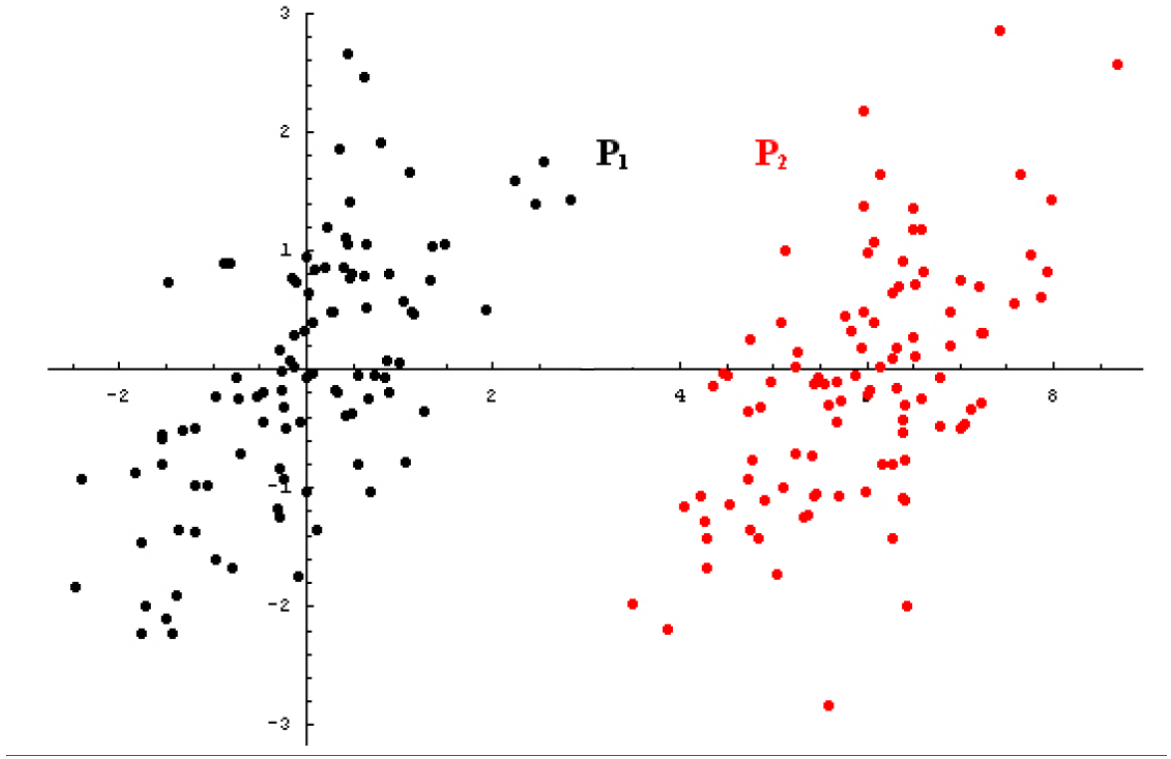
Linear boundaries



Notation

- $\mu_i = \mathbb{E}(X|Y = i)$: mean vector of X under P_i.
- Σ_i: covariance matrix of $X|Y = i$.
- Σ: under homoscedasticity $\Sigma_0 = \Sigma_1 = \Sigma$.
- n_i: number of elements of class i.
- $\bar{x}_i$: estimator of μ_i.
- S_i: estimator of Σ_i.
- $S_W = \frac{(n_0-1)S_0+(n_1-1)S_1}{n_0+n_1+2}$: pooled estimator of Σ.

We want to find the "best linear classifier". In geometrical terms, we wish to find the equation of a hyperplane leaving μ_0 and μ_1 in different subspaces, defined in order to classify (with the lowest possible error) the observations into either P_0 or P_1 according to the half-space they belong.



Fisher's method: Assume $\Sigma_0 = \Sigma_1 = \Sigma$. We want to find the direction $w \in \mathbb{R}^d$ that maximizes the separation between μ_0 y μ_1 with some restriction about the variance of the projections:

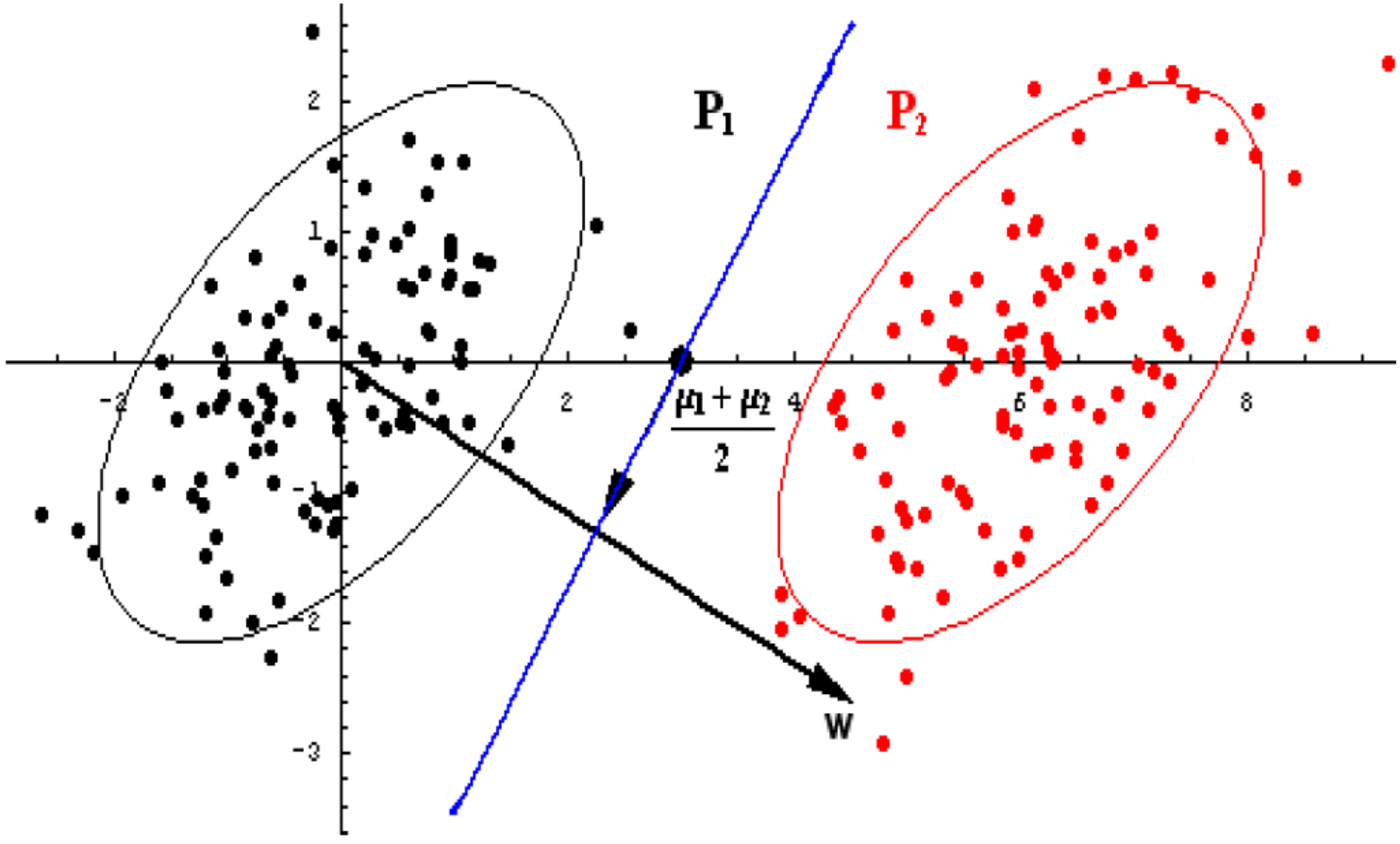
$$w^* = \underset{w}{argmax}(w'(\mu_0 - \mu_1))^2 \quad \text{sujeto a} \quad w'\Sigma w = 1.$$

This is equivalent to maximize

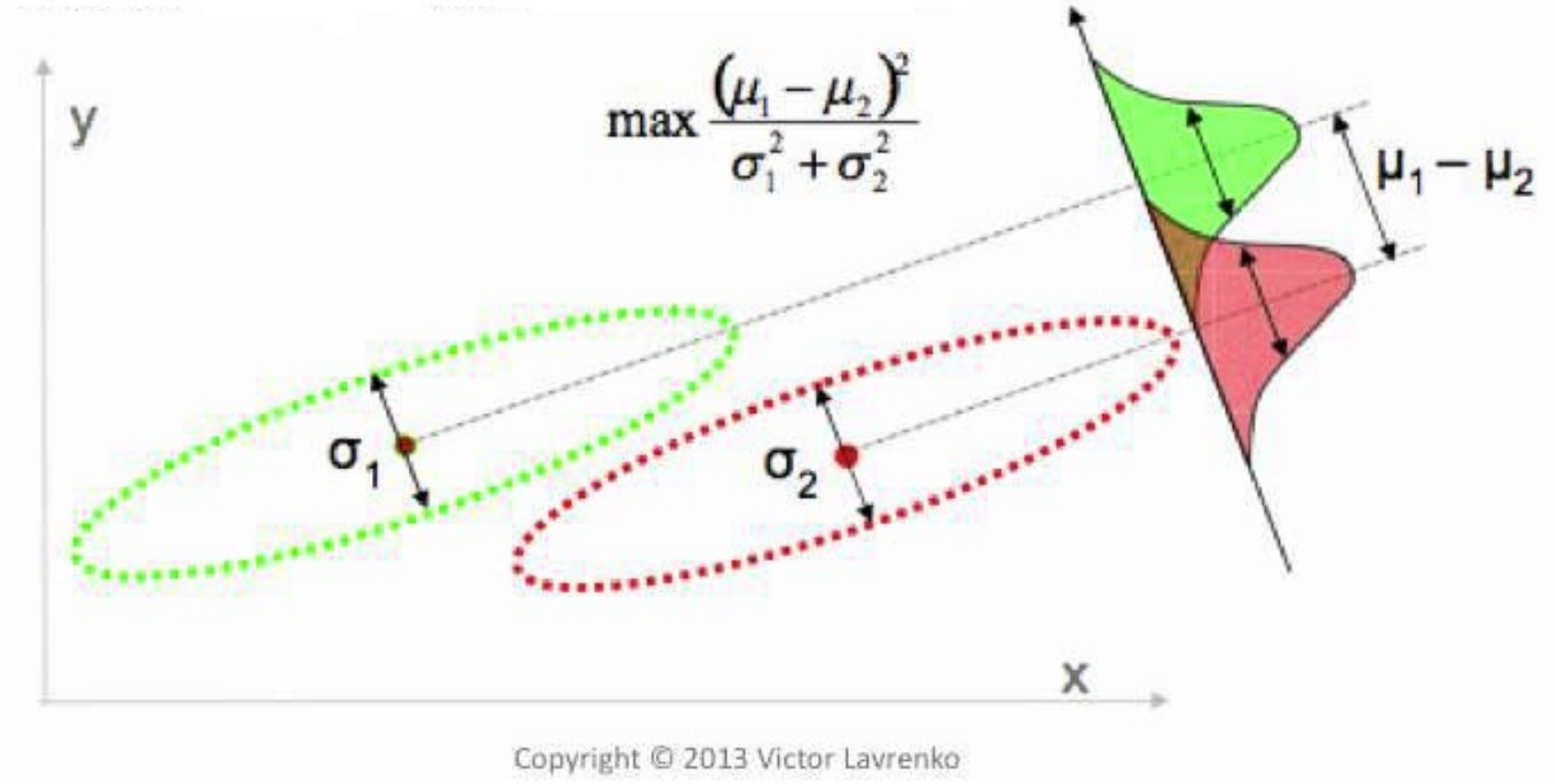
$$\frac{(w'(\mu_0 - \mu_1))^2}{w'\Sigma w}.$$

The solution is proportional to $w = \Sigma^{-1}(\mu_1 - \mu_0)$.

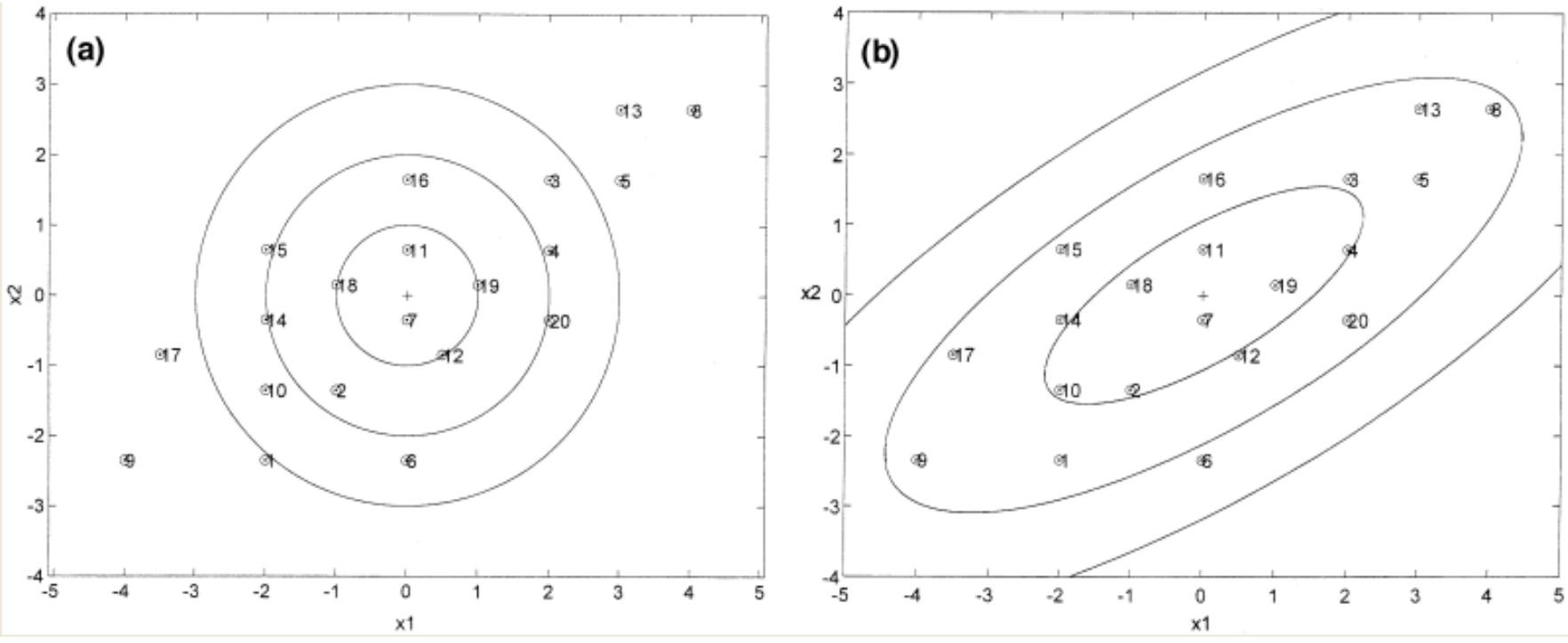
Given a new observation x, it is classified in the class which minimizes the distance between $w'x$ and $w'\mu_i$.


P1
P2
W
μ1 + μ2
2
2
4
6
8
2
1
-1
-2
-3

Interpretation

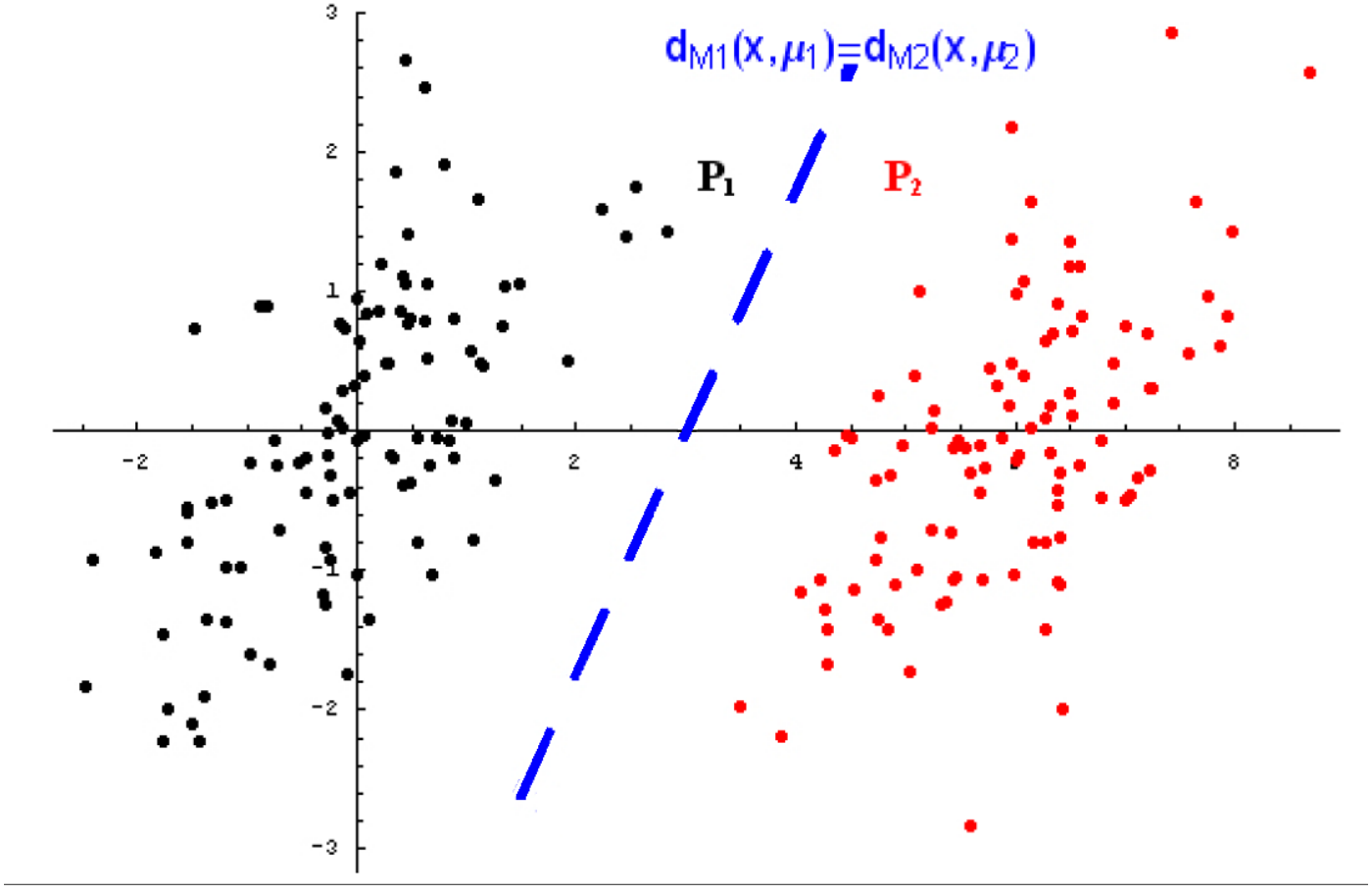


Mahalanobis distance



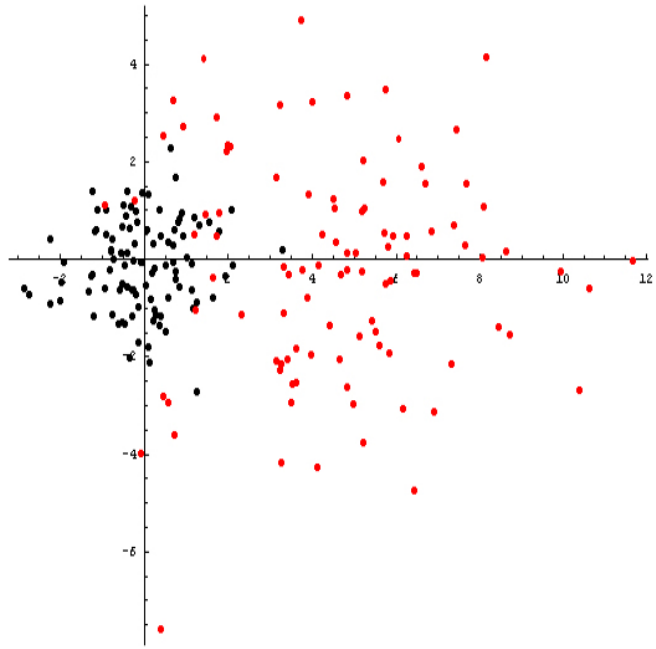
Alternative approach: Assume $\Sigma_0 = \Sigma_1 = \Sigma$. The classifier which assigns the new observation x to the population which minimizes the Mahalanobis distance between x and μ_i.

$$g(x) = \begin{cases} 0 & , \text{ if } (x-\mu_0)'\Sigma^{-1}(x-\mu_0) < (x-\mu_1)'\Sigma^{-1}(x-\mu_1) \\ 1 & , \text{ otherwise} \end{cases}$$



Several issues

- In practice, Σ, μ_0 y μ_1 are replaced with their corresponding estimators S_W, $\bar{x}_0$ y $\bar{x}_1$.
- In the Gaussian case, when $X|Y = i \sim N(\mu_i, \Sigma)$, the Fisher's methods is optimal.
- It is robust under heteroscedasticity, $\Sigma_0 \neq \Sigma_1$.



On the linear rule in the functional case

The functional counterpart of the linear rule would be as follows: assume that $\mathcal{X} \subset L^2[a,b]$ for some closed finite interval $[a,b] \subset \mathbb{R}$. A linear classifier for the functional measurement $x \in \mathcal{X}$ would be obtained by projecting the infinite dimensional x onto the real line and comparing such projection with those corresponding to the mean functions $\mu_j = \mathbb{E}(X|Y=j)$ for $j=0,1$. The projection "direction" β would be selected as the maximizer of the distance between the projected class means $\langle \beta, \mu_0 \rangle$ and $\langle \beta, \mu_1 \rangle$. The optimization is subject to the restriction $\int_a^b \beta(t)\langle \beta, \gamma(t,\cdot)\rangle dt = 1$, where $\gamma(s,t) = \mathrm{Cov}(X(t), X(s)|Y=j)$ for $j=0,1$. As in the finite-dimensional case this would lead to maximizing

$$\frac{\mathrm{Var}(\mathbb{E}(\langle \beta, X\rangle | Y))}{\mathbb{E}(\mathrm{Var}(\langle \beta, X\rangle | Y))}.$$

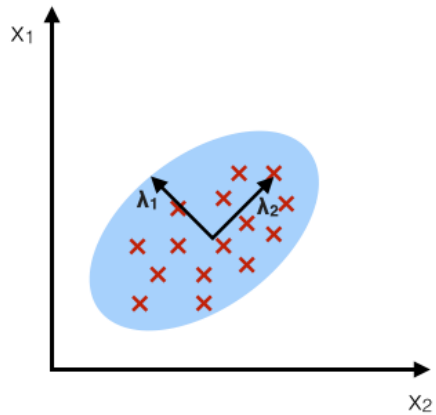
Unfortunately, the analogy with the finite-dimensional case must stop here since the covariance operator associated with the kernel $\gamma(s,t)$ is not in general invertible, so the above maximization problem has no solution. Analogously, the Mahalanobis distance cannot be extended directly to the functional setting. James and Hastie JRSS-B(2001), Berrendero, Bueno and Cuevas JMLR (2020)

LDA for dimensionality reduction

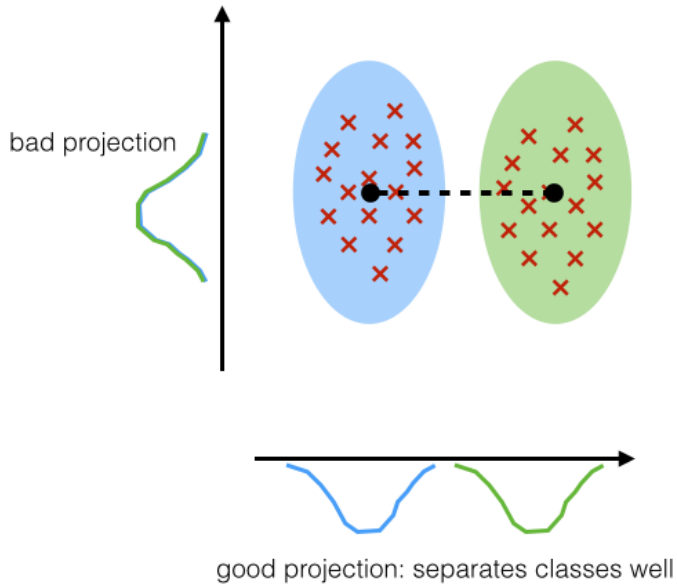
LDA can also be used as a method for dimensionality reduction:

- The best projection is calculated taking into account the class.
- The maximum dimension of the space we project onto is $C - 1$ (C denotes the number of classes).

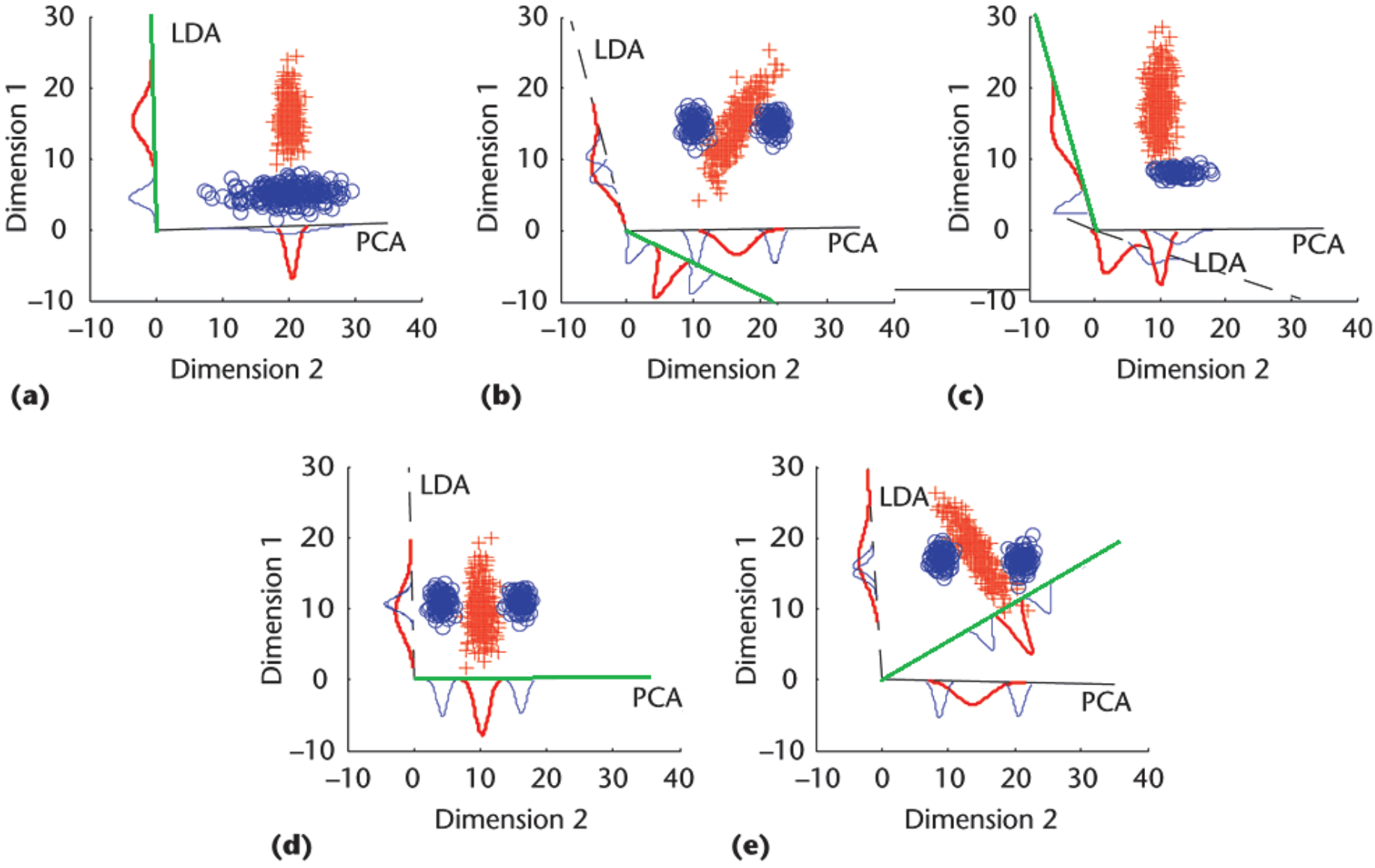
PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



LDA for dimensionality reduction



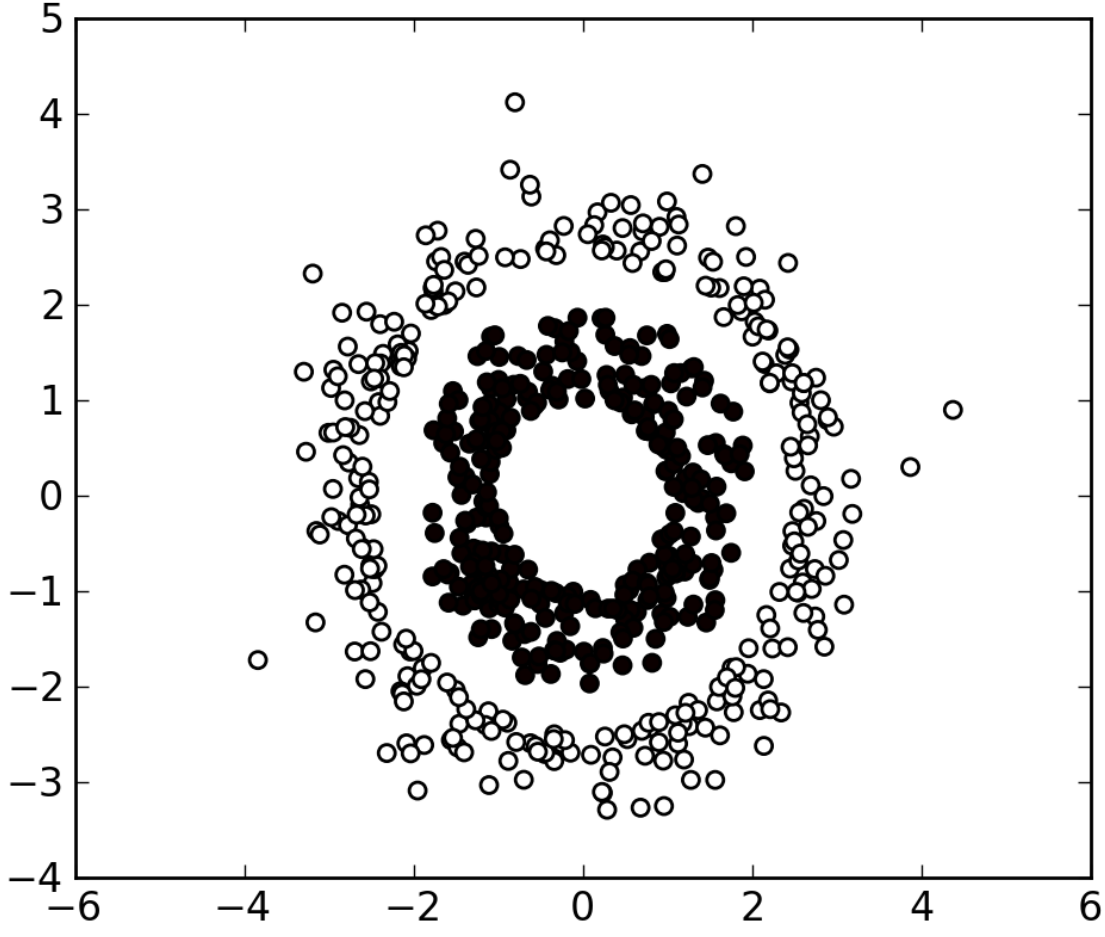
LDA summary

Advantages

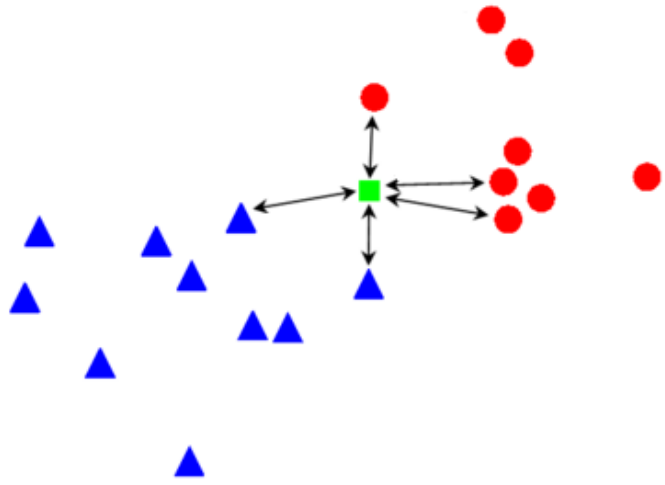
- Simple, easy to understand and interpret (better linear separation).
- Fast and easy to compute.
- Optimal under its assumptions.
- Allows for dimension reduction.
- Very popular, easy to find, many variants.

Disadvantages

- Problems if assumptions are not met.
- Does not handle non-linearities (linear).
- Problems in high dimensions (inversion of covariance matrix).


5
4
3
2
1
0
−1
−2
−3
−4
−6
−4
−2
0
2
4
6

The idea behind k-NN is very simple: what is close is most similar and belongs to the same class. Given an observation x, the k closest observations to x (k-neighbors) are searched among the X_i in the training set. x is assigned to the majority class among the k-neighbors.



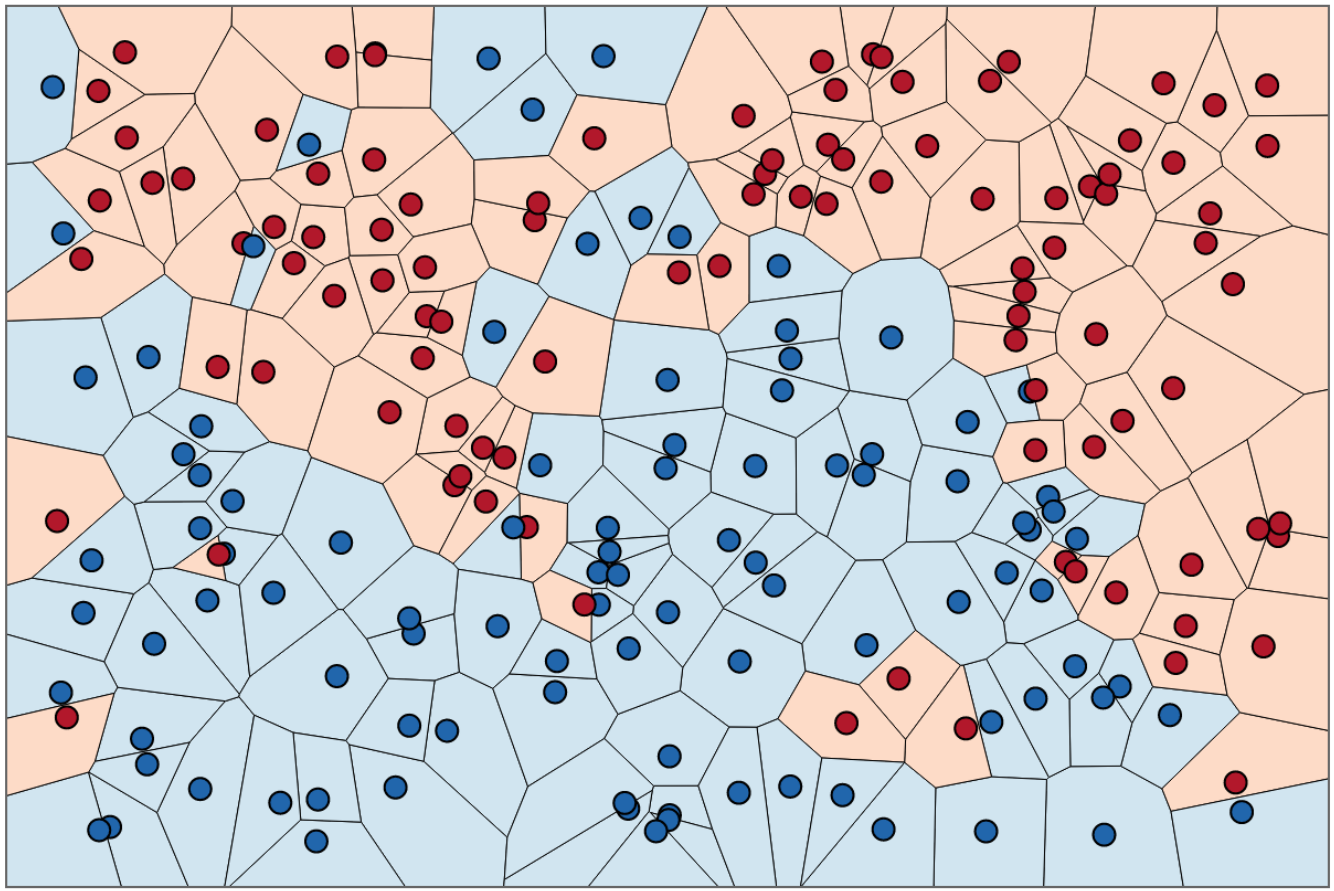
x is classified based on the k closest neighbors. It can be considered as a plug-in method taking

$$\eta(x) = \eta_n(x) = \frac{1}{k} \sum_{i=1}^{n} I_{\{x_i \in k(x)\}} Y_i.$$

Then,

$$g_n(x) = \mathbb{I}_{\{\eta_n(x) > 1/2\}}$$

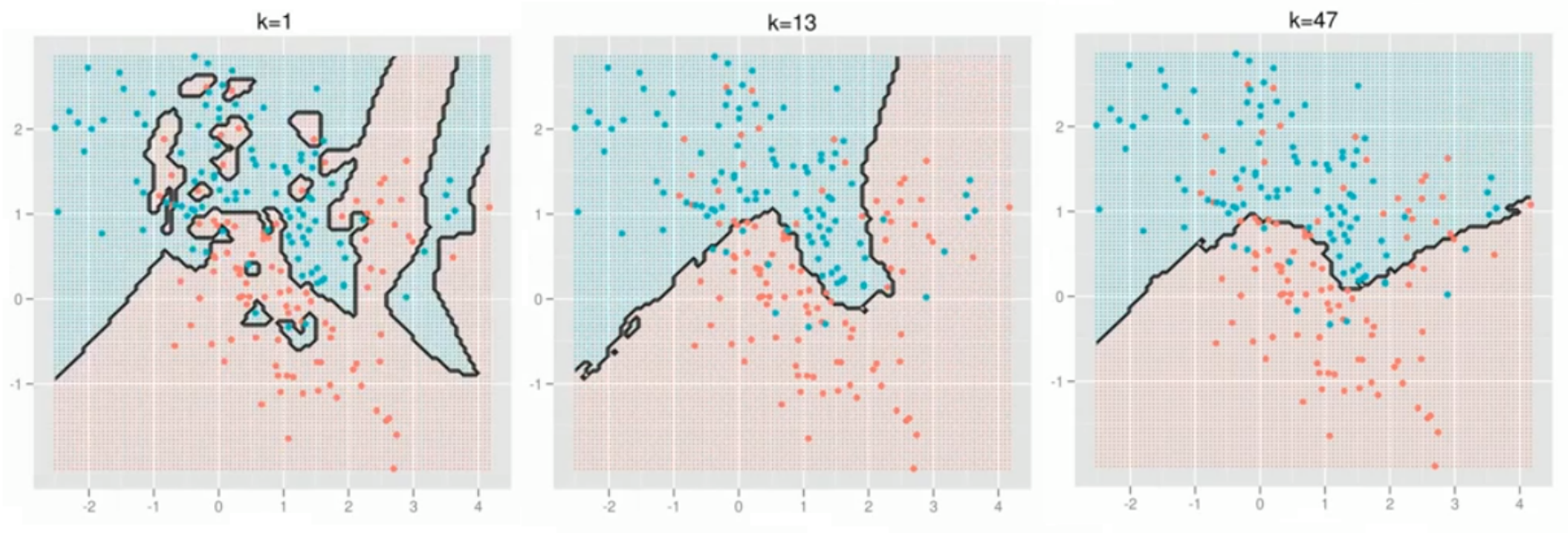
One only neighbor



The k-NN method: Universal consistency

The applicability of this classifier does not depend on any assumption on the distribution of (X, Y). In fact, it is universally consistent (i.e., $L_n \to L$ with probability one, for all distribution of (X, Y), provided that $k \to \infty$ and $k/n \to 0$ (Stone, 1977, Devroye et al. (2013))

This universal consistency property only holds for the finite-dimensional case. **It is no longer true in the functional case**, where strong additional assumptions are required. Cérou and Guyader (2006)



Summary of k-NN

Advantages

- Simple and easy to understand.
- Captures non-linearities and local variations.
- Valid for any distribution (universally consistent).
- Good empirical results, a good benchmark.
- Very popular, easy to find, many variations.
- The training cost is zero.

Disadvantages

- Being non-parametric, it can have problems in high dimensions if we do not have many data points.
- We need to tune the parameter k, and there can be ties.
- We need to decide the metric.
- The classification cost is high if the sample size is large.

Other classifiers

- Kernel-based classifiers.
- Decision trees.
- Support vector machines (SVM).
- Neural networks (MLP, recurrent, deep...).
- Regression-based classifiers.
- Evolutionary algorithms.
- Rule-based classifiers.
- Naive Bayes.
- Deep learning-based classifiers.

...

There are countless comparisons between algorithms, for example https://docs.microsoft.com/es-es/azure/machine-learning/studio/algorithm-choice

Mixture of experts

A mixture of classifiers is a set $g_1(x), \ldots, g_k(x)$ whose estimated class y is a combination of the estimates of each of the classifiers. Independence is necessary to take into account several requirements.

- Hierarchical.
- Bagging: unweighted voting.
- Boosting: weighted voting.
- Networks.
- ...

Depth-based classifiers

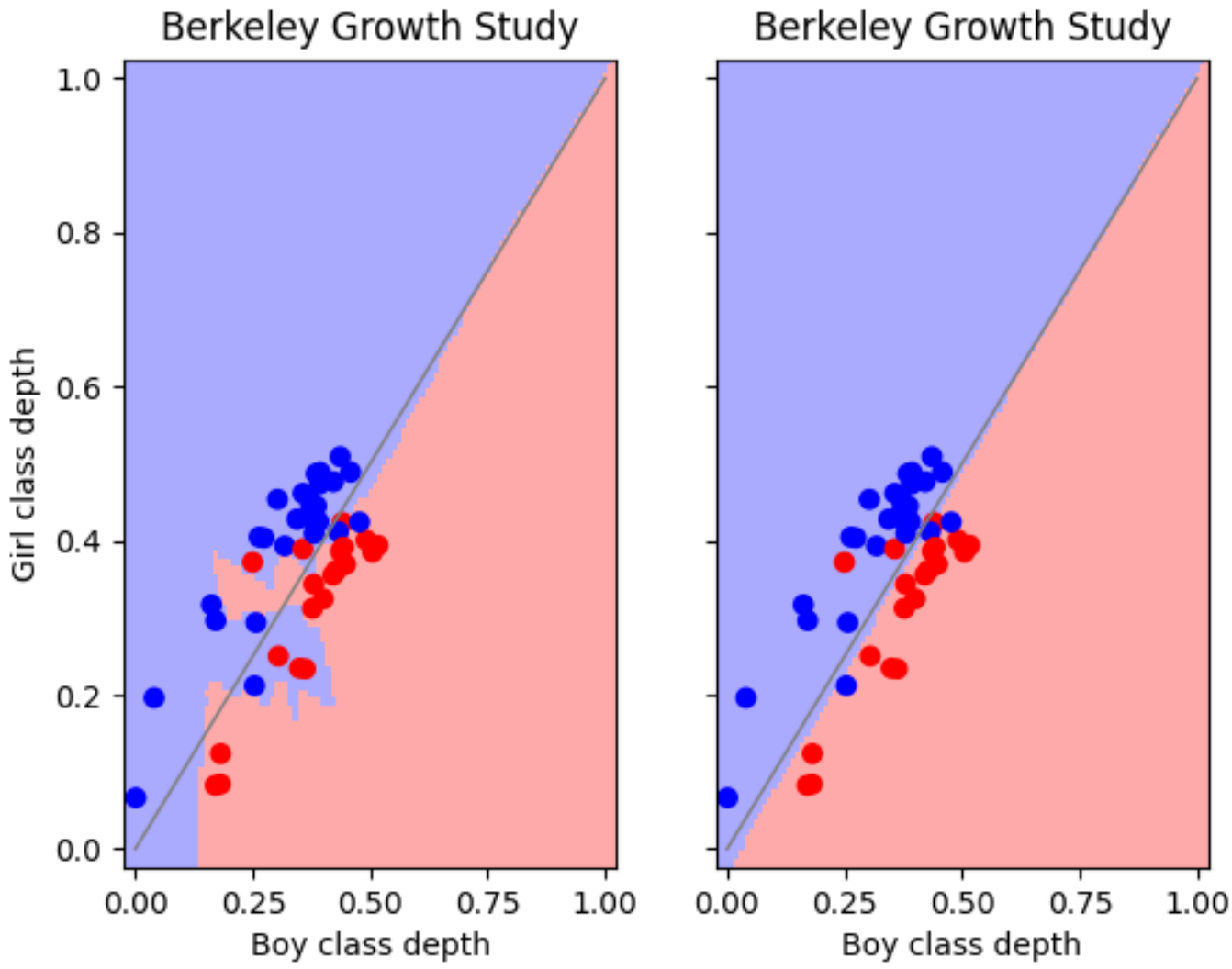
Given a depth measure $D(\mathbb{P}_{\ltimes}, x)$, we can assign a new incoming x_0 to the population where it is more deeply placed.

Typically these methods fail when the populations are "nested" or extremely heteroscedastic.

DD-classifier Li et al. (2012).

Sguera et al. (2014)

DD-Classifier for functional data



Cuesta-Albertos et al. (2015).

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

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

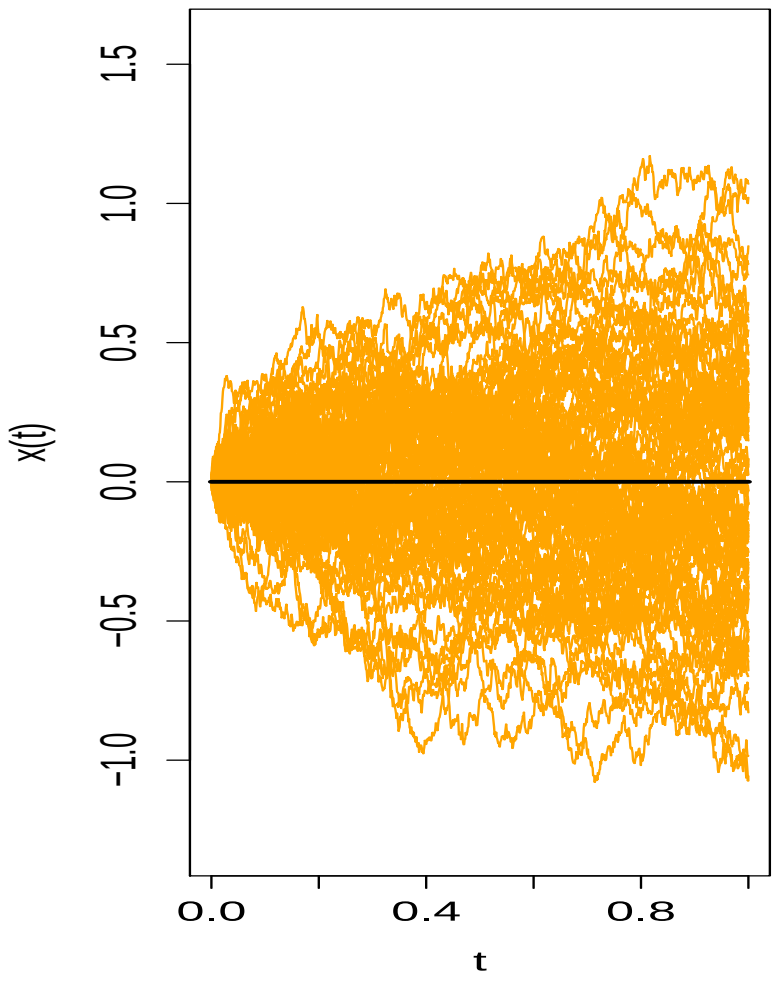
Bayes error

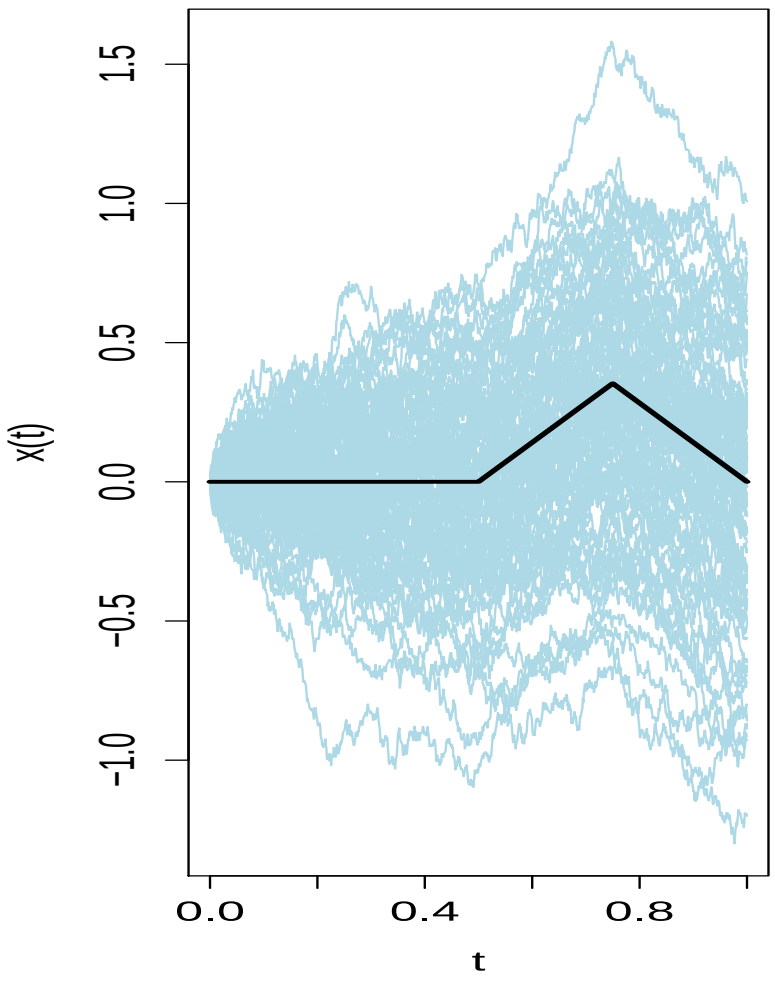
(1) $\eta^*(X)|Y = 0 \sim N\left(-\frac{1}{2} \parallel m \parallel_K^2, \parallel m \parallel_K\right).$

(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 $N \to \infty$,

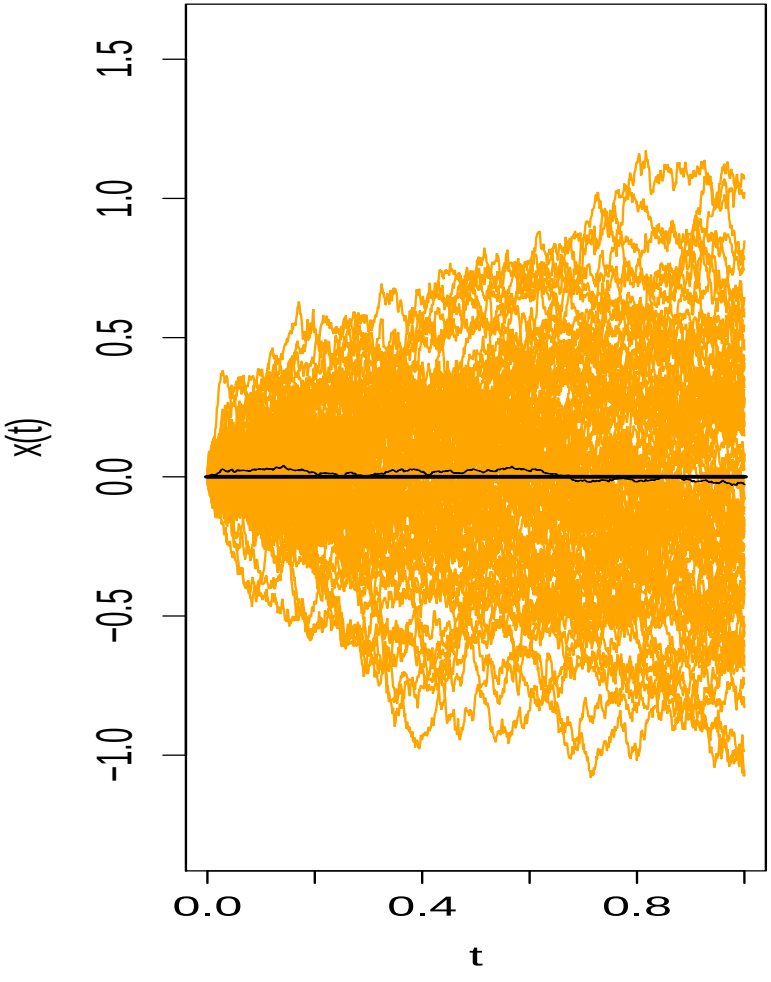
$$2^N \sum_{j=1}^{2^N} (\Delta_j m)(\Delta_j X) - 2^{N-1} \sum_{j=1}^{2^N} (\Delta_j m)^2 \to \eta(X)$$

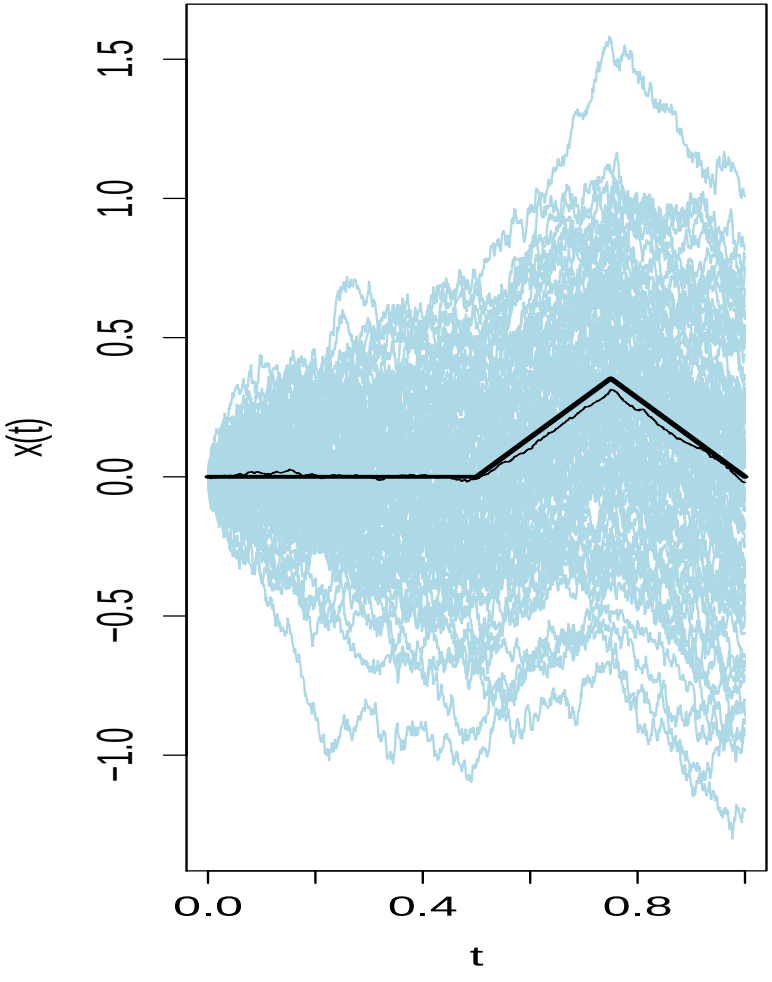
m is unknown

Naive choice: we can replace it with

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

Example: estimated Bayes rule

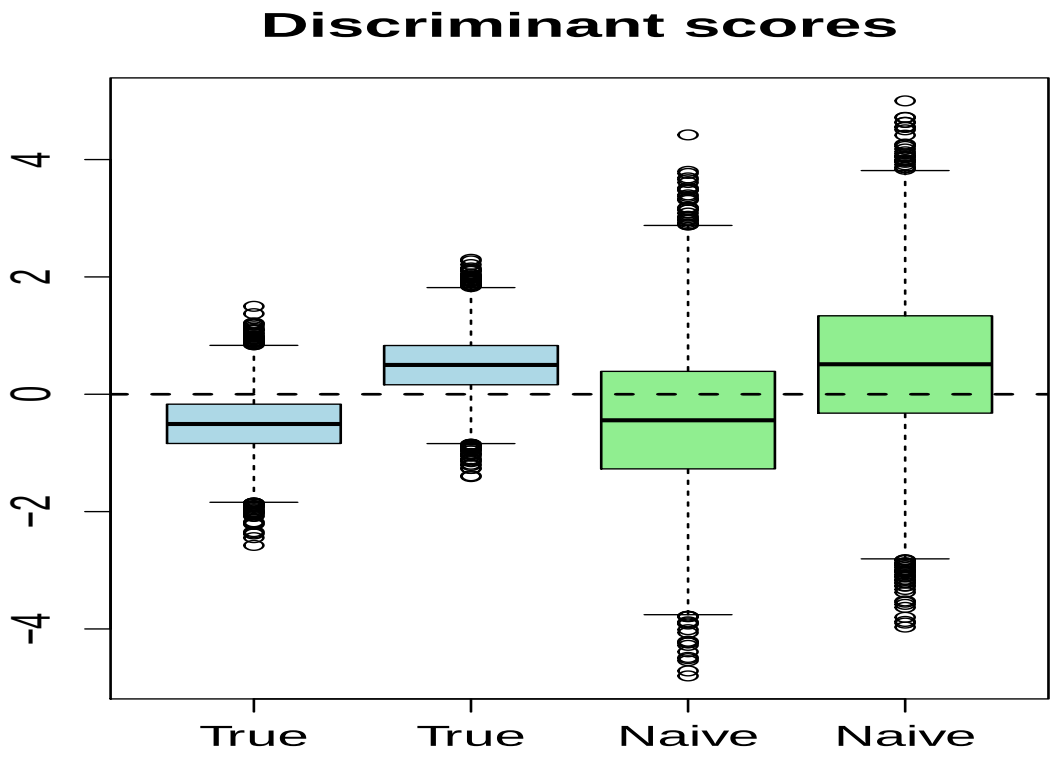




Discriminant scores and classification errors

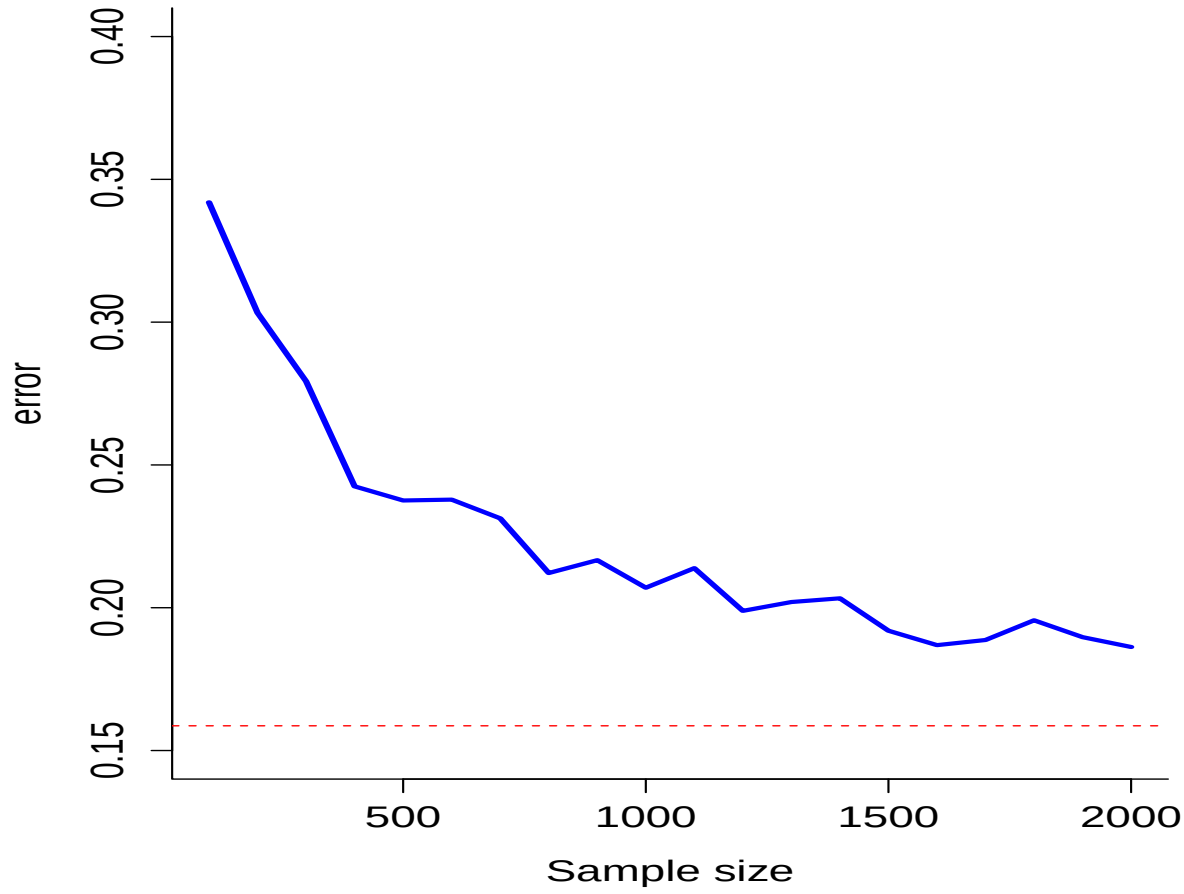
Based on 10000 test trajectories of each model ($n_0 = n_1 = 100$)

Rule	Bayes	True	Naive
Error	15.87%	15.98%	34.88%



Classification error as n increases

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



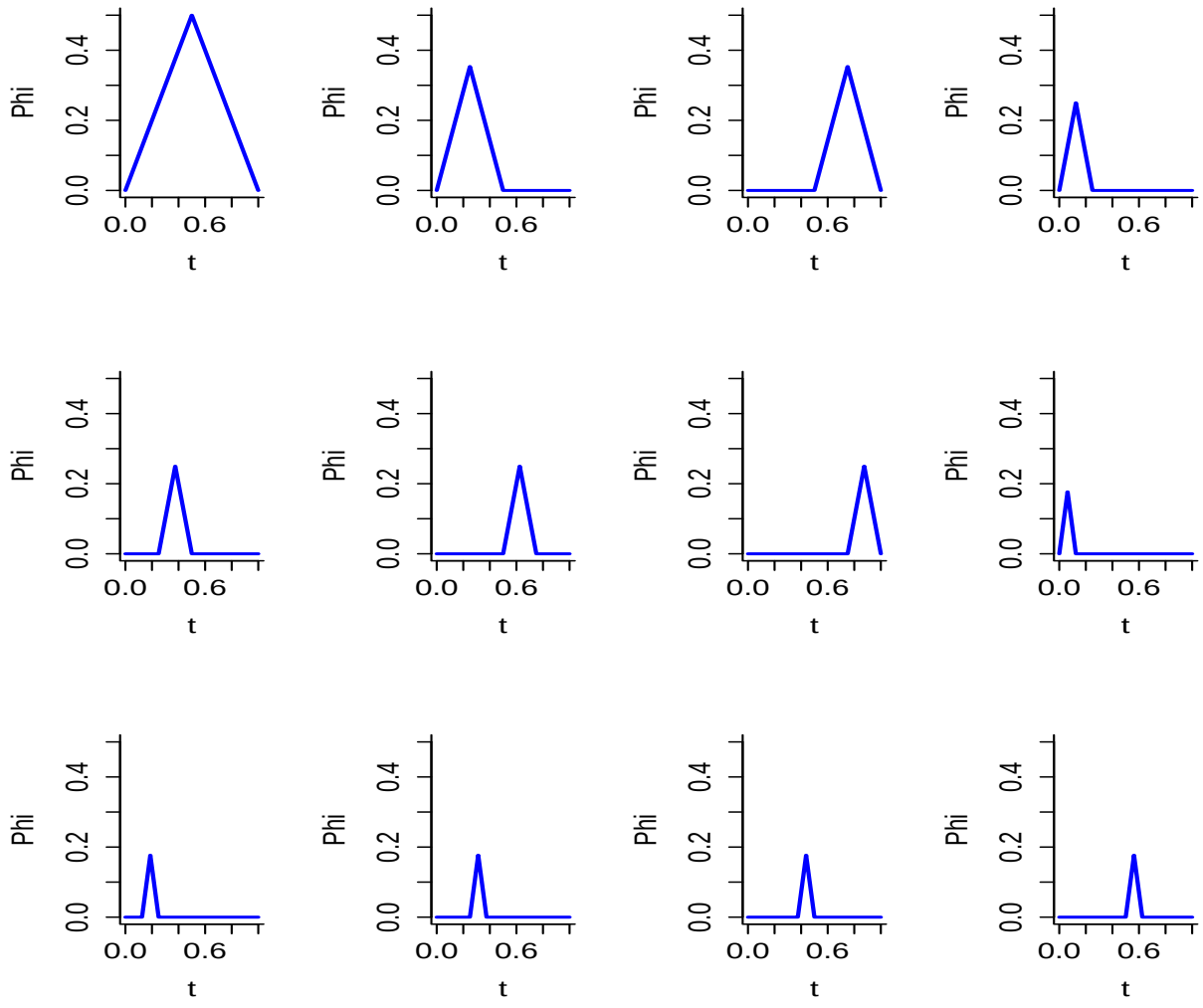
Alternatives

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

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

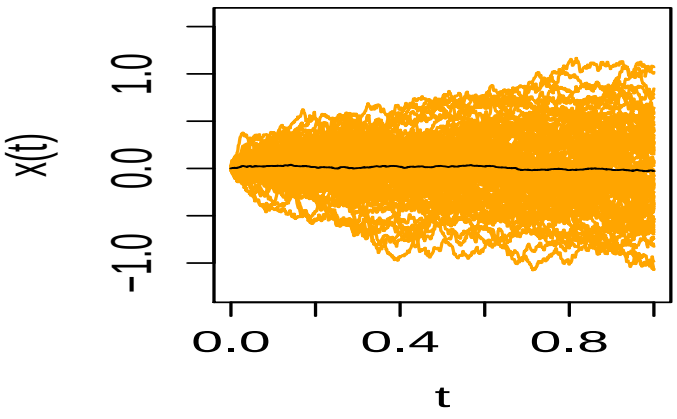
Natural estimators are obtained replacing $m(\cdot)$ with $\bar{X}(\cdot)$.

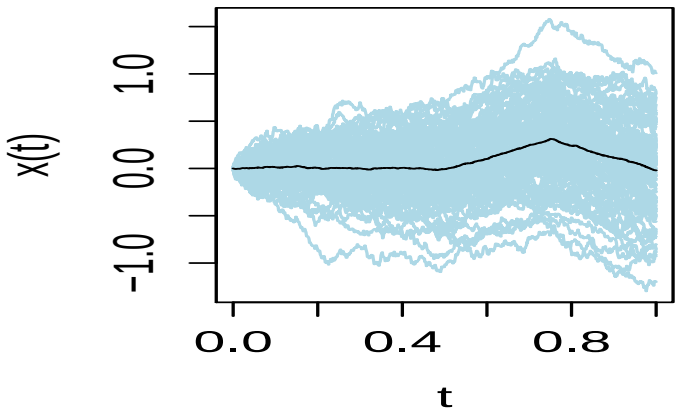
For an appropriate K,

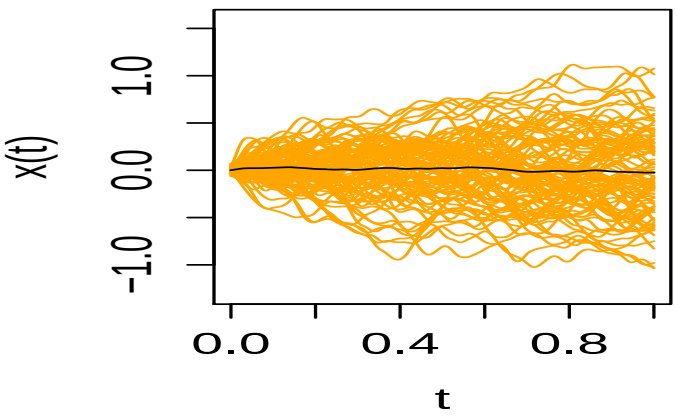
$$\hat{m}(t) = \hat{c}_{0,0}\, t + \sum_{k=1}^{K} \sum_{j=1}^{2^{k-1}} \hat{c}_{k,j} \Phi_{k,j}(t)$$

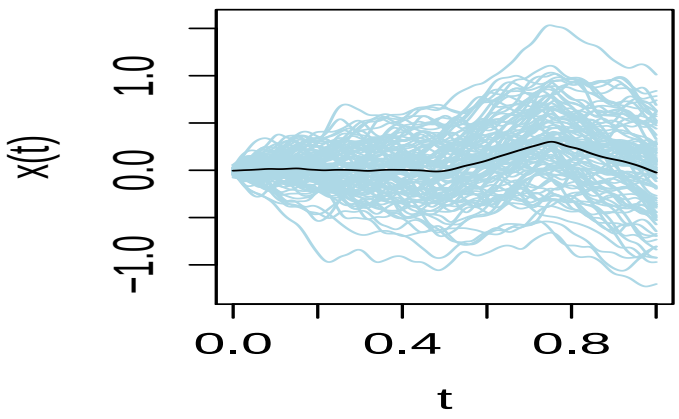
Smoothed trajectories

Using splines





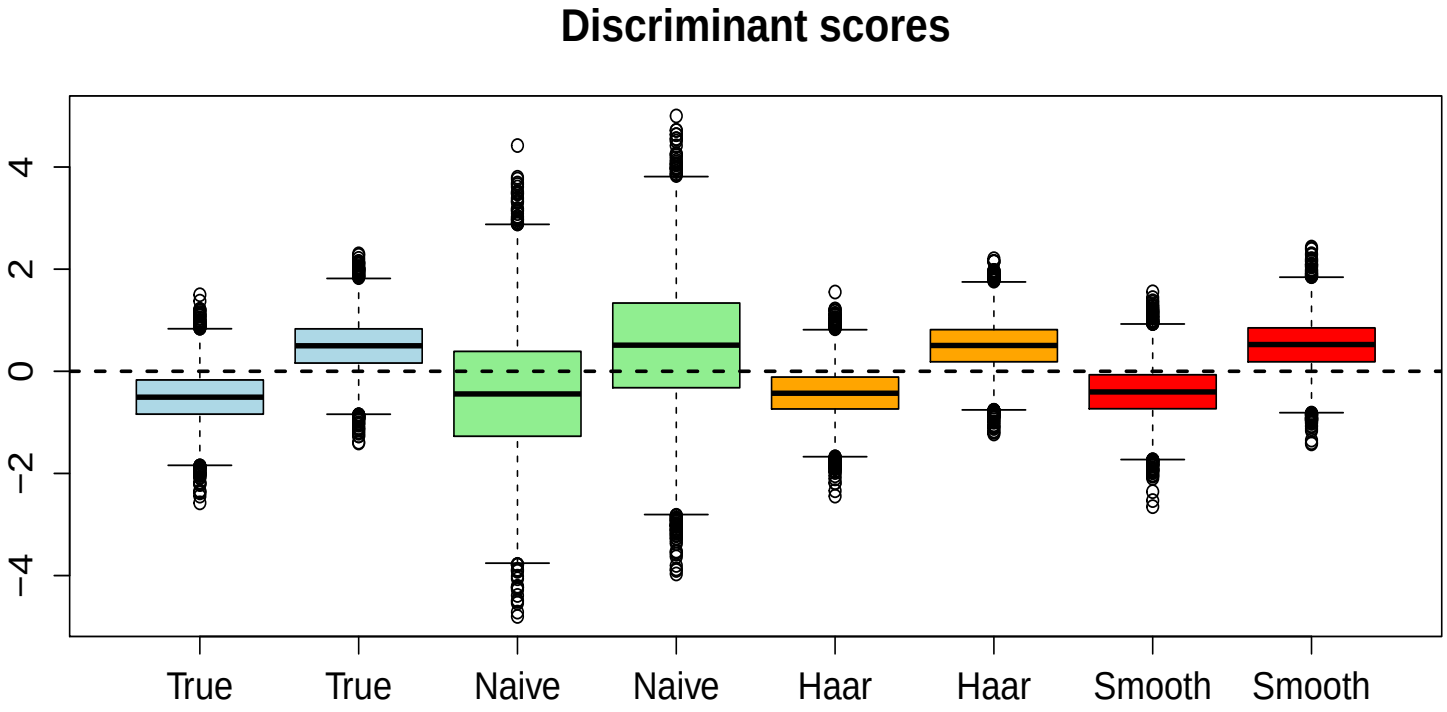




Discriminant scores and classification errors

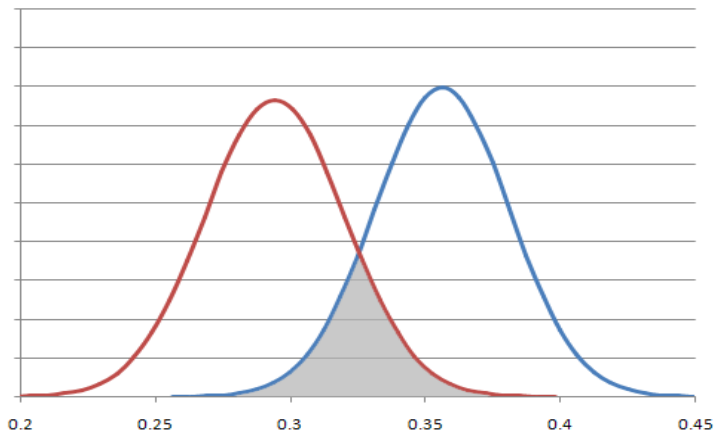
Based on 10000 test trajectories of each model ($n_0 = n_1 = 100$)

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.

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

Multivariate setting

$g^*(X) = 1$ if and only if

$$\mathbf{X}^\dagger \mathbf{K}^{-1} \mathbf{m} - \frac{1}{2} \mathbf{m}^\dagger \mathbf{K}^{-1} \mathbf{m} > 0$$

Functional setting (RKHS)

$g^*(X) = 1$ if and only if

$$\langle X, m \rangle_K - \frac{1}{2} \parallel m \parallel_K^2 > 0.$$

- RKHS approach is an useful tool for classification of Gaussian processes.
- R-N 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

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

Bibliography I


Antonio Arauzo-Azofra, José Luis Aznarte, and José M Benítez.
Empirical study of feature selection methods based on individual feature evaluation for classification problems.
Expert Systems with Applications, 38(7):8170–8177, 2011.


Amparo Baíllo, Antonio Cuevas, and Juan Antonio Cuesta-Albertos.
Supervised classification for a family of Gaussian functional models.
Scandinavian Journal of Statistics, 38(3):480–498, 2011.


I. Barba, E. Miró-Casas, E. Pladevall, R. Sebastián, J. R. Berrendero, J.L. Torrecilla, A. Cuevas, and D. García-Dorado.
High fat diet induces metabolic changes associated to increased oxidative stress in male hearts.
Draft, 2015.


J R Berrendero, A Cuevas, and J L Torrecilla.
The mRMR variable selection method: a comparative study for functional data.
Journal of Statistical Computation and Simulation, 86(5):891–907, 2016.


J R Berrendero, A Cuevas, and J L Torrecilla.
Variable selection in functional data classification: a maxima hunting proposal.
Statistica Sinica, 26(2):619–638, 2016.

Bibliography II


José R Berrendero, Antonio Cuevas, and José L Torrecilla.
On near perfect classification and functional Fisher rules via reproducing kernels.
arXiv:1507.04398, pages 1–27, 2015.


Gérard Biau, Benoît Cadre, and Quentin Paris.
Cox process functional learning.
Stat. Inference Stoch. Process., 18(3):257–277, 2015.


Gérard Biau, Frédéric Cérou, and Arnaud Guyader.
On the rate of convergence of the bagged nearest neighbor estimate.
Journal of Machine Learning Research, 11(Feb):687–712, 2010.


Denis Bosq.
Linear processes in function spaces: theory and applications, volume 149.
Springer Science & Business Media, 2012.


Frédéric Cérou and Arnaud Guyader.
Nearest neighbor classification in infinite dimension.
ESAIM: Probability and Statistics, 10:340–355, 2006.


JA Cuesta and C Matrán.
The strong law of large numbers for k-means and best possible nets of banach valued random variables.
Probability theory and related fields, 78(4):523–534, 1988.

Bibliography III


Juan A Cuesta-Albertos, Manuel Febrero-Bande, and Manuel Oviedo de la Fuente.
The DD G-classifier in the functional setting.
arXiv:1501.00372, 2015.


Antonio Cuevas.
A partial overview of the theory of statistics with functional data.
Journal of Statistical Planning and Inference, 147:1–23, 2014.


Antonio Cuevas, Manuel Febrero, and Ricardo Fraiman.
An anova test for functional data.
Computational statistics & data analysis, 47(1):111–122, 2004.


Aurore Delaigle and Peter Hall.
Achieving near perfect classification for functional data.
Journal of the Royal Statistical Society B, 74(2):267–286, 2012.


Aurore Delaigle and Peter Hall.
Methodology and theory for partial least squares applied to functional data.
The Annals of Statistics, 40(1):322–352, 2012.


Aurore Delaigle, Peter Hall, and N Bathia.
Componentwise classification and clustering of functional data.
Biometrika, 99(2):299–313, 2012.

Bibliography IV


Luc Devroye, László Györfi, and Gábor Lugosi.
A Probabilistic Theory of Pattern Recognition, volume 31.
Springer Science & Business Media, 2013.


Ramón Díaz-Uriarte and Sara Alvarez de Andrés.
Gene selection and classification of microarray data using random forest.
BMC Bioinformatics, 7:3, 2006.


C Ding and H Peng.
Minimum redundancy feature selection from microarray gene expression data.
Journal of Bioinformatics and Computational Biology, 3(2):185–205, 2005.


Jianqing Fan and Jinchi Lv.
A selective overview of variable selection in high dimensional feature space.
Statistica Sinica, 20(1):101, 2010.


Manuel Febrero-Bande, Pedro Galeano, and Wenceslao González-Manteiga.
Measures of influence for the functional linear model with scalar response.
Journal of Multivariate Analysis, 101(2):327–339, 2010.


Manuel Febrero-Bande and Manuel Oviedo de la Fuente.
Statistical computing in functional data analysis: the R package fda. usc.
Journal of Statistical Software, 51(4):1–28, 2012.

Bibliography V


Frédéric Ferraty and Philippe Vieu.
Nonparametric Functional Data Analysis: Theory and Practice.
Springer, 2006.


Liliana Forzani, Ricardo Fraiman, and Pamela Llop.
Consistent nonparametric regression for functional data under the stone–besicovitch conditions.
IEEE Transactions on Information Theory, 58(11):6697–6708, 2012.


Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti A Zadeh.
Feature Extraction: Foundations and Applications.
Springer, 2006.


Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik.
Gene selection for cancer classification using support vector machines.
Machine Learning, 46(1-3):389–422, 2002.


Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin.
The elements of statistical learning: data mining, inference and prediction.
The Mathematical Intelligencer, 27(2):83–85, 2009.


Lajos Horváth and Piotr Kokoszka.
Inference for Functional Data with Applications.
Springer, 2012.

Bibliography VI


Lajos Horváth and Ron Reeder.
Detecting changes in functional linear models.
Journal of Multivariate Analysis, 111:310–334, 2012.


T. Kailath.
RKHS approach to detection and estimation problems I: Deterministic signals in Gaussian noise.
IEEE Transactions on Information Theory, 17(5):530–549, 1971.


Jun Li, Juan A Cuesta-Albertos, and Regina Y Liu.
Dd-classifier: Nonparametric classification procedure based on dd-plot.
Journal of the American Statistical Association, 107(498):737–753, 2012.


Martin A. Lindquist and Ian W. McKeague.
Logistic regression with Brownian-like predictors.
Journal of the American Statistical Association, 104(488):1575–1585, 2009.


Robert Liptser and Albert N Shiryaev.
Statistics of Random Processes: I. General Theory.
Springer, 2013.


Huan Liu and Hiroshi Motoda.
Feature Selection for Knowledge Discovery and Data Mining.
Springer, 2012.

Bibliography VII


Peter Mörters and Yuval Peres.
Brownian Motion.
Cambridge University Press, 2010.


Emanuel Parzen.
An Approach to Time Series Analysis.
The Annals of Mathematical Statistics, 32(4):951–989, 1961.


Emanuel Parzen.
Extraction and detection problems and reproducing kernel Hilbert spaces.
Journal of the Society for Industrial & Applied Mathematics, Series A: Control, 1(1):35–62, 1962.


Hanchuan Peng, Fuhui Long, and Chris Ding.
Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8):1226–1238, 2005.


David Pollard.
A central limit theorem for k-means clustering.
The Annals of Probability, pages 919–926, 1982.

Bibliography VIII


David Pollard et al.
Strong consistency of k-means clustering.
The Annals of Statistics, 9(1):135–140, 1981.


Cristian Preda, Gilbert Saporta, and Caroline Lévéder.
PLS classification of functional data.
Computational Statistics, 22(2):223–235, 2007.


J O Ramsay and B W Silverman.
Functional Data Analysis.
Springer, 2005.


Yvan Saeys, Iñaki Inza, and Pedro Larrañaga.
A review of feature selection techniques in bioinformatics.
Bioinformatics, 23(19):2507–2517, 2007.


Carlo Sguera, Pedro Galeano, and Rosa Lillo.
Spatial depth-based classification for functional data.
Test, 23(4):725–750, 2014.


La Shepp.
Radon-Nikodym Derivatives of Gaussian Measures.
37(2):321–354, 1966.

Bibliography IX


Charles J Stone.
Consistent nonparametric regression.
The annals of statistics, pages 595–620, 1977.


Gábor J Székely and Maria L Rizzo.
Brownian distance covariance.
The Annals of Applied Statistics, 3(4):1236–1265, 2009.


Gábor J Székely and Maria L Rizzo.
On the uniqueness of distance covariance.
Statistics & Probability Letters, 82(12):2278–2282, 2012.


Gábor J Székely and Maria L Rizzo.
Energy statistics: A class of statistics based on distances.
Journal of Statistical Planning and Inference, 143(8):1249–1272, 2013.


Gábor J Székely, Maria L Rizzo, and Nail K Bakirov.
Measuring and testing dependence by correlation of distances.
The Annals of Statistics, 35(6):2769–2794, 2007.


Dale Varberg.
On equivalence of Gaussian measures.
Pacific Journal of Mathematics, 11(2):751–762, 1961.


D.E. Varberg.
On Gaussian measures equivalent to Wiener measure.
Transactions of the American Mathematical Society, 113:262–273, 1964.