Curve learning

Gérard Biau

UNIVERSITÉ MONTPELLIER II

Summary

- The problem
- The mathematical model
- Functional classification
  1. Fourier filtering
  2. Wavelet filtering
- Applications

The problem

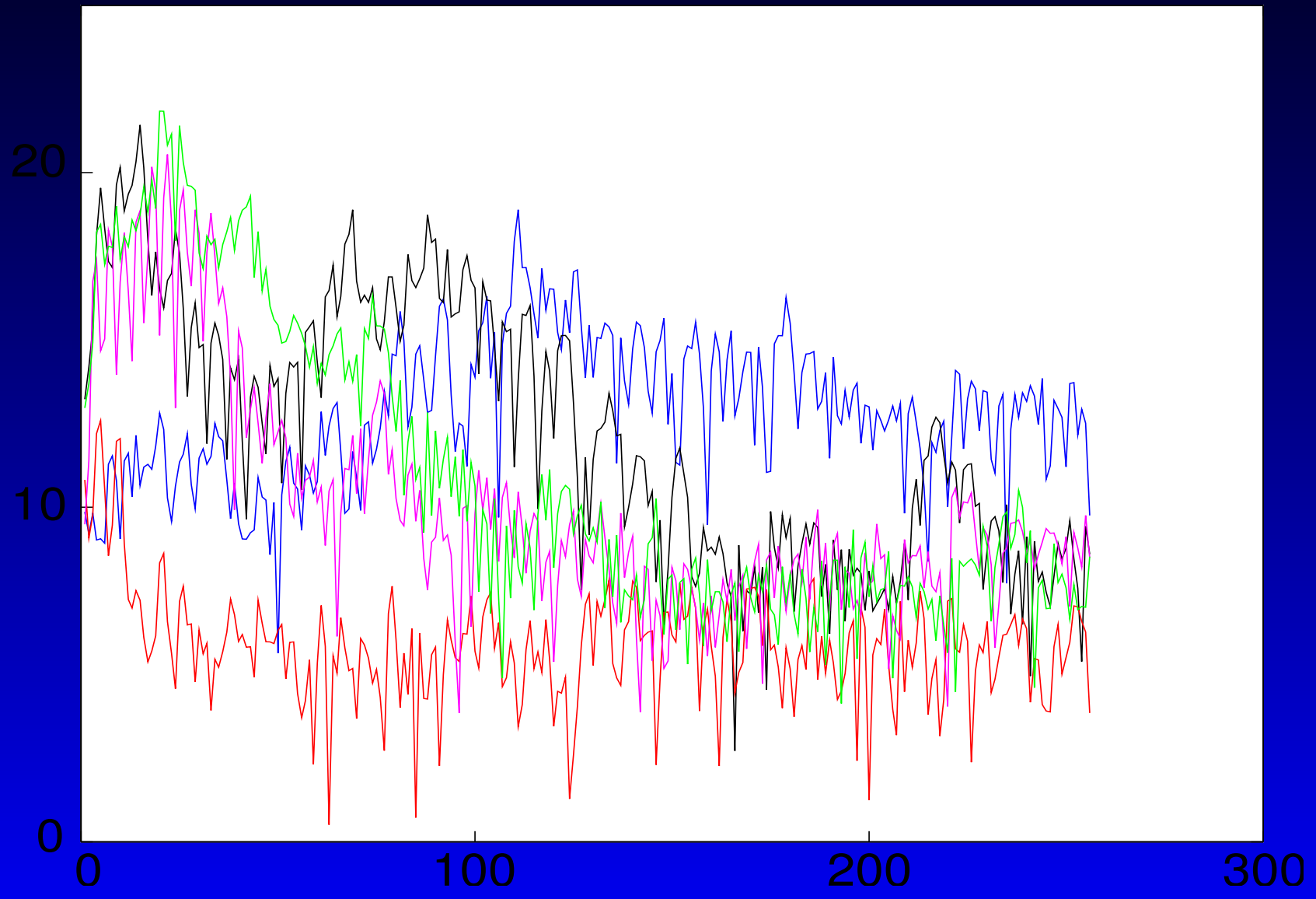
- In many experiments, practitioners collect samples of curves and other functional observations.

The problem

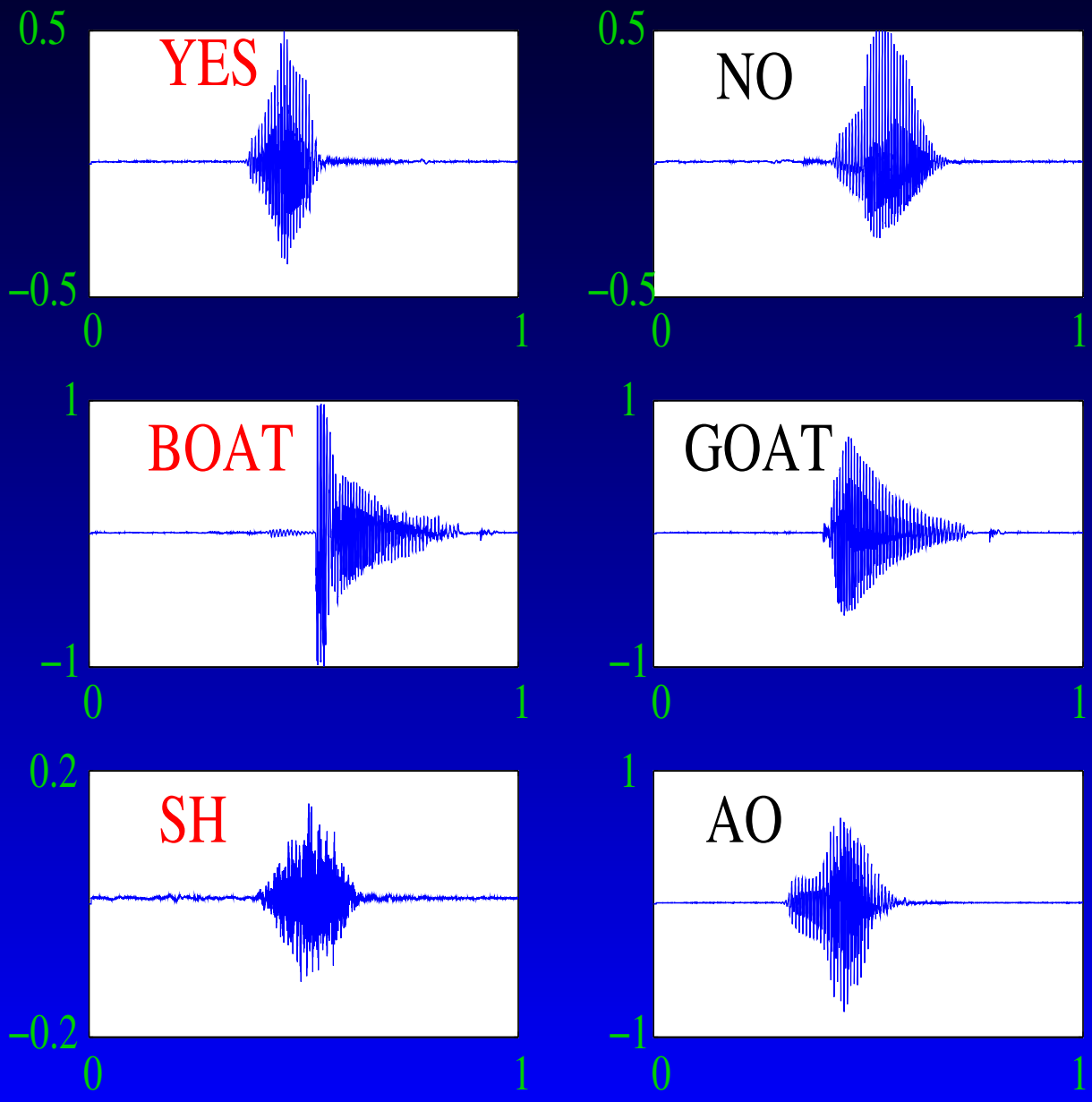
- In many experiments, practitioners collect samples of curves and other functional observations.

- We wish to investigate how classical learning rules can be extended to handle functions.

A typical problem



A speech recognition problem



Classification

- Given a random pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$, the problem of classification is to predict the unknown nature Y of a new observation X.

Classification

- Given a random pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$, the problem of classification is to predict the unknown nature Y of a new observation X.
- The statistician creates a classifier

$$g : \mathbb{R}^d \to \{0, 1\}$$

which represents her guess of the label of X.

Classification

- Given a random pair $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$, the problem of classification is to predict the unknown nature Y of a new observation X.
- The statistician creates a classifier

$$g : \mathbb{R}^d \to \{0, 1\}$$

  which represents her guess of the label of X.
- An error occurs if $g(X) \neq Y$, and the probability of error for a particular classifier g is

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

Classification

- The Bayes rule

$$g^*(x) = \begin{cases} 0 & \text{if } \mathbf{P}\{Y = 0 | X = x\} \geq \mathbf{P}\{Y = 1 | X = x\} \\ 1 & \text{otherwise,} \end{cases}$$

is the optimal decision. That is, for any decision function $g : \mathbb{R}^d \to \{0, 1\}$,

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

Classification

- The Bayes rule

$$g^*(x) = \begin{cases} 0 & \text{if } \mathbf{P}\{Y=0|X=x\} \geq \mathbf{P}\{Y=1|X=x\} \\ 1 & \text{otherwise,} \end{cases}$$

  is the optimal decision. That is, for any decision function $g : \mathbb{R}^d \to \{0,1\}$,

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

- The problem is thus to construct a reasonable classifier g_n based on i.i.d. observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Error criteria

- The goal is to make the error probability

$$\mathbf{P}\{g_n(X) \neq Y\}$$

close to L^*.

Error criteria

- The goal is to make the error probability

$$\mathbf{P}\{g_n(X) \neq Y\}$$

close to L^*.

- Definition. *A decision rule g_n is called consistent if*

$$\mathbf{P}\{g_n(X) \neq Y\} \to L^* \quad \text{as } n \to \infty.$$

Error criteria

- The goal is to make the error probability

$$\mathbf{P}\{g_n(X) \neq Y\}$$

close to L^*.

- Definition. *A decision rule g_n is called consistent if*

$$\mathbf{P}\{g_n(X) \neq Y\} \to L^* \quad \text{as } n \to \infty.$$

- Definition. *A decision rule is called universally consistent if it is consistent for any distribution of the pair (X, Y).*

Two simple and popular rules

- The moving window rule:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \mathbf{1}_{\{Y_i=0, X_i \in B_{x,h_n}\}} \geq \sum_{i=1}^n \mathbf{1}_{\{Y_i=1, X_i \in B_{x,h_n}\}} \\ 1 & \text{otherwise.} \end{cases}$$

Two simple and popular rules

- The moving window rule:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \mathbf{1}_{\{Y_i=0, X_i \in B_{x,h_n}\}} \geq \sum_{i=1}^n \mathbf{1}_{\{Y_i=1, X_i \in B_{x,h_n}\}} \\ 1 & \text{otherwise.} \end{cases}$$

- The nearest neighbor rule:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^{k_n} \mathbf{1}_{\{Y_{(i)}(x)=0\}} \geq \sum_{i=1}^{k_n} \mathbf{1}_{\{Y_{(i)}(x)=1\}} \\ 1 & \text{otherwise.} \end{cases}$$

Two simple and popular rules

- The moving window rule:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \mathbf{1}_{\{Y_i=0, X_i \in B_{x,h_n}\}} \geq \sum_{i=1}^n \mathbf{1}_{\{Y_i=1, X_i \in B_{x,h_n}\}} \\ 1 & \text{otherwise.} \end{cases}$$

- The nearest neighbor rule:

$$g_n(x) = \begin{cases} 0 & \text{if } \sum_{i=1}^{k_n} \mathbf{1}_{\{Y_{(i)}(x)=0\}} \geq \sum_{i=1}^{k_n} \mathbf{1}_{\{Y_{(i)}(x)=1\}} \\ 1 & \text{otherwise.} \end{cases}$$

- *Devroye, Györfi and Lugosi (1996). A Probabilistic Theory of Pattern Recognition.*

Notation

- The model. The random variable X takes values in a metric space $(\mathcal{F}, \rho)$. The distribution of the pair (X, Y) is completely specified by μ and η:

$$\mu(A) = \mathbf{P}\{X \in A\} \quad \text{and} \quad \eta(x) = \mathbf{P}\{Y = 1 | X = x\}.$$

Notation

- The model. The random variable X takes values in a metric space $(\mathcal{F}, \rho)$. The distribution of the pair (X, Y) is completely specified by μ and η:

$$\mu(A) = \mathbf{P}\{X \in A\} \quad \text{and} \quad \eta(x) = \mathbf{P}\{Y = 1 | X = x\}.$$

- And all the definitions remain the same...

Assumptions

- (H1) The space $\mathcal{F}$ is complete.

Assumptions

- (H1) The space $\mathcal{F}$ is complete.
- (H2) The following differentiation result holds:

$$\lim_{h \to 0} \frac{1}{\mu(B_{x,h})} \int_{B_{x,h}} \eta \, \mathrm{d}\mu = \eta(x) \quad \text{in } \mu\text{-probability.}$$

Assumptions

- (H1) The space $\mathcal{F}$ is complete.
- (H2) The following differentiation result holds:

$$\lim_{h \to 0} \frac{1}{\mu(B_{x,h})} \int_{B_{x,h}} \eta \, \mathrm{d}\mu = \eta(x) \quad \text{in } \mu\text{-probability.}$$

- Aïe aïe aïe...

Results

- Theorem [consistency]. *Assume that* (H1) *and* (H2) *hold. If* $h_n \to 0$ *and, for every* $k \geq 1$, $\mathcal{N}_k(h_n/2)/n \to 0$ *as* $n \to \infty$, *then*

$$\lim_{n \to \infty} L(g_n) = L^* .$$

Results

- Theorem [consistency]. *Assume that* (H1) *and* (H2) *hold. If* $h_n \to 0$ *and, for every* $k \geq 1$, $\mathcal{N}_k(h_n/2)/n \to 0$ *as* $n \to \infty$, *then*

$$\lim_{n \to \infty} L(g_n) = L^* \, .$$

- $h_n \simeq 1/\log\log n$: curse of dimensionality.

Curse of dimensionality?

- Consider the hypercube $[-a, a]^d$ and an inscribed hypersphere with radius $r = a$. Then

$$f_d = \frac{\text{volume sphere}}{\text{volume cube}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)}.$$

Curse of dimensionality?

- Consider the hypercube $[-a, a]^d$ and an inscribed hypersphere with radius $r = a$. Then

$$f_d = \frac{\text{volume sphere}}{\text{volume cube}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)}.$$

d	1	2	3	4	5	6	7
f_d	1	0.785	0.524	0.308	0.164	0.081	0.037

Curse of dimensionality?

- Consider the hypercube $[-a, a]^d$ and an inscribed hypersphere with radius $r = a$. Then

$$f_d = \frac{\text{volume sphere}}{\text{volume cube}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)}.$$

d	1	2	3	4	5	6	7
f_d	1	0.785	0.524	0.308	0.164	0.081	0.037

- As the dimension increases, most spherical neighborhoods will be empty!

Classification in Hilbert spaces

- Here, the observations X_i take values in the infinite dimensional space $L_2\big([0,1]\big)$.

Classification in Hilbert spaces

- Here, the observations X_i take values in the infinite dimensional space $L_2\big([0,1]\big)$.
- The trigonometric basis, formed by the functions

$$\psi_1(t) = 1, \quad \psi_{2j}(t) = \sqrt{2}\cos(2\pi jt),$$

and $\quad \psi_{2j+1}(t) = \sqrt{2}\sin(2\pi jt)\,, \quad j = 1, 2, \ldots,$

is a complete orthonormal system in $L_2\big([0,1]\big)$.

Classification in Hilbert spaces

- Here, the observations X_i take values in the infinite dimensional space $L_2\big([0,1]\big)$.
- The trigonometric basis, formed by the functions

$$\psi_1(t) = 1, \quad \psi_{2j}(t) = \sqrt{2}\cos(2\pi jt),$$

and $\quad \psi_{2j+1}(t) = \sqrt{2}\sin(2\pi jt)\,, \quad j = 1, 2, \ldots,$

is a complete orthonormal system in $L_2\big([0,1]\big)$.

- We let the Fourier representation of X_i be

$$X_i(t) = \sum_{j=1}^{\infty} X_{ij}\psi_j(t).$$

Dimension reduction

- Select the effective dimension d to approximate each X_i by

$$X_i(t) \approx \sum_{j=1}^{d} X_{ij}\psi_j(t).$$

Dimension reduction

- Select the effective dimension d to approximate each X_i by

$$X_i(t) \approx \sum_{j=1}^{d} X_{ij}\psi_j(t).$$

- Perform nearest neighbor classification based on the coefficient vector

$$X_i^{(d)} = (X_{i1}, \ldots, X_{id})$$

for the selected dimension d.

Data-splitting

- We split the data into

Data-splitting

- We split the data into
  - a training sequence

$$(X_1, Y_1), \ldots, (X_\ell, Y_\ell)$$

Data-splitting

- We split the data into
  - a training sequence

$$(X_1, Y_1), \ldots, (X_\ell, Y_\ell)$$

  - a testing sequence

$$(X_{\ell+1}, Y_{\ell+1}), \ldots, (X_{m+\ell}, Y_{m+\ell}).$$

Data-splitting

- We split the data into
  - a training sequence
    $$(X_1, Y_1), \ldots, (X_\ell, Y_\ell)$$
  - a testing sequence
    $$(X_{\ell+1}, Y_{\ell+1}), \ldots, (X_{m+\ell}, Y_{m+\ell}).$$
- The training sequence defines a class of nearest neighbor classifiers with members $g_{\ell,k,d}$.

Empirical risk minimization

- The testing sequence is used to select a particular classifier by minimizing the penalized empirical risk

$$\frac{1}{m}\sum_{i\in\mathcal{J}_m} 1_{\left\{g_{\ell,k,d}(X_i^{(d)})\neq Y_i\right\}} + \frac{\lambda_d}{\sqrt{m}}.$$

Empirical risk minimization

- The testing sequence is used to select a particular classifier by minimizing the penalized empirical risk

$$\frac{1}{m}\sum_{i\in\mathcal{J}_m} 1_{\left\{g_{\ell,k,d}(X_i^{(d)})\neq Y_i\right\}} + \frac{\lambda_d}{\sqrt{m}}.$$

- In practice, the regularization penalty may be formally replaced by

$$\lambda_d = 0 \quad \text{for } d \leq d_n \quad \text{and } \lambda_d = +\infty \quad \text{for } d \geq d_n+1.$$

Result

- Theorem [oracle inequality]. *Assume that*

$$\sum_{d=1}^{\infty} e^{-2\lambda_d^2} < \infty. \qquad (1)$$

Then there exists a constant $C > 0$ *such that*

$$L(\hat{g}_n) - L^*$$

$$\leq \inf_{d \geq 1} \left[(L_d^* - L^*) + \inf_{1 \leq k \leq \ell} \left(L(g_{\ell,k,d}) - L_d^* \right) + \frac{\lambda_d}{\sqrt{m}} \right]$$

$$+ C \sqrt{\frac{\log \ell}{m}}.$$

Result

- Theorem [oracle inequality]. *Assume that*

$$\sum_{d=1}^{\infty} e^{-2\lambda_d^2} < \infty. \qquad (1)$$

Then there exists a constant $C > 0$ *such that*

$$L(\hat{g}_n) - L^*$$

$$\leq \inf_{d \geq 1} \left[(L_d^* - L^*) + \inf_{1 \leq k \leq \ell} \left(L(g_{\ell,k,d}) - L_d^* \right) + \frac{\lambda_d}{\sqrt{m}} \right]$$

$$+ C \sqrt{\frac{\log \ell}{m}}.$$

Result

- Theorem [oracle inequality]. *Assume that*

$$\sum_{d=1}^{\infty} e^{-2\lambda_d^2} < \infty. \qquad (1)$$

Then there exists a constant $C > 0$ *such that*

$$L(\hat{g}_n) - L^*$$

$$\leq \inf_{d \geq 1} \left[(L_d^* - L^*) + \inf_{1 \leq k \leq \ell} \left(L(g_{\ell,k,d}) - L_d^* \right) + \frac{\lambda_d}{\sqrt{m}} \right]$$

$$+ C \sqrt{\frac{\log \ell}{m}}.$$

Result

- Theorem [oracle inequality]. *Assume that*

$$\sum_{d=1}^{\infty} e^{-2\lambda_d^2} < \infty. \qquad (1)$$

Then there exists a constant $C > 0$ *such that*

$$L(\hat{g}_n) - L^*$$

$$\leq \inf_{d \geq 1} \left[(L_d^* - L^*) + \inf_{1 \leq k \leq \ell} \left(L(g_{\ell,k,d}) - L_d^* \right) + \frac{\lambda_d}{\sqrt{m}} \right]$$

$$+ C \sqrt{\frac{\log \ell}{m}}.$$

Results

- Theorem [consistency]. *Under the assumption (1) and*

$$\lim_{n\to\infty} \ell = \infty, \ \lim_{n\to\infty} m = \infty, \ \lim_{n\to\infty} \frac{\log \ell}{m} = 0,$$

the classifier $\hat{g}_n$ *is universally consistent.*

Results

- Theorem [consistency]. *Under the assumption (1) and*

$$\lim_{n\to\infty} \ell = \infty, \quad \lim_{n\to\infty} m = \infty, \quad \lim_{n\to\infty} \frac{\log \ell}{m} = 0,$$

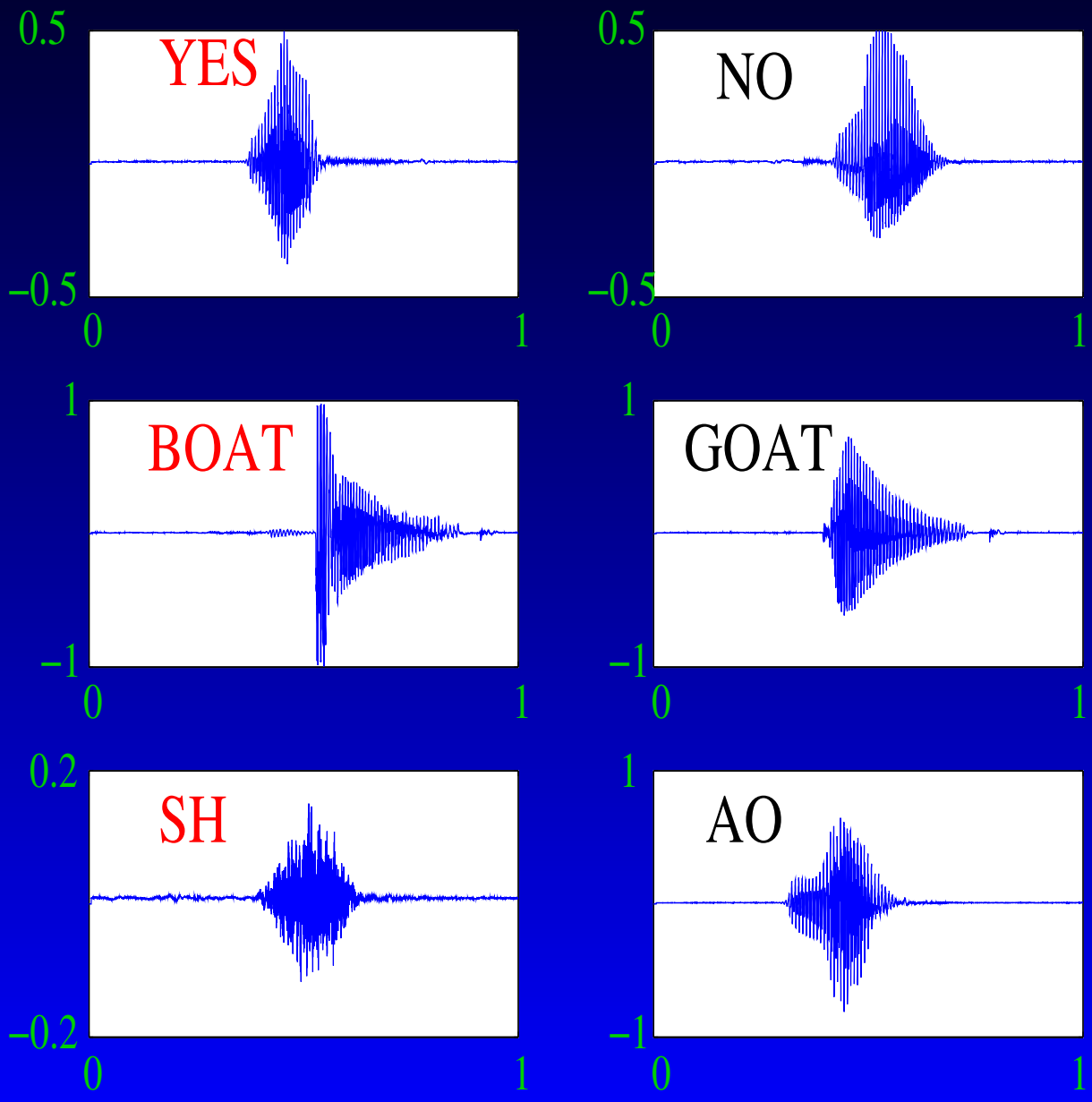
the classifier $\hat{g}_n$ *is universally consistent.*

- Under stronger assumptions, the classifier $\hat{g}_n$ is strongly consistent, that is

$$\mathbf{P}\{\hat{g}_n(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n)\} \to L^*$$

with probability one.

Illustration



Results

Method	Error rates
FOURIER	0.10 0.21 0.16
NN-DIRECT	0.36 0.42 0.42
QDA	0.07 0.35 0.19

Multiresolution analysis

- $V_0 \subset V_1 \subset V_2 \subset \ldots \subset L_2\big([0,1]\big)$.

Multiresolution analysis

- $V_0 \subset V_1 \subset V_2 \subset \ldots \subset L_2\big([0,1]\big)$.
- ϕ: the father wavelet such that $\{\phi_{j,k} : k = 0, \ldots, 2^j - 1\}$ is an orthornormal basis of V_j.

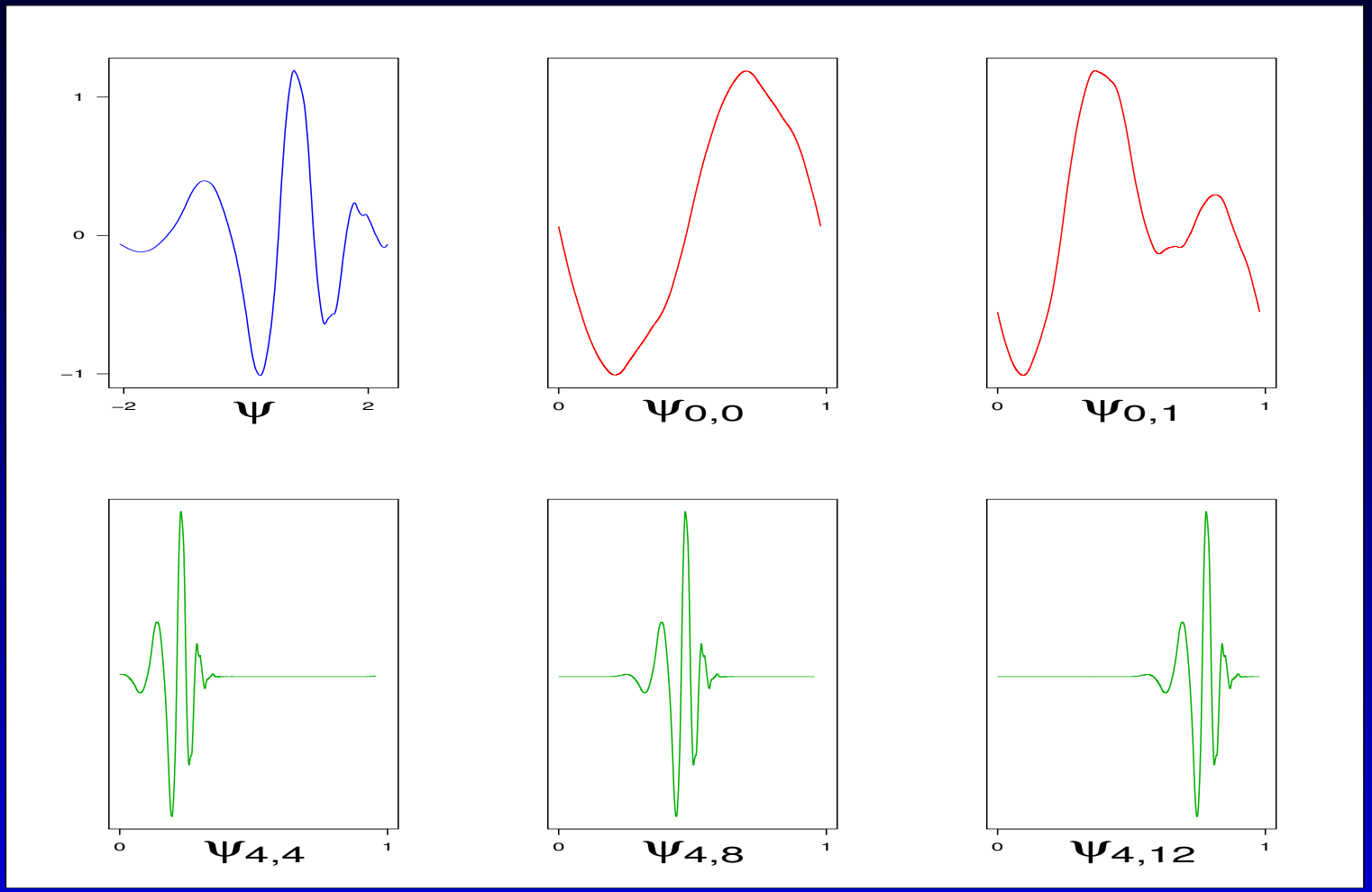
Multiresolution analysis

- $V_0 \subset V_1 \subset V_2 \subset \ldots \subset L_2\big([0,1]\big)$.
- ϕ: the father wavelet such that $\{\phi_{j,k} : k = 0, \ldots, 2^j - 1\}$ is an orthornormal basis of V_j.
- For each j, one constructs an orthornormal basis $\{\psi_{j,k} : k = 0, \ldots, 2^j - 1\}$ of W_j such that $V_{j+1} = V_j \bigoplus W_j$. ψ is the mother wavelet.

Multiresolution analysis

- $V_0 \subset V_1 \subset V_2 \subset \ldots \subset L_2\big([0,1]\big)$.
- ϕ: the father wavelet such that $\{\phi_{j,k} : k = 0, \ldots, 2^j - 1\}$ is an orthornormal basis of V_j.
- For each j, one constructs an orthornormal basis $\{\psi_{j,k} : k = 0, \ldots, 2^j - 1\}$ of W_j such that $V_{j+1} = V_j \bigoplus W_j$. ψ is the mother wavelet.
- Any function X_i of $L_2\big([0,1]\big)$ reads

$$X_i(t) = \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} \xi^i_{j,k} \psi_{j,k}(t) + \eta^i \phi_{0,0}(t) \, .$$


1
0
−1
−2
Ψ
2
0
Ψ0,0
1
0
Ψ0,1
1
0
Ψ4,4
1
0
Ψ4,8
1
0
Ψ4,12
1

- Fix a maximum resolution level J and expand each observation as

$$X_i(t) \approx \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{j,k}^i \psi_{j,k}(t) + \eta^i \phi_{0,0}(t) \, .$$

- Fix a maximum resolution level J and expand each observation as

$$X_i(t) \approx \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{j,k}^i \psi_{j,k}(t) + \eta^i \phi_{0,0}(t) \, .$$

- Rewrite each X_i as a linear combination of 2^J basis functions ψ_j :

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij} \psi_j(t) \, .$$

- Fix a maximum resolution level J and expand each observation as

$$X_i(t) \approx \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{j,k}^i \psi_{j,k}(t) + \eta^i \phi_{0,0}(t) \, .$$

- Rewrite each X_i as a linear combination of 2^J basis functions ψ_j :

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij} \psi_j(t) \, .$$

- How to select the most pertinent basis functions?

Thresholding

- Reorder the first 2^J basis functions $\{\psi_1, \ldots, \psi_{2^J}\}$ into $\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\}$ via the scheme

$$\sum_{i=1}^{\ell} {X_{ij_1}}^2 \geq \sum_{i=1}^{\ell} {X_{ij_2}}^2 \geq \ldots \geq \sum_{i=1}^{\ell} {X_{ij_{2^J}}}^2 .$$

Thresholding

- Reorder the first 2^J basis functions $\{\psi_1, \ldots, \psi_{2^J}\}$ into $\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\}$ via the scheme

$$\sum_{i=1}^{\ell} {X_{ij_1}}^2 \geq \sum_{i=1}^{\ell} {X_{ij_2}}^2 \geq \ldots \geq \sum_{i=1}^{\ell} {X_{ij_{2^J}}}^2 .$$

- Wavelet coefficients are ranked using a global thresholding on the mean of the square coefficients.

Classification

- For each $d = 1, \ldots, 2^J$, let $D_\ell^{(d)}$ be a collection of rules $g^{(d)} : \mathbb{R}^d \to \{0, 1\}$.

Classification

- For each $d = 1, \ldots, 2^J$, let $D_\ell^{(d)}$ be a collection of rules $g^{(d)} : \mathbb{R}^d \to \{0, 1\}$.

- Let $\mathbb{S}_{C_\ell^{(d)}}(m)$ denote the corresponding shatter coefficients, and let $\mathbb{S}_{C_\ell}^{(J)}(m)$ be the shatter coefficient of all rules $\{g^{(d)} : d = 1, \ldots, 2^J\}$.

Selection step

- We select both d and $g^{(d)}$ optimally by minimizing the empirical probability of error:

$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) = \arg \min_{1 \leq d \leq 2^J,\, g \in D_\ell^{(d)}} \left[\frac{1}{m} \sum_{i \in \mathcal{J}_m} 1_{[g^{(d)}(X_i^{(d)}) \neq Y_i]} \right].$$

Selection step

- We select both d and $g^{(d)}$ optimally by minimizing the empirical probability of error:

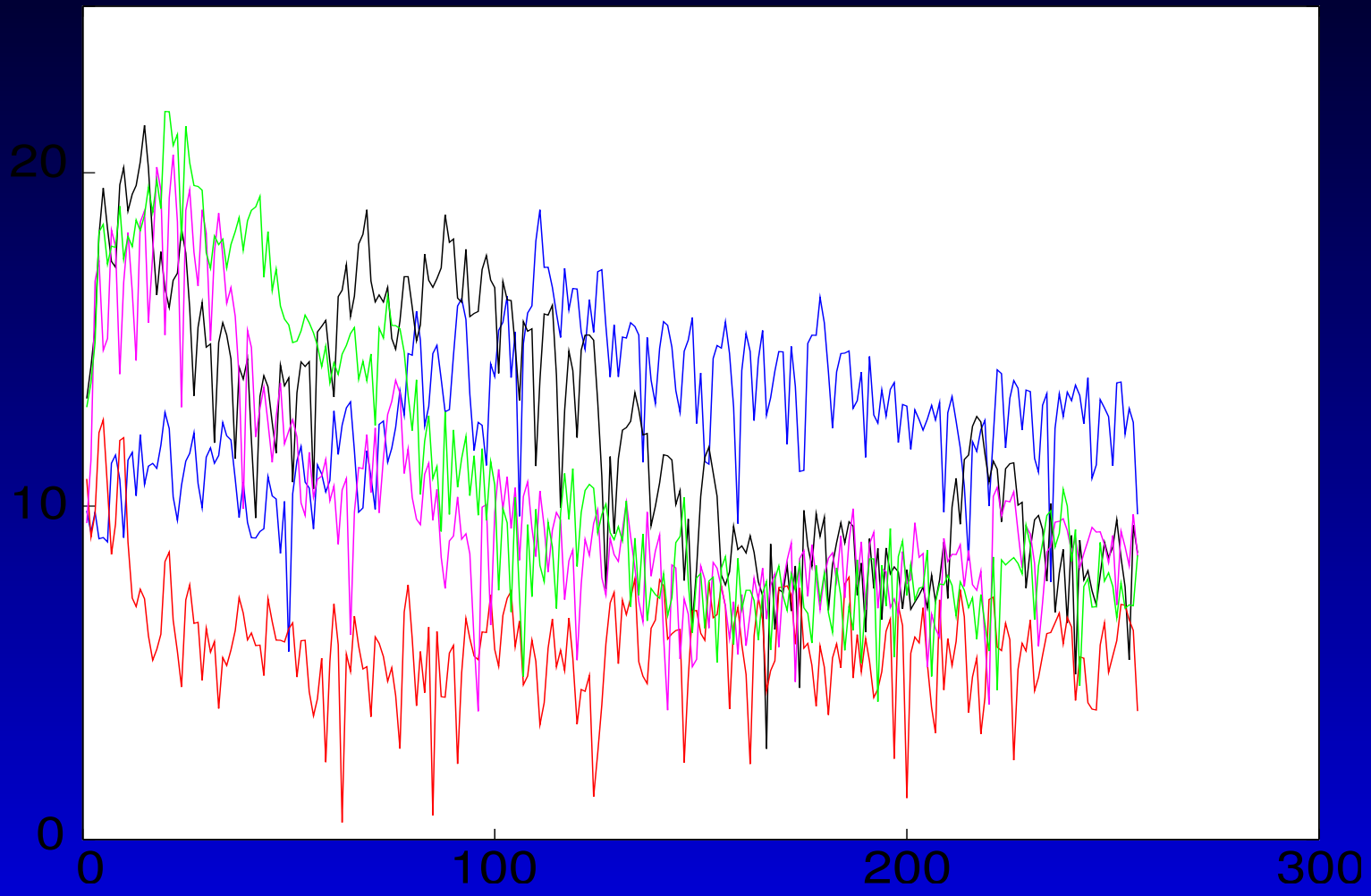
$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) = \arg \min_{1 \leq d \leq 2^J,\, g \in D_\ell^{(d)}} \left[\frac{1}{m} \sum_{i \in \mathcal{J}_m} \mathbf{1}_{[g^{(d)}(X_i^{(d)}) \neq Y_i]} \right].$$

- Theorem.

$$L(\hat{g}) - L^* \leq L_{2^J}^* - L^* + \mathbf{E}\left\{ \inf_{1 \leq d \leq 2^J,\, g^{(d)} \in D_\ell^{(d)}} L_\ell(g^{(d)}) \right\} - L_{2^J}^*$$

$$+ C\mathbf{E}\left\{ \sqrt{\frac{\log\left(\mathbb{S}_{C_\ell}^{(J)}(m)\right)}{m}} \right\}.$$

Illustration



"sh", "iy", "dcl", "ao", "aa", $\ell = 250, m = 250$.

Methods

- W-QDA: $D_\ell^{(d)}$ contains quadratic discriminant rules.
- W-NN: $D_\ell^{(d)}$ contains nearest-neighbor rules.
- W-T: $D_\ell^{(d)}$ contains binary trees.

Methods

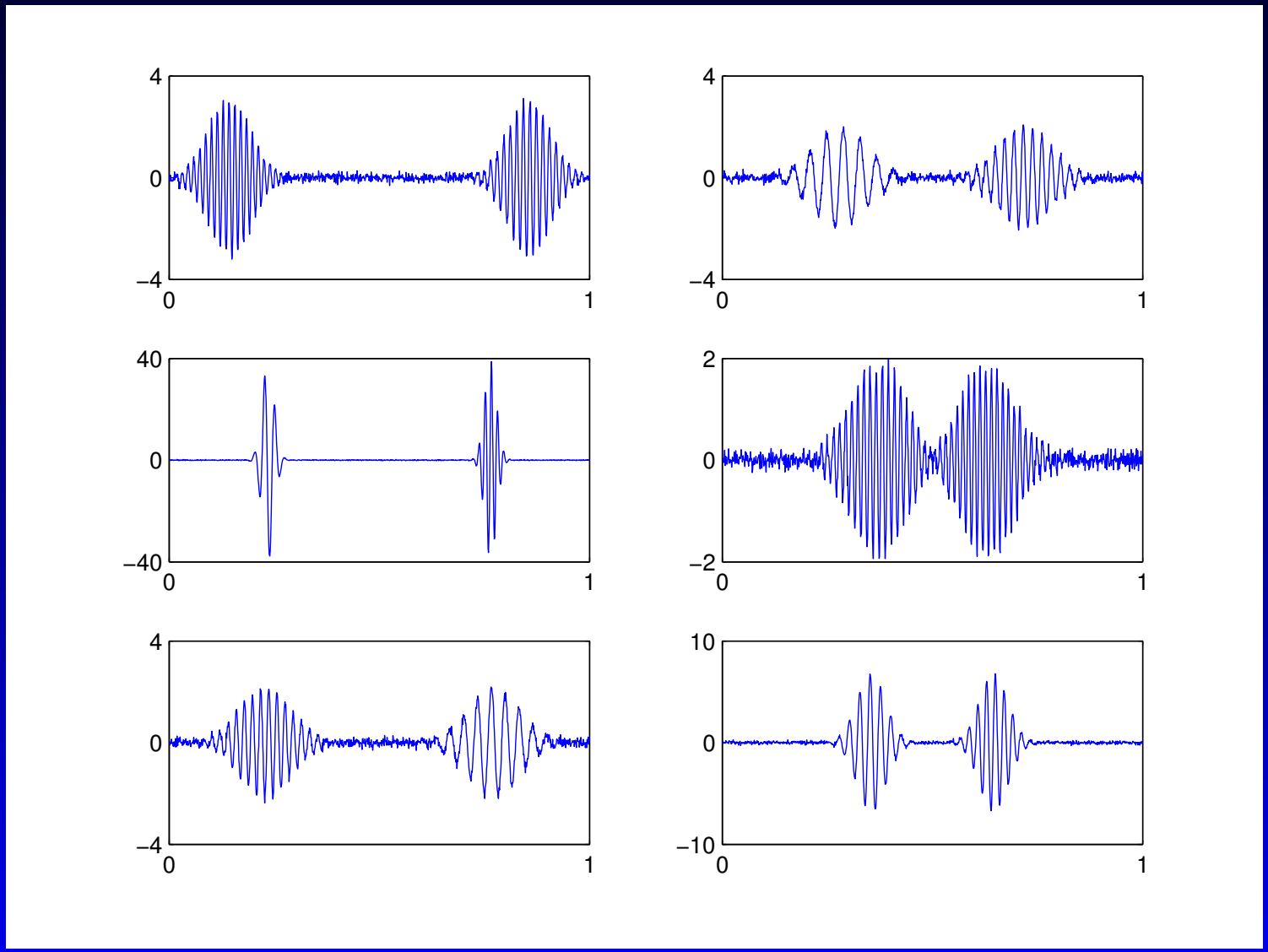
- W-QDA: $D_\ell^{(d)}$ contains quadratic discriminant rules.
- W-NN: $D_\ell^{(d)}$ contains nearest-neighbor rules.
- W-T: $D_\ell^{(d)}$ contains binary trees.

- F-NN: Fourier filtering approach.
- MPLSR: Multivariate Partial Least Square regression.

Results

Method	*Error rates*	$\hat{d}$
W-QDA	0.1042	7.30
W-NN	0.1096	19.52
W-T	0.1253	9.10
F-NN	0.1277	48.76
MPLSR	0.0904	5.96

A simulated example



A simulated example

- Pairs $(X_i(t), Y_i)$ are generated by:

$$X_i(t) = \frac{1}{50}\Big(\sin(F_i^1 \pi t) f_{\mu_i,\sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i,\sigma_i}(t)\Big) + \varepsilon_i$$

A simulated example

- Pairs $(X_i(t), Y_i)$ are generated by:

$$X_i(t) = \frac{1}{50}\left(\sin(F_i^1 \pi t) f_{\mu_i,\sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i,\sigma_i}(t)\right) + \varepsilon_i$$

  - $f_{\mu,\sigma} = \mathcal{N}(\mu, \sigma^2)$ where $\mu \sim \mathcal{U}(0.1, 0.4)$ and $\sigma \sim \mathcal{U}(0, 0.005)$

A simulated example

- Pairs $(X_i(t), Y_i)$ are generated by:

$$X_i(t) = \frac{1}{50}\left(\sin(F_i^1 \pi t) f_{\mu_i,\sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i,\sigma_i}(t)\right) + \varepsilon_i$$

  - $f_{\mu,\sigma} = \mathcal{N}(\mu, \sigma^2)$ where $\mu \sim \mathcal{U}(0.1, 0.4)$ and $\sigma \sim \mathcal{U}(0, 0.005)$
  - F_i^1 and $F_i^2 \sim \mathcal{U}(50, 150)$

A simulated example

- Pairs $(X_i(t), Y_i)$ are generated by:

$$X_i(t) = \frac{1}{50}\Big(\sin(F_i^1 \pi t) f_{\mu_i,\sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i,\sigma_i}(t)\Big) + \varepsilon_i$$

  - $f_{\mu,\sigma} = \mathcal{N}(\mu, \sigma^2)$ where $\mu \sim \mathcal{U}(0.1, 0.4)$ and $\sigma \sim \mathcal{U}(0, 0.005)$
  - F_i^1 and $F_i^2 \sim \mathcal{U}(50, 150)$
  - $\varepsilon_i \sim \mathcal{N}(0, 0.5)$.

A simulated example

- Pairs $(X_i(t), Y_i)$ are generated by:

$$X_i(t) = \frac{1}{50}\left(\sin(F_i^1 \pi t) f_{\mu_i,\sigma_i}(t) + \sin(F_i^2 \pi t) f_{1-\mu_i,\sigma_i}(t)\right) + \varepsilon_i$$

  - $f_{\mu,\sigma} = \mathcal{N}(\mu, \sigma^2)$ where $\mu \sim \mathcal{U}(0.1, 0.4)$ and $\sigma \sim \mathcal{U}(0, 0.005)$
  - F_i^1 and $F_i^2 \sim \mathcal{U}(50, 150)$
  - $\varepsilon_i \sim \mathcal{N}(0, 0.5)$.
- For $i = 1, \ldots, n$

$$Y_i = \begin{cases} 0 & \text{if } \mu_{i,1} \leq 0.25 \\ 1 & \text{otherwise.} \end{cases}$$

Results

Method	*Error rates*	$\hat{d}$
W-QDA	0.0843	7.36
W-NN	0.1255	23.92
W-T	0.1275	20.48
F-NN	0.3493	75.88
MPLSR	0.4413	3.68

Results

Method	*Error rates*	$\hat{d}$
W-QDA	0.0843	7.36
W-NN	0.1255	23.92
W-T	0.1275	20.48
F-NN	0.3493	75.88
MPLSR	0.4413	3.68

Sampling

- In practice, the curves are always observed at discrete time points, say t_p, $p = 1, \ldots, P$.

Sampling

- In practice, the curves are always observed at discrete time points, say t_p, $p = 1, \ldots, P$.

- There is only an estimation of the wavelet coefficients available:

$$X_i(t) \approx \sum_{j=1}^{2^J} \hat{X}_{ij} \psi_j(t),$$

where

$$\hat{X}_{ij} = \frac{1}{P} \sum_{p=1}^{P} X_i(t_p) \psi_j(t_p).$$

Sampling consistency

- Theorem. *Let $D_\ell^{(d)}$ be the family of all k-nearest neighbor rules, and suppose that*

$$\lim_{n\to\infty} \ell = \infty, \quad \lim_{n\to\infty} m = \infty, \quad \lim_{n\to\infty} \frac{\log \ell}{m} = 0.$$

Then the selected rule $\hat{g}$ is universally consistent, that is

$$\lim_{J\to\infty} \lim_{n\to\infty} \lim_{P\to\infty} L(\hat{g}) = L^*.$$