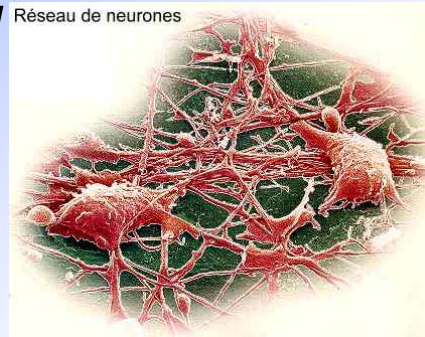


Réseaux de neurones à entrées fonctionnelles

Nathalie Villa (GRIMM - SMASH)

Université Toulouse Le Mirail

Réseau de neurones



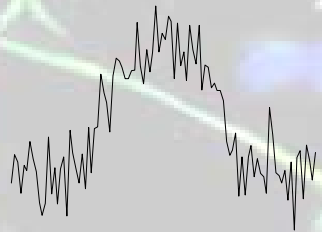
BUT DU TRAVAIL

Ou comment utiliser des réseaux de neurones en statistique fonctionnelle

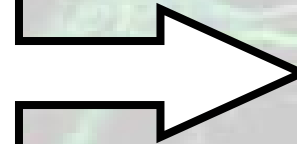
???

Réseaux fonctionnels : Mode d'emploi

But :  Régression

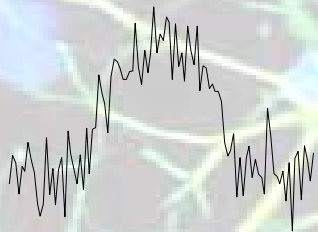


Perceptron
multi-couches

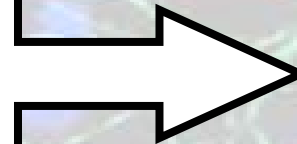


Y réel

 Discrimination



Perceptron
multi-couches



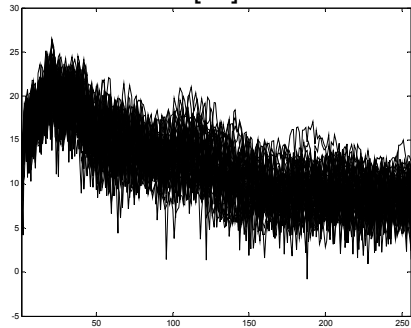
$Y =$

$$\begin{pmatrix} 1 \\ c_1 \\ \vdots \\ 1 \\ c_K \end{pmatrix}$$

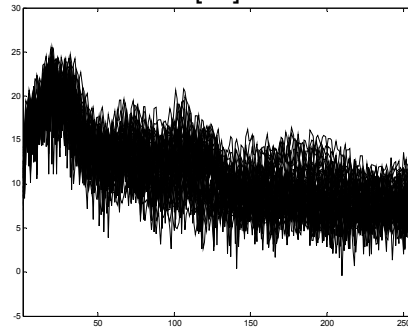
Exemples

1) Données de phonèmes (discrimination)

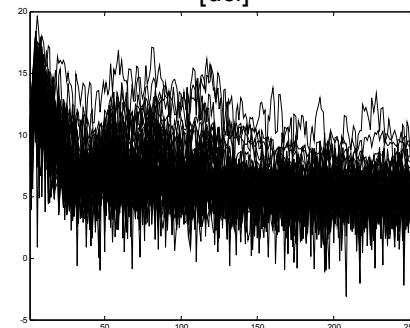
[aa]



[ao]

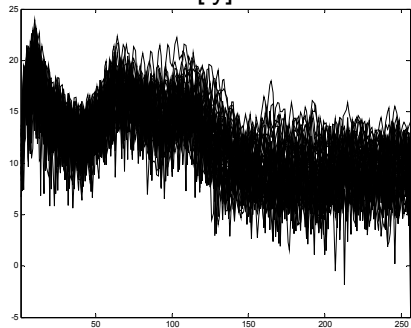


[dcl]

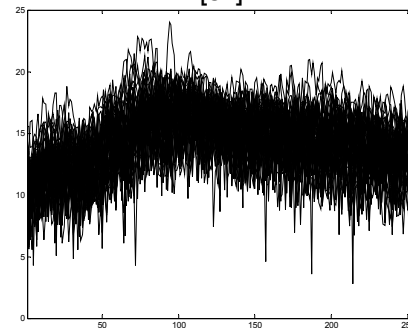


Enregistrements de voix

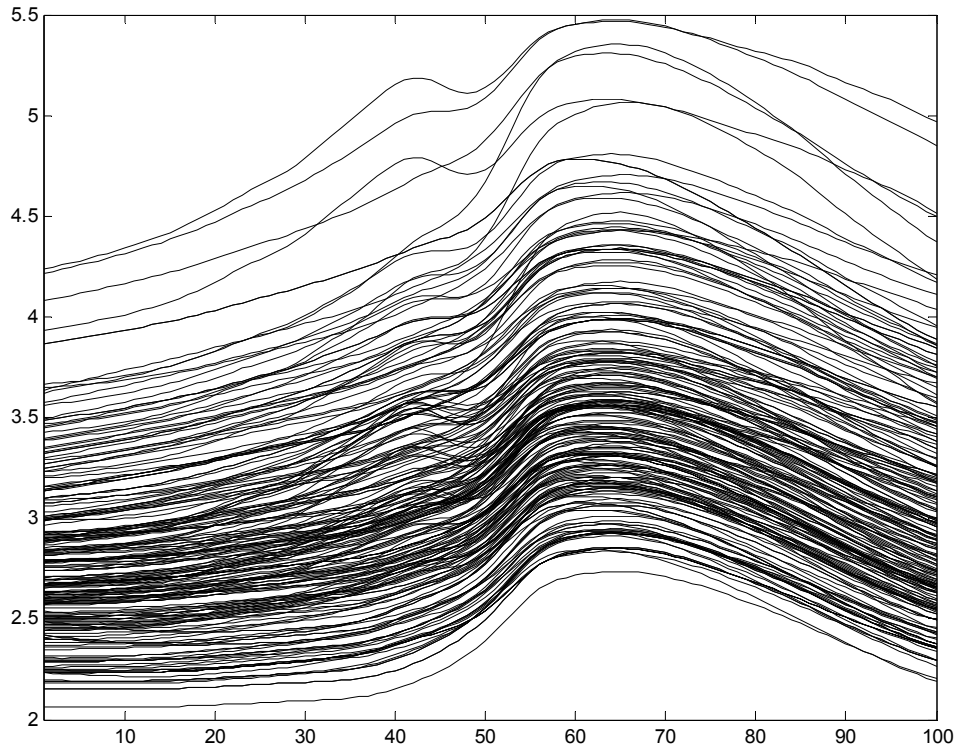
[iy]



[sh]



2) Données de spectrométrie (régression)



Spectres d'absorbance



Masse de
matière
grasse

Le programme...

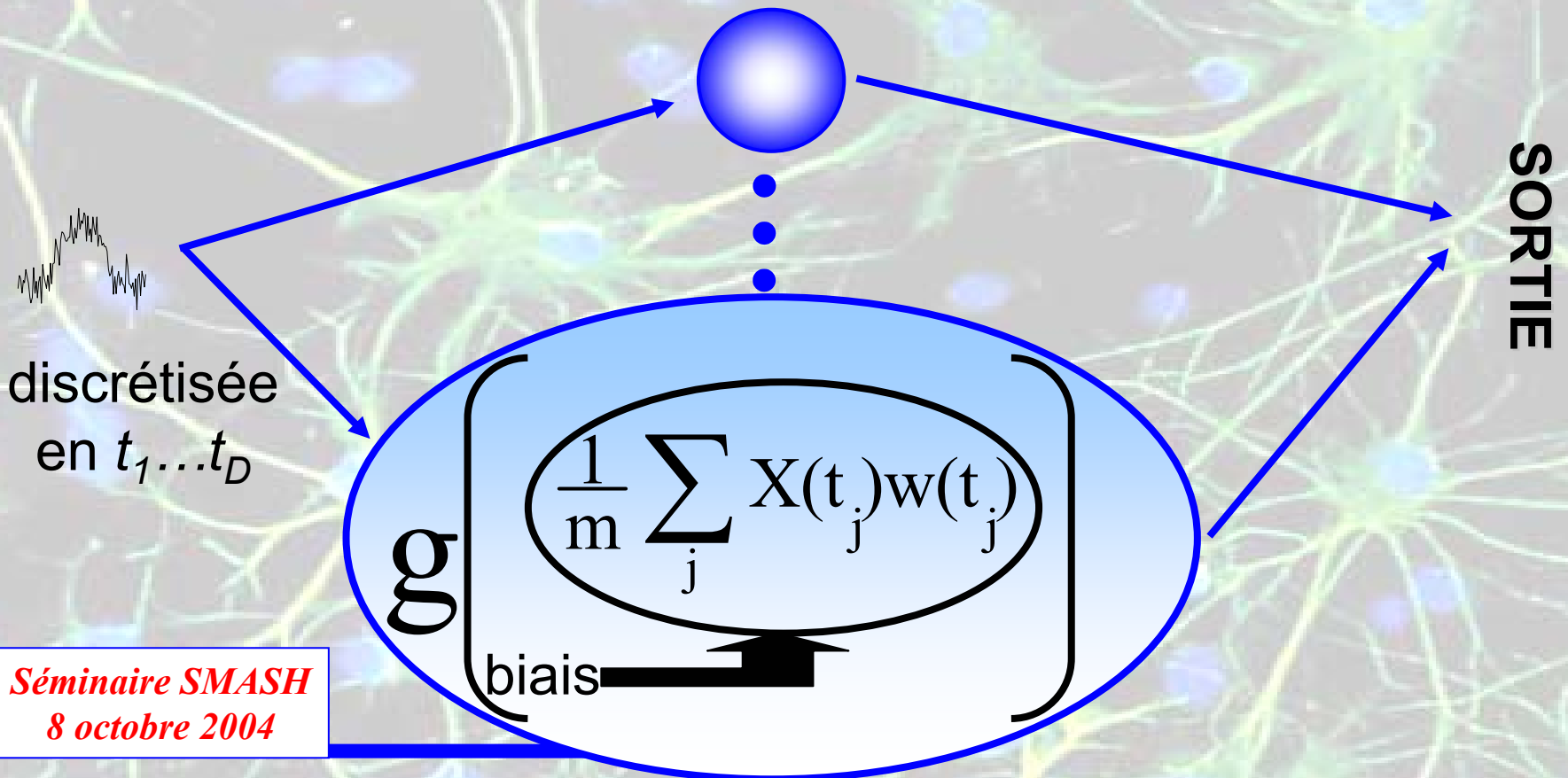
- *Etat des lieux en réseaux de neurones fonctionnels*
- *SIR*
- *SIR-NN*

ETAT DES LIEUX EN RESEAUX DE NEURONES FONCTIONNELS

Rossi, Conan-Guez (2002)

1) Approche directe

$$X \quad w \quad \boxed{\approx \langle X, w \rangle + b} \quad \rightarrow \quad \boxed{\text{Sigmoid}} \quad a \quad Y$$



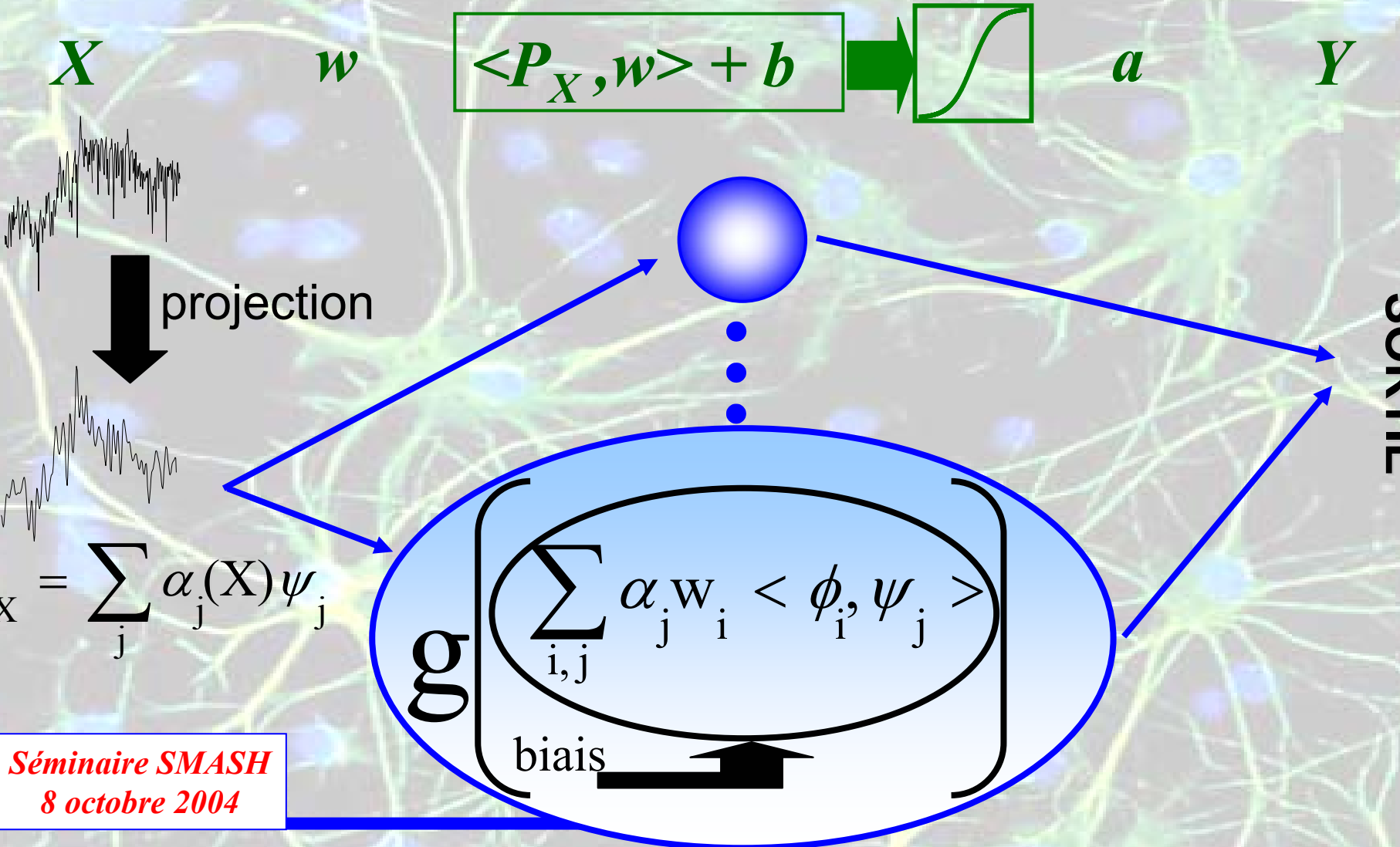
Représentation des fonctions de poids w :

- *Représentation linéaire* des poids par rapport à une base de B-Splines, d'ondelettes, de fonctions trigonométriques... :

$$w(t) = \sum_i w_i \phi_i(t)$$

- *Représentation non linéaire* des poids par un perceptron multicouche multidimensionnel.

2) Approche par projection



Résultats

- *Approximation universelle* : il existe un perceptron fonctionnel qui approche avec la précision voulue n'importe quelle application allant d'un compact de l'espace L^2 dans \mathbf{R} .
- *Consistance* : les paramètres (w) et (a) qui minimisent l'erreur empirique construite à partir d'un nombre fini d'observations discrétisées en un nombre fini de points convergent p_s vers les paramètres optimaux théoriques lorsque le nombre d'observations et le nombre de points de discrétisation tendent vers l'infini.

Limites

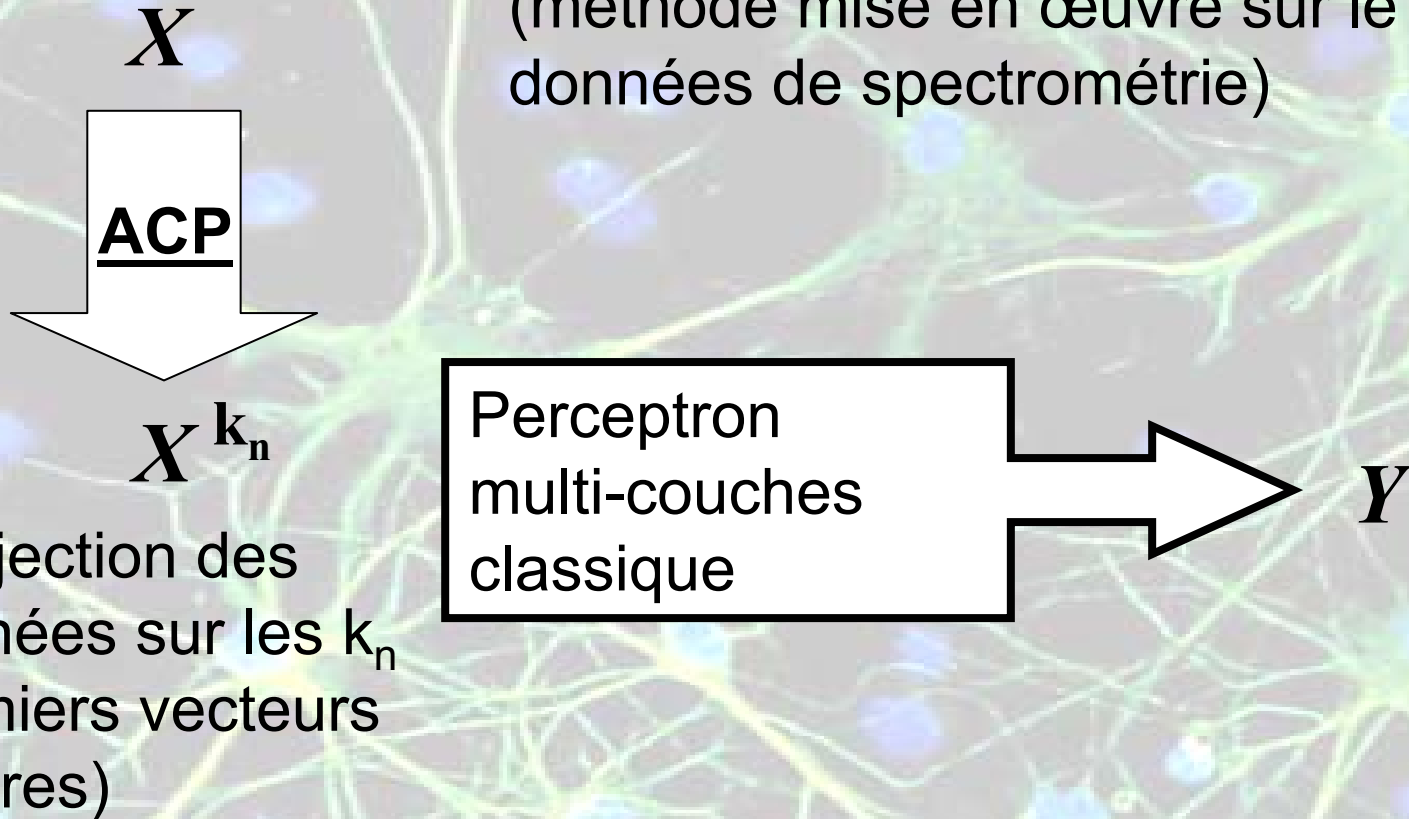
- *Approche directe* : la détermination des minima locaux peut devenir lourde lorsque le nombre de points de discrétisation augmente, particulièrement si la fonction de poids est représentée par un perceptron multicouche.
- *Approche par projection* : trouver une base de projection adaptée au problème ; le choix du type de la base ainsi que du nombre de fonctions à y introduire n'est pas évident à priori.

Risque de perte d'informations pertinentes.

Thodberg (1996)

Base de projection qui dépend des données

(méthode mise en œuvre sur le jeu de données de spectrométrie)



Avantages

- Le jeu de données est simplifié ;
- La base de projection dépend des données (procédure automatique de détermination).

Inconvénients

- La base de projection ne dépend pas de la cible mais uniquement des variables explicatives (base de projection non optimisée).

Risque de perte d'informations pertinentes.

- Pas de résultat de convergence démontré (méthode empirique).

SIR

*Déterminer une base de projection
pertinente*

Sliced Inverse Regression : Le modèle

Li (1991)

Pour X multidimensionnel

$$Y = f(a'_1 X, \dots, a'_q X, \varepsilon)$$

- ε centrée et indépendante de X
- f inconnue
- $(a_j)_j$ linéairement indépendants

Idée : Estimer par des méthodes d'algèbre linéaire l'espace EDR ($\text{Vect}\{a_j\}$) : **SIR** ;

Estimer la fonction f (méthodes non paramétriques, réseaux de neurones...).

SIR Fonctionnelle (FIR)

Ferré, Yao (2003)

Dauxois, Ferré, Yao (2003)

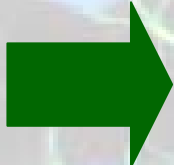
Pour X fonctionnel, $Y = f(\langle a_1, X \rangle, \dots, \langle a_q, X \rangle, \varepsilon)$

Théorème : (Condition de Li)

Notons $A = (\langle a_1, X \rangle, \dots, \langle a_q, X \rangle)^T$; si

$$\forall u \in L^2, \exists v \in \mathbf{R}^q : \mathbf{E}(\langle u, X \rangle / A) = v^T A$$

alors, $\mathbf{E}(X / Y)$ appartient à $\text{Vect} \{ \Gamma_X a_j \}$ où $\Gamma_X = \mathbf{E}(X \otimes X)$.

 L'espace EDR s'obtient par décomposition spectrale de l'opérateur $\Gamma_X^{-1} \Gamma_{\mathbf{E}(X/Y)}$.

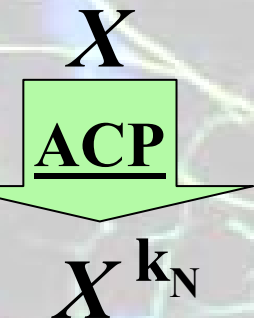
Problème

Γ_X n'est pas un opérateur borné !

➔ Γ_X^N est mal conditionné ;

➔ Les vecteurs propres de $(\Gamma_X^N)^{-1} \Gamma_{E(X/Y)}^N$ ne convergent pas vers les vecteurs propres de $\Gamma_X^{-1} \Gamma_{E(X/Y)}$.

Solution



Sous certaines hypothèses techniques, les vecteurs propres de $(\Gamma_X^{kN})^{-1} \Gamma_{E(X/Y)}^N$ convergent vers (a_j) .

FIR régularisée

D'après les travaux de

Tihonov (1963) ; Leurgans, Moyeed et Silverman (1993)

Idée : On part de l'hypothèse que X fait partie d'un ensemble de fonctions « lisses » (S) et on contraint les vecteurs propres à appartenir également à cet espace en pénalisant Γ_X par une fonctionnelle de régularisation.

Concrètement

On estime $\forall f, g \in S$,

$$\langle \Gamma_X f, g \rangle \text{ par } Q_\alpha^N(f, g) = \langle \Gamma_X^N f, g \rangle + \alpha [f, g]$$

$$\text{où } [f, g] = \int_\tau D^2 f(t) D^2 g(t) dt$$

Théorème : (Consistance)

Sous l'hypothèse de Li et des hypothèses techniques,

$$\frac{\langle \Gamma_{E(X/Y)}^N(a, a) \rangle}{Q_\alpha^N(a, a)}$$

atteint son maximum sur S avec une probabilité qui tend vers 1 lorsque N tend vers $+\infty$.

De plus, si a_1^N est le maximum de cette fonction sur S alors

$$\langle \Gamma_X^N(a_1^N - a_1^N), a_1^N - a_1^N \rangle \xrightarrow{P, N \rightarrow +\infty} 0$$

Remarques

- *Condition de Li* : Li démontre que cette condition est peu restrictive pour des vecteurs X de grande dimension ;
- *Pénalisation* : L'hypothèse de régularité sur X est faite au travers du choix de $[\cdot, \cdot]$: d'autres choix conduiraient au même résultat de consistance ;
- *Estimation de $\Gamma_{E(X/Y)}$* : L'estimateur de $\Gamma_{E(X/Y)}$ doit converger à une vitesse \sqrt{N} . Plusieurs choix sont possibles suivant les buts poursuivis...

Estimation de $\Gamma_{E(X/Y)}$

But :  Régression

➤ Estimateur par tranchage du support : pour une partition $(I_h)_h$ du support de Y ,

$$\Gamma_{E(X/Y)}^N = \sum_h \frac{N_h}{N} \mu_h \otimes \mu_h$$

où $N_h = \sum_n I_{\left\{ Y^n \in I_h \right\}}$ et $\mu_h = \frac{1}{N_h} \sum_n X^n I_{\left\{ Y^n \in I_h \right\}}$

➤ Estimateur à noyau :

$$\hat{\Gamma}_{E(X/Y)}^N = \frac{1}{N} \sum_n E(X / \hat{Y} = Y^n) \otimes E(X / \hat{Y} = Y^n)$$

$$\text{où } E(X / \hat{Y} = y) = \frac{\sum_n X^n K\left(\frac{Y^n - y}{h}\right)}{\sum_m K\left(\frac{Y^m - y}{h}\right)}$$

↳ Discrimination

$$\Gamma_{E(X/Y)}^N = \frac{1}{N} \sum_k N_k E(X / \hat{Y} = k) \otimes E(X / \hat{Y} = k)$$

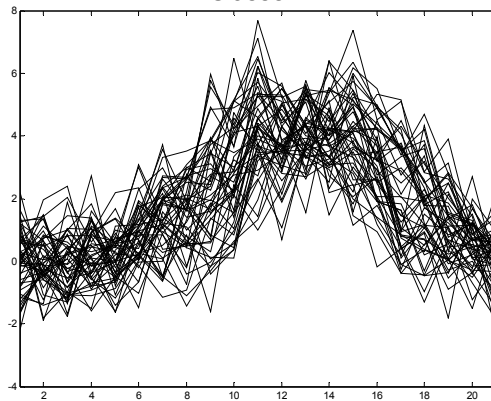
où $N_k = \sum_n I_{\{Y^n = k\}}$ et $E(X / \hat{Y} = k) = \frac{1}{N_k} \sum_n X^n I_{\{Y^n = k\}}$

Exemples (en discrimination)

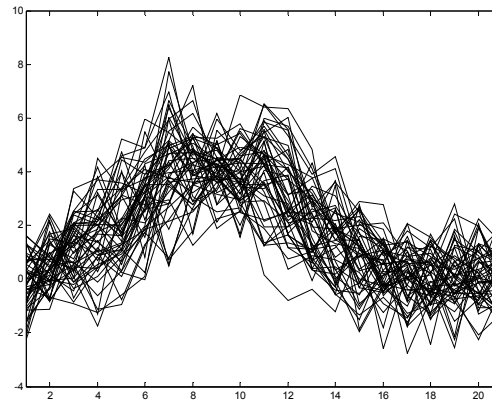
Ferré, Villa (2005)

1) Données simulées : « waveform data »

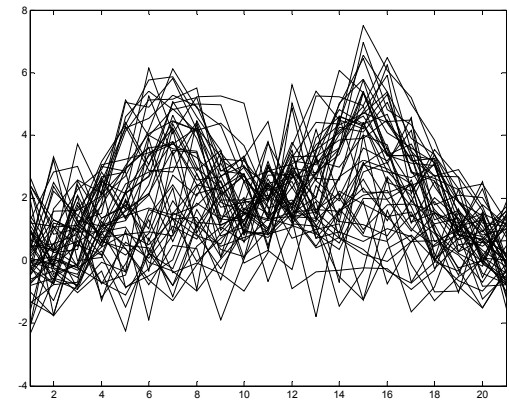
Classe 1



Classe 2



Classe 3



$$uh_1(t) + (1-u)h_2(t) + \varepsilon(t)$$

$$uh_1(t) + (1-u)h_3(t) + \varepsilon(t)$$

$$uh_2(t) + (1-u)h_3(t) + \varepsilon(t)$$

$$\bullet h_1(t) = \max(6 - |t - 11|, 0)$$

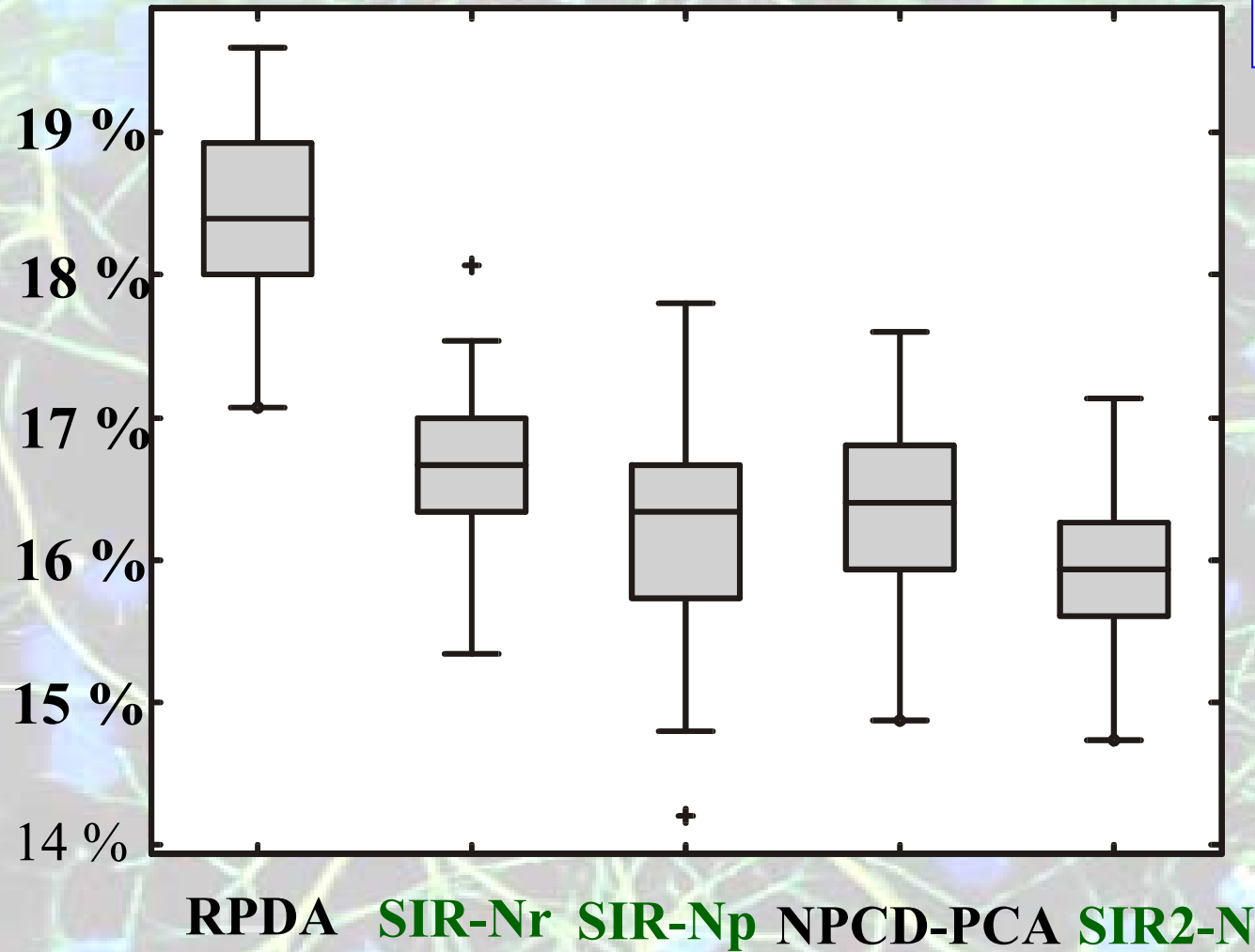
$$\bullet h_2(t) = h_1(t - 4)$$

$$\bullet h_3(t) = h_1(t + 4)$$

Méthodes comparées :

- ✓ **SIR régularisée + Noyau**
- ✓ **SIR projetée + Noyau**
- ✓ **SIR inverse généralisé (Ferré, Yao 2004) + Noyau**
- ✓ **Ridge-PDA (*Hastie, Buja, Tibschirani*)**
- ✓ **NPCD – PCA (*Ferraty, Vieu*)**

Protocole expérimental : **Sur 50 échantillons aléatoires, on effectue la discrimination sur un échantillon d'apprentissage et on calcule le taux d'erreur sur un échantillon de test.**

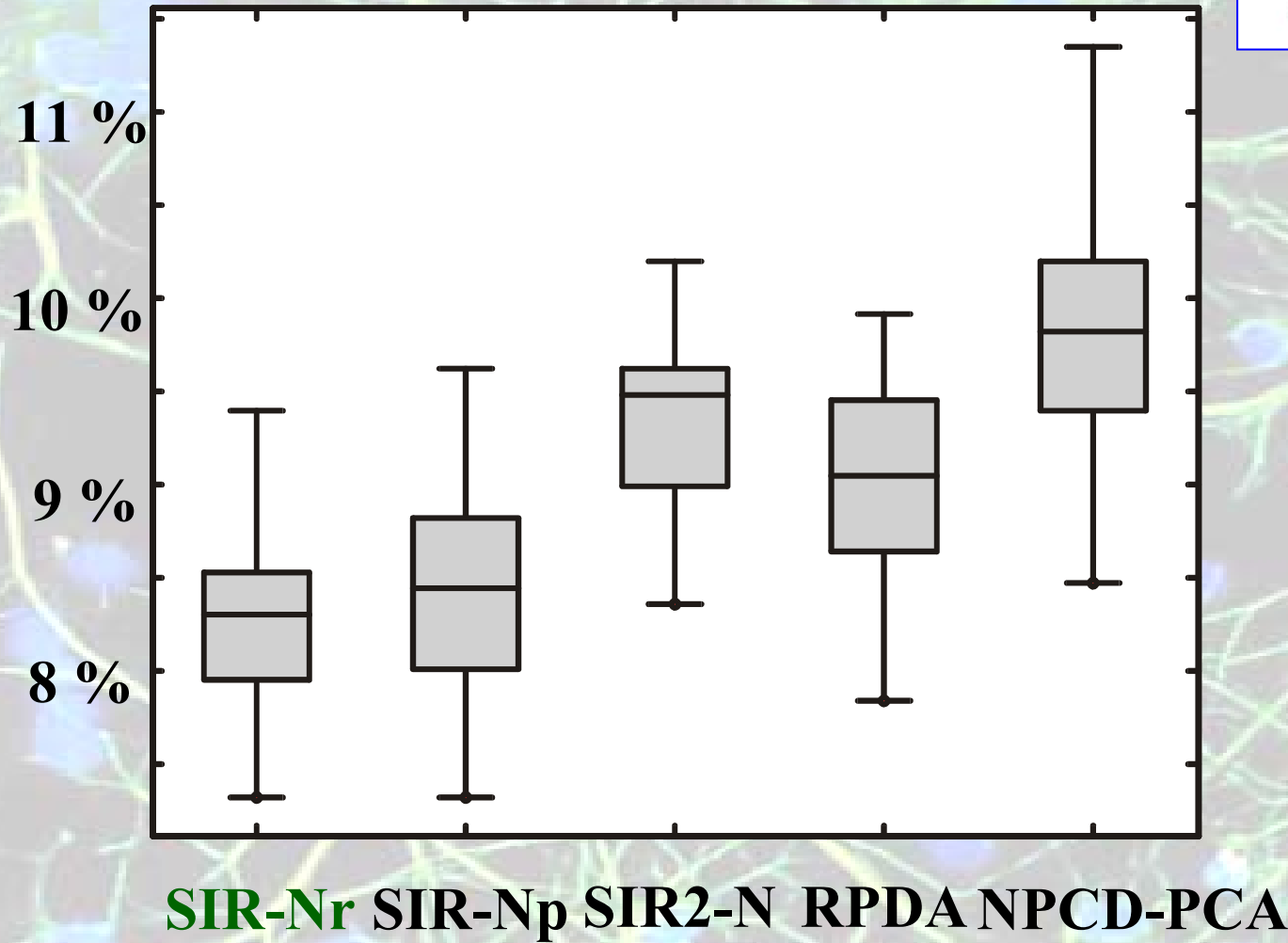


2) Données réelles : les phonèmes

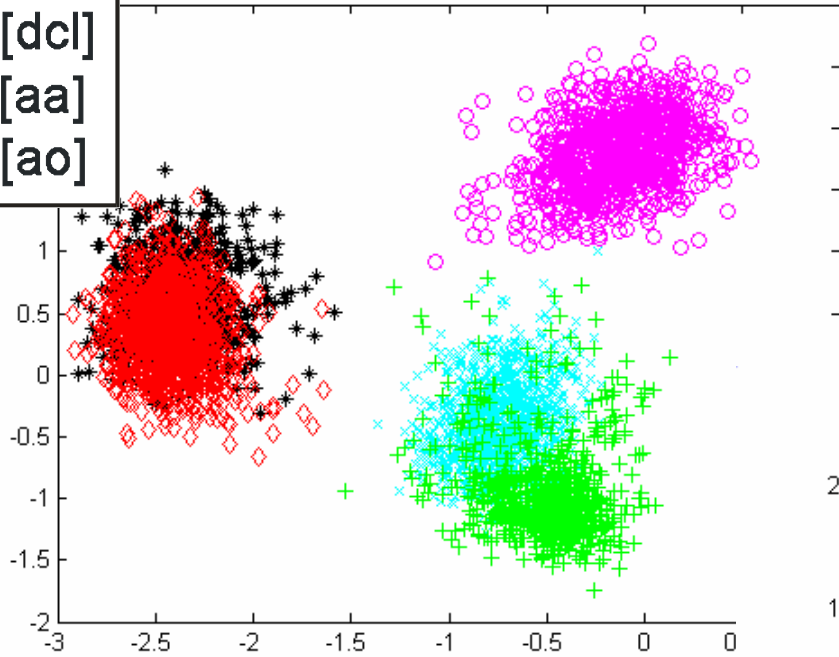
Méthodes comparées :

- ✓ **SIR régularisée + Noyau**
- ✓ **SIR projetée + Noyau**
- ✓ **SIR pseudo-inverse (Ferré, Yao 2004) + Noyau**
- ✓ **Ridge-PDA (Hastie, Buja, Tibschirani)**
- ✓ **NPCD – PCA (Ferraty, Vieu)**

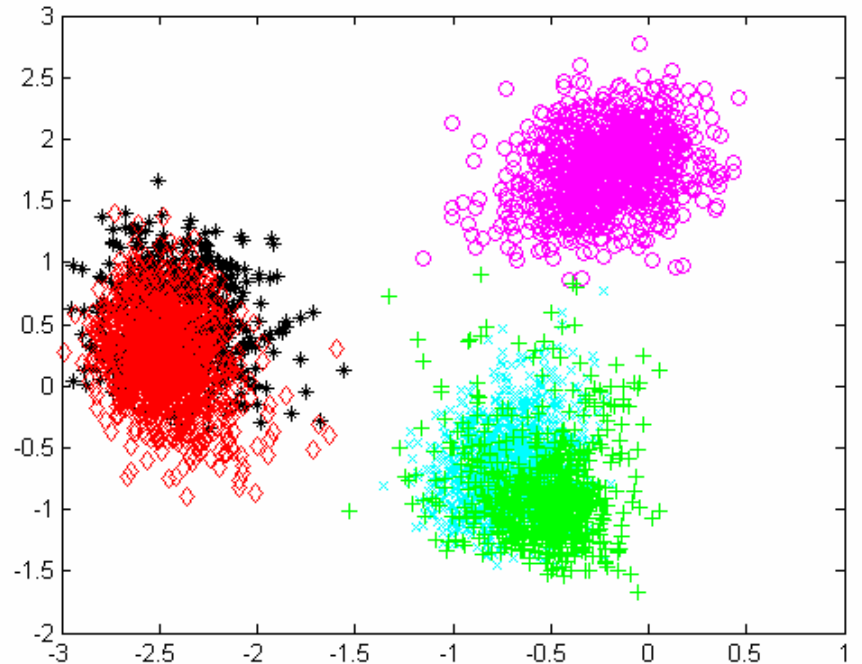
Protocole expérimental : **Idem.**



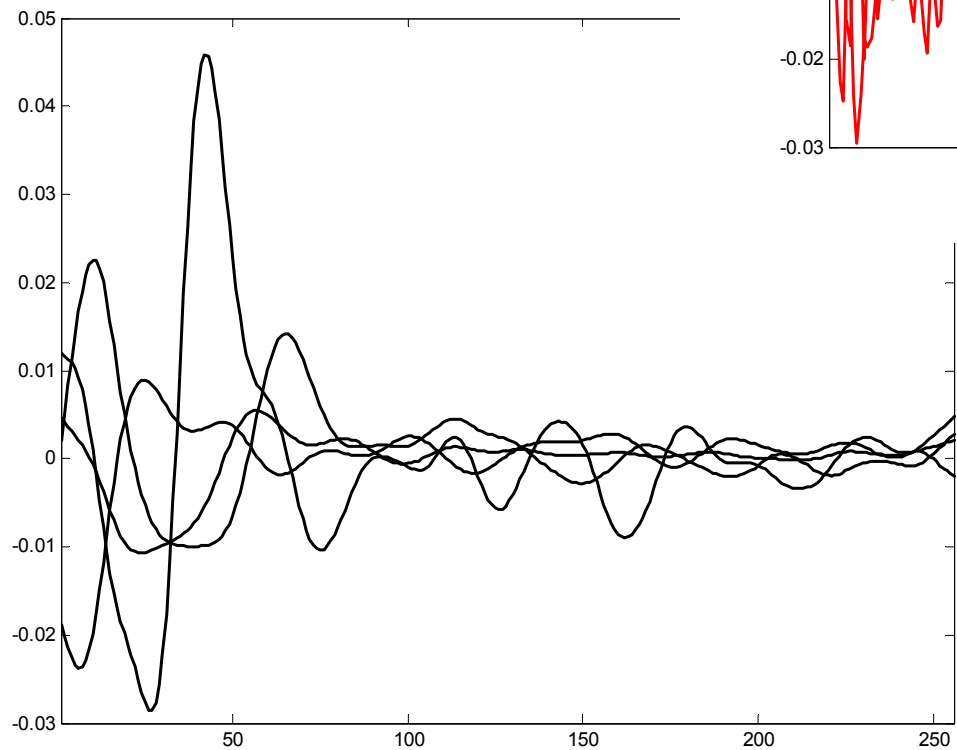
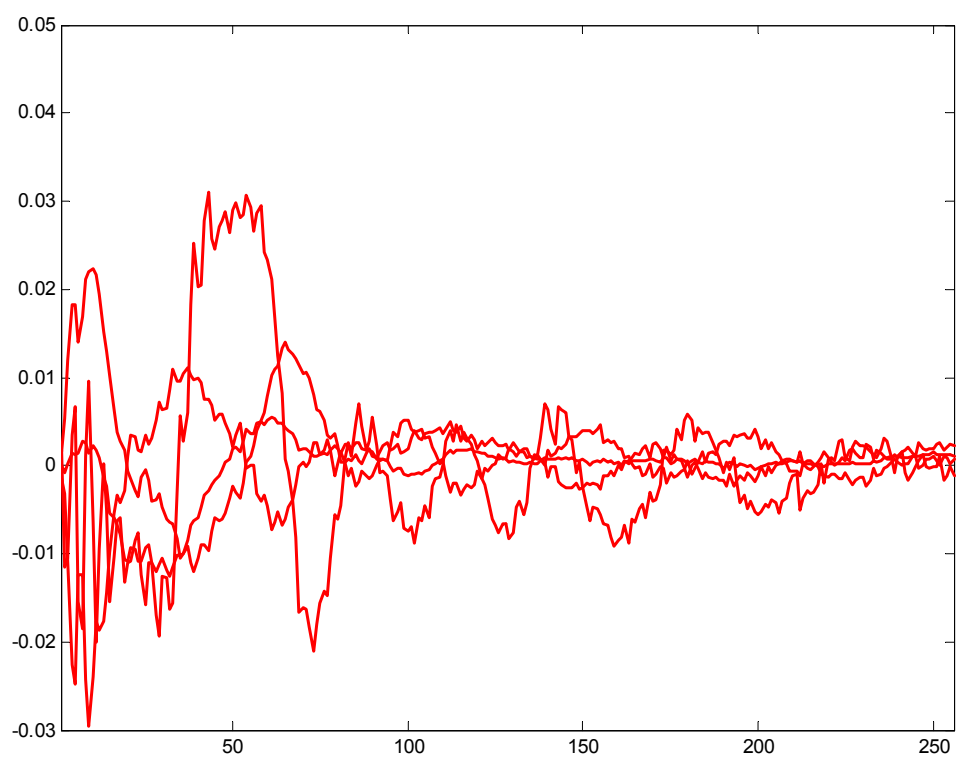
- △ [sh]
- × [iy]
- + [dcl]
- * [aa]
- ◇ [ao]



Projection sur les deux
premiers vecteurs
propres SIR projetée



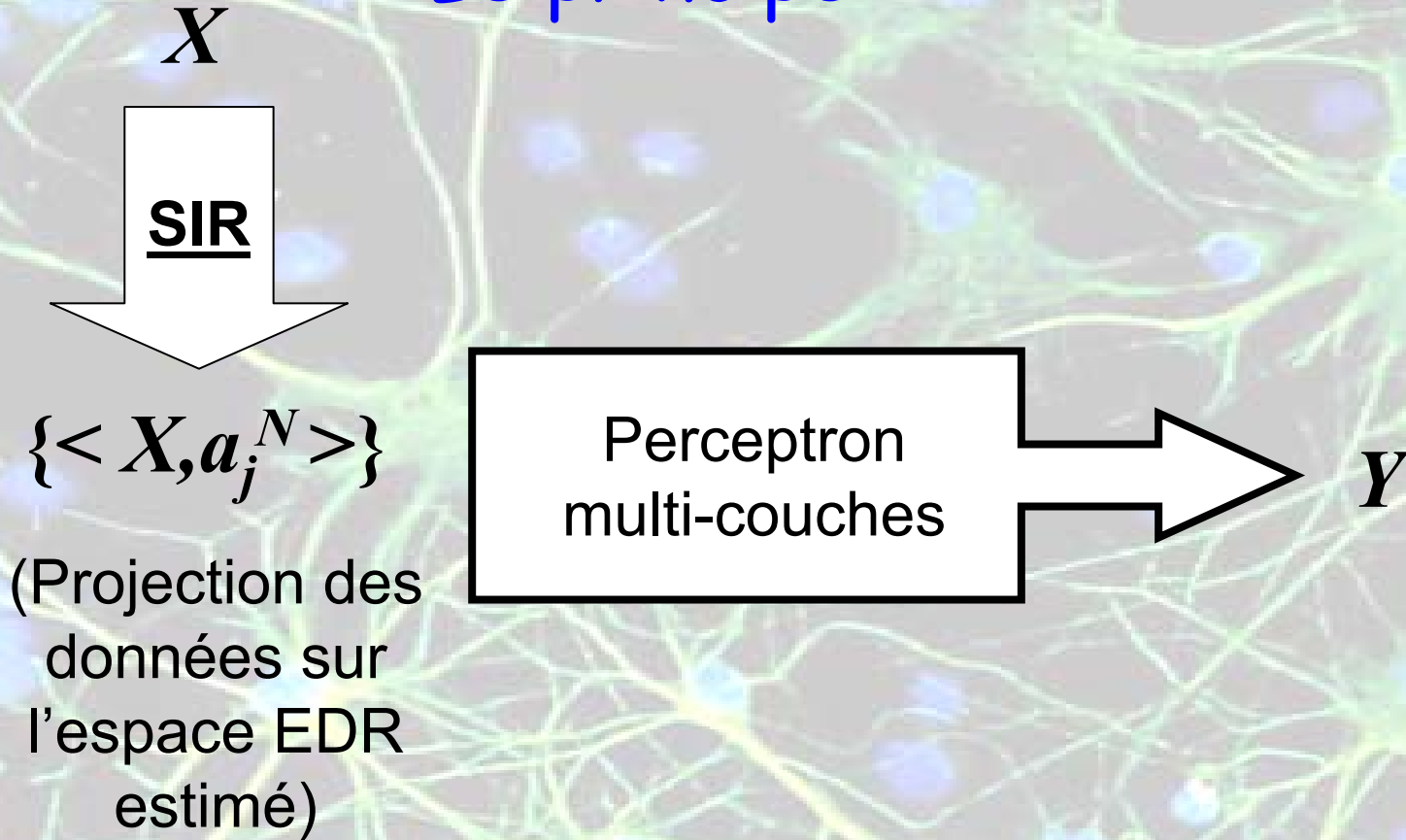
Projection sur les deux
premiers vecteurs
propres SIR
régularisée



SIR-NN

Où l'on reparle de réseau de neurones...

SIR-NN : Le principe



Théorème : (Consistance)

Sous les hypothèses du théorème précédent et un certain nombre d'hypothèses techniques

(qui sont, par exemple, vérifiées par un perceptron avec comme fonction de transfert sur la couche cachée la fonction sigmoïde et comme fonction d'erreur, l'erreur quadratique moyenne),

les poids permettant d'obtenir l'erreur empirique minimum convergent en probabilité vers les poids théoriques optimaux lorsque le nombre d'observations tend vers $+\infty$.

Avantages

- Le jeu de données est simplifié ;
- La base de projection dépend des données (procédure automatique de détermination) ;
- La base de projection tient compte de la cible : c'est la projection optimale des données pour le problème de discrimination ;
- Un résultat de convergence est démontré pour l'estimation de la base (FIR) et pour l'estimation des poids du réseau.

Simulations et exemples

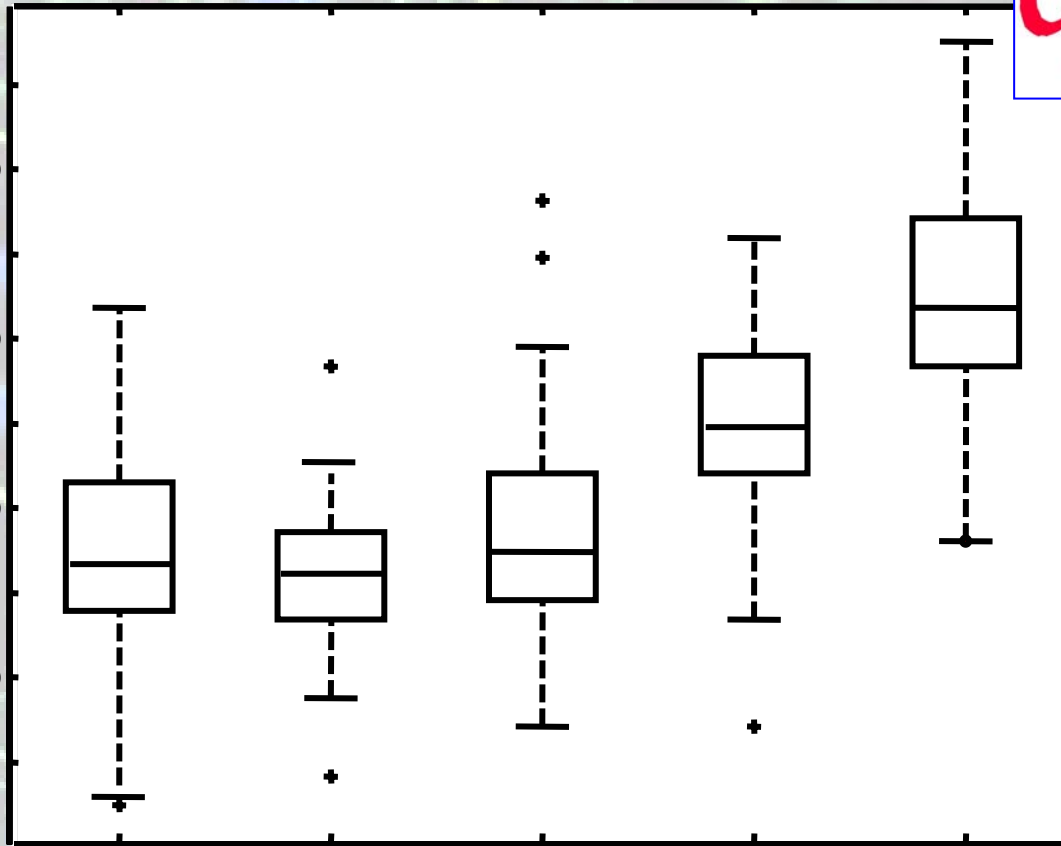
1) *Données de phonèmes*

Méthodes comparées :

- ✓ **SIR régularisée + NN**
- ✓ **SIR régularisée + Noyau**
- ✓ **SIR projetée + NN**
- ✓ **Ridge-PDA** (*Hastie, Buja, Tibschirani*)
- ✓ **NPCD – PCA** (*Ferraty, Vieu*)

Protocole expérimental : **Idem.**

0.11
0.105
0.1
0.095
0.09
0.085
0.08
0.075
0.07



SIRr-NN

SIRp-NN

NPCD-PCA

SIR-Noyau

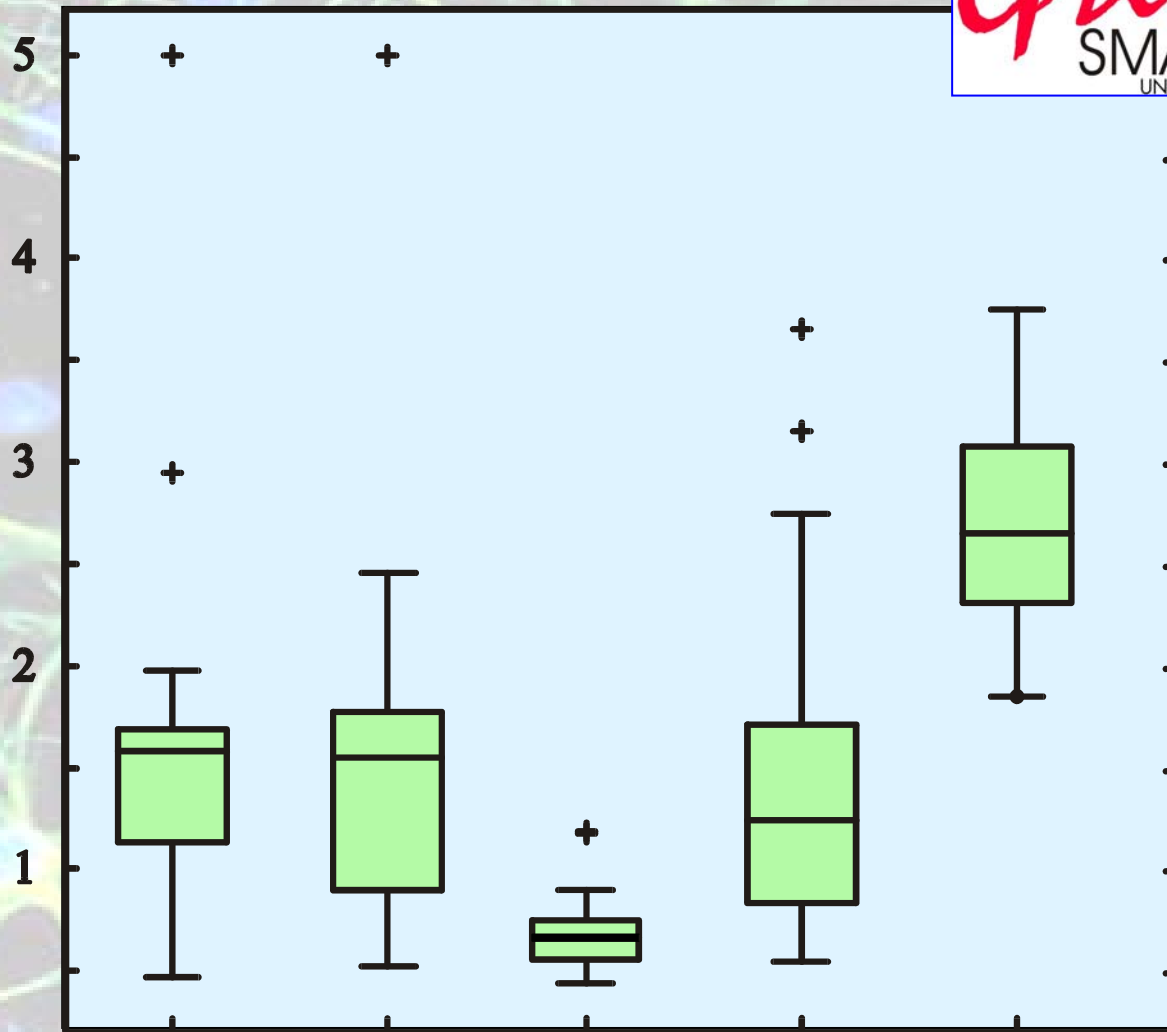
R-PDA

2) Données de spectrométrie

Méthodes comparées :

- ✓ **SIR régularisée + NN**
- ✓ **SIR pseudo-inverse + NN**
- ✓ **ACP + NN** (\approx Thodberg)
- ✓ **NNf** (Rossi, méthode projection sur Spline)
- ✓ **SIR + Linéaire**

Protocole expérimental : **Idem.**



ACP-NN NNf **SIR-NNr** SIR-NNn SIR-l

Merci de votre attention

