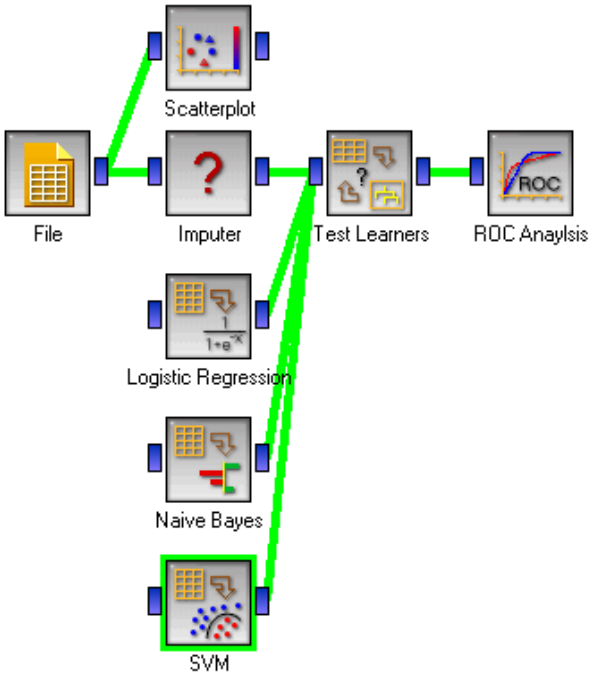
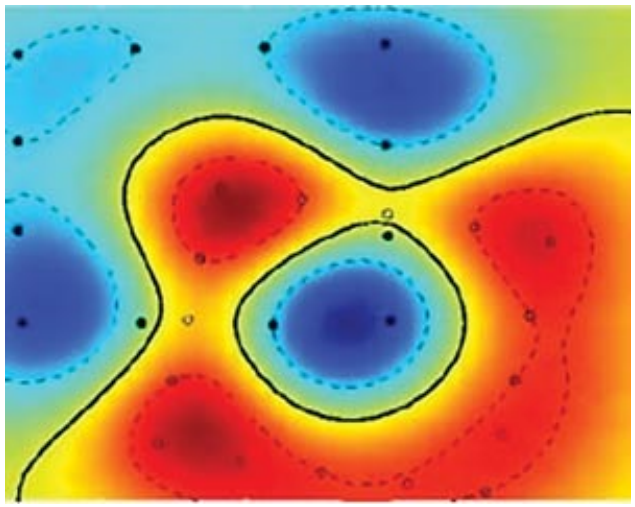
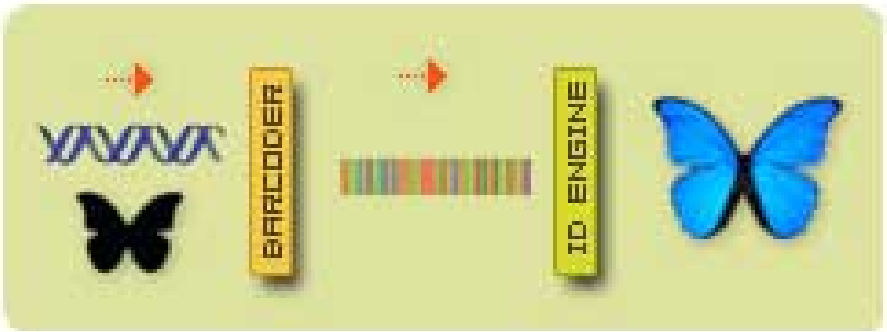
Statistique & Machine Learning







Intervenants

- **Stéphan Clémençon** (Telecom ParisTech - Département TSI)
  - Contact: `stephan.clemencon@telecom-paristech.fr`
  - Profil: Enseignement/Recherche/Conseil/Industrie
  - Mots-clés: processus stochastiques (markoviens, empiriques, etc.), apprentissage statistique, applications: finance, high tech, biosciences
- **Aurélien Garivier** (CNRS - Département TSI)
  - Contact: `aurelien.garivier@telecom-paristech.fr`
  - Profil: Enseignement/Recherche
  - Mots-clés: théorie de l'information, apprentissage on-line et par renforcement, processus de décision markoviens

Data mining = Fouille de données

Motivations pour la fouille de données

- Explosion des capacités de stockage
- Bases de données massives
  - finance, génomique, marketing, industrie ...
- Les données sont partout !
  - de grande dimension, hétérogènes, structurées
- Il existe des approches génériques et automatisables

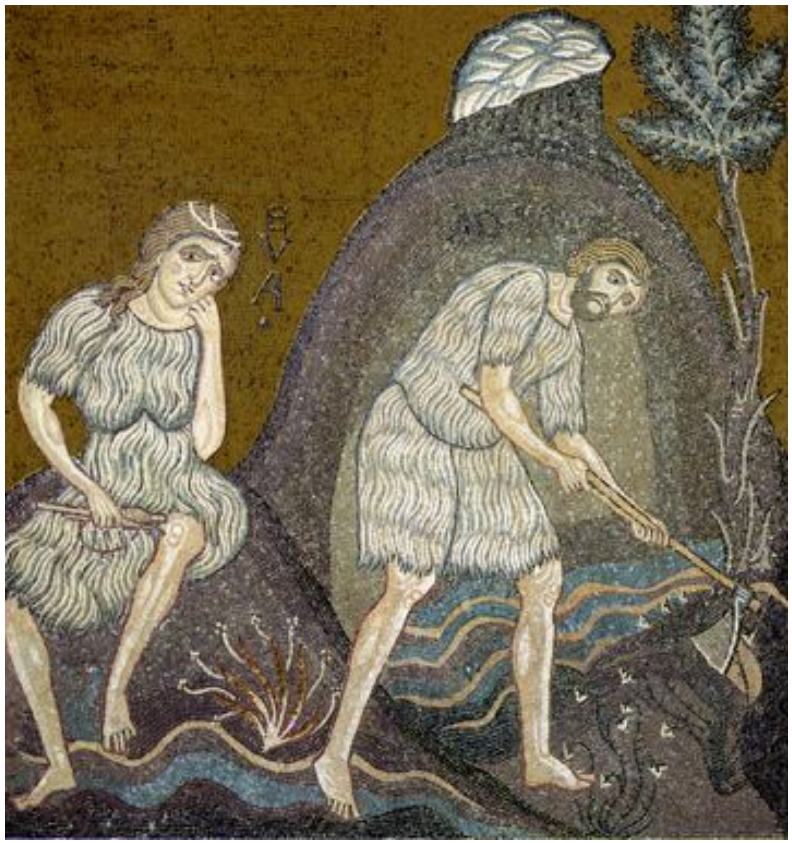
Motivations pour la fouille de données

- Explosion des capacités de stockage
- Bases de données massives
  - finance, génomique, marketing, industrie ...
- Les données sont partout !
  - de grande dimension, hétérogènes, structurées
- Il existe des approches génériques et automatisables

Le but de ce cours : les découvrir !

Les données aujourd'hui

Les chiffres du travail (1)



Les chiffres du travail (2)

Taux d'activité par tranche d'âge hommes vs. femmes

	A	B	C	D	E	F	G	H	I
1									
2	**Taux d'activité par tranche d'âge de 1975 à 2005**								
3	**En %**								
4		**1975**	**1976**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**
5	**Femmes**								
6	15-24 ans	45,5	45,7	45,2	43,9	44,2	42,9	42,1	41,87
7	25-49 ans	58,6	60,3	62,1	62,8	64,7	65,4	66,2	67,55
8	50 ans et plus	42,9	43,1	44,4	43,9	44,8	45,9	45,2	43,47
9	Ensemble	51,5	52,5	53,6	53,6	54,8	55,1	55,1	55,29
10	**Hommes**								
11	15-24 ans	55,6	54,7	53,7	52,2	52,5	52,0	50,4	45,02
12	25-49 ans	97,0	97,1	96,9	96,9	96,9	97,1	96,9	96,75
13	50 ans et plus	79,5	78,8	79,5	78,8	79,4	78,3	75,4	71,65
14	Ensemble	82,5	82,2	82,1	81,6	81,8	81,5	80,4	78,14

Les chiffres du travail (2)

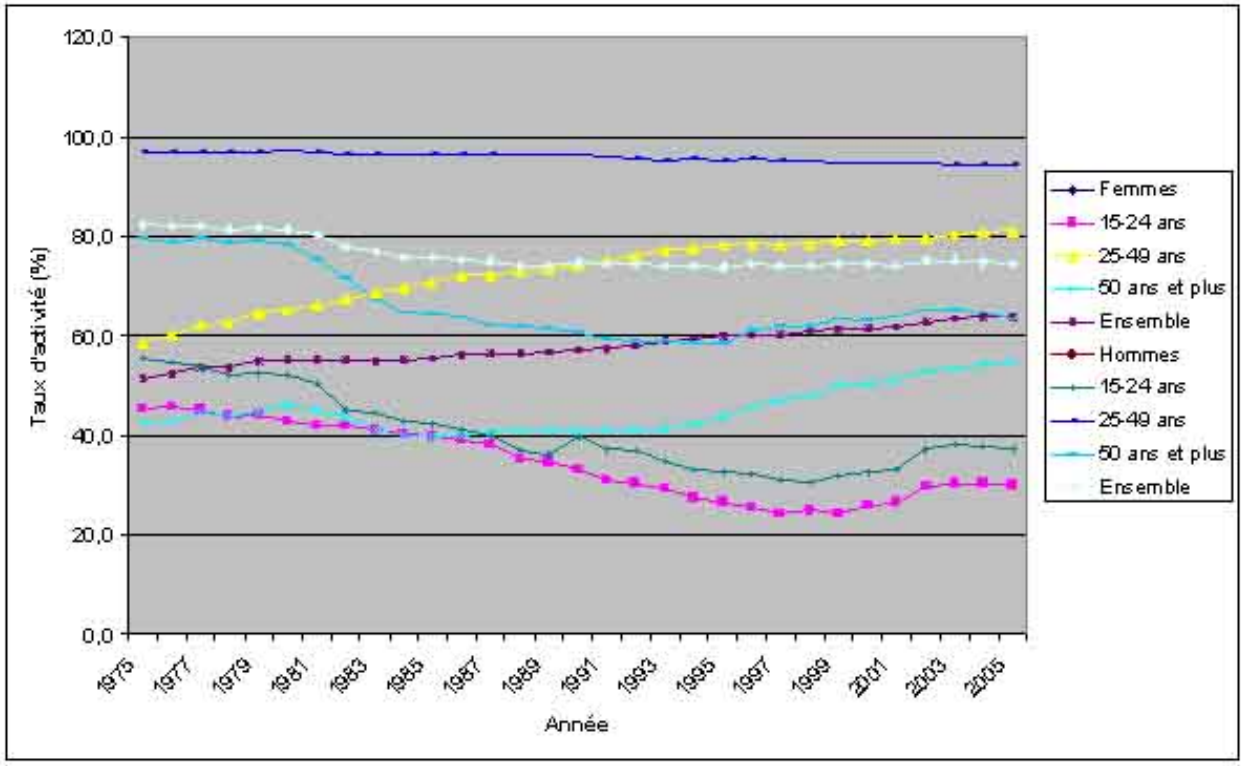
Taux d'activité par tranche d'âge hommes vs. femmes

	A	B	C	D	E	F	G	H	I
1									
2	**Taux d'activité par tranche d'âge de 1975 à 2005**								
3	**En %**								
4		**1975**	**1976**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**
5	**Femmes**								
6	15-24 ans	45,5	45,7	45,2	43,9	44,2	42,9	42,1	41,87
7	25-49 ans	58,6	60,3	62,1	62,8	64,7	65,4	66,2	67,55
8	50 ans et plus	42,9	43,1	44,4	43,9	44,8	45,9	45,2	43,47
9	Ensemble	51,5	52,5	53,6	53,6	54,8	55,1	55,1	55,29
10	**Hommes**								
11	15-24 ans	55,6	54,7	53,7	52,2	52,5	52,0	50,4	45,02
12	25-49 ans	97,0	97,1	96,9	96,9	96,9	97,1	96,9	96,75
13	50 ans et plus	79,5	78,8	79,5	78,8	79,4	78,3	75,4	71,65
14	Ensemble	82,5	82,2	82,1	81,6	81,8	81,5	80,4	78,14

http://www.insee.fr/

Les chiffres du travail (3)

Taux d'activité par tranche d'âge hommes vs. femmes

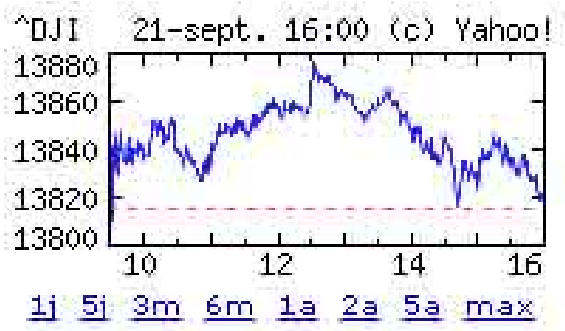


Le monde de la finance (1)

Wall Street à la clotûre, un lundi...

Le monde de la finance (2)

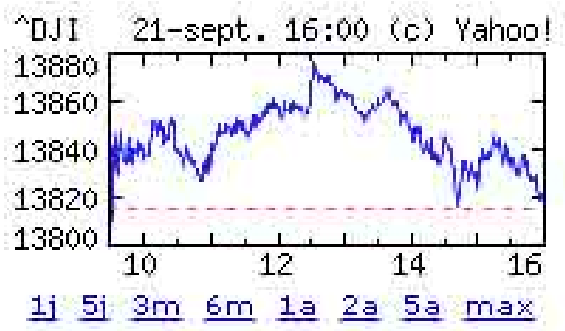
DOW JONES INDUSTRIAL AVERAGE IN (DJI:^DJI)	
Dern. Cours:	13.820,19
Heure:	21 sept.
Variation:	⬆ 53,49 (0,39%)
Clôture Préc.:	13.766,70
Ouverture:	13.768,33
Var. Journalière:	13.768,25 - 13.877,17
Var. sur 1 an:	11.926,80 - 14.121,00
Volume:	419.389.397



Le monde de la finance (2)

DOW JONES INDUSTRIAL AVERAGE IN (DJI:^DJI)

Dern. Cours:	13.820,19
Heure:	21 sept.
Variation:	↑ 53,49 (0,39%)
Clôture Préc.:	13.766,70
Ouverture:	13.768,33
Var. Journalière:	13.768,25 - 13.877,17
Var. sur 1 an:	11.926,80 - 14.121,00
Volume:	419.389.397



http://fr.finance.yahoo.com/

L'imagerie médicale (1)

L'imagerie médicale (2)

Internet (1)

Internet (2)

Netscape Proxy format:

```
format=%Ses->client.ip% 146.127.62.22 %Req->vars.pauth-user% [%SYSDATE%] "%Req->reqpb.proxy-request%
%Req->srvhdrs.clf-status% %Req->vars.p2c-cl% %Req->vars.remote-status% %Req->vars.r2p-cl%
%Req->headers.content-length% %Req->vars.p2r-cl% %Req->vars.c2p-hl% %Req->vars.p2c-hl%
%Req->vars.p2r-hl% %Req->vars.r2p-hl% %Req->vars.xfer-time% %Req->vars.actual-route%
%Req->vars.cli-status% %Req->vars.svr-status% %Req->vars.cch-status%
146.127.123.16 146.127.62.22 - [10/Dec/1997:00:30:09 -0500] "GET http://www.nba.com/bulls/ HTTP/1.0" 200 881
200 8816 - - 321 164 359 164 1 SOCKS(146.127.11.3:1080) FIN FIN NON-CACHEABLE
146.127.253.84 146.127.62.22 - [10/Dec/1997:00:30:12 -0500] "GET http://www.pathfinder.com/NY1/bug.html
HTTP/1.0" 200 377 200 377 - - 392 203 418 203 1 SOCKS(146.127.11.3:1080) FIN FIN REFRESHED
146.127.253.84 146.127.62.22 - [10/Dec/1997:00:30:12 -0500] "GET
http://www.pathfinder.com/NY1/images/steel.gif HTTP/1.0" 304 - 304 - - - 443 142 468 142 0
SOCKS(146.127.11.3:1080) FIN FIN UP-TO-DATE
```

Séquençage du génome humain (1)

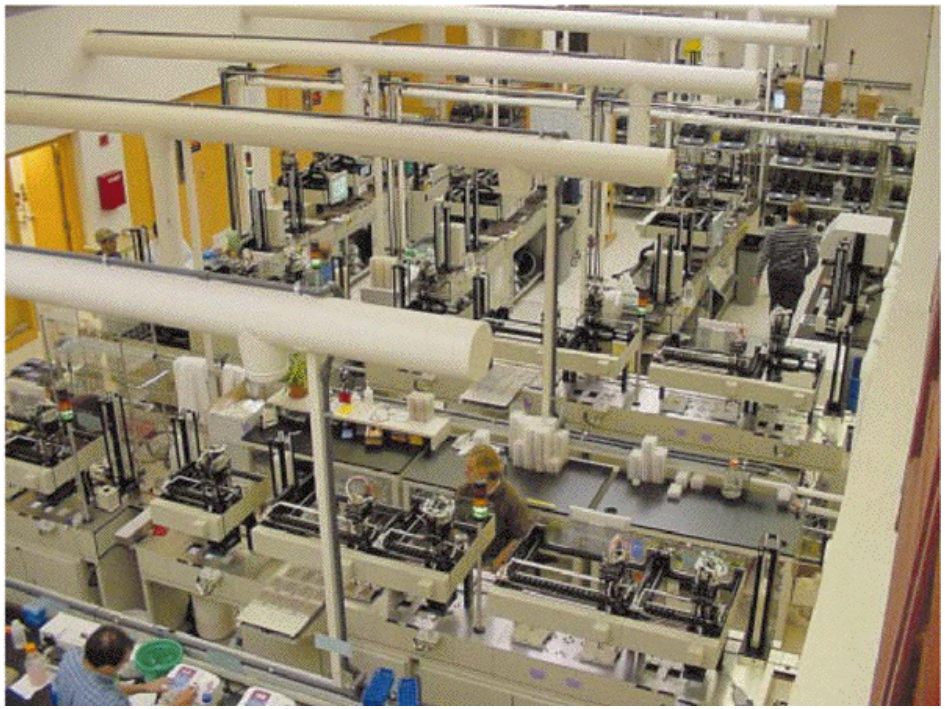
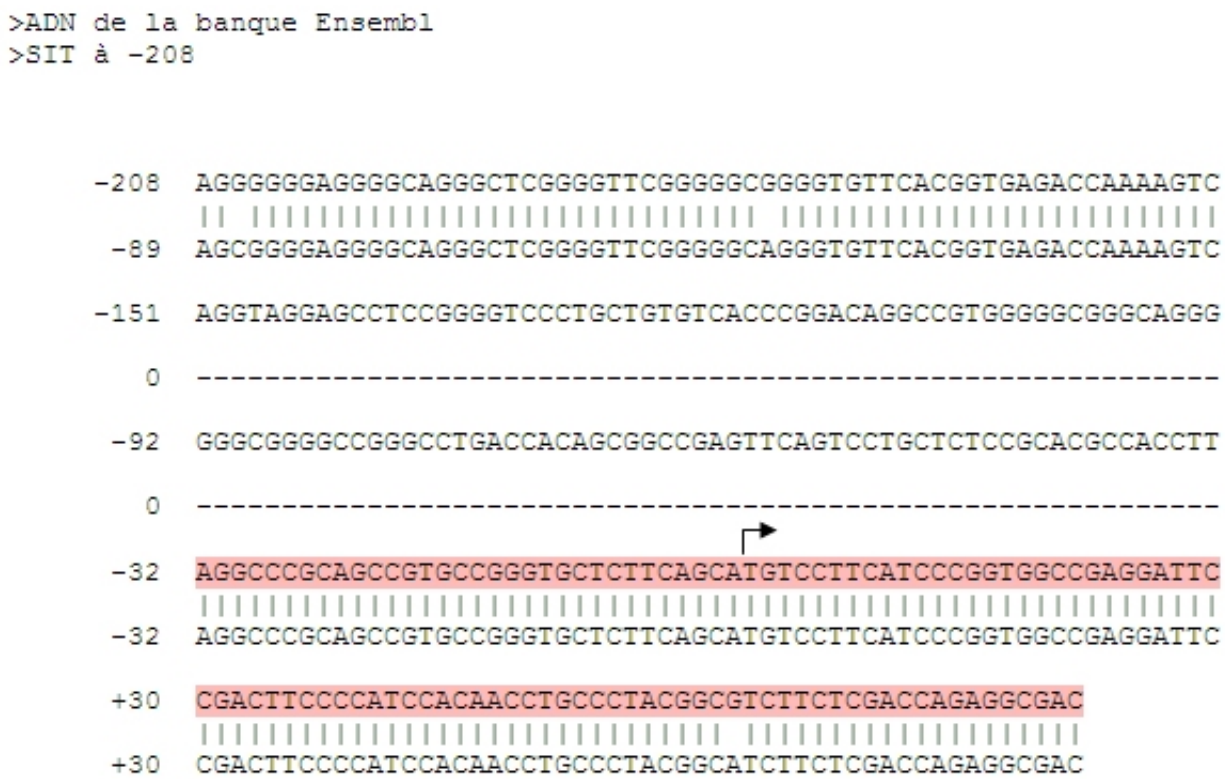


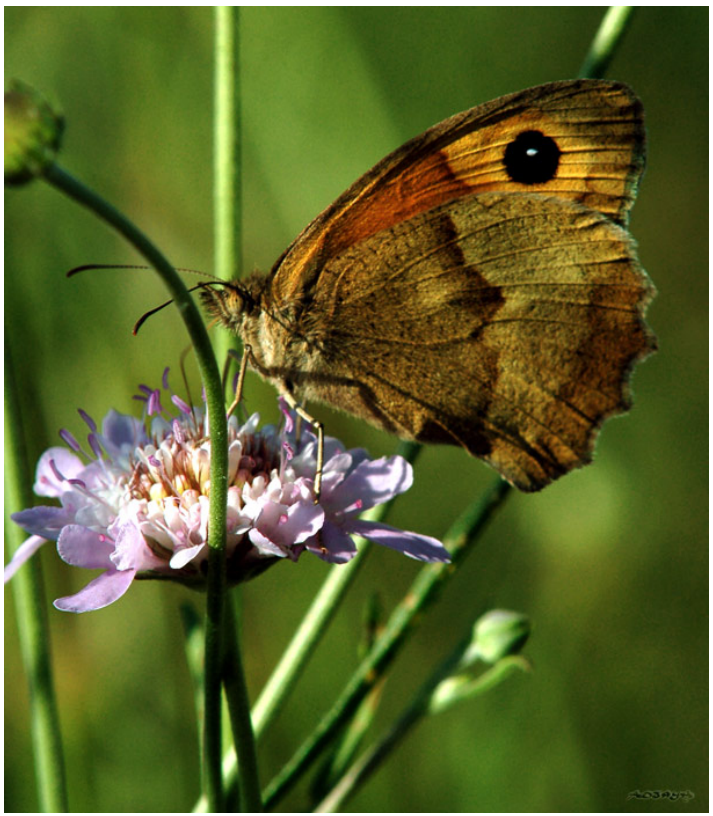
Plate-forme de séquençage génotypage OUEST-genopole

Séquençage du génome humain (2)



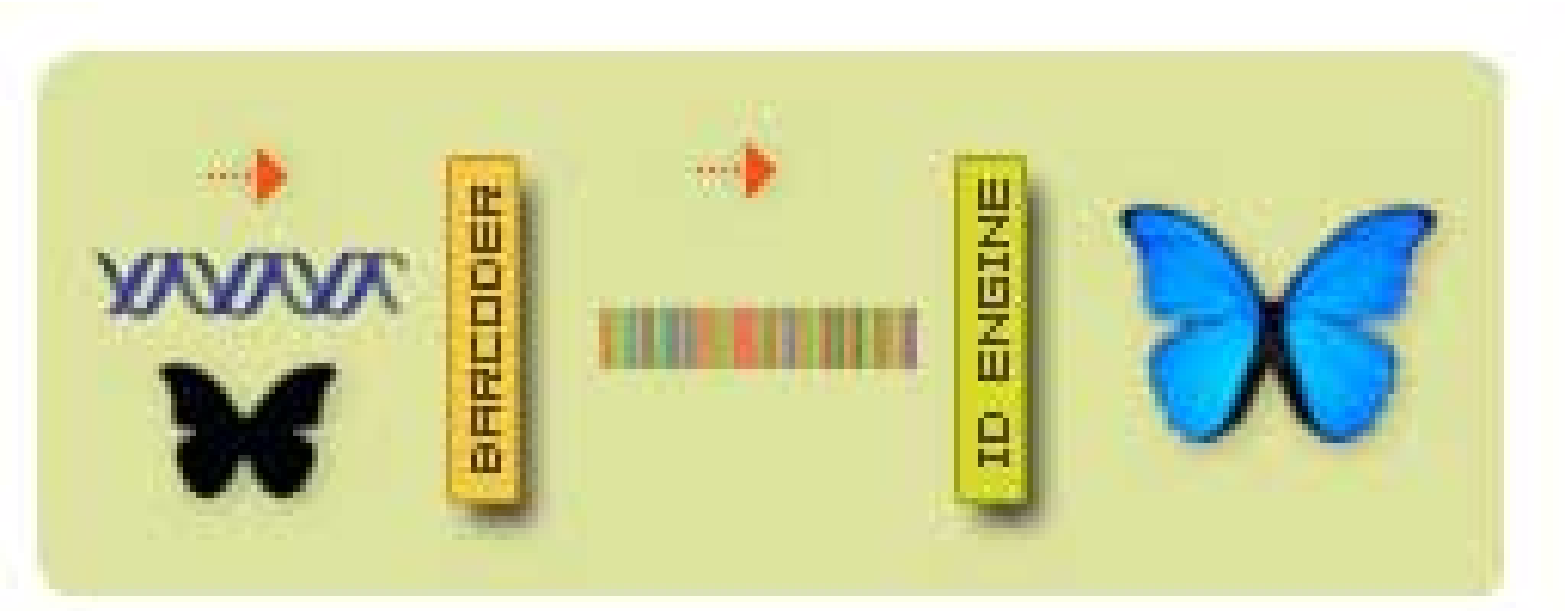
Alignement d'une séquence identifiée dans le foie humain avec l'ADN génomique de la banque Ensembl.

Barcoding Of Life Data Systems (1)



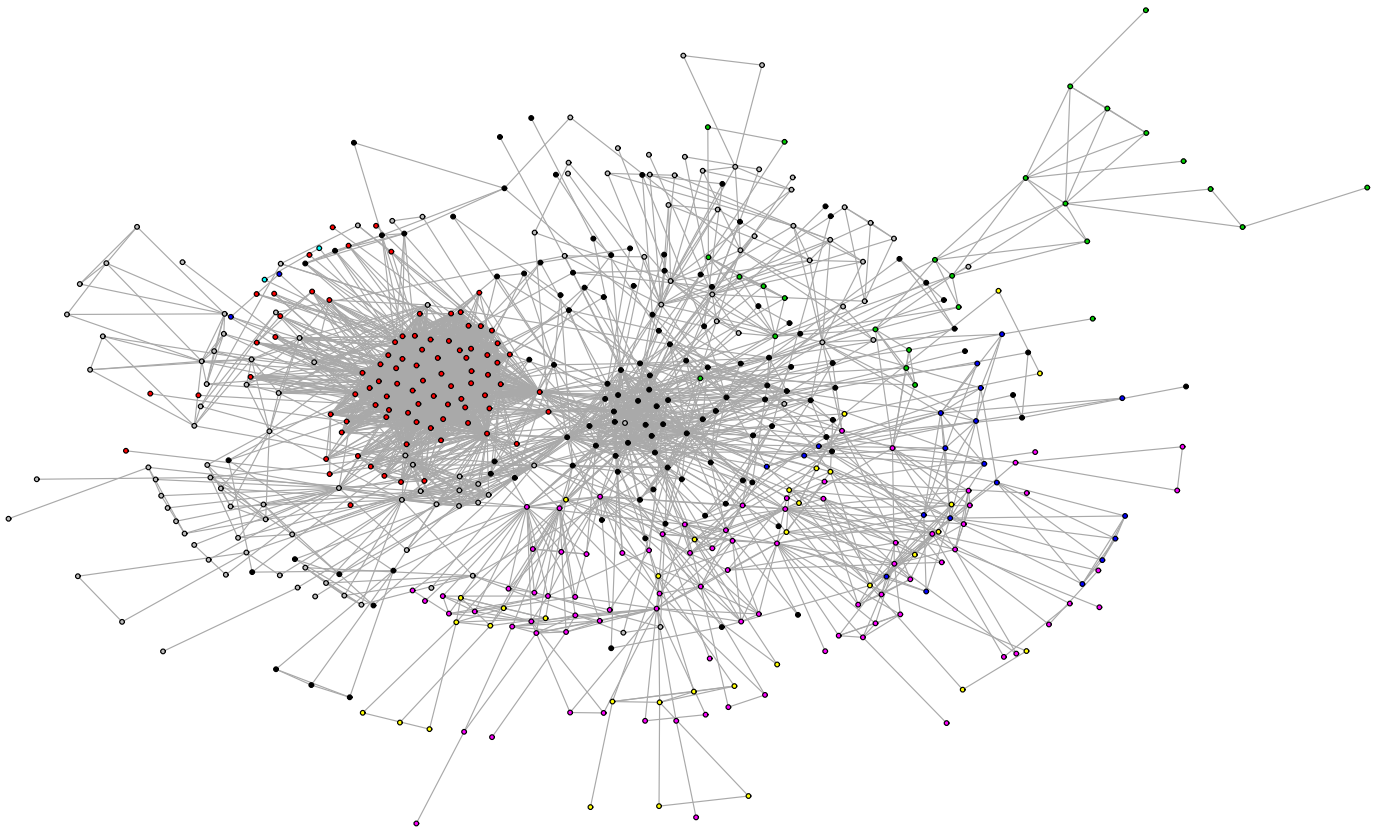
Barcoding of Life Data Systems (2)

Barcoding of Life Data Systems (2)

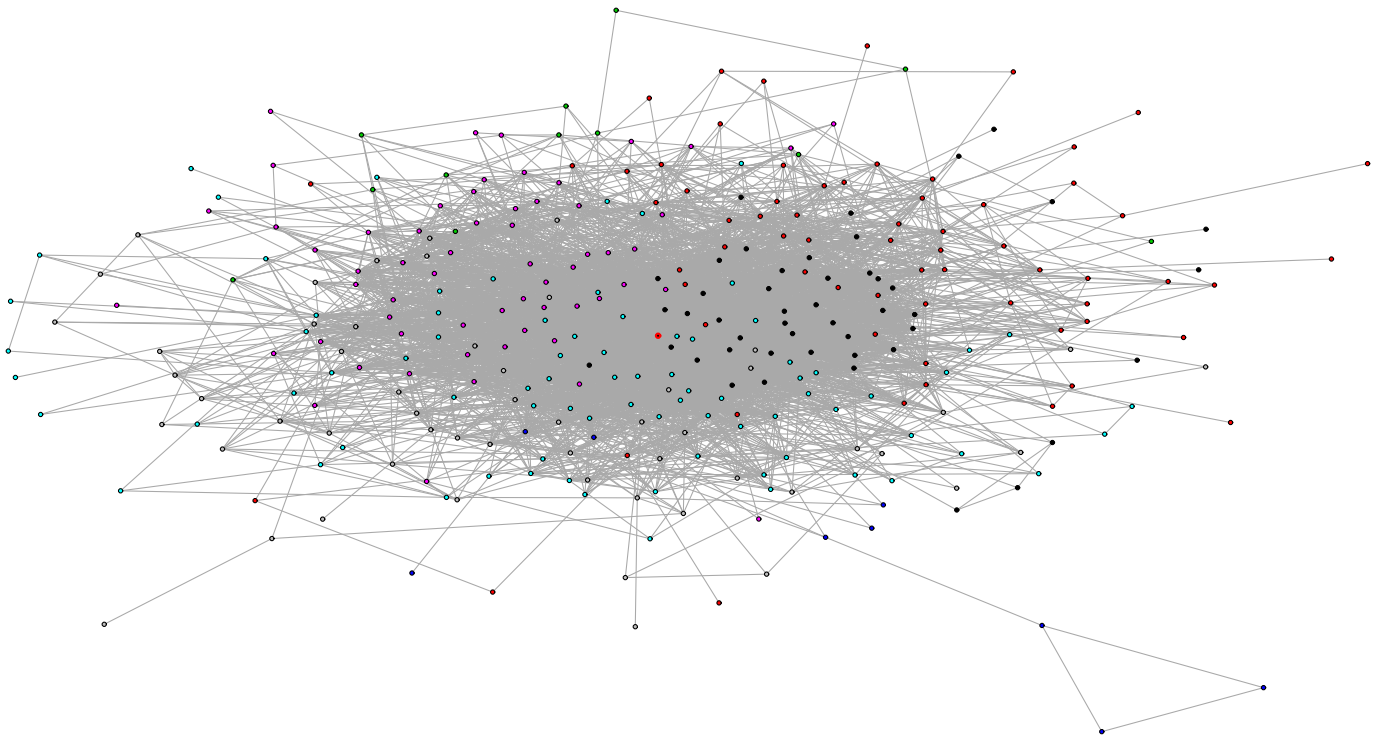


http://www.barcodinglife.org/

E-marketing (1)- Livres



E-marketing (1)- Jeux vidéos



Nature des données

- Vecteurs/Matrices
- Chaînes de caractères
- Graphes/Réseaux
- Fonctions/Séries temporelles

Les questions de data mining

- Identification de dépendances
- Segmentation/Clustering/Classification
- Détection d'anomalies
- Réduction de la dimension
- Sélection de variables
- Interprétation/Parcimonie
- Visualisation

Le preprocessing, une étape critique

- Nettoyage ou filtrage des données
- Données incomplètes
- Données aberrantes
- Données hétérogènes ou multi-échelles
- Indexation

Le désert du réel...

Les outils

Domaines afférents

- Informatique :
  - BDD
  - algorithmique
- **Machine Learning :**
  - méthodes effectives pour la grande dimension
- **Mathématiques :**
  - algèbre linéaire
  - **modélisation aléatoire**,
  - probabilités / **statistique**
  - **apprentissage statistique**
  - optimisation
  - traitement du signal

Cours de statistique "typique"

- Estimation paramétrique
- Intervalles/Domaines de confiance
- Tests d'hypothèses
- Régression
- Analyse en composantes principales

Aspects non abordés dans ce type de cours

- Classification
- Méthodes non-paramétriques
- Statistique bayésienne
- Sélection de modèles
- Théorie de la décision

Pourquoi faire appel à l'apprentissage statistique?

- Typologie des problèmes
- No Free Lunch !
- Choix des critères de performance
- Notion de risque
- Contrôle de la complexité
- Validation des règles de décision
- Rôle du rééchantillonnage
- Monitoring des modèles de prévision

Machine Learning... un peu plus que des stats

- Méthodes non-paramétriques opérationnelles
- Traitement de données complexes / de grande dimension
- Diversité des contextes
  - supervisé, non-supervisé, semi-supervisé, séquentiel, one-pass, ...
- Couplage des principes inférentiels avec des algorithmes !

Programme des premières séances

- Séances 1 - 2 : **Introduction - Contexte**
  Introduction
  Eléments de statistique (Rappels)
  Nomenclature des problèmes rencontrés
  Applications (exemples)
  Réduction de la dimension - ACP & co.

- Séance 3 : **Un peu de théorie: classification binaire**
  Le principe de la minimisation du risque empirique
  Théorie de Vapnik-Chervonenkis (complexité combinatoire)
  Une solution statistique.... un problème informatique!

Programme des premières séances (2)

- Séance 4 : **Algorithmes "classiques" pour la classification**
  Analyse discriminante linéaire et régression logistique
  Les "plus proches voisins" et variantes
  Méthodes de partitionnement - l'algorithme CART
  Le perceptron - méthodes linéaires
  Réseaux de Neurones
- Séance 5: **Algorithmes "avancés" pour la classification**
  SVM
  Boosting
  Random Forest
- Séances 6 et 7: **D'autres problèmes supervisés**
  Ranking/scoring
  Classification multi-label
  Régression ordinale et Régression

Statistical learning - Historical milestones

- 1943: Artificial neuron model - McCullough, Pitts
- 1958: Perceptron algorithm - Rosenblatt
- 60's: Data-mining - John Tukey
- 1971: Uniform laws of large numbers - Vapnik, Chervonenkis
- 1974, 1986: Backpropagation algorithm
- 1984: CART - Breiman, Friedman, Stone, Olshen
- 1984: Theory of the learnable - Valiant
- 1995: Statistical learning theory - Vapnik

Documentation

- Livres:
  - Pattern classification (2001) - Wiley-Interscience
    par R. Duda, P. Hart, D. Stork
  - The Elements of Statistical Learning (2001) - Springer
    par T. Hastie, R. Tibshirani, J. Friedman
  - All of Statistics (2004) - Springer
    par L. Wasserman
  - Matrix Methods in Data Mining and Pattern Recognition (2007) - SIAM
    par L. Eldén
- Article :
  - "The curse and blessings of dimensionnality" D. Donoho - IMS

Logiciels

- Librairie "state-of-the-art" :
  - **The R Project for Statistical Computing**
  - `http://www.r-project.org/`
- Autres applications (logiciels libres) :
  - WEKA
  - Orange
  - RapidMiner

Machine-Learning: les acteurs

- Monde académique:
  - Départements: Maths (Appli), Informatique, Bioinformatique, etc.
    Un savoir fondamental selon le panorama dressé par Carnegie Mellon
  - Journaux: JMLR, Machine Learning, Data-Mining Knowledge Discovery, etc.
  - Conférences: NIPS, ICML, COLT, UAI, etc.

- Industrie:
  - High-tech: google labs, yahoo labs, Exalead, biotech
  - CRM
  - Finance, credit-scoring
  - Signal, image or speech processing, automatic anomaly detection

Rappels de statistique

Détendez-vous...

Détendez-vous...
le film va commencer !

Modèle statistique

- Observation comme réalisation de X variable aléatoire de loi inconnue P^*
- On suppose X à valeurs dans $(E, \mathbb{E})$
- Modèle statistique = triplet $\mathcal{M} = (E, \mathbb{E}, \mathcal{P})$

  où $\mathcal{P} = \{P_\theta \ : \ \theta \in \Theta\}$ famille de lois candidates pour P^*
- Θ est un paramétrage de $\mathcal{P}$, on note $P^* = P_{\theta^*}$
- Le modèle est paramétrique si Θ est un sev d'un espace euclidien
- Le modèle est dit non-paramétrique sinon (dim ∞).
- Modèle identifiable : $\theta \mapsto P_\theta$ est injective

Vraisemblance du paramètre

- On représente $\mathcal{P}$ par la classe des densités associées

$$\{f(x,\theta) \ : \ \theta \in \Theta\}$$

- Vraisemblance : pour x fixé,

$$L_x(\theta) = f(x,\theta) \ .$$

- Exemple : $X = (X_1, \ldots, X_n)$ i.i.d. de loi de Bernoulli $\mathcal{B}(\theta)$

$$L(\theta) = \prod_{i=1}^{n} \left(\theta^{X_i}(1-\theta)^{1-X_i}\right) = \theta^{S_n}(1-\theta)^{n-S_n}$$

où $S_n = \sum_{i=1}^{n} X_i$.

Notion de statistique

- Soit X une observation/ un échantillon. Une statistique est une fonction mesurable $T : E \longrightarrow \mathbb{R}^k$ de X. On dira que $T(X)$ ou $T(X_1, \ldots, X_n)$ est une statistique de l'échantillon.
- Exemple : Moyenne empirique

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- Exemple : Variance empirique

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- Estimation de paramètres $g(\theta^*)$

Estimation

- Exemple d'estimateur = Maximum de vraisemblance
- Dans le modèle de Bernoulli $\mathcal{B}(\theta)$ avec $\theta \in [0,1]$:

$$\hat{\theta}_n = \bar{X}$$

- Risque quadratique et décomposition biais-variance :

$$R(\hat{\theta}_n, \theta^*) = \mathbb{E}_{\theta^*}\left((\hat{\theta}_n - \theta^*)^2 \right)$$
$$= \left(\mathbb{E}(\hat{\theta}_n) - \theta^* \right)^2 + \mathbb{V}_{\theta^*}(\hat{\theta}_n) = \frac{\theta^*(1-\theta^*)}{n} \leq \frac{1}{4n}$$

- Propriétés : consistance, normalité asymptotique (vitesse)
- Et si $\theta^* \notin \Theta$? Et si le modèle est faux ?

Intervalle de confiance - paramètre d'une Bernoulli

- Intervalle aléatoire $I(n,\alpha)$ t.q. $P(\theta^* \in I(n,\alpha)) \geq 1-\alpha$
- Par l'inégalité de Bienaymé-Tchebychev :

$$I(n,\alpha) = \left[\bar{X} - \frac{1}{\sqrt{4n\alpha}}, \bar{X} + \frac{1}{\sqrt{4n\alpha}}\right] .$$

- Par l'inégalité de Hoeffding :

$$I(n,\alpha) = \left[\bar{X} - \sqrt{\frac{\log(2/\alpha)}{2n}}, \bar{X} + \sqrt{\frac{\log(2/\alpha)}{2n}}\right] .$$

- Par la loi limite (Φ fdr de la loi $\mathcal{N}(0,1)$) : $I_\infty(n,\alpha) =$

$$\left[\bar{X} - \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right]$$

Régression

- Modèle linéaire gaussien

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \ .$$

où $\mathbf{Y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$ sont les données
et $\beta \in \mathbb{R}^n$, $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$

- On suppose : $\mathbf{X}^T\mathbf{X}$ inversible (identifiabilité)
- Estimateur des moindres carrés :

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T Y$$

$$\hat{s}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2$$

Questions autour de l'estimateur des moindres carrés

Problèmes :

- Précision de la prédiction : biais faible - grande variance
- Interprétabilité si p est grand

Solutions :

- Réduction de la dimension de la matrice $\mathbf{X}$
- Méthodes pénalisées (shrinkage des coefficients)
- Estimation vs. Prédiction

Les problèmes statistiques revisités

Generic setup for supervised learning

- Random pair $= (X, Y) \sim P$ unknown
- $X =$ observation vector in $\mathcal{X}(\mathbb{R}^d)$, ici $d \gg 1$
- $Y =$ univariate label in $\mathcal{Y} \subset \mathbb{R}$
- Predictor: $g : \mathcal{X} \to \mathcal{Y}$ in a class $\mathcal{G}$
- Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$
- Risk functional (unknown!) $=$ Generalization error

$$L(g) = \mathbb{E}\left(\ell(Y, g(X))\right)$$

  to minimize over $g \in \mathcal{G}$.

- Data $= D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ i.i.d. as P

Example 1 - Regression

- Example: Prediction of a stock price
- $\mathcal{X}$ = vector of descriptors (financial information, macro-economic indicators, ...)
- $\mathcal{Y} = \mathbb{R}$
- Loss function = quadratic error

$$\ell(y, z) = (y - z)^2$$

- Optimal solution: $g^*(x) = \mathbb{E}(Y \mid X = x)$

Example 2 - Scoring

- Classification data: $\mathcal{Y} = \{0, 1\}$
- Set $\eta(x) = \mathbb{E}(Y \mid X = x) = \mathbb{P}\{Y = 1 \mid X = x\}$
- Logistic regression

$$f(x) = \log\left(\frac{\eta(x)}{1 - \eta(x)}\right)$$

- Additive logistic model

  $\longrightarrow$ back to linear regression

Example 3 - Binary classification

- Example: Prediction of the state of a system
- $\mathcal{Y} = \{-1, +1\}$
- Loss function:

$$\ell(y, z) = \mathbb{I}\{y \neq z\}$$

- Risk functional:

$$\begin{aligned} L(g) &= \mathbb{P}\{Y \neq g(X)\} \\ &= \mathbb{P}\{Y \cdot g(X) < 0\} = \mathbb{E}\left(\mathbb{I}_{\mathbb{R}^+}(-Y \cdot g(X))\right) \end{aligned}$$

Example 4 - Multiclass Classification

- Example: handwritten character recognition
- $\mathcal{Y} = \{1, \ldots, M\}$
- Loss function

$$\ell(y, z) = \mathbb{I}\{y \neq z\}$$

- In practice:
  - One Against All
  - One vs. One
  - Error-Correcting Output Coding

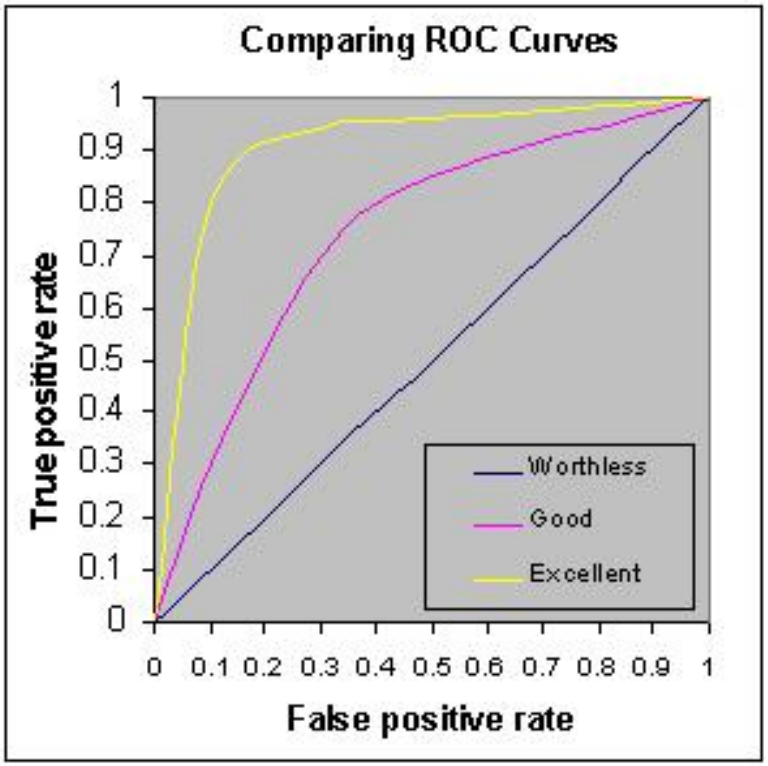
Example 5 - Unsupervised learning

- No label information Y
- Statistical model: $\{p(x, \theta) \ : \ \theta \in \Theta\}$
- Recover the density function of X based on D_n
- Loss function:
$$\ell(x, \theta) = -\log p(x, \theta)$$
- Applications: clustering, modes vs. anomaly detection
- Subproblem: Level set estimation

Example 6 - Ranking and scoring

- Classification data
- Set $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$
- Prediction based on scoring rules $s : \mathcal{X} \to \mathbb{R}$
- Goal: find s which ranks as η

Example 6 - Scoring and ROC Curves



- True positive rate:

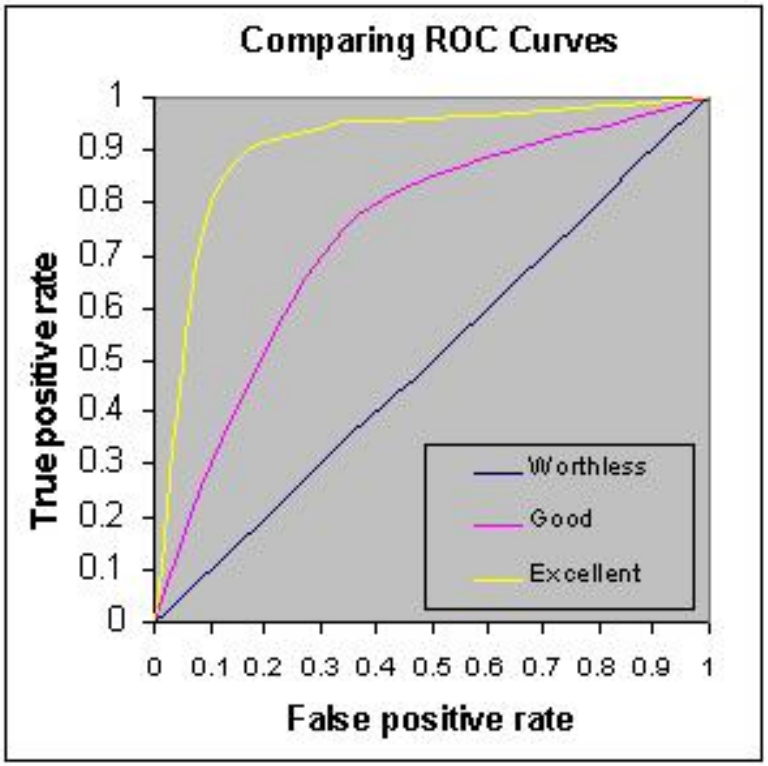
$$\mathrm{TPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $x \mapsto \left(\mathrm{FPR}_s(x), \mathrm{TPR}_s(x)\right)$

Example 6 - Scoring and ROC Curves



- True positive rate:

$$\mathrm{TPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = 1\right)$$

- False positive rate:

$$\mathrm{FPR}_s(x) = \mathbb{P}\left(s(X) \geq x \mid Y = -1\right)$$

Receiving Operator Characteristic curve: $x \mapsto \left(\mathrm{FPR}_s(x), \mathrm{TPR}_s(x)\right)$

AUC = Area Under an ROC Curve

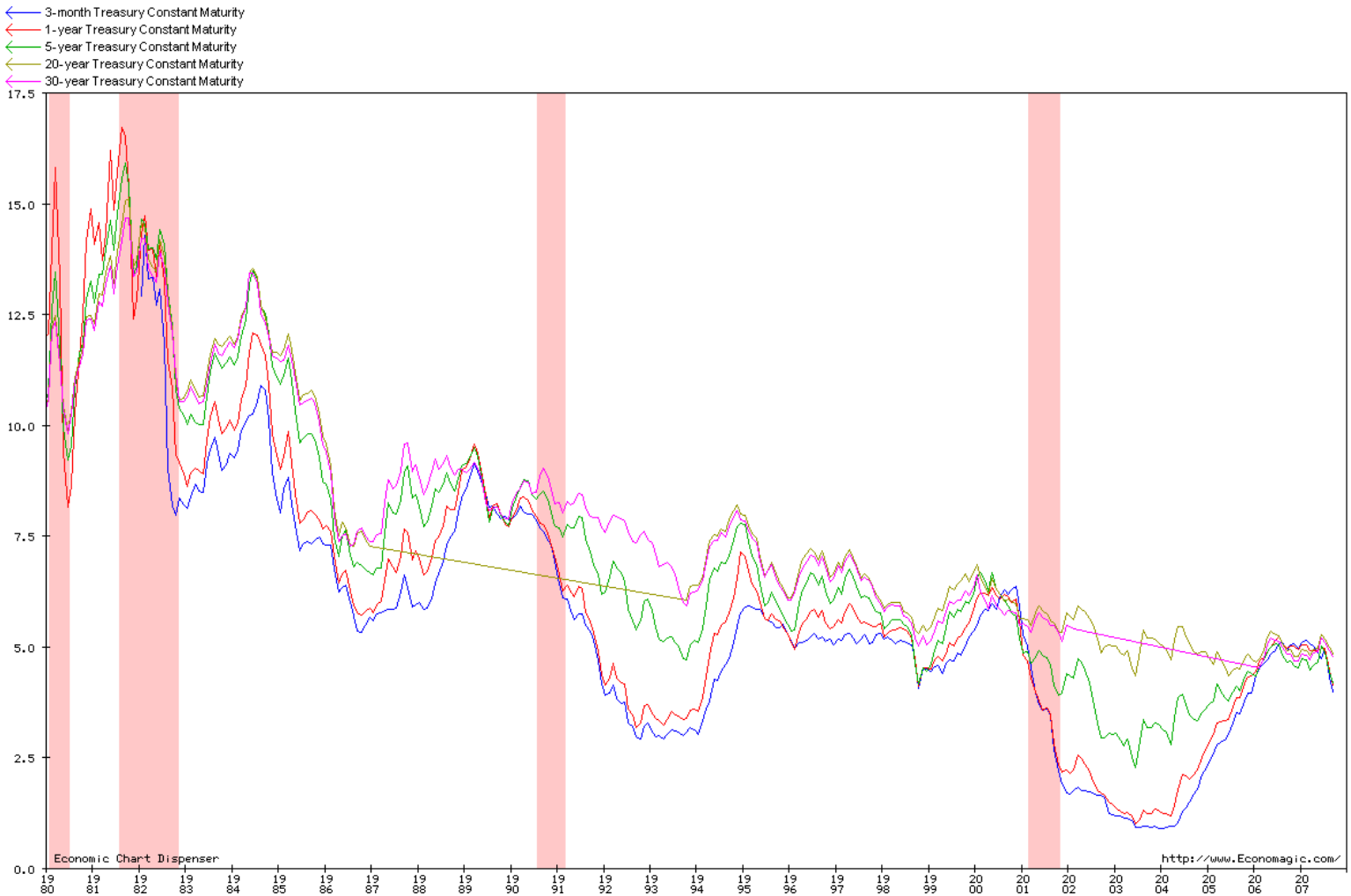
Data analysis

- Standard tools revisited
- Nonlinear PCA, kernel PCA
- Sparse PCA
- Independent Component Analysis

Réduction de la dimension

Exemple 1 - Finance

Analyse des taux d'intérêt



Exemple 1 - Finance (2)

- Variables = 18 maturités
  = 1M, 3M, 6M, 9M, 1Y, 2Y, ..., 30Y

- Observations = Historique mensuel sur 8 ans
  = 96 valeurs

Exemple 2 - Web 2.0

Last-FM - webradio de type collaboratif

Exemple 2 - Web 2.0 (2)

- 28302 artistes et leurs "tags"
- Variables = 735 tags
  = trance, techno, ambient, alternative,
  rap metal, rock, ...
- Observations = 2840 utilisateurs

Exemple 3 - Reconnaissance de visages

Exemple 3 - Reconnaissance de visages (2)

- Variables = 256 x 256 pixels
- Observations = 64 images

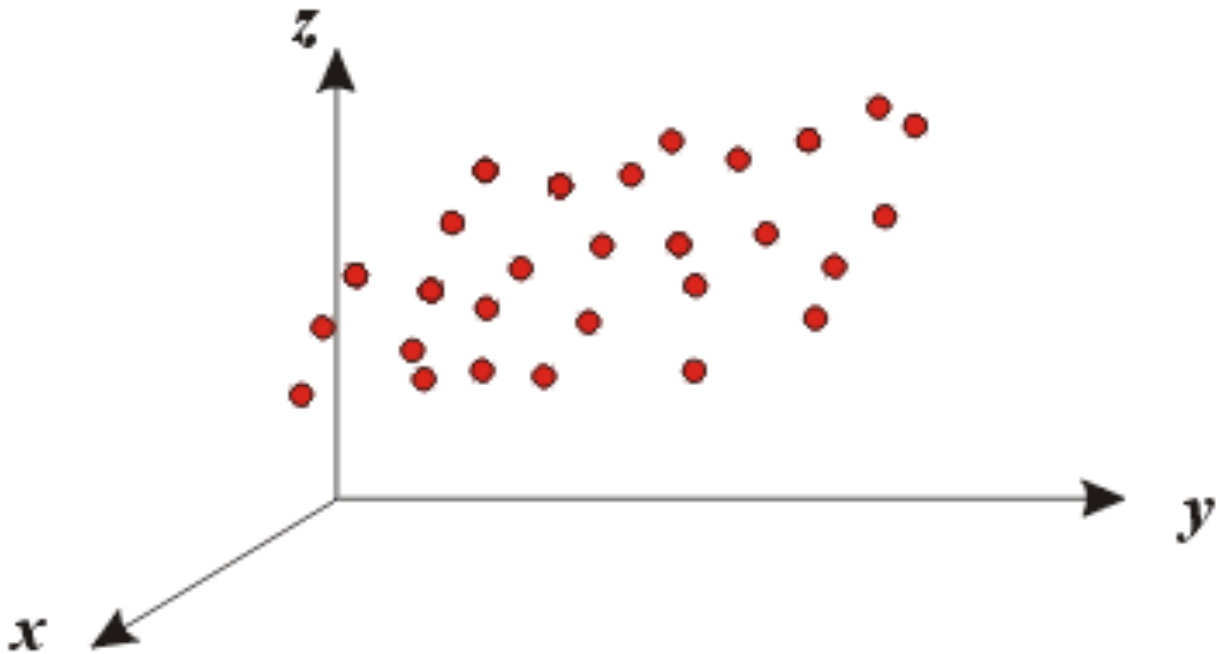
Traits communs

- Données multivariées
- Besoin d'interprétation
- Variabilité expliquée par des combinaisons de variables

Données

- Dimension = nombre de variables = p
- Taille de l'échantillon = nombre d'observations = n
- Tableau $n \times p$ de variables quantitatives

Représentation graphique



$\Rightarrow$ Nuage de n points dans $\mathbb{R}^p$

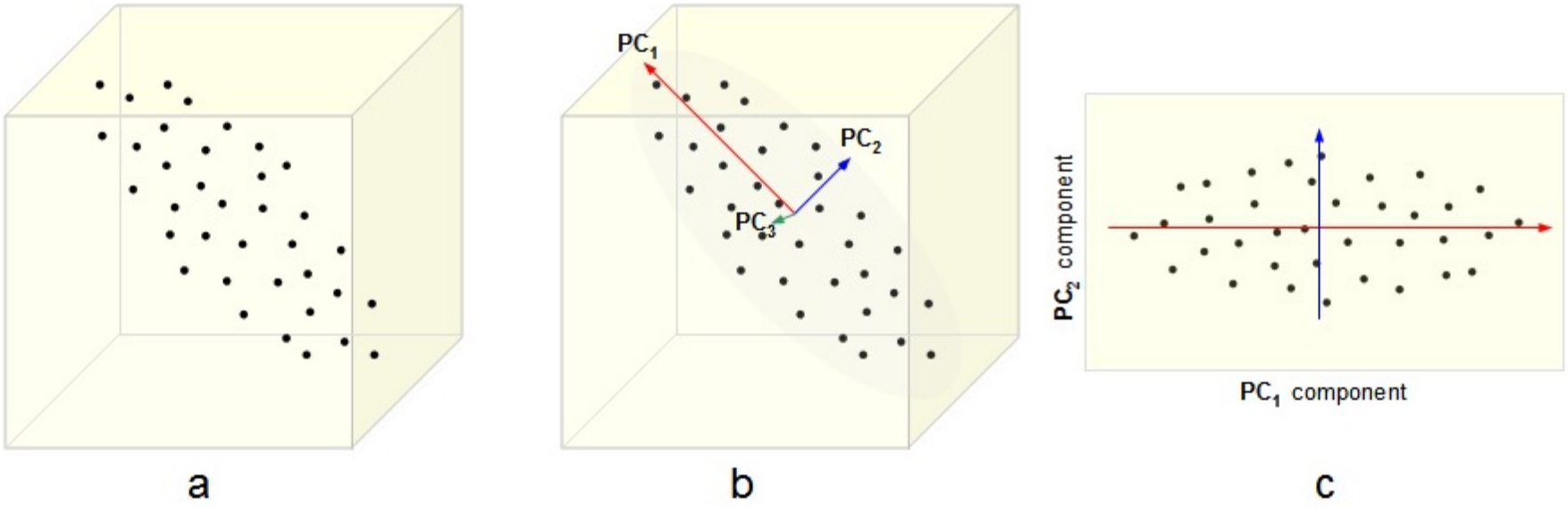
Objectifs

- Réduction de la dimension
- Visualisation du nuage en 2D ou 3D
- Explication de la variabilité

Analyse en Composantes Principales (ACP)

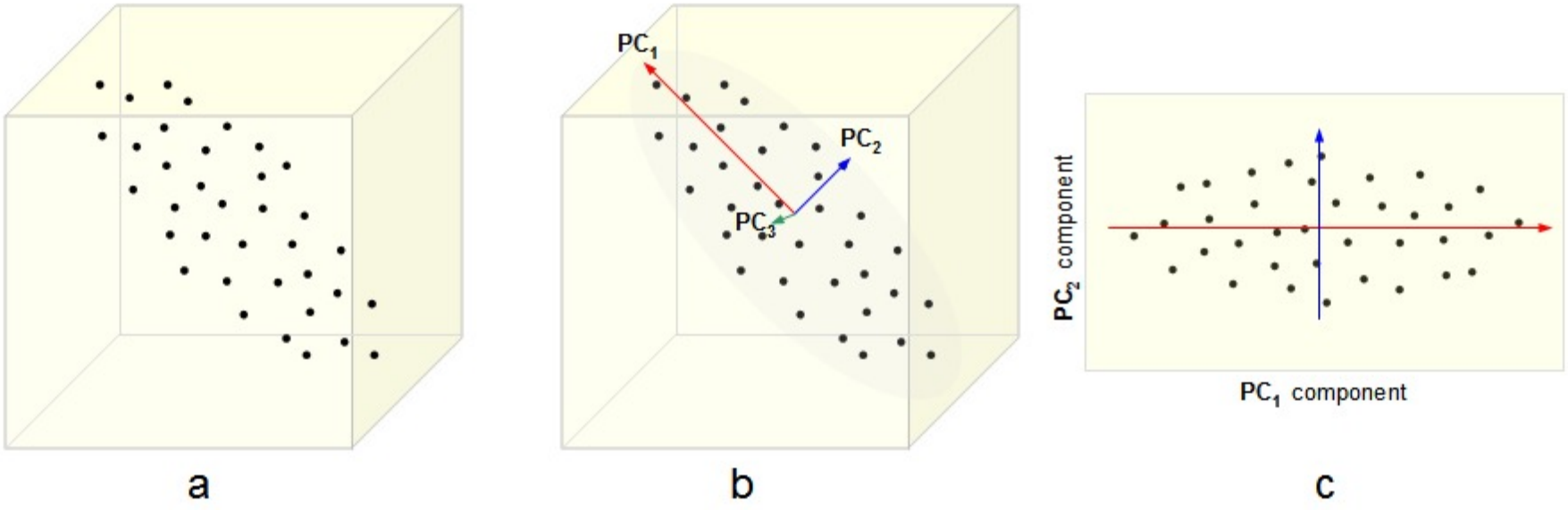
Philosophie de l'ACP

$\longrightarrow$ Projeter le nuage selon la "bonne" direction



Philosophie de l'ACP

$\longrightarrow$ Projeter le nuage selon la "bonne" direction



Idée : maximiser la dispersion

Cadre statistique : Tableau de données

- Observations : $X_i \in \mathbb{R}^p \quad , \quad 1 \leq i \leq n$
- Variable j : $X_{1j}, \ldots, X_{nj}$
- Matrice $n \times p$ de données $X = (X_1, \ldots, X_n)^T$

$$X = (X_{ij})_{i,j} = \begin{pmatrix} X_{11} & \ldots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \ldots & X_{np} \end{pmatrix}$$

Matrice de covariance empirique

- Barycentre

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \in \mathbb{R}^p$$

- Matrice de covariance empirique $(p \times p)$

$$S = (s_{kj})_{k,j} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T - \bar{X} \bar{X}^T$$

Meilleure direction

- Direction de projection $a \in \mathbb{R}^p$
- Echantillon (1D) $= (a^T X_1, \ldots, a^T X_n)$
- Maximiser la variance empirique en a :

$$s_a^2 = a^T S a$$

- Solution :

  vecteur propre $g_{(1)}$ de la plus grande valeur propre l_1

Diagonalisation de S symétrique réelle

- Valeurs propres : $l_1 \geq \ldots \geq l_p$
- Vecteurs propres orthonormés $g_{(1)}, \ldots, g_{(p)}$
- Réduction de la matrice $S = GLG^T$ où
  - $L = diag(l_1, \ldots, l_p)$ matrice diagonale $p \times p$
  - G matrice orthogonale $p \times p$

$$G = (g_{(1)}, \ldots, g_{(p)}) = (g_{kj})_{k,j}$$

Composantes principales (CP)

- Composantes principales : pour tout vecteur $z \in \mathbb{R}^p$

$$y_j(z) = g_{(j)}^T(z - \bar{X}) \ , \quad 1 \leq j \leq p$$

- La matrice $n \times p$

$$Y = (y_j(X_i))_{1 \leq i \leq n,\ 1 \leq j \leq p}$$

remplace la matrice X des données initiales.

Corrélation empirique "Variable vs. CP"

- Corrélations empiriques entre la variable k et la CP y_j :

$$\tilde{r}_{kj} = g_{kj}\sqrt{\frac{l_j}{s_{kk}}} \qquad \text{(définition)}$$

- Propriété :

$$\sum_{j=1}^{p} \tilde{r}_{kj}^2 = 1$$

Variance empirique de la k-ème variable

- Part de la variance empirique de la k-ème variable expliquée par les 2 premières CP (y_1, y_2) :

$$\tilde{r}_{k1}^2 + \tilde{r}_{k2}^2$$

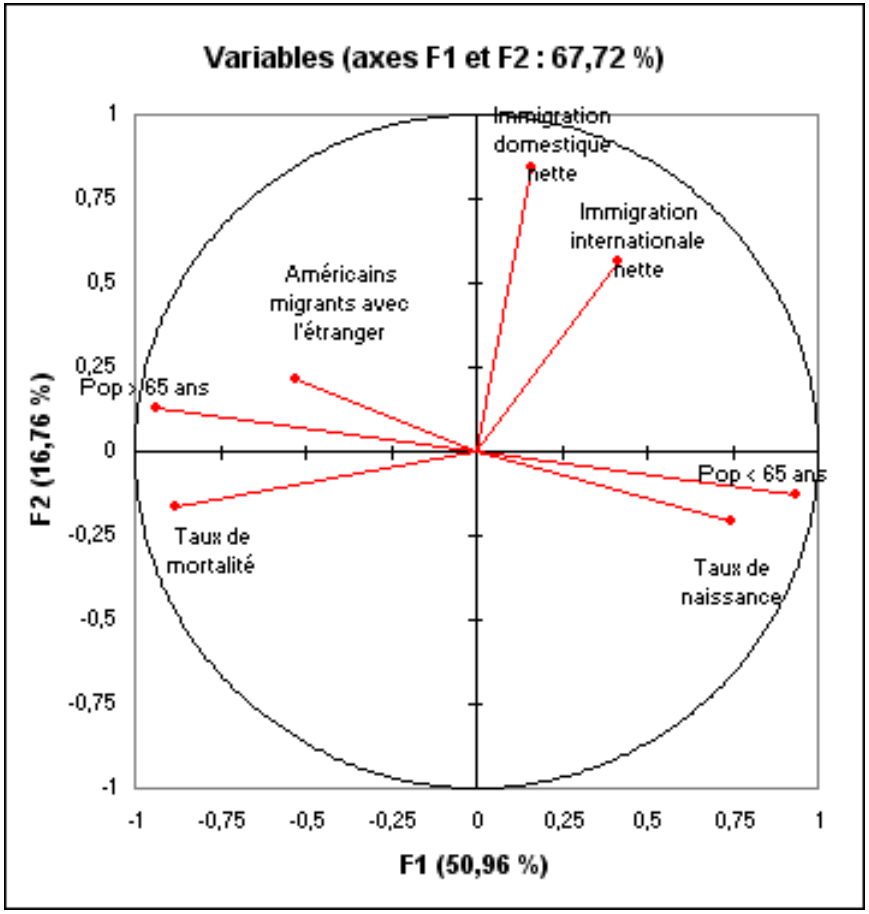
- On a :

$$I_1 + I_2 = \sum_{k=1}^{p} s_{kk}(\tilde{r}_{k1}^2 + \tilde{r}_{k2}^2)$$

- Visualisation 2D : Disque des corrélations

Disque des corrélations

- Point $(\tilde{r}_{k1}, \tilde{r}_{k2})$ correspond la variable k



Variance empirique des données

- Part de la variance empirique du nuage de points expliquée par la CP y_j:

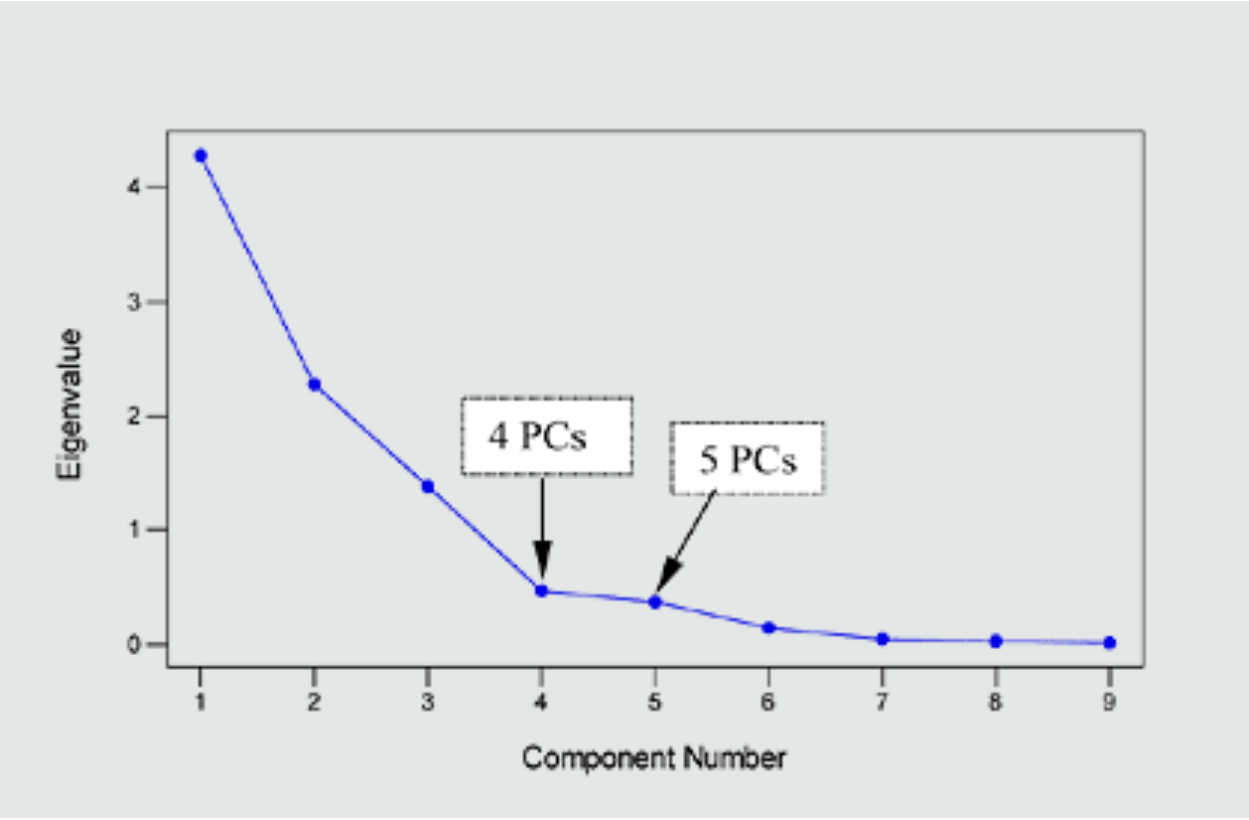
$$v_j = \frac{l_j}{Tr(S)}$$

où $Tr(S) = \sum_{j=1}^{p} l_j$.

- Visualisation : scree-graph

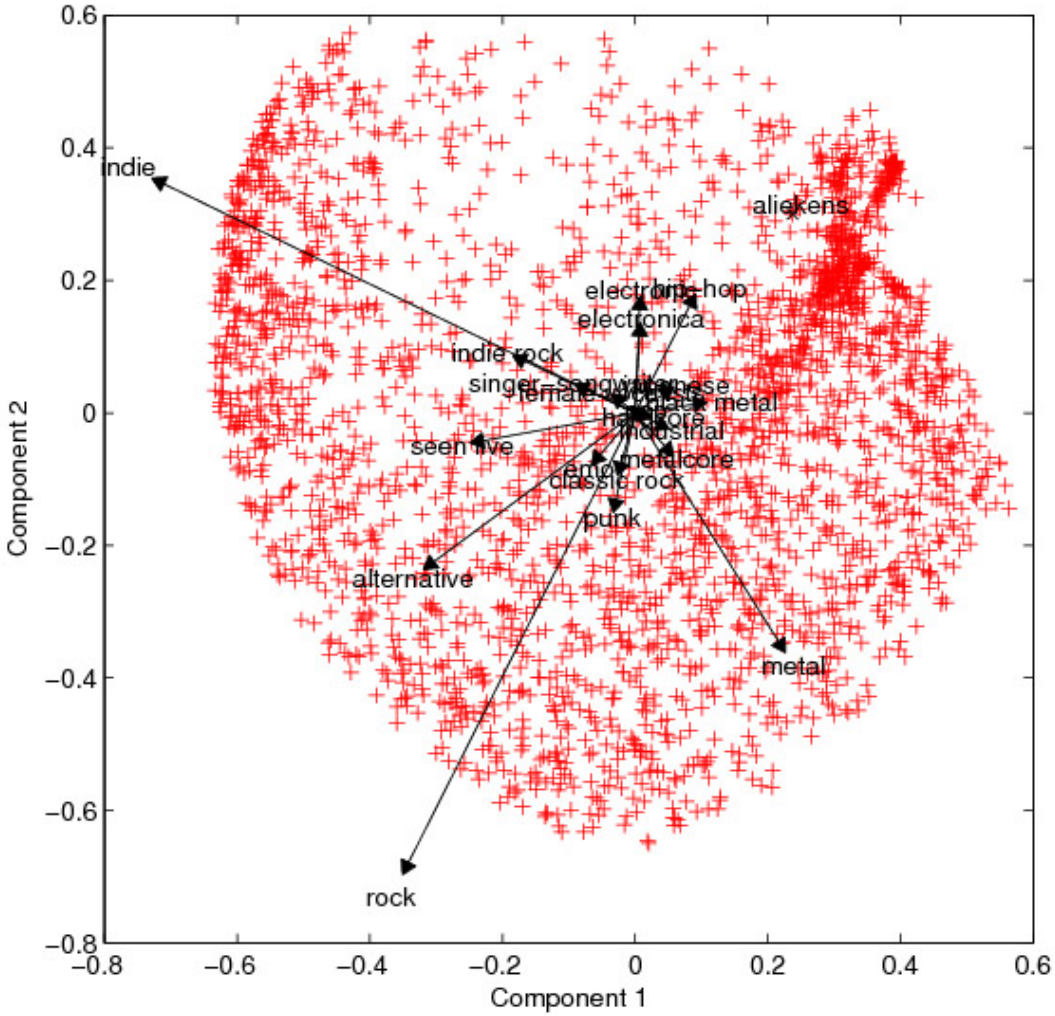
Scree-graph

- Axes = indice j de la CP et part de variance v_j



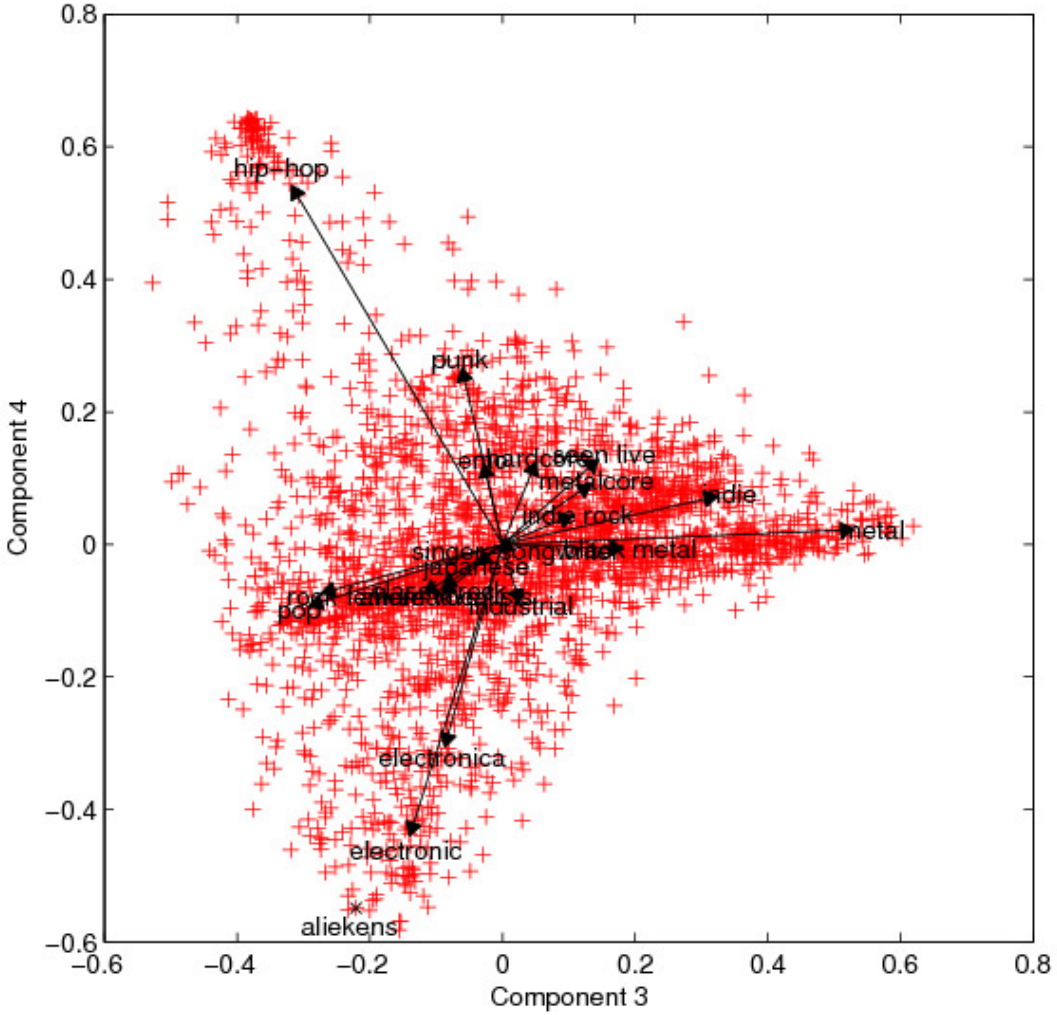
Résultats de l'ACP - Last-FM (1)

Projection du nuage de points sur (CP1, CP2)



Résultats de l'ACP - Last-FM (2)

Projection du nuage de points sur (CP3, CP4)

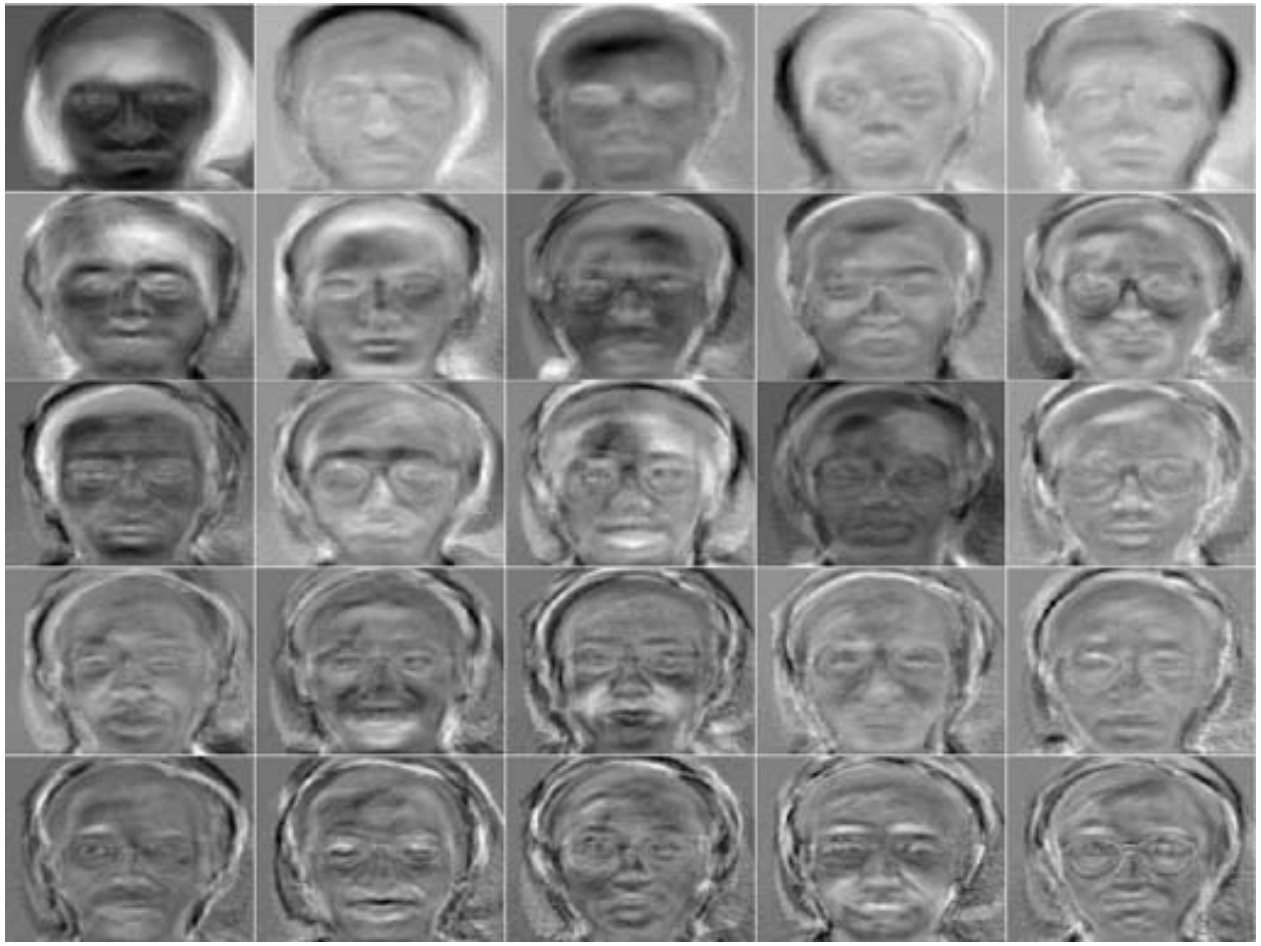


Résultats de l'ACP - Visages (1)

Données

Résultats de l'ACP - Visages (2)

"Visages propres"



Résultats de l'ACP - Visages (3)

Reconstruction partielle (sous-colonne de la matrice Y)

Résultats de l'ACP - Visages (4)

Projection d'autres images

Quelques remarques

- ACP = outil linéaire
- Orthogonalité des composantes principales
- **En pratique :**

  Réduction de la matrice $R = (r_{kj})_{k,j}$ des corrélations

$$r_{kj} = \frac{s_{kj}}{\sqrt{s_{kk} s_{jj}}}$$

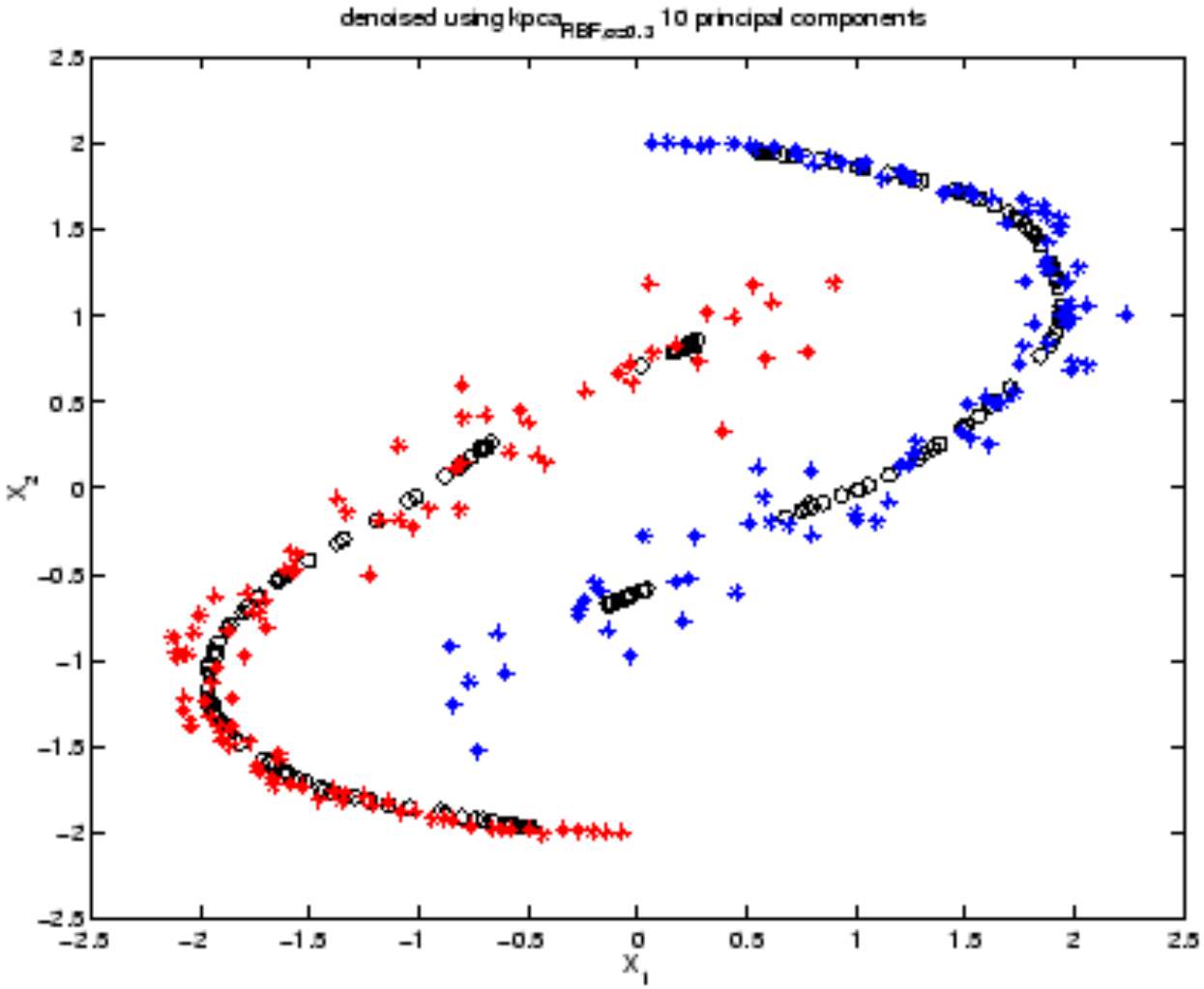
- **Obstacle numérique :**

  Réduction de S en très grande dimension

Quand est-ce que ça marche ?

- Nuages de points ellipsoïdaux
- Modèle implicite = modèle gaussien
- Information portée par les statistiques d'ordre 2
- Absence de valeurs aberrantes

Echec de l'ACP



$\Rightarrow$ **Extension :** ACP non-linéaire (à noyau)

Noyaux positifs

Définition

Soit $\mathcal{X}$ l'espace où vivent les observations.

Noyau positif

Une fonction $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau positif si et seulement si

1. k est symétrique: $k(x, x') = k(x', x)$, $\forall x, x' \in \mathcal{X}$
2. k est positive:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \quad \forall c_i \in \mathbb{R}, \quad \forall x_i \in \mathcal{X}, \quad \forall n \geq 1$$

Un théorème d'analyse

Théorème de Mercer

Pour tout noyau positif k sur $\mathcal{X}$ il existe un espace de Hilbert $\mathcal{H}$ et une application Φ tels que:

$$k(x, x') = < \Phi(x), \Phi(x') >, \quad \forall x, x' \in \mathcal{X}$$

où $< , >$ représente le produit scalaire sur $\mathcal{H}$.

Commentaires

- Le théorème de Mercer est non constructif: il ne fournit ni $\mathcal{H}$, ni Φ

Commentaires

- Le théorème de Mercer est non constructif: il ne fournit ni $\mathcal{H}$, ni Φ
- En pratique:

Commentaires

- Le théorème de Mercer est non constructif: il ne fournit ni $\mathcal{H}$, ni Φ
- En pratique:
  - $\mathcal{H}$ est un espace de grande dimension

Commentaires

- Le théorème de Mercer est non constructif: il ne fournit ni $\mathcal{H}$, ni Φ
- En pratique:
  - $\mathcal{H}$ est un espace de grande dimension
  - Φ est une application non-linéaire

Commentaires

- Le théorème de Mercer est non constructif: il ne fournit ni $\mathcal{H}$, ni Φ
- En pratique:
  - $\mathcal{H}$ est un espace de grande dimension
  - Φ est une application non-linéaire
- $\mathcal{H}$ est un espace de représentation des données, connu sous le nom de "feature space" ou espace de caractéristiques

Commentaires

- Le théorème de Mercer est non constructif: il ne fournit ni $\mathcal{H}$, ni Φ
- En pratique:
  - $\mathcal{H}$ est un espace de grande dimension
  - Φ est une application non-linéaire
- $\mathcal{H}$ est un espace de représentation des données, connu sous le nom de "feature space" ou espace de caractéristiques
- L'astuce du noyau consiste à faire l'impasse sur $\mathcal{H}$ et Φ si on sait qu'ils existent!

Distances images

- Norme euclidienne sur $\mathbb{R}^m$: $\forall u \in \mathbb{R}^m,\ \|u\| = \sqrt{< u, u >}$ où $< , >$ produit scalaire sur $\mathbb{R}^m$

Distances images

- Norme euclidienne sur $\mathbb{R}^m$: $\forall u \in \mathbb{R}^m,\ \|u\| = \sqrt{< u, u >}$ où $< , >$ produit scalaire sur $\mathbb{R}^m$
- Distance euclidienne:
  $d(u, v) = \|u - v\| = \sqrt{< u, u > + < v, v > -2 < u, v >}$

Distances images

- Norme euclidienne sur $\mathbb{R}^m$: $\forall u \in \mathbb{R}^m,\ \|u\| = \sqrt{< u, u >}$ où $< , >$ produit scalaire sur $\mathbb{R}^m$
- Distance euclidienne:
  $d(u, v) = \|u - v\| = \sqrt{< u, u > + < v, v > -2 < u, v >}$
- Transformation non-linéaire : $\Phi\ :\ \mathbb{R}^d \rightarrow \mathbb{R}^m$ avec $m > d$

Distances images

- Norme euclidienne sur $\mathbb{R}^m$: $\forall u \in \mathbb{R}^m,\ \|u\| = \sqrt{< u, u >}$ où $< , >$ produit scalaire sur $\mathbb{R}^m$
- Distance euclidienne:
  $d(u, v) = \|u - v\| = \sqrt{< u, u > + < v, v > -2 < u, v >}$
- Transformation non-linéaire : $\Phi\ :\ \mathbb{R}^d \rightarrow \mathbb{R}^m$ avec $m > d$
- Noyau: $k(x, x') =< \Phi(x), \Phi(x') >$

Distances images

- Norme euclidienne sur $\mathbb{R}^m$: $\forall u \in \mathbb{R}^m,\ \|u\| = \sqrt{< u, u >}$ où $< , >$ produit scalaire sur $\mathbb{R}^m$
- Distance euclidienne:
  $d(u, v) = \|u - v\| = \sqrt{< u, u > + < v, v > -2 < u, v >}$
- Transformation non-linéaire : $\Phi\ :\ \mathbb{R}^d \rightarrow \mathbb{R}^m$ avec $m > d$
- Noyau: $k(x, x') =< \Phi(x), \Phi(x') >$
- Distance image:

$$d_\Phi(x, x') = \|\Phi(x) - \Phi(x')\| = \sqrt{k(x, x) + k(x', x') - 2k(x, x')}$$

$\Rightarrow$ la distance induite par Φ ne fait intervenir que le noyau

Intérêt pour la classification

- Aucune complication algorithmique en remplaçant le produit scalaire par une autre mesure de similarité

Intérêt pour la classification

- Aucune complication algorithmique en remplaçant le produit scalaire par une autre mesure de similarité
- Transformer un problème initialement non-linéaire en un problème linéaire en envoyant les données dans un espace plus grand

Intérêt pour la classification

- Aucune complication algorithmique en remplaçant le produit scalaire par une autre mesure de similarité
- Transformer un problème initialement non-linéaire en un problème linéaire en envoyant les données dans un espace plus grand

Intérêt pour la classification

- Aucune complication algorithmique en remplaçant le produit scalaire par une autre mesure de similarité
- Transformer un problème initialement non-linéaire en un problème linéaire en envoyant les données dans un espace plus grand

Exemple

Soit $f(x, y) = ax^2 + bx + c - y = 0$ une surface de décision polynomiale (parabole dans $\mathbb{R}^2$).

Intérêt pour la classification

- Aucune complication algorithmique en remplaçant le produit scalaire par une autre mesure de similarité
- Transformer un problème initialement non-linéaire en un problème linéaire en envoyant les données dans un espace plus grand

Exemple

Soit $f(x, y) = ax^2 + bx + c - y = 0$ une surface de décision polynomiale (parabole dans $\mathbb{R}^2$).

Rôle clé de la transformation:

$$\begin{array}{rccc} \Phi & : & \mathbb{R}^2 & \rightarrow & \mathbb{R}^4 \\ & & x & \mapsto & \left(x^2, x, 1, y\right)^T \end{array}$$

Du non-linéaire au linéaire

Exemple (suite)

On peut écrire:

$$g(x^2, x, 1, y) = ax^2 + bx + c - y = 0$$

où $g(u, v, w, y) = au + bv + cw - y$.

L'équation $g(u, v, w, y) = 0$ définit une surface de décision linéaire dans $\mathbb{R}^4$.

Du non-linéaire au linéaire

Exemple (suite)

On peut écrire:

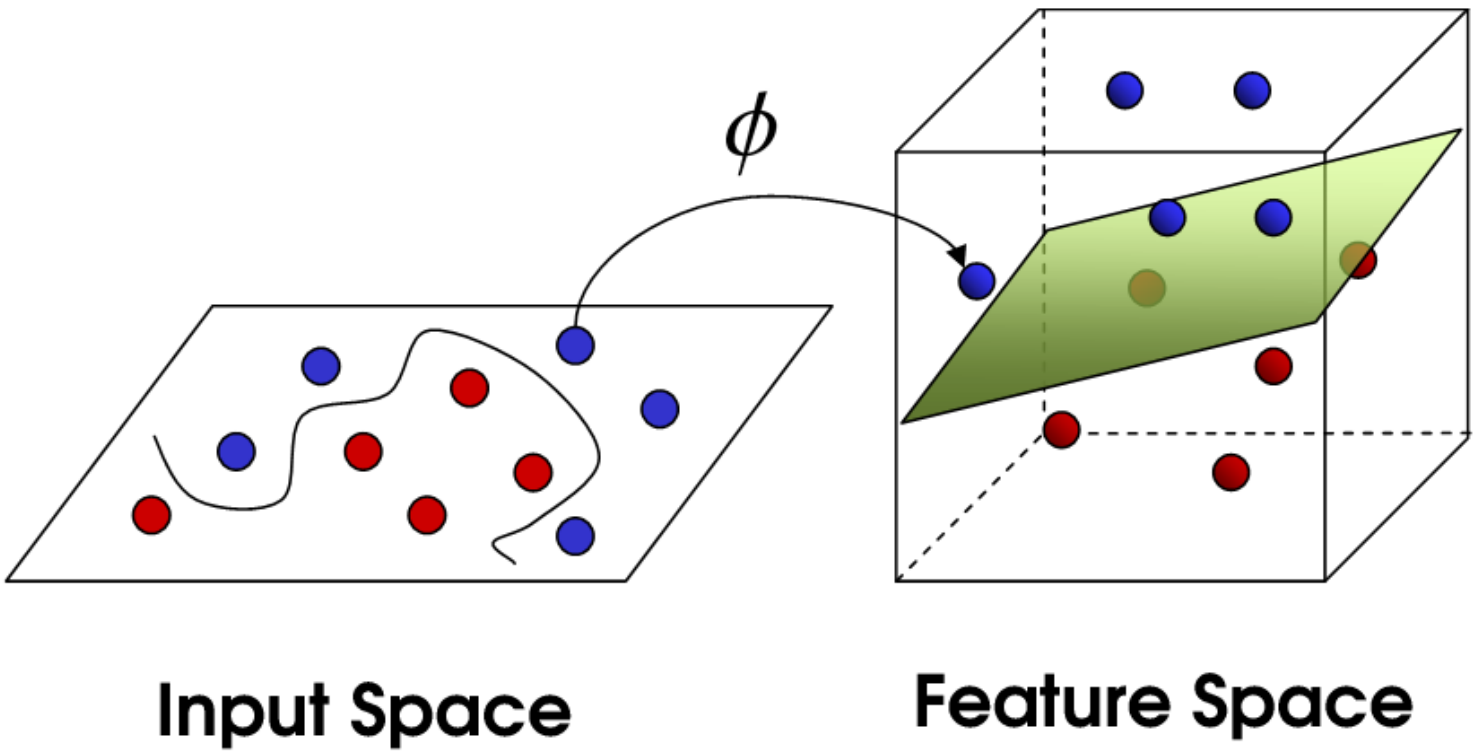
$$g(x^2, x, 1, y) = ax^2 + bx + c - y = 0$$

où $g(u, v, w, y) = au + bv + cw - y$.

L'équation $g(u, v, w, y) = 0$ définit une surface de décision linéaire dans $\mathbb{R}^4$.

Un problème non-linéaire dans un certain espace peut parfois se formuler comme un problème linéaire dans un espace plus grand.

Du non-linéaire au linéaire



ACP à noyau

ACP classique

On considère un nuage de points $x_1, \ldots, x_n$ centrés en l'origine.

ACP classique

On considère un nuage de points $x_1, \ldots, x_n$ centrés en l'origine.

Buts de l'ACP

- méthode de visualisation des données
- réduction de la dimension effective des données

ACP classique

On considère un nuage de points $x_1, \ldots, x_n$ centrés en l'origine.

Buts de l'ACP

- méthode de visualisation des données
- réduction de la dimension effective des données

L'ACP consiste à identifier les composantes principales de l'échantillon constituées par

1. la meilleure direction de projection du nuage de points
   i.e. celle de variance maximale

ACP classique

On considère un nuage de points $x_1, \ldots, x_n$ centrés en l'origine.

Buts de l'ACP

- méthode de visualisation des données
- réduction de la dimension effective des données

L'ACP consiste à identifier les composantes principales de l'échantillon constituées par

1. la meilleure direction de projection du nuage de points
   i.e. celle de variance maximale
2. puis, la meilleure direction de projection orthogonale à la première

ACP classique

On considère un nuage de points $x_1, \ldots, x_n$ centrés en l'origine.

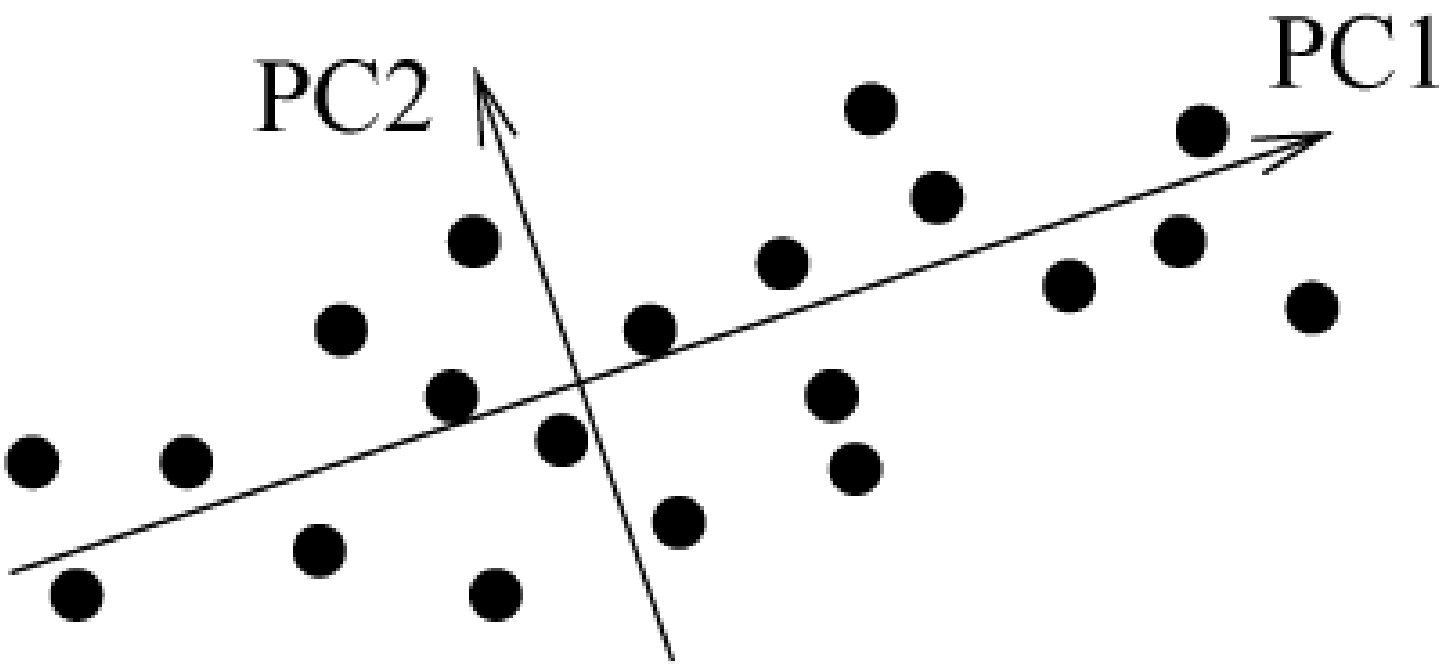
Buts de l'ACP

- méthode de visualisation des données
- réduction de la dimension effective des données

L'ACP consiste à identifier les composantes principales de l'échantillon constituées par

1. la meilleure direction de projection du nuage de points
   i.e. celle de variance maximale
2. puis, la meilleure direction de projection orthogonale à la première
3. et, ainsi de suite, jusqu'à la n-ième

ACP



ACP (suite)

- Projection orthogonale d'un vecteur x sur la direction $w \in \mathbb{R}^d$:

$$p_w(x) = \frac{< x, w >}{\|w\|}$$

ACP (suite)

- Projection orthogonale d'un vecteur x sur la direction $w \in \mathbb{R}^d$:

$$p_w(x) = \frac{< x, w >}{\|w\|}$$

- Variance empirique du nuage de points selon la direction w:

$$\mathbb{V}(p_w) = \frac{1}{n} \sum_{i=1}^{n} \frac{< x_i, w >^2}{\|w\|^2}$$

ACP (suite)

- Projection orthogonale d'un vecteur x sur la direction $w \in \mathbb{R}^d$:

$$p_w(x) = \frac{< x, w >}{\|w\|}$$

- Variance empirique du nuage de points selon la direction w:

$$\mathbb{V}(p_w) = \frac{1}{n} \sum_{i=1}^{n} \frac{< x_i, w >^2}{\|w\|^2}$$

- Matrice de covariance empirique $\Sigma = \frac{1}{n} \sum_{i=1}^{n} x_i \; x_i^T$

ACP (suite)

- Projection orthogonale d'un vecteur x sur la direction $w \in \mathbb{R}^d$:

$$p_w(x) = \frac{< x, w >}{\|w\|}$$

- Variance empirique du nuage de points selon la direction w:

$$\mathbb{V}(p_w) = \frac{1}{n} \sum_{i=1}^{n} \frac{< x_i, w >^2}{\|w\|^2}$$

- Matrice de covariance empirique $\Sigma = \frac{1}{n} \sum_{i=1}^{n} x_i \; x_i^T$
- On a donc :

$$\mathbb{V}(p_w) = \frac{w^T \Sigma w}{\|w\|^2}$$

Problème d'optimisation

Première composante principale

$$\arg\max_{w} \mathbb{V}(p_w) = \frac{w^T \Sigma w}{\|w\|^2}$$

Problème d'optimisation

Première composante principale

$$\arg\max_{w} \mathbb{V}(p_w) = \frac{w^T \Sigma w}{\|w\|^2}$$

Solution

Les composantes principales sont les vecteurs propres de la Σ rangés selon la décroissance des valeurs propres correspondantes.

Problème d'optimisation

Première composante principale

$$\arg\max_{w} \mathbb{V}(p_w) = \frac{w^T \Sigma w}{\|w\|^2}$$

Solution

Les composantes principales sont les vecteurs propres de la Σ rangés selon la décroissance des valeurs propres correspondantes.

Remarque : la matrice Σ est symétrique réelle donc diagonalisable dans une base orthonormée.

ACP (suite)

On cherche un vecteur v et un réel λ tels que:

$$\Sigma v = \lambda v$$

Or, on a :

$$\Sigma v = \frac{1}{n} \sum_{i=1}^{n} < x_i, v > \ x_i$$

D'où:

$$v = \sum_{i=1}^{n} \left(\frac{< x_i, v >}{n\lambda} \right) x_i = \sum_{i=1}^{n} \alpha_i x_i$$

ACP (suite)

On cherche un vecteur v et un réel λ tels que:

$$\Sigma v = \lambda v$$

Or, on a :

$$\Sigma v = \frac{1}{n} \sum_{i=1}^{n} < x_i, v > \ x_i$$

D'où:

$$v = \sum_{i=1}^{n} \left(\frac{< x_i, v >}{n\lambda} \right) x_i = \sum_{i=1}^{n} \alpha_i x_i$$

On utilise

$$x_j^T \Sigma v = \lambda < x_j, v >, \quad \forall j$$

et on y substitue les expressions de Σ et v:

$$\frac{1}{n} \sum_{i=1}^{n} \alpha_i \left\langle x_j, \sum_{k=1}^{n} < x_k, x_i > \ x_k \right\rangle = \lambda \sum_{i=1}^{n} \alpha_i < x_j, x_i >$$

ACP (suite)

- On note $K = (< x_i, x_j >)_{i,j}$ la matrice de Gram

ACP (suite)

- On note $K = (< x_i, x_j >)_{i,j}$ la matrice de Gram
- On peut écrire alors le système:

$$K^2\alpha = n\lambda K\alpha$$

ACP (suite)

- On note $K = (< x_i, x_j >)_{i,j}$ la matrice de Gram
- On peut écrire alors le système:

$$K^2\alpha = n\lambda K\alpha$$

- Pour résoudre en α, on résout donc le problème aux éléments propres

$$K\alpha = n\lambda\alpha$$

Défauts de l'ACP classique

- elle est adaptée surtout au cas de réalisations de gaussiennes multivariées

Défauts de l'ACP classique

- elle est adaptée surtout au cas de réalisations de gaussiennes multivariées
  - en général, la non-corrélation n'implique pas l'indépendance des directions principales

Défauts de l'ACP classique

- elle est adaptée surtout au cas de réalisations de gaussiennes multivariées
  - en général, la non-corrélation n'implique pas l'indépendance des directions principales
  - alternative : Analyse en Composantes Indépendantes (plutôt que Principales)

Défauts de l'ACP classique

- elle est adaptée surtout au cas de réalisations de gaussiennes multivariées
  - en général, la non-corrélation n'implique pas l'indépendance des directions principales
  - alternative : Analyse en Composantes Indépendantes (plutôt que Principales)
- elle est adaptée aux structures linéaires

Défauts de l'ACP classique

- elle est adaptée surtout au cas de réalisations de gaussiennes multivariées
  - en général, la non-corrélation n'implique pas l'indépendance des directions principales
  - alternative : Analyse en Composantes Indépendantes (plutôt que Principales)
- elle est adaptée aux structures linéaires
  - les nuages de points ne sont pas tous ellipsoidaux!!

Défauts de l'ACP classique

- elle est adaptée surtout au cas de réalisations de gaussiennes multivariées
  - en général, la non-corrélation n'implique pas l'indépendance des directions principales
  - alternative : Analyse en Composantes Indépendantes (plutôt que Principales)
- elle est adaptée aux structures linéaires
  - les nuages de points ne sont pas tous ellipsoidaux!!
  - alternative : Kernel PCA

ACP à noyau

- On applique une transformation Φ qui envoie le nuage de points X dans un espace où la structure est linéaire

ACP à noyau

- On applique une transformation Φ qui envoie le nuage de points X dans un espace où la structure est linéaire
- La matrice de covariance de $\Phi(X) = (\Phi(x_1), \ldots, \Phi(x_n))^T$ est alors

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \Phi(x_i)^T$$

ACP à noyau

- On applique une transformation Φ qui envoie le nuage de points X dans un espace où la structure est linéaire
- La matrice de covariance de $\Phi(X) = (\Phi(x_1), \ldots, \Phi(x_n))^T$ est alors

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)\Phi(x_i)^T$$

- **Astuce du noyau :** $K = \left(k(x_i, x_j)\right)_{i,j} = \left(\Phi(x_i)^T \Phi(x_j)\right)_{i,j}$

ACP à noyau (suite)

- "Directions" principales de la forme:

$$p_i(x) = \sum_{j=1}^{n} \alpha_{i,j} k(x_j, x)$$

ACP à noyau (suite)

- "Directions" principales de la forme:

$$p_i(x) = \sum_{j=1}^{n} \alpha_{i,j} k(x_j, x)$$

- le vecteur $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,n})$ est solution du problème d'optimisation:

$$\max_{\alpha} \frac{\alpha^T K^2 \alpha}{\alpha^T K \alpha}$$

sous les contraintes: $\alpha_i^T K \alpha_j$ pour $j = 1, \ldots, i-1$

ACP à noyau (suite)

- "Directions" principales de la forme:

$$p_i(x) = \sum_{j=1}^{n} \alpha_{i,j} k(x_j, x)$$

- le vecteur $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,n})$ est solution du problème d'optimisation:

$$\max_{\alpha} \frac{\alpha^T K^2 \alpha}{\alpha^T K \alpha}$$

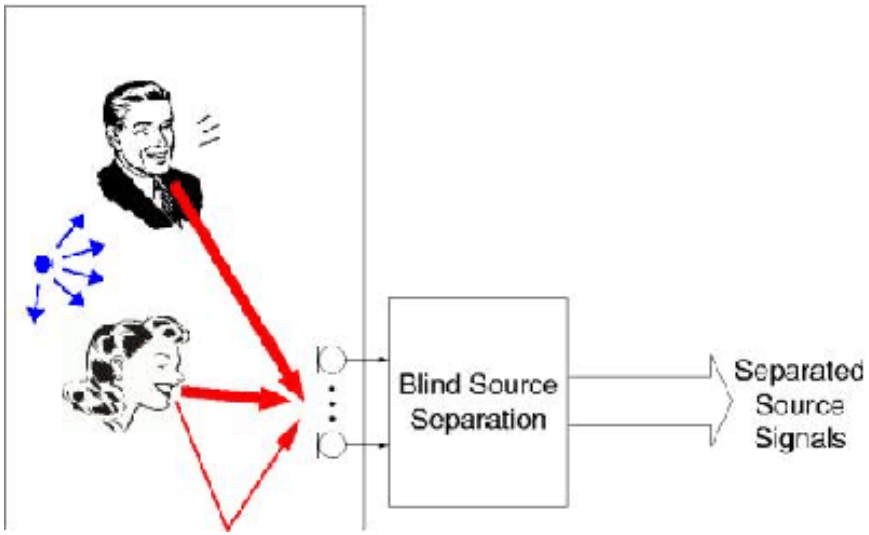
  sous les contraintes: $\alpha_i^T K \alpha_j$ pour $j = 1, \ldots, i-1$

- on résout le problème aux éléments propres:

$$K\alpha = n\lambda\alpha$$

Analyse en Composantes Indépendantes (ACI)

Problème du "cocktail-party"



- ACP = fondée sur la notion de corrélation
- Bonne notion = notion d'indépendance

Corrélation vs. Indépendance

- Or : X et Y indépendants $\Rightarrow \mathrm{cov}(X, Y) = 0$
- Réciproque fausse en général, sauf cas gaussien...
- De l'ACP vers l'ACI... (beaucoup plus difficile !)

Formulation du problème

- $S = (S_1, \ldots, S_d)^T$ sources indépendantes et non-gaussiennes inconnues
- $\mathbf{A}$ matrice de mélange $d \times d$ inconnue
- $X = (X_1, \ldots, X_d)^T$ observations (capteurs), on suppose $\mathrm{Cov}(X) = \mathbf{I}$
- On a le système : $X = \mathbf{A}S$
- On cherche $\mathbf{A}$ orthogonale telle que :

$$S = \mathbf{A}^T X \qquad \text{ait des composantes indépendantes}$$

Théorie de l'information

- Entropie d'une v.a. $Z \sim p(z)$:

$$H(Z) = -\mathbb{E}(\log(p(Z)))$$

- Considérons les v.a. T de variance v, alors

$$Z \sim \mathcal{N}(0,1) \quad \rightarrow \quad \max_{T} H(T)$$

- Information mutuelle pour $S = (S_1, \ldots, S_d)^T$:

$$I(S) = \sum_{i=1}^{d} H(S_i) - H(S)$$

ACI par méthode entropique

- Propriété de l'entropie : si $S = \mathbf{A}^{\mathbf{T}} X$

$$H(S) = H(X) + \log(|\det(\mathbf{A})|)$$

- On a donc le problème d'optimisation suivant :

$$\rightarrow \quad \min_{\mathbf{A}:\mathbf{A}^T\mathbf{A}=\mathbf{I}} I(\mathbf{A}^T X) = \sum_{i=1}^{d} H(S_i) - H(X)$$

- Interprétation : écart du comportement gaussien (minimisation de l'entropie des composantes)