

Analyse de données et Traitement de l'Information: Analyse en Composantes Principales

Stéphane Marchand-Maillet

Département d'Informatique
Université de Genève
`stephane.marchand-maillet@unige.ch`

Master en Sciences Informatiques - Semestre d'Automne

Introduction

Qu'est ce que l'Analyse de Données ?

- ▶ Analyse **statistique** de données
 - ▶ données multivariées (variables aléatoires multidimensionnelles)
 - ▶ analyses de la **distribution** des données
- ▶ Visualisation
- ▶ Compréhension
- ▶ Prévion
- ▶ Reconnaissance

Les données

Une **population** d'**individus** ou d'**éléments** décrite par des **variables**, **caractéristiques** ou **descripteurs**

- ▶ variables qualitatives (symboliques) :
diplôme, pays, occurrence
- ▶ variables quantitatives (numériques) :
age, chiffre d'affaire, intensité lumineuse

Les données sont des mesures physiques, sociologiques, informatiques,...

- ▶ données statiques
- ▶ données temporelles

Les données

Cardinalité et dimension

L'analyse de données s'applique lorsque :

- ▶ population importante
- ▶ nombreuses variables

Les données

Cardinalité et dimension

L'analyse de données s'applique lorsque :

- ▶ population importante
- ▶ nombreuses variables

Exemple : analyse d'images



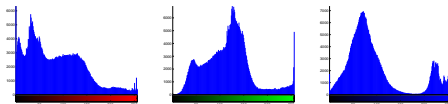
Les données

Cardinalité et dimension

L'analyse de données s'applique lorsque :

- ▶ population importante
- ▶ nombreuses variables

Exemple : analyse d'images



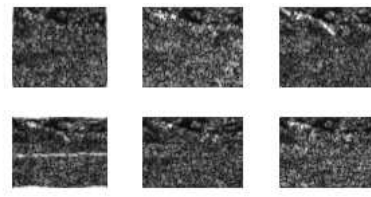
Les données

Cardinalité et dimension

L'analyse de données s'applique lorsque :

- ▶ population importante
- ▶ nombreuses variables

Exemple : analyse d'images



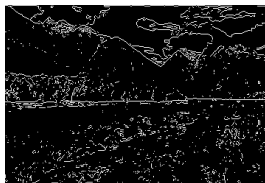
Les données

Cardinalité et dimension

L'analyse de données s'applique lorsque :

- ▶ population importante
- ▶ nombreuses variables

Exemple : analyse d'images



Les données

Imprécision de la mesure

Selon la procédure d'acquisition, les mesures peuvent être entachées d'erreur.

$$z = x + \eta$$

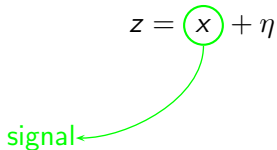
Les données

Imprécision de la mesure

Selon la procédure d'acquisition, les mesures peuvent être entachées d'erreur.

$$z = \textcircled{x} + \eta$$

signal ←



Les données

Imprécision de la mesure

Selon la procédure d'acquisition, les mesures peuvent être entachées d'erreur.

$$z = \underbrace{x}_{\text{signal}} + \underbrace{\eta}_{\text{bruit}}$$

Les données

Imprécision de la mesure

Selon la procédure d'acquisition, les mesures peuvent être entachées d'erreur.

$$z = \underbrace{x}_{\text{signal}} + \underbrace{\eta}_{\text{bruit}}$$

- ▶ x est la signal contenant l'information
- ▶ η est appelé *bruit*, et ne contient pas d'information

Les données

Estimation et régression

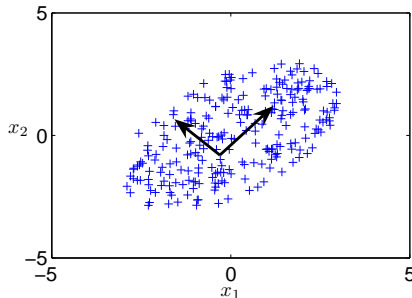
Du fait du caractère aléatoire et imprécis des données, le problème consiste à identifier des paramètres **statistiques** sur population étudiée.

Les données

Estimation et régression

Du fait du caractère aléatoire et imprécis des données, le problème consiste à identifier des paramètres **statistiques** sur population étudiée.

Les variables sont elles corrélées ?

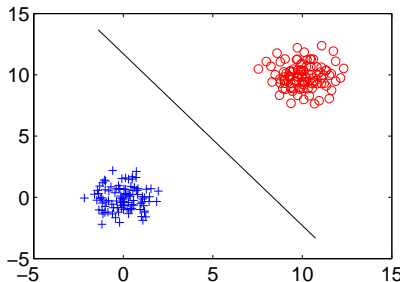


Les données

Estimation et régression

Du fait du caractère aléatoire et imprécis des données, le problème consiste à identifier des paramètres **statistiques** sur population étudiée.

Les données sont elles séparables ?

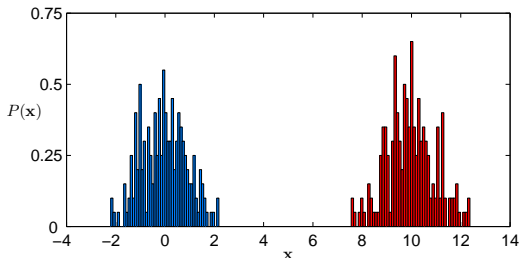


Les données

Estimation et régression

Du fait du caractère aléatoire et imprécis des données, le problème consiste à identifier des paramètres **statistiques** sur population étudiée.

Quelle est la distribution des données ?

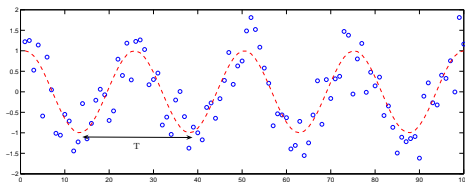


Les données

Estimation et régression

Du fait du caractère aléatoire et imprécis des données, le problème consiste à identifier des paramètres **statistiques** sur population étudiée.

Y a t'il des caractéristiques particulières aux mesures ?

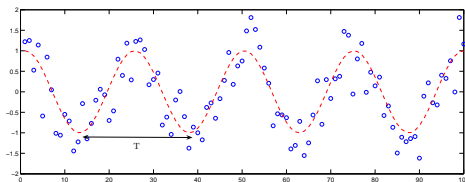


Les données

Estimation et régression

Du fait du caractère aléatoire et imprécis des données, le problème consiste à identifier des paramètres **statistiques** sur population étudiée.

Y a t'il des caractéristiques particulières aux mesures ?



⇒ L'analyse de donnée repose sur des techniques d'estimation et d'optimisation statistique

Les données

Analyse supervisée/non-supervisée

Selon les problèmes, il existe ou pas (ou partiellement) des données d'**entraînement**

- ▶ Problème supervisé : un label est associé à un sous-ensemble de données permettant l'**apprentissage** d'algorithmes de traitement des données (automates d'apprentissages).
- ▶ Problème non-supervisé : aucune information n'est disponible autre que les variables issues des données (clusterisation automatique, réduction de dimensions,...)

Applications

- ▶ **Reconnaissance de Formes → recherche et reconnaissance de motifs (pattern)**
 - ▶ Vision
 - ▶ Parole
 - ▶ Robotique
 - ▶ ...

Applications

- ▶ **Reconnaissance de Formes → recherche et reconnaissance de motifs (pattern)**
 - ▶ Vision
 - ▶ Parole
 - ▶ Robotique
 - ▶ ...
- ▶ **Prévisions**
 - ▶ Statistiques financières
 - ▶ Télécom
 - ▶ ...

Applications

- ▶ Recherche d'information
 - ▶ Web, Web 2.0
 - ▶ Multimedia
 - ▶ ...

Applications

- ▶ Recherche d'information
 - ▶ Web, Web 2.0
 - ▶ Multimedia
 - ▶ ...
- ▶ Datamining
 - ▶ Bio-informatique
 - ▶ Business intelligence
 - ▶ Sciences humaines, économiques

Quelques livres

- ▶ Pattern Classification, Richard O. Duda, Peter E. Hart, David G. Stork.
- ▶ Pattern Recognition and Machine Learning, Christopher M. Bishop
- ▶ The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome H. Friedman

Représentation des données, espaces de données

Définitions

- ▶ données : éléments, individus, objets
→ constituent une population Ω
- ▶ variables : mesures, descripteurs, caractères
→ projettent Ω dans un espace \mathcal{F}

$$p_i \in \Omega \rightarrow \mathbf{x}_i \in \mathcal{F}$$

- ▶ \mathcal{F} est un "espace de description/représentation" de Ω

Espace de représentation

La topologie de \mathcal{F} dépend des caractères mesurés

- ▶ $\mathcal{F} \subset \mathbb{R}^N, \mathcal{F} \subset \mathbb{Z}^N$: variables numériques
 - quantitatif mesurable : revenu, poids, ...*
 - quantitatif d'ordre : note, rang, ...*
 - quantitatif de comptage : fréquence, contingence, ...*
 - quantitatif binaire : succès-échec, présence-absence, ...*
- ▶ $\mathcal{F} \subset S = \{A, B, C, \dots\}$: variables symboliques
 - qualitatif nominal : lieu géographique, catégorie socioprofessionnelle, ...*
 - qualitatif ordinal : pas d'accord, sans opinion, ...*
 - qualitatif textuel : titre de film, nom d'auteur, ...*

Espace de représentation

Mesures de similarités

L'espace \mathcal{F} est muni d'une métrique

A tout couple $\{p_i, p_j\} \in \Omega^2$ est associée une mesure $d(\mathbf{x}_i, \mathbf{x}_j)$ qui indique la **proximité** des deux éléments correspondants dans \mathcal{F} .

Exemples de mesures

- ▶ Euclidienne (L_2), $d = \sqrt{\sum_k^N (x_i^k - x_j^k)^2}$: dissimilarité
- ▶ Dirac $\delta(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & \text{si } \mathbf{x}_i = \mathbf{x}_j \\ 0 & \text{sinon} \end{cases}$: similarité
- ▶ divergence de Kullback, distance de Mahalanobis, test du χ^2 ...

⇒ Il est possible de définir une infinité de (dis)similarités

Espace de représentation

Tableau quantitatif

Tableaux explicitant les variables des éléments

$X =$

VARIABLES	ELEMENTS			
	p_1	p_2	\cdots	p_M
	x^1	\cdots	\cdots	\cdots
	x^2	\vdots	\ddots	
	\vdots	\vdots	\ddots	
	x^N	\vdots		\ddots

Espace de représentation

Tableau quantitatif

Tableaux explicitant les variables des éléments

- Matrice $N \times M$ de variables x_i^k

$X =$

VARIABLES	ELEMENTS			
	p_1	p_2	\cdots	p_M
	x^1	\dots	\dots	\dots
	x^2	\vdots	\ddots	
	\vdots	\vdots	\ddots	
x^N	\vdots			\ddots

Espace de représentation

Tableau quantitatif

Tableaux explicitant les variables des éléments

$X =$

	ELEMENTS			
	p_1	p_2	\cdots	p_M
x^1	\cdots	\cdots	\cdots	\cdots
x^2	\vdots	\ddots		
\vdots	\vdots		\ddots	
x^N	\vdots			\ddots

VARIABLES

- ▶ Matrice $N \times M$ de variables x_i^k
- ▶ Chaque colonne contient le vecteur \mathbf{x}_i décrivant p_i par ses N variables

Espace de représentation

Tableau quantitatif de variables indicatrices

Comment représenter les variables **symboliques** ?

$$X =$$

VARIABLES	ELEMENTS				
		p_1	p_2	\cdots	p_M
	x^1	0	1	\dots	1
	x^2	1	\ddots		
	\vdots	\vdots		0	
	x^N	1			\ddots

Espace de représentation

Tableau quantitatif de variables indicatrices

Comment représenter les variables **symboliques** ?

$X =$

	ELEMENTS			
	p_1	p_2	\dots	p_M
x^1	0	1	\dots	1
x^2	1	\ddots		
\vdots	\vdots		0	
x^N	1			\ddots

VARIABLES

- **Occurrence** d'un symbole s dans un élément p_i
- Matrice binaire
- *Vector Space Model*

Espace de représentation

Tableau de contingence - matrice de covariance

Relation entre les variables

La matrice $C = XX^T$ est appelée **tableau de contingence** (données symboliques). Elle est liée à la **matrice de covariance** (données numériques centrées)

VARIABLES

	VARIABLES			
	x^1	x^2	\dots	x^N
x^1				
x^2	\ddots	\vdots		
\vdots	\dots	$\sum_k x_k^i x_k^j$	\dots	
x^N		\vdots	\ddots	

C =

Espace de représentation

Tableau de contingence - matrice de covariance

Relation entre les variables

La matrice $C = XX^T$ est appelée **tableau de contingence** (données symboliques). Elle est liée à la **matrice de covariance** (données numériques centrées)

VARIABLES

	VARIABLES			
	x^1	x^2	\dots	x^N
x^1				
x^2	\ddots	\vdots		
\vdots	\dots	$\sum_k x_k^i x_k^j$	\dots	
x^N		\vdots	\ddots	

C =

- ▶ symbolique : c_{ij} = nbre d'élts possédant **à la fois** les symboles i et j
- ▶ numérique : c_{ij} mesure la corrélation entre les variables i et j

Espace de représentation

Tableau des similarités

Proximité entre les éléments

La matrice D est appelée **matrice des distances**

$D =$

		ELEMENTS			
ELEMENTS		p_1	p_2	\cdots	p_M
	p_1				
	p_2	\ddots	\vdots		
	\vdots	\cdots	$d(\mathbf{x}_i, \mathbf{x}_j)$	\cdots	
	p_M		\vdots	\ddots	

Espace de représentation

Tableau des similarités

Proximité entre les éléments

La matrice D est appelée **matrice des distances**

$D =$

		ELEMENTS			
ELEMENTS		p_1	p_2	\cdots	p_M
	p_1				
	p_2	\ddots	\vdots		
	\vdots	\cdots	$d(\mathbf{x}_i, \mathbf{x}_j)$	\cdots	
	p_M		\vdots	\ddots	

- ▶ Tableau $M \times M$, occupation mémoire très importante
- ▶ En général symétrique et défini-positif

Outils de description des données

- Une fois la population Ω et l'espace de représentation \mathcal{F} définis, nous allons nous intéresser aux outils fondamentaux pour **décrire** le couple $\{\Omega, \mathcal{F}\}$

Outils de description des données

- ▶ Une fois la population Ω et l'espace de représentation \mathcal{F} définis, nous allons nous intéresser aux outils fondamentaux pour **décrire** le couple $\{\Omega, \mathcal{F}\}$
- ▶ Description géométrique

Outils de description des données

- ▶ Une fois la population Ω et l'espace de représentation \mathcal{F} définis, nous allons nous intéresser aux outils fondamentaux pour **décrire** le couple $\{\Omega, \mathcal{F}\}$
- ▶ Description géométrique
- ▶ Description stochastique

Géométrie des données

L'espace \mathbb{R}^N

- ▶ Espace vectoriel de dimension N
- ▶ Sous-espace, affine, vectoriel
- ▶ Espace Euclidien
 - ▶ Produit scalaire : $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \sum_k x^k y^k$
 - ▶ $\mathbf{x} \perp \mathbf{y} \equiv \langle \mathbf{x}, \mathbf{y} \rangle = 0$
 - ▶ Norme $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$
 - ▶ Distance $d^2(\mathbf{x} - \mathbf{y}) = \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle$
- ▶ Généralisation de la distance euclidienne
 - ▶ produit scalaire $\langle \mathbf{x}, \mathbf{y} \rangle_S = \mathbf{x}^T S \mathbf{y}$
 - ▶ S matrice symétrique définie positive
($\forall \mathbf{w} \in \mathbb{R}^N, \|\mathbf{w}\| > 0, \mathbf{w}^T S \mathbf{w} > 0$)

Inertie d'un nuage de points

- ▶ Ensemble de points $\Omega = \{\mathbf{x}_i\}$, $i = 1 \dots, M$, $\mathbf{x} \in \mathbb{R}^N$
- ▶ Centre de gravité :

$$\mathbf{g} = \frac{1}{M} \sum_{\mathbf{x} \in \Omega} \mathbf{x}$$

- ▶ Inertie de Ω par rapport à un point \mathbf{a} :

$$I_{\mathbf{a}} = \sum_{\mathbf{x} \in \Omega} d(\mathbf{x}, \mathbf{a})^2$$

- ▶ On note $I = I_{\mathbf{g}}$ l'inertie de Ω par rapport à son centre de gravité \mathbf{g}

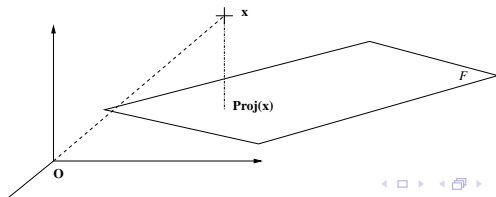
Inertie d'un nuage de points (2)

- Inertie de Ω par rapport à un sous-espace \mathcal{F} :

$$I_{\mathcal{F}} = \sum_{\mathbf{x} \in \Omega} d^2(\mathbf{x}, \mathcal{F})$$

- soit l'opérateur $\text{Proj}_{\mathcal{F}}(\mathbf{x})$ la projection orthogonale de \mathbf{x} dans \mathcal{F} , alors :

$$I_{\mathcal{F}} = \sum_{\mathbf{x} \in \Omega} d^2(\mathbf{x}, \text{Proj}_{\mathcal{F}}(\mathbf{x}))$$



Décomposition de l'inertie (th. de Huygens)

- Pour un point :

$$\forall \mathbf{a} \in \mathbb{R}^N, l_{\mathbf{a}} = l_{\mathbf{g}} + d^2(\mathbf{a}, \mathbf{g})$$

→ \mathbf{g} point **d'inertie minimum**

- Pour un sous espace \mathcal{F} : soit $\mathcal{F}_{\mathbf{g}}$ s.e. parallèle à \mathcal{F} passant par \mathbf{g} , alors

$$l_{\mathcal{F}} = l_{\mathcal{F}_{\mathbf{g}}} + d^2(\mathcal{F}, \mathcal{F}_{\mathbf{g}})$$

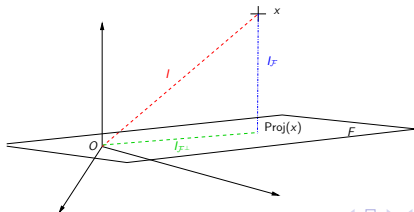
→ $\mathcal{F}_{\mathbf{g}}$ sous espace **d'inertie minimum** // à \mathcal{F}

Décomposition de l'inertie : Inertie expliquée

- ▶ Nuage centré $\Rightarrow \mathbf{g} = O$, (O origine de l'espace)
- ▶ Soit \mathcal{F} un sous espace passant par l'origine (sous espace vectoriel, s.e.v), alors

$$I = I_{\mathcal{F}} + I_{\mathcal{F}^{\perp}}, \text{ avec } I_{\mathcal{F}^{\perp}} \text{ inertie des points projetés dans } \mathcal{F}^{\perp}$$

- ▶ $I_{\mathcal{F}}$ est appelée l'**Inertie Expliquée** par \mathcal{F}
- ▶ $I_{\mathcal{F}^{\perp}}$ est appelée l'**Inertie Résiduelle** de \mathcal{F}



Expression matricielle

- ▶ Données centrées \Rightarrow

$$I = \sum_{\mathbf{x} \in \Omega} \langle \mathbf{x}, \mathbf{x} \rangle \Rightarrow I = \sum_i^M \sum_j^N (x_i^j)^2 \Rightarrow I = \sum_j^N \sum_i^M (x_i^j)^2$$

- ▶ Si X tableau *variables/éléments* de taille $N \times M$
- ▶ alors

$$I = \text{trace}(XX^T)$$

- ▶ la matrice XX^T est la **matrice d'inertie**, ou $N \times$ la matrice de **covariance**

Conclusion

Comme nous le verrons par la suite, l'inertie est une grandeur fondamentale pour l'analyse de données. Selon les algorithmes (ACP, ADL,...), on va chercher les sous-espaces minimisant ou maximisant l'inertie des éléments considérés.

Interprétation statistique

- ▶ Jusqu'à présent, nous avons décrit les données par une formulation déterministe : à un point p est associé un vecteur de variables $\mathbf{x} \in \mathbb{R}^N$.
- ▶ Dans de nombreux problèmes, il est également intéressant de modéliser \mathbf{x} comme une variable aléatoire décrite par une loi de probabilité
- ▶ L'analyse des données s'effectue alors par l'analyse statistique de la population

Rappel sur la théorie probabiliste

Variable Aléatoire Discrète

- ▶ x est une variable aléatoire discrète prenant une valeur dans l'ensemble $\mathcal{X} = \{v_1, \dots, v_m\}$.
- ▶ p_i est la probabilité que x prenne la valeur v_i

$$p_i = \Pr[x = v_i], \quad i = 1, \dots, m$$

- ▶ L'ensemble $\{p_1, \dots, p_m\}$ est exprimé par la fonction de probabilité $P(x)$, tq.,

$$P(x) \geq 0, \quad \text{et} \quad \sum_{x \in \mathcal{X}} P(x) = 1$$

- ▶ Histogramme des occurrences de x normalisé

Espérances

- Espérance ou moyenne

$$E[x] = \mu = \sum_{x \in \mathcal{X}} xP(x) = \sum_i^m v_i p_i$$

- Moment d'ordre 2

$$E[x^2] = \sum_{x \in \mathcal{X}} x^2 P(x)$$

- Variance

$$\text{Var}[x] = \sigma^2 = E[(x - \mu)^2] = \sum_{x \in \mathcal{X}} (x - \mu)^2 P(x)$$

σ : écart-type

Variables centrées normalisées

- ▶ Inégalité de Chebichev

$$\Pr[|x - \mu| > n\sigma] \leq \frac{1}{n^2}$$

- ▶ Relation entre l'écart-type et la dispersion des valeurs autour de la moyenne
- ▶ $\frac{x - \mu}{\sigma} \Rightarrow$ **normalisation** de la dispersion autour de la moyenne

Paires de v.a. discrètes

- ▶ x et y v.a. prenant valeurs dans $\mathcal{X} = \{v_1, \dots, v_m\}$ et $\mathcal{Y} = \{w_1, \dots, w_n\}$ respectivement.
- ▶ Pour tout couple (v_i, w_j) , il existe une **probabilité jointe**

$$p_{ij} = \Pr[x = v_i, y = w_j]$$

- ▶ Fonction de probabilité jointe $P(x, y)$ pour laquelle

$$P(x, y) \geq 0, \quad \text{et} \quad \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y) = 1$$

- ▶ Histogramme des co-occurrences de (x, y) normalisé

Probabilité marginale, indépendance

► Distributions marginales

$$P_x(x) = \sum_{y \in \mathcal{Y}} P(x, y)$$

$$P_y(y) = \sum_{x \in \mathcal{X}} P(x, y)$$

► x et y sont dites **statistiquement indépendantes** si

$$P(x, y) = P_x(x)P_y(y)$$

Espérance de deux variables aléatoires

- D'une manière générale

$$E[f(x, y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x, y) P(x, y)$$

- Moyenne

$$\mu_x = E[x] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x P(x, y)$$

$$\mu_y = E[y] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} y P(x, y)$$

- Variance

$$\sigma_x^2 = \text{Var}[x] = E[(x - \mu_x)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x - \mu_x)^2 P(x, y)$$

$$\sigma_y^2 = \text{Var}[y] = E[(y - \mu_y)^2] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (y - \mu_y)^2 P(x, y)$$

Coefficient de corrélation

- Covariance

$$\sigma_{xy} = \text{Cov}(x, y) = E[(x - \mu_x)(y - \mu_y)]$$

- Inégalité de Cauchy Schwartz

$$\sigma_{xy} \leq \sigma_x \sigma_y$$

- Coefficient de corrélation

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \in [-1, 1]$$

- $\rho = 1$, les variables sont corrélées positivement
- $\rho = 0$, les variables sont décorrélées
- $\rho = -1$, les variables sont corrélées négativement

Probabilité conditionnelle

v.a. statistiquement dépendante

$P(x|y) = \Pr[x = v_i | y = w_j]$: probabilité de connaître x sachant y .

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

Règles de Bayes

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

ou

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x \in \mathcal{X}} P(x, y)}$$

Vraisemblance

Supposons que x est la *cause* d'un événement qui a pour *effet mesurable* y .

La règle de Bayes permet de déterminer la probabilité *a posteriori* $P(x|y)$ de connaître x sachant y à partir de la *vraisemblance* $P(y|x)$ que y est réellement induit par x , et de la probabilité *a priori* $P(x)$ de l'occurrence de x .

$$a \text{ posteriori} = \frac{\text{vraisemblance} \times a \text{ priori}}{\text{evidence}}$$

Le dénominateur $P(y) = \sum_{x \in \mathcal{X}} P(x, y)$, appelé *évidence*, est un terme de normalisation (tq. $\sum_x P(x|y) = 1$).

Variables Aléatoires Continues

- ▶ Lorsque x prend ses valeurs dans un continuum, on définit une fonction de *densité de probabilité* $p(x)$

$$p(x) = \Pr[x \in (a, b)] = \int_a^b p(x) dx$$

- ▶ $p(x) \geq 0$ et $\int_{-\infty}^{\infty} p(x) dx = 1$
- ▶ La plupart des définitions pour les v.a. discrètes se transforment dans le cas continu en remplaçant l'opérateur \sum par \int .

Variables aléatoire multivariées

Extension du problème à N variables aléatoires x^1, x^2, \dots, x^N

- ▶ Notation vectorielle $\mathbf{x} = [x^1, x^2, \dots, x^N]^T, \in \mathbb{R}^N$
- ▶ Fonction de probabilité jointe $P(\mathbf{x}), P(\mathbf{x}) \geq 0, \sum P(\mathbf{x}) = 1$
- ▶ $P(\mathbf{x})$ fonction multidimensionnelle qui peut être très complexe
- ▶ Les règles énoncées pour x sont valables pour \mathbf{x}
- ▶ En particulier

$$\text{v.a. indépendante } P(\mathbf{x}) = \prod_i P_{x^i}(x^i)$$

$$\text{Proba. cond. } P(x^1, x^2 | x^3) = \frac{P(x^1, x^2, x^3)}{P(x^3)}$$

Espérance, matrice de covariance

- ▶ De manière générale
 $E[\mathbf{f}(\mathbf{x})] = \sum_{\mathbf{x}} \mathbf{f}(\mathbf{x})P(\mathbf{x}), \mathbf{f} : [0, 1]^N \rightarrow [0, 1]^N$
- ▶ Vecteur moyen $\boldsymbol{\mu} = \sum \mathbf{x}P(\mathbf{x})$
- ▶ Matrice de covariance

$$\Sigma = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1N} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1} & \sigma_{N2} & \dots & \sigma_{NN} \end{pmatrix}$$

- ▶ Σ symétrique, diagonale = variances, hors-diagonale = co-variances
- ▶ Variables indépendantes $\Rightarrow \Sigma$ diagonale (pas réciproque!)

Expression Matricielle

- ▶ X tableau *variables/éléments*

$$\Sigma = \frac{1}{N} X X^T$$

- ▶ Σ est dite **semi défini positive**

$$\forall \mathbf{w} \in \mathbb{R}^N, \mathbf{w}^T \Sigma \mathbf{w} \geq 0$$

- ▶ μ et Σ décrivent à **l'ordre 2** la distribution statistique multidimensionnelle d'une population Ω .

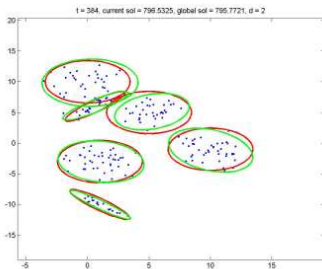
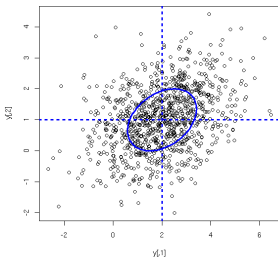
Distribution Gaussienne Multivariée

- ▶ **Théorème Central-Limite** : La distribution de la somme de d variables aléatoires tend vers la loi Normale (Gaussienne)
- ▶ Distribution Gaussienne est complètement décrite à l'ordre 2

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} \exp \left[-0.5(\mathbf{x} - \boldsymbol{\mu}) \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})^T \right]$$

- ▶ Hypothèse très fréquente : la distribution de Ω est de la forme $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Modélisation Gaussienne

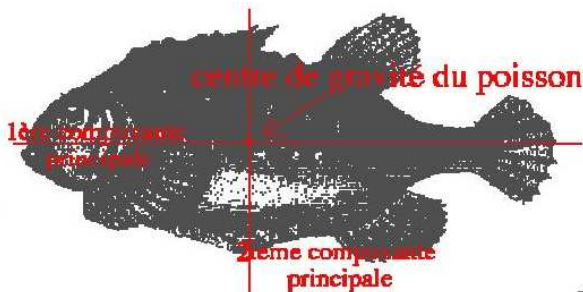


Analyse en Composantes

- ▶ Analyse de données multivariées
 - ▶ Nuage Ω composé de M points définis par N variables x^i
 - ▶ $\mathbf{x} \in \mathbb{R}^N$
- ▶ Nécessité de comprendre la distribution "spatiale" de ces données
 - ▶ Possibilité de visualiser les données
 - ▶ Extraire les caractéristiques les plus importantes
- ▶ Compression de la représentation
 - ▶ Qualité de la reconstruction
 - ▶ Manipulations facilitées des données
- ▶ Séparation des données selon leurs propriétés principales

Analyse en Composantes Principales (ACP)

L'ACP a pour but de déterminer pour un nuage de points Ω un **sous-espace** dans lequel les données seront représentées de manière **compacte** par des variables **décorélées**. Ces nouvelles variables font apparaître des **propriétés géométriques intrinsèques** d'importance **décroissante** de Ω .



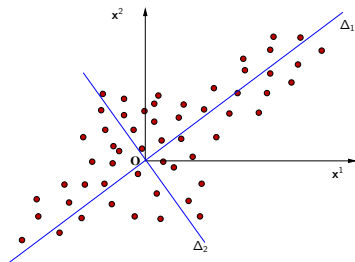
Rappel sur l'Inertie

- ▶ Soit $\mathbb{R}^N = \Delta_1 \oplus \Delta_2 \oplus \cdots \oplus \Delta_N$ la décomposition en s.e. orthogonaux de dimension 1 (axe Δ)
- ▶ Pour un nuage de points Ω , l'inertie totale se décompose

$$I = I_{\Delta_1} + I_{\Delta_2} + \cdots + I_{\Delta_N}$$

- ▶ L'ACP recherche tous les axes Δ_i tq. $I_{\Delta_i} \geq I_{\Delta_{i+1}}$
- ▶ Les projections sur les axes *expliquant* le maximum d'inertie globale conservent le maximum d'information sur les données
- ▶ Ω contenu dans un s.e $\iff I_{\Delta_i} = 0, \forall i > d$

Interprétation géométrique



Recherche des axes maximisant
l'inertie expliquée



recherche des axes de variance maximale

Analyse de la matrice de covariance

Recherche de l'axe \mathbf{u} minimisant l'erreur quadratique

- ▶ $\sum_i \|\mathbf{x}_i - \langle \mathbf{x}_i, \mathbf{u} \rangle \mathbf{u}\|^2 \simeq \sum_i \mathbf{u}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{u} = \text{Trace}(\mathbf{u}^T \Sigma \mathbf{u})$
- ▶ avec la contrainte $\mathbf{u}^T \mathbf{u} = 1$ (pour éviter la solution $\|\mathbf{u}\| = 0$)

Minimisation

- ▶ Par Lagrange, pour une dimension, minimization de $J = \mathbf{u}^T \Sigma \mathbf{u} - \lambda(1 - \mathbf{u}^T \mathbf{u}) \Leftrightarrow \frac{\partial J}{\partial \mathbf{u}} = 0 \Leftrightarrow \Sigma \mathbf{u} - \lambda \mathbf{u} = 0$
- ▶ $\Leftrightarrow \Sigma \mathbf{u} = \lambda \mathbf{u} \Leftrightarrow \mathbf{u}$ est un vecteur propre de Σ

Diagonalisation de la matrice de covariance

Décomposition spectrale

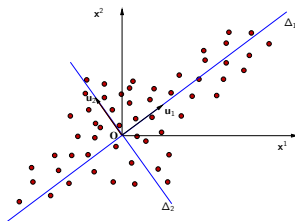
- ▶ On résout toutes les dimensions en un seul système
- ▶ \Leftrightarrow diagonalisation de Σ par recherche de vecteurs propres

$$\Sigma = U\Lambda U^T, \quad U, \Lambda \in \mathbb{R}^{N \times N}$$

- ▶ Les colonnes de la matrice de rotation U contiennent les **vecteurs propres** \mathbf{u}_i unitaires et orthogonaux de Σ .
- ▶ $\Lambda = \text{diag}[\lambda_1, \dots, \lambda_N]$ **valeurs propres** de Σ .

Composantes Principales

- ▶ Les valeurs propres ordonnées $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_N$ sont égales aux Inerties $I_{\Delta_1} \geq I_{\Delta_2} \geq \dots \geq I_{\Delta_N}$
- ▶ Les N vecteurs propres ordonnés \mathbf{u}_i définissent les axes Δ_i et sont appelés les **composantes principales**
- ▶ La nouvelle base de représentation est maintenant $\{\mathbf{u}_i\}_{i=1,\dots,N}$



Contribution des axes à l'inertie totale

$$I_{\Delta_i} = \lambda_i, \text{ et } I = \sum_i^N I_{\Delta_i}$$

Définition

1. Contribution Absolue de Δ_i à I : $ca(\Delta_i/I) = \lambda_i$
2. Contribution Relative : $cr(\Delta_i/I) = \frac{\lambda_i}{\lambda_1 + \dots + \lambda_N}$

⇒ Pourcentage d'Inertie Expliquée par Δ_i

3. Pourcentage des d premiers axes :

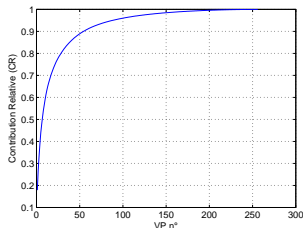
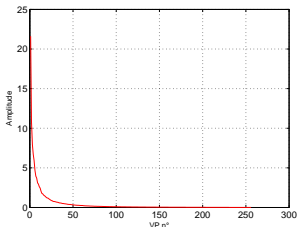
$$cr(\Delta_1 \oplus \Delta_2 \cdots \oplus \Delta_d) = \frac{\lambda_1 + \lambda_2 \cdots + \lambda_d}{\lambda_1 + \lambda_2 \cdots + \lambda_N}$$

Décroissance de la contribution des composantes

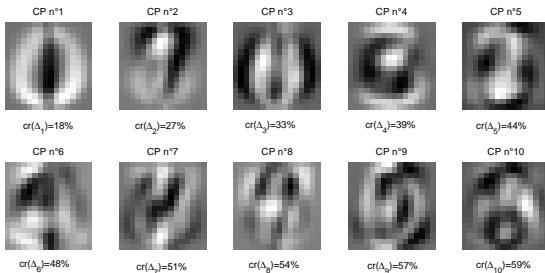
Données Digit



7291 images 16×16 (8 bits)
 $\Rightarrow \mathbf{x}_i \in \mathbb{R}^{256}, i = 1 \dots 7291$



Pourcentage d'Inertie Expliquée



Projection sur les composantes principales

La définition d'un nouvel espace implique de nouvelles coordonnées décrivant les données

$$y_i^j = \langle \mathbf{x}_i, \mathbf{u}_j \rangle$$

⇒ La j ième composante des nouvelles coordonnées \mathbf{y}_i d'un point i s'obtient en projetant le vecteur \mathbf{x}_i sur la j ième composante principale \mathbf{u}_j

$$\mathbf{y}_i = U^T \mathbf{x}_i$$

Approximation des données

On peut ne retenir que les d premières composantes (par exemple si $\text{cr}(\Delta_d, I) \geq 90\%$). Dans ce cas, les données sont approximées dans le nouvel espace de **dimension réduite**

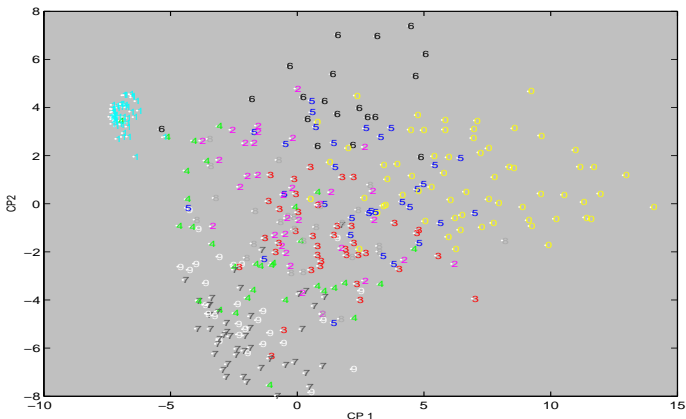
$$\tilde{\mathbf{y}}_i^2 = U_d^T \mathbf{x}_i, \quad \tilde{\mathbf{y}} \in \mathbb{R}^d$$

$\Rightarrow U_d \in \mathbb{R}^{N \times d}$ matrice des d premières composantes

- ▶ Si $d = 2$ ou $3 \rightarrow$ possibilité de visualiser les données
- ▶ Si $d \ll N \rightarrow$ compression des données
- ▶ Expressivité des d premières composantes

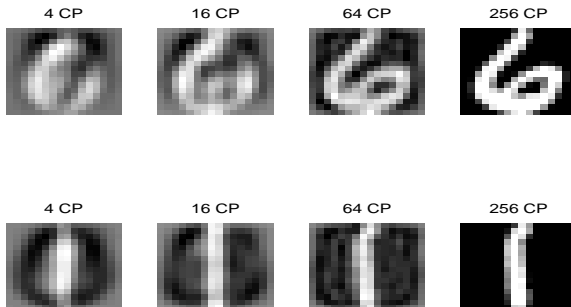
Visualisation des données

$$\tilde{\mathbf{y}}_i^2 = [\mathbf{u}_1^T \mathbf{x}_i, \mathbf{u}_2^T \mathbf{x}_i]$$



Reconstruction

$$\tilde{\mathbf{x}}_i = \sum_{j=1}^d y_i^j \mathbf{u}_j = {}^t \tilde{\mathbf{y}}_i^d U_d$$



Qualité de la représentation des éléments

- ▶ Soit 2 points projetés sur un axe Δ_k
 - ▶ si ils sont éloignés sur $\Delta_k \Rightarrow$, éloignés dans l'espace original
 - ▶ si ils sont proches sur Δ_k , pas de conclusion...
- ▶ La qualité de représentation de \mathbf{x}_i par Δ_k se mesure par

$$Q_{\Delta_k}(\mathbf{x}_i) = \cos^2(\mathbf{x}_i, \mathbf{u}_k) = \frac{\langle \mathbf{x}_i, \mathbf{u}_k \rangle^2}{\|\mathbf{x}_i\|^2}$$

- ▶ Sur un sous-espace $E = \Delta_k \oplus \Delta_q \oplus \dots \oplus \Delta_p$

$$Q_E(\mathbf{x}_i) = \cos^2(\mathbf{x}_i, \mathbf{u}_k) + \cos^2(\mathbf{x}_i, \mathbf{u}_q) + \dots + \cos^2(\mathbf{x}_i, \mathbf{u}_p)$$

Exemple

Projection sur le plan principal ($\Delta_1 \oplus \Delta_2$)

Label	0	1	2	3	4	5	6	7	8	9
Qualité	0.7	1.5	0.4	0.2	0.7	0.2	0.5	0.9	0.4	0.8

Contribution d'un élément à la définition des axes

- ▶ Contribution absolue d'un point i à Δ_k

$$ca(\mathbf{x}_i, \Delta_k) = \frac{1}{N} \langle \mathbf{x}_i, \mathbf{u}_k \rangle^2$$

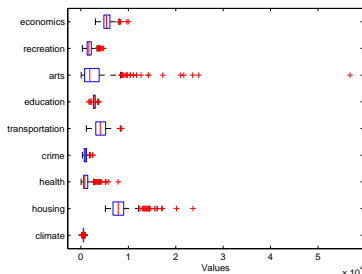
⇒ Plus la projection d'un point sur un axe est importante, plus cet élément contribue à l'existence de l'axe

- ▶ Contribution relative d'un élément à l'Inertie de Δ_k

$$cr(\mathbf{x}_i, \Delta_k) = \frac{ca(\mathbf{x}_i, \Delta_k)}{I_{\Delta_k}} = \frac{\langle \mathbf{x}_i, \mathbf{u}_k \rangle^2}{\lambda_k}$$

ACP sur des données réduites

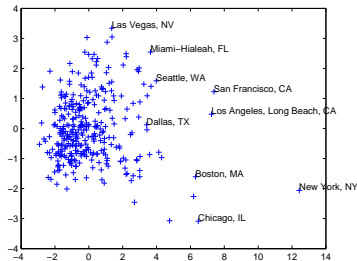
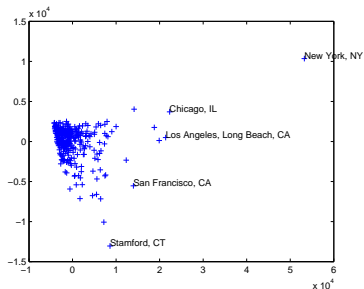
- ▶ Variables initiales hétérogènes \rightarrow homogénéité des combinaisons linéaires ?
- ▶ Exemple : Notation de villes américaines



- ▶ Echelle de notes : selon les catégories 100 \rightarrow 10000
- ▶ Nécessité de normaliser les données par leur variances

Réduction des données

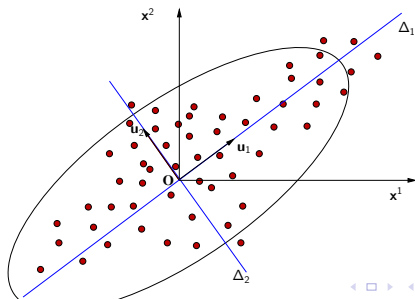
- ▶ Définir la métrique $\langle ., . \rangle_V$, avec $V = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$
- ▶ La matrice de covariance Σ_V des données réduites est égale à la matrice de **corrélacion** R des données initiales
- ▶ Décomposition spectrale de R (à la place de Σ)



Optimalité de l'ACP

Distribution gaussienne

- ▶ L'ACP décompose la matrice de covariance selon ces valeurs propres/vecteur propres
- ▶ Meilleure base pour représenter $\mathcal{N}(\mu, \Sigma)$



Limitations de l'ACP

1. L'ACP étudie les corrélations entre variables \rightarrow relations linéaires
 \Rightarrow pas de prise en compte des relations non-linéaires !
2. L'ACP optimise un critère quadratique (matrice de covariance)
 \Rightarrow sensible aux valeurs extrêmes
3. L'ACP est optimale pour des données Gaussiennes
 \Rightarrow mal adaptée aux données clusterisées