Analyse des données

Chapitre 1: Préliminaires

Mohamed Essaied Hamrita
mhamrita@gmail.com
hamrita.e-monsite.com

Institut Supèrieur de Mathématiques Appliqués & Informatique - Kairouan

Janvier 2016

M1: Ingénierie Financière

Plan du cours

Chapitre 1 : Préliminaires
Chapitre 2 : Analyse en Composantes Principales (ACP)
Chapitre 3 : Analyse Factorielle des Correspondances (AFC)
Chapitre 4 : Analyse Discriminante

Pré-acquis

Statistique descriptive.
Théorie des probabilités.
Algèbre linéaire (produit scalaire, decomposition selon une base, matrices, valeurs et vecteurs propres, métriques).
Software : Tout au long de ce cours, on utilisera le logiciel **R** pour les manipulations. (C'est un logiciel libre et donc, gratuit téléchargeable depuis www.r-project.org).

1. Vocabulaires
2. But et domaines d'application de l'AD
3. Description des données quantitatives
4. Mesures de liaison entre deux variables
5. Tableaux des données
6. La matrice des poids
7. Point moyen et tableau centré
8. Matrice de variance-covariance

Vocabulaires

Vocabulaires

Population (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de champ de l'étude.

Vocabulaires

Population (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de champ de l'étude.

Individu (ou unité statistique) : tout élément de la population.

Vocabulaires

Population (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de champ de l'étude.
Individu (ou unité statistique) : tout élément de la population.
Échantillon : sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

Vocabulaires

Population (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de champ de l'étude.

Individu (ou unité statistique) : tout élément de la population.

Échantillon : sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

Taille de l'échantillon $\mathbf{n}$: cardinal du sous-ensemble correspondant.

Enquête (statistique) : opération consistant à observer (ou mesurer, ou questionner...) l'ensemble des individus d'un échantillon.

Vocabulaires

Population (ou population statistique) : ensemble (au sens mathématique du terme) concerné par une étude statistique. On parle parfois de champ de l'étude.
Individu (ou unité statistique) : tout élément de la population.
Échantillon : sous-ensemble de la population sur lequel sont effectivement réalisées les observations.
Taille de l'échantillon $\mathbf{n}$: cardinal du sous-ensemble correspondant.
Enquête (statistique) : opération consistant à observer (ou mesurer, ou questionner...) l'ensemble des individus d'un échantillon.
Recensement : enquête dans laquelle l'échantillon observé est la population tout entière (enquête exhaustive).

Vocabulaires

Sondage : enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population (enquête non exhaustive).

Vocabulaires

Sondage : enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population (enquête non exhaustive).
Variable (statistique) : ensemble de caractéristiques d'une population. On distingue deux types de variables statistiques : variables quantitatives(si elle est représentée par un nombre : age, taille, prix,...) et variables qualitatives (si elle n'est pas mesurable). Une variable qualitative peut être nominale (si elle appartient à une catégorie donnée. Exemple : sexe, couleur, CSP) ou ordinale (si les catégories sont ordonnées. Exemple : passable, assez bien, bien, très bien)

Vocabulaires

Sondage : enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population (enquête non exhaustive).
Variable (statistique) : ensemble de caractéristiques d'une population. On distingue deux types de variables statistiques : variables quantitatives(si elle est représentée par un nombre : age, taille, prix,...) et variables qualitatives (si elle n'est pas mesurable). Une variable qualitative peut être nominale (si elle appartient à une catégorie donnée. Exemple : sexe, couleur, CSP) ou ordinale (si les catégories sont ordonnées. Exemple : passable, assez bien, bien, très bien)
Analyse des données : ensemble de méthodes statistiques qui permettent de résumer l'information contenue dans de grands tableaux de données.

L'analyse des données

But : L'AD a pour objectifs de synthétiser, structurer l'information contenue dans des données multi-dimentionnelles (n individus, p variables).

Domaines d'application : L'analyse des données est utilisée dans tous les domaines (Marketing quantitatif (études de marchés, enquête de satisfaction,...), économie, sciences sociales,...).

Description des données quantitatives

On appelle **variable** un vecteur $\mathbf{x}$ de taille n. Chaque coordonnée x_i correspond à un individu. On s'intéresse ici à des valeurs numériques (variable quantitative).

Description des données quantitatives

On appelle **variable** un vecteur $\mathbf{x}$ de taille n. Chaque coordonnée x_i correspond à un individu. On s'intéresse ici à des valeurs numériques (variable quantitative).

Poids : Chaque individu peut avoir un poids p_i, tel que $\sum_{i=1}^{n} p_i = 1$.

On a souvent $p_i = 1/n$.

Description des données quantitatives

On appelle **variable** un vecteur $\mathbf{x}$ de taille n. Chaque coordonnée x_i correspond à un individu. On s'intéresse ici à des valeurs numériques (variable quantitative).

Poids : Chaque individu peut avoir un poids p_i, tel que $\sum_{i=1}^{n} p_i = 1$.

On a souvent $p_i = 1/n$.

Résumés : on dispose d'une série d'indicateurs qui ne donne qu'une vue partielle des données : effectif, moyenne, médiane, variance, écart type, minimum, maximum, étendue, 1er quartile (25% inférieurs), 4ème quartile (25% supérieurs),. . .Ces indicateurs mesurent principalement la tendance centrale et la dispersion. On utilisera principalement la moyenne, la variance et l'écart type.

La moyenne et l'écart type

La **moyenne** est définie par : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ou $\overline{x} = \sum_{i=1}^{n} p_i x_i$.

La moyenne est une mesure de tendance centrale et est sensible aux valeurs extrêmes.

La **variance** est définie par : $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$ ou

$$\sigma^2 = \sum_{i=1}^{n} p_i (x_i - \overline{x})^2.$$

On peut montrer que $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$ ou $\sigma^2 = \sum_{i=1}^{n} p_i x_i^2 - \overline{x}^2$.

L'**écart type** σ est la racine carré de la variance.

Mesures de liaison entre deux variables

La **covariance** entre deux variables x et y est donnée par :

$$cov(x, y) = S_{xy} = \sum_{i=1}^{n} p_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} p_i x_i y_i - \overline{x}\,\overline{y}$$

et le coefficient de corrélation est donné par :

$$r_{xy} = \frac{S_{xy}}{\sigma_x \sigma_y}.$$

Le coefficient de corrélation possède les propriétés suivantes :
$|r_{xy}| \leq 1$.
Si r_{xy} est très proche de 1, on dit que x et y sont fortement liées.
Si $r_{xy} = 0$, on dit que x et y sont dé-corrélées. (Cela ne veut pas dire qu'elles sont indépendantes).

Corrélation et liaison significative

Problème : A partir de quelle valeur de r_{xy} peut-on considérer que les variable x et y sont liées ?

Corrélation et liaison significative

Problème : A partir de quelle valeur de r_{xy} peut-on considérer que les variable x et y sont liées ?

Méthode : On se place dans le cas où le nombre d'observations est $n > 30$. On peut montrer que : $F^{ob} = \frac{(n-2)r_{xy}^2}{1-r_{xy}^2} \sim F(1, n-2)$ où F est la loi de Ficher-Snédécor.

Test : On veut tester :

$$\begin{cases} H_0 : & x \text{ et } y \text{ sont indépendantes} \\ H_1 : & x \text{ et } y \text{ sont corrélées} \end{cases}$$

Règle de décision : On se fixe un risque d'erreur α (1% ou 5% en général) et on calcul la probabilité $p(F(1, n-2) > F^{ob}) = \pi$. Si $\alpha > \pi$, on rejette H_0 au seuil α.

Exemple

On veut mesurer la liason entre deux variables aléatoires X et Y. Depuis 120 observations de ces deux variables, on a tiré les statistiques suivantes :
$\sum x_i = 360$, $\sum y_i = -240$, $\sum x_i y_i = -244$, $\sum x_i^2 = 2508$ et $\sum y_i^2 = 1551$.
1) Déterminer la covariance du couple aléatoire (X, Y).
2) En déduire la corrélation de ce couple aléatoire.
3) Tester la significativité de la corrélation entre X et Y. (On prendra $\alpha = 5\%$ et $F_{1-\alpha/2}(1, 118) = 5.15455$).

Exemple

1) La covariance du couple aléatoire (X, Y) est définie par :

$$cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y}$$

Avec $\overline{X} = \sum X_i/n = 360/120 = 3$ et
$\overline{Y} = \sum Y_i/n = -240/120 = -2.$
$cov(X, Y) = 1/120 \times (-244) - (-6) = 3.966 \simeq 4.$

Exemple

2) La corrélation est $r_{xy} = \dfrac{cov(X, Y)}{\sigma_x \sigma_y}$ avec :

$$\sigma_x^2 = 1/n \sum X_i^2 - \overline{X}^2 = 1/120 \times 2508 - 3^2 = 12 \text{ et } \sigma_y^2 = 9$$

D'où $r_{x,y} = \dfrac{4}{\sqrt{12 \times 9}} = 0.3849.$

Exemple

3) La statistique du test est $F^{ob} = \dfrac{(n-2)r_{xy}^2}{1-r_{xy}^2} \sim F(1, n-2)$.

$F^{ob} = \dfrac{118 \times 0.3849^2}{1-0.3849^2} = 20.52 > F_{0.975}(1, 118) = 5.1545$, donc on rejette H_0.

On conclut qu'au seuil $\alpha = 5\%$, il existe une corrélation significative entre X et Y.

Tableaux des données

Le tableau de données de n individus et p variables, noté X est donné par :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{i1} & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Tableaux des données

Le tableau de données de n individus et p variables, noté X est donné par :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ x_{i1} & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Le vecteur variable est le vecteur colonne de la matrice X.
Le vecteur individu est le vecteur ligne de la matrice X.

La matrice des poids

Lorsque les individus n'ont pas la même importance, on fait recours aux poids qui sont définis comme suit : $p_1 + p_2 + \ldots + p_n = 1$.
La matrice des poids est représentée par une matrice diagonale de taille n :

$$\begin{pmatrix} p_1 & 0 & \cdots & 0 \\ 0 & p_1 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & p_1 \end{pmatrix}$$

Dans le cas où les individus ont le même poids $p_i = 1/n$, la matrice D sera $D = \frac{1}{n} I_n$.

Point moyen et tableau centré

Le **point moyen** est le vecteur g des moyennes arithmétiques de chaque **variable** :$g' = (\overline{x}_{.1}, \overline{x}_{.2}, \ldots, \overline{x}_{.p})$ où $\overline{x}_{.j} = \sum_{i=1}^{n} p_i x_{ij}$.

En notation matricielle, le point g s'écrit : $g = X'D\mathbf{1}_n$ où $\mathbf{1}_n = (1, 1, \ldots, 1)'$.

Point moyen et tableau centré

Le **point moyen** est le vecteur g des moyennes arithmétiques de chaque **variable** :$g' = (\overline{x}_{.1}, \overline{x}_{.2}, \ldots, \overline{x}_{.p})$ où $\overline{x}_{.j} = \sum_{i=1}^{n} p_i x_{ij}$.

En notation matricielle, le point g s'écrit : $g = X'D\mathbf{1}_n$ où $\mathbf{1}_n = (1, 1, \ldots, 1)'$.

Le **tableau centré** est obtenu en centrant les variables autour de leur moyennes$y_{ij} = x_{ij} - \overline{x}_{.j}$. Soit $Y = X - \mathbf{1}_n g' = (I_n - \mathbf{1}_n'\mathbf{1}_n D)X$.

Matrice de variance-covariance

La matrice des variances-covariance est une matrice carré de dimension p définie par :

$$V = \begin{pmatrix} \sigma_1^2 & S_{12} & \cdots & S_{1p} \\ S_{21} & \sigma_2^2 & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p11} & S_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

où S_{kl} est la covariance entre les variables x_k et x_l et σ_j^2 est la variance de la variable x_j.
La matrice V peut être calculée comme suit :

$$V = X'DX - gg' = Y'DY$$

Matrice de corrélation

La matrice de corrélation est donnée par :

$$R = \begin{pmatrix} 1 & r_{12}^2 & \cdots & r_{1p}^2 \\ r_{21}^2 & 1 & \vdots & r_{2p}^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1}^2 & \cdots & \ldots & 1 \end{pmatrix}$$

La matrice de corrélation se calcule de la manière suivante : $R = D_{1/\sigma} V D_{1/\sigma}$ où $D_{1/\sigma}$ est :

$$D_{1/\sigma} = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \frac{1}{\sigma_p} \end{pmatrix}$$

Exemple

Onconsidère le tableau des données dans lequel on a observé les notes de 9 élèves pour 4 matières.

```
##         Maths Physique Français Anglais
## Fatma     6.0      6.0      5.0     5.5
## Ali       8.0      8.0      8.0     8.0
## Kawther   6.0      7.0     11.0     9.5
## Nidhal   14.5     14.5     15.5    15.0
## Nabiha   14.0     14.0     12.0    12.5
## Wiem     11.0     10.0      5.5     7.0
## Youssef   5.5      7.0     14.0    11.5
## Sarah    13.0     12.5      8.5     9.5
## Wafa      9.0      9.5     12.5    12.0
```

Exemple

Onconsidère le tableau des données dans lequel on a observé les notes de 9 élèves pour 4 matières.

```
##          Maths Physique Français Anglais
## Fatma      6.0      6.0      5.0     5.5
## Ali        8.0      8.0      8.0     8.0
## Kawther    6.0      7.0     11.0     9.5
## Nidhal    14.5     14.5     15.5    15.0
## Nabiha    14.0     14.0     12.0    12.5
## Wiem      11.0     10.0      5.5     7.0
## Youssef    5.5      7.0     14.0    11.5
## Sarah     13.0     12.5      8.5     9.5
## Wafa       9.0      9.5     12.5    12.0
```

1) Calculer le point moyen g de ce tableau.
2) Donner le tableau centré.
3) Calculer la matrice des variances-covariances de ce tableau.
4) En déduire la matrice des corrélations.

Exemple

```
# 1) On peut calculer le point moyen par 3 méthodes:
(g<-colMeans(notes)) # Calculer la moyenne des colonnes.

##     Maths  Physique  Français   Anglais
##  9.666667  9.833333 10.222222 10.055556

(g<-apply(notes,2,mean))  # appliquer la moyenne aux colonnes.

##     Maths  Physique  Français   Anglais
##  9.666667  9.833333 10.222222 10.055556

n<-nrow(notes) ; p<-ncol(notes)
D1<-diag(1/n,n,n)
In<-matrix(1,1,n)
g<-t(notes)%*%D1%*%t(In)  # calcul matriciel
t(g)

##         Maths Physique Français  Anglais
## [1,] 9.666667 9.833333 10.22222 10.05556
```