UNIVERSITÉ DE VERSAILLES SAINT-QUENTIN-EN-YVELINES

MODULE METHODOLOGIQUE M4

Statistique Descriptive :
Description multidimensionnelle

UVED
Université Virtuelle Environnement & Développement Durable

30 AVRIL 2014
VERSION 300414

STATISTIQUE DESCRIPTIVE : DESCRIPTION MULTIDIMENSIONNELLE

Cécile Mallet

SOMMAIRE

4. Description multidimensionnelle

Dans la plupart des applications on observe non pas une ou deux variables par individu mais un nombre p souvent élevé. L'étude séparée de chacune de ces variables est une phase indispensable mais qui n'est pas suffisante car elle laisse de côté l'étude des liaisons qui peuvent exister entre les variables. Les méthodes descriptives ont pour objectif d'organiser, de simplifier et d'aider à comprendre l'information sous-jacente d'un ensemble important de données. Un ensemble de données peut être « important » soit parce que le nombre de variables est important soit parce que le nombre d'individus est important. Il existe deux grandes familles de méthodes dédiées l'une est dédiée à la représentation des individus avec un nombre de variable plus réduit (méthodes d'analyses factorielles) en utilisant les proximités entre variable l'autre est dédiée au regroupement des individus similaires (méthodes de classification).

Matrices d'observations

Soit un vecteur X composé de p variables $X=(X_1,X_2,...,X_p)$

Soit un échantillon de n individus extraits de la population

La matrice d'observation se définit par :

$$X=\begin{pmatrix} x_{11} & x_{12} & x_{1j} & x_{1p} \\ x_{21} & x_{22} & x_{2j} & x_{2p} \\ & & & \\ x_{i1} & x_{i2} & x_{ij} & x_{ip} \\ & & & \\ x_{n1} & x_{n2} & x_{nj} & x_{np} \end{pmatrix}$$ *en général n >> p*

$x_i=\begin{pmatrix} x_{i1} & x_{i2} & x_{ij} & x_{ip} \end{pmatrix}$ *sont les p variables observées sur le* i^{eme} *individu, c'est à dire le vecteur X observé sur le* i^{eme} *individu*

$$X_j=\begin{pmatrix} x_{1j} \\ x_{ij} \\ \\ x_{nj} \end{pmatrix}$$ *est la* j^{eme} *variable* X_j *observée sur l'ensemble des n individus*

x_{ij} est l'observation de la j^{eme} variable X_j sur le i^{eme} individu

*Exemple. Données extraites du modèle de prévision météorologique du centre européen. Les variables sont la température, la pression, et l'humidité au sol, l'altitude de l'isotherme, les contenus intégrés en vapeur d'eau et en eau liquide. Les individus sont les points de grille du modèle. L'atmosphère terrestre est ainsi représentée par 351 137 individus (points de grille) décrits chacun par un vecteur en dimension 485(60*8+5)*

indice	***T (K)***	***P (Pa)***	***H (%)***	***Iso (km)***	***…***	***…***	***IWV (mg/cm2)***	***ICLW (g/cm2)***
1	*267*	*100899*	*81*	*NaN*			*2.5*	*0.71*
2	*280*	*101458*	*68*	*1.64*			*1.99*	*1.43*
…				*…*				
258203	*285*	*100756*	*82*	*1.74*			*6.79*	*1.59*
258204	*287*	*100354*	*80*	*1.88*			*1.28*	*1.32*

Vecteur moyen

Le vecteur moyen $\overline{X}$ est le centre de gravité G du nuage de points, il décrit où se trouve le nuage de points :

$$\overline{X} = \frac{1}{n}\left(\sum_{i=1}^{n} x_{i1} \quad \sum_{i=1}^{n} x_{i2} \quad \cdots \quad \sum_{i=1}^{n} x_{ip}\right)$$

c'est le vecteur des moyennes des p variables $\overline{X} = \left(\overline{X}_1 \quad \overline{X}_2 \quad \cdots \quad \overline{X}_j \quad \cdots \quad \overline{X}_p\right)$

Exemple des données météorologiques :

T (K)	***P (Pa)***	***H (%)***	***Iso (km)***	***IWV (mg/cm2)***	***ICLW (g/cm2)***
290	*98267*	*69*	*NaN*	*5.8*	*3.1*

Matrice de dispersion ou matrice de variances-covariances S

La matrice de variance covariance S décrit la dispersion du nuage de points dans l'espace de dimension p.

Pour obtenir S il faut calculer les variances de chaque variable :

$$S_{jj} = C_{X_j X_j} = s_{X_j}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{X}_j)^2 = \frac{1}{n}\sum_{i=1}^{n} x_{ij}{}^2 - \overline{X}_j{}^2$$

ainsi que les p(p-1)/2 covariances des variables prises 2 à 2. La covariance entre Xj et Xk :

$$S_{jk} = C_{X_j X_k} = \frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \overline{X}_k)(x_{ij} - \overline{X}_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ik}x_{ij} - \overline{X}_k\overline{X}_j$$

Avec ces éléments on peut construire la matrice de dispersion S(carrée et symétrique $S_{jk}=S_{kj}$) aussi appelée matrice de variance covariance. Les variances sont sur la diagonale et les covariances à l'extérieures de cette diagonale. S n'a pas une interprétation directe.

$$S = \begin{pmatrix} S_{11} & \cdots & S_{1p} \\ \vdots & & \vdots \\ S_{p1} & \cdots & S_{pp} \end{pmatrix}$$

Matrice des corrélations R

De même la matrice des corrélations décrit les corrélations de toutes les variables 2 à 2. Elle comporte des 1 sur la diagonale et les coefficients de corrélation linéaire entre les variables prises 2 à 2 en dehors de la diagonale. Elle caractérise la structure des corrélations entre variables.

$$R = \begin{pmatrix} r_{11} & & r_{1p} \\ & & \\ r_{p1} & & r_{pp} \end{pmatrix} ou\ r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj}S_{kk}}} = \frac{C_{X_jX_k}}{s_{X_j}s_{X_k}}$$

Si j=k le coefficient de corrélation d'une variable avec elle même est égal à 1

Tous les coefficients sont dans [-1, 1], c'est un nombre sans unité

Si r >> 0 corrélation croissante si r << 0 corrélation décroissante

Exemple des données météorologiques

1.0000	-0.1113	-0.2720	NaN	0.2131	0.8085
-0.1113	1.0000	0.0439	NaN	0.0348	-0.0514
-0.2720	0.0439	1.0000	NaN	-0.0625	0.1141
NaN	NaN	NaN	NaN	NaN	NaN
0.2131	0.0348	-0.0625	NaN	1.0000	0.2632
0.8085	-0.0514	0.1141	NaN	0.2632	1.0000

Représentation géométrique

Le tableau de données dispose d'une masse d'information sous forme rectangulaire. Le principe des méthodes de statistique exploratoire multidimensionnelle repose sur la représentation géométrique des n lignes (individus) et des p colonnes (variables) par des points. Deux représentations sont possibles :

- nuage de n points (individus) dans un espace euclidien de dimension p (généralisation du diagramme de dispersion), dit ***espace des individus***. Chaque ligne du tableau de données est un point qui a p coordonnées.
- nuage de p points (variables) dans un espace euclidien de dimension n, dit ***espace des variables***. Chaque colonne du tableau de données est un point qui a n coordonnées.

On peut ainsi définir des distances entre individus ou des distances entre variables suivant le cas. La proximité de deux points (individus) ou de deux points (variables) traduit des associations statistiques.

Distance de Mahalanobis

En physique la distance entre deux points se calcule généralement par la formule de Pythagore : carré de la distance est la somme des carrés des différences des coordonnées, car les dimensions sont de même nature.

$$d^2(x_i,x_k) = (x_i - x_k)^t (x_i - x_k) = \sum_j (x_{ij} - x_{kj})^2$$

En statistique chaque dimension correspond à une variable qui a une unité particulière. Comment calculer la distance entre 2 atmosphères décrites par 6 variables (cf exemple 1). Il faudrait alors pondérer chaque variable par un coefficient en fonction de son importance. De plus le caractère perpendiculaire des différentes directions est une convention de représentation totalement arbitraire on pourrai représenter les données avec des axes obliques. Utiliser la distance euclidienne simple revient en quelque sorte à dire que la température est perpendiculaire à l'humidité.

On utilisera donc une distance euclidienne généralisée de la forme :

$$d_M{}^2(x_i,x_k) = \left(x_i - x_k\right)^t M\left(x_i - x_k\right)$$

M est une matrice symétrique de taille p définie positive appelée métrique. Le choix de M dépend de l'utilisateur. En pratique

- $M = I$ revient à la distance usuelle
- $M = D_{\frac{1}{s^2}} = \begin{pmatrix} 1/S_1^2 & 0 & 0 & 0 \\ 0 & & 0 & 0 \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & 1/S_p^2 \end{pmatrix}$

  matrice diagonale des inverses des variances revient à diviser chaque variable par son écart type.

Ainsi la distance entre individus ne dépend plus des unités. Cette métrique donne à chaque caractère la même importance quelle que soit sa dispersion, alors que M = I donne plus d'importance aux variables les plus dispersées, pour lesquelles les différences entre individus sont plus grandes.

- $M = S^{-1}$ est appelée distance de Mahalanobis.

la distance de Mahalanobis est en particulier utile pour calculer la distance entre un individu x_i (un point) et le centre du nuage de point $\overline{X}$. Elle permet de prendre en compte la dispersion plus ou moins grande du nuage de point dans certaines directions

$$d_{S^{-1}}{}^2(x_i,\overline{X}) = \left(x_i - \overline{X}\right)^t S^{-1}\left(x_i - \overline{X}\right)$$

Elle fournira par exemple un critère pour la détection de valeurs extrêmes.

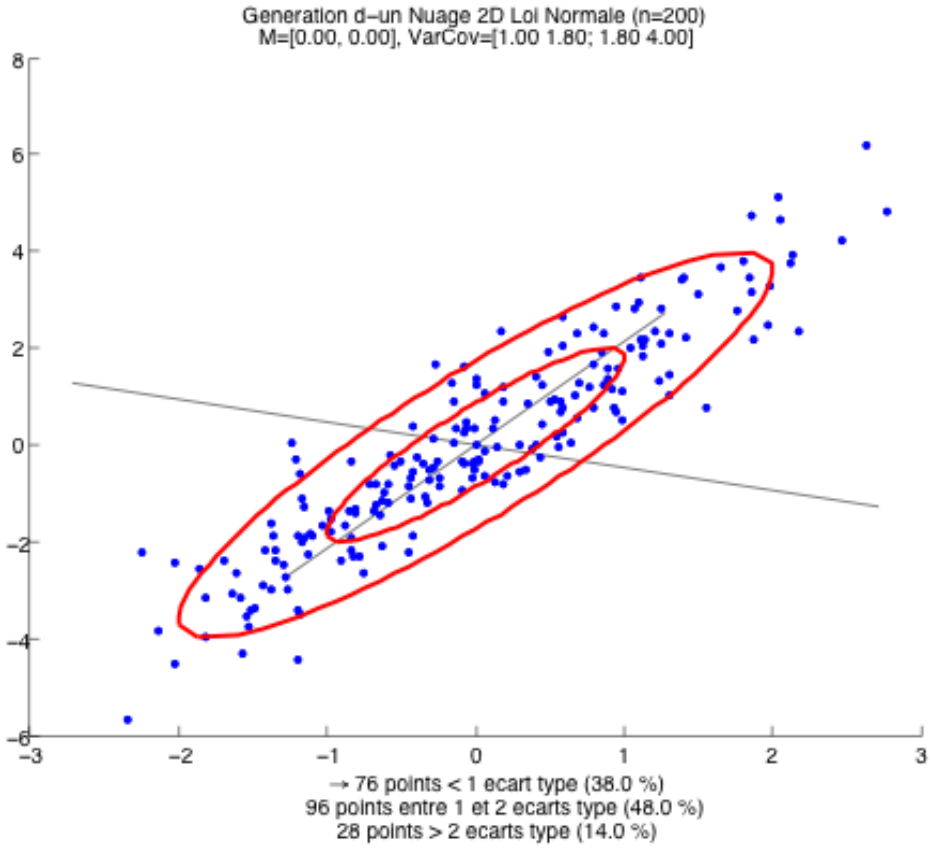

Figure 10 : Ellipsoïdes de Mahalanobis