MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 1: Introduction, Multivariate Location and Scatter

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Practical Things

Practical Things

- Lecturer: Pauliina Ilmonen, pauliina.ilmonen(a)aalto.fi
- The first lecture is on Monday January 8th at 12.15-14.00
- Exercises: Jaakko Pere, jaakko.pere(a)aalto.fi
- There are four exercise groups, choose the one that fits best to your schedule

Self Study

Before the course starts, make sure that you know how to calculate the univariate means, medians, variances, and max and min values. Familiarize yourself with the correlation coefficients and common graphical presentations (boxplots, scatter plots, histograms, bar plots, pie charts) of data. Learn to calculate the multivariate mean vector and covariance matrix. Make sure that you know what is a cumulative distribution function, a probability density function, and a probability mass function. Make sure that you know what is the expected value of a random variable. Read about univariate and multivariate normal distributions and elliptical distributions. Make sure that you know what is meant by central symmetric distributions and skew distributions. Recall what are the determinant, eigenvectors and eigenvalues of a matrix and make sure that you know what is meant by a symmetric matrix and a positive definite matrix.

How to pass this course?

You are expected to

- Attend the lectures and be active - not compulsory, no points, but highly recommended.
- Submit your project work on time - THIS IS COMPULSORY - max 6 points.
- Take the exam - max 24 points. (The course examinations is on Friday 19.4.)
- Participate to weekly exercises (group 1, group 2, group 3 OR group 4) - not compulsory, but highly recommended - max 3 points.
- Be ready to present your homework solutions in the exercise group - not compulsory, but highly recommended - max 3 points.

Max total points $= 6 + 24 + 3 + 3 = 36$. You need at least 16 points in order to pass the course.

Exercises

Participate to weekly exercises (group 1, group 2, group 3 OR group 4) - not compulsory, but highly recommended - max 3 points. If you attend 2-3 times, you get 1 point. If you attend 4-5 times, you get 2 points. If you attend at least 6 times (out of 11 times), you get 3 points.

In order to earn the exercise points, you have to arrive on time to the exercise session. The names of the participants are collected at the beginning of each exercise class. You can not get any exercise points without attending the exercises.

Exercise session 11 is reserved for the project work and for summarizing the contents of the course.

Attending all the exercise sessions, including the last one, is highly recommended!

Homework

Solve the homework problems and be ready to present your solutions in the exercise group - not compulsory, but highly recommended - max 3 points. Note that your solution does not have to be perfect or even correct — trying your very best is enough! If you solve your homework assignments 2-3 times, you get 1 point. If you solve your homework assignments 4-5 times, you get 2 points. If you solve your homework assignments at least 6 times (out of 10 times), you get 3 points.

In order to earn the homework points, you have to arrive on time to the exercise session and write your name to the homework list. You can not get any homework points without attending the exercises.

Project Work

Find a multivariate (at least 3-variate) dataset (Statistics Finland (=Tilastokeskus), OECD, collect yourself, ...), set a research question, and perform multivariate analysis. Write a report (max 10 pages), and submit it in MyCourses before Monday 15.4. at 12.00 (midday).

Goals of the project work:

- Description of the research questions
- Description of the dataset
- Univariate and bivariate statistical analysis to present the variables
- Application of your chosen multivariate statistical methods to answer research questions (justification and output)
- Conclusions and answers to the question raised at the beginning
- Critical evaluation of the analysis

Remember that No findings is a finding! Note that you will automatically get 0 points from the exam if you will not submit your project work on time!

How to get a good grade?

- Attend the lectures and be active!
- Work hard on your project work.
- Be active in exercises!
- Study for the exam!

Grading is based on the total points as follows: 16p -> 1, 20p -> 2, 24p -> 3, 28p -> 4, 32p -> 5.

Introduction

Introduction

The first step of all statistical analysis is the univariate and bivariate analysis. First calculate the univariate means, medians, variances, max and min values. Then calculate the correlation coefficients. And take a look at your data — literally! Make histograms of continuous variables and pie charts of categorical variables. Make boxplots to detect univariate outliers, and make scatter plots to detect bivariate structures.

Note that visualization is not always easy when the data contains a large number of individuals, but do not skip plotting your data! It is very important that you get familiar with your data before you conduct any large multivariate analysis.

Multivariate Location and Scatter

Multivariate Data

Let x denote a p-variate random vector with a cumulative distribution function F_x. Let X denote a $n \times p$ data matrix of independent and identically distributed (i.i.d.) observations $x_1, x_2, ..., x_n$ from the distribution F_x.

Location Functionals

Definition

A $p \times 1$ vector-valued functional $T(F_x)$, which is affine equivariant in the sense that

$$T(F_{Ax+b}) = AT(F_x) + b$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b, is called a location functional.

Scatter Functionals

Definition

A $p \times p$ matrix-valued functional $S(F_x)$ which is positive definite and affine equivariant in the sense that

$$S(F_{Ax+b}) = AS(F_x)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b, is called a scatter functional.

Location and Scatter Estimates

The corresponding sample statistics are obtained if the functionals are applied to the empirical cumulative distribution F_n based on a sample $x_1, x_2, \ldots, x_n$. Notation $T(F_n)$ and $S(F_n)$ or $T(X)$ and $S(X)$ is used for the sample statistics. The location and scatter sample statistics then also satisfy

$$T(XA^T + 1_n b^T) = AT(X) + b$$

and

$$S(XA^T + 1_n b^T) = AS(X)A^T$$

for all nonsingular $p \times p$ matrices A and for all p-vectors b.

Scatter Functionals

Scatter matrix functionals are usually standardized such that in the case of standard multivariate normal distribution $S(F_x) = I$.

Shape Functionals

Definition

If a positive definite $p \times p$ matrix-valued functional $S(F_x)$ satisfies that $S(F_{Ax+b})$ is proportional to $AS(F_x)A^T$ for all nonsingular $p \times p$ matrices A and for all p-vectors b, then $S(F_x)$ is called a shape functional.

Traditional Functionals

The first examples of location and scatter functionals are the mean vector and the regular covariance matrix:

$$T_1(F_x) = E(x) \text{ and } S_1(F_x) = Cov(F_x) = E\left((x - E(x))(x - E(x))^T\right).$$

The Sample Mean Vector and the Sample Covariance Matrix

Traditional estimates of the mean vector and the covariance matrix are calculated as follows:

$$T_1(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and

$$S_1(X) = Cov(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left((x_i - T_1(X))(x_i - T_1(X))^T\right).$$

Why do we need other location and scatter measures???

Scatter Functionals

There are several other location and scatter functionals, even families of them, having different desirable properties (robustness, efficiency, limiting multivariate normality, fast computations, etc).

Moment Based Functionals

Location and scatter functionals can be based on the third and fourth moments as well. A location functional based on third moments is

$$T_2(F_x) = \frac{1}{p} E\left((x - E(x))^T Cov(F_x)^{-1} (x - E(x)) x\right)$$

and a scatter matrix functional based on fourth moments is

$$S_2(F_x) = \frac{1}{p+2} E\left((x - E(x))(x - E(x))^T Cov(F_x)^{-1} (x - E(x))(x - E(x))^T\right).$$

Example 1: Bivariate Normal Distribution

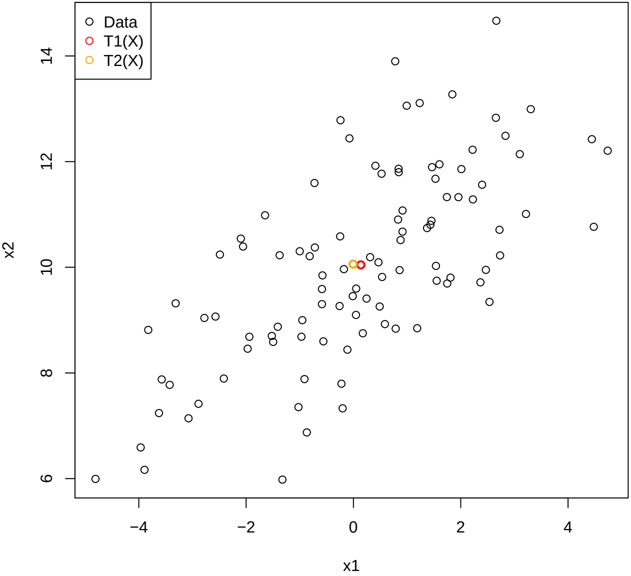
In this example we consider bivariate normal distribution $N(\mu, \mathrm{A})$, where

$$\mathrm{A} = \begin{bmatrix} 4 & 2 \\ 2 & 3 \end{bmatrix}$$

and

$$\mu = \begin{bmatrix} 0 & 10 \end{bmatrix}.$$

Example 1: Bivariate Normal Distribution



Example 1: Bivariate Normal Distribution

We simulated 100 samples from $N(\mu, \mathrm{A})$ and we then calculated the sample mean vector $T_1(X)$, the location vector based on third moments $T_2(X)$, the sample covariance matrix $S_1(X)$ and the scatter matrix based on fourth moments $S_2(X)$ of each sample. In order to compare $T_1(X)$, $T_2(X)$, $S_1(X)$, and $S_2(X)$, we calculated the means of the estimates.

$$T_1(X): \begin{bmatrix} 0.006703295 \\ 10.001765054 \end{bmatrix} \qquad T_2(X): \begin{bmatrix} 0.01626947 \\ 9.99082058 \end{bmatrix}$$

$$S_1(X): \begin{bmatrix} 4.029396 & 2.034711 \\ 2.034711 & 2.968536 \end{bmatrix} \qquad S_2(X): \begin{bmatrix} 3.9197916 & 2.003406 \\ 2.003406 & 2.924344 \end{bmatrix}$$

Both location estimates seem to estimate the parameter μ and both scatter estimates seem to estimate the parameter A.

Example 2: Independent Components, Skewed Distributions

In this example we consider $\mathrm{Gamma}(\alpha, \beta)$ and $\chi^2(k)$ distributions, where

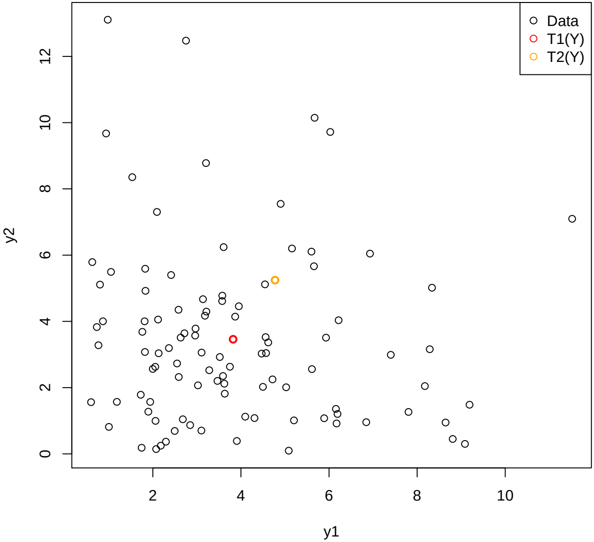
$$\alpha = 2,$$

$$\beta = 0.5$$

and

$$k = 3.$$

Example 2: Independent Components, Skewed Distributions



Example 2: Independent Components, Skewed Distribution

As in Example 1, we ran the simulation 100 times and calculated the means.

$$T_1(Y): \begin{bmatrix} 4.031022 \\ 2.964918 \end{bmatrix} \qquad T_2(Y): \begin{bmatrix} 5.944029 \\ 4.740199 \end{bmatrix}$$

$$S_1(Y): \begin{bmatrix} 8.16111692 & 0.04234064 \\ 0.04234064 & 5.76640662 \end{bmatrix} \qquad S_2(Y): \begin{bmatrix} 13.4080726 & 0.1142734 \\ 0.1142734 & 9.8194396 \end{bmatrix}$$

Here the location estimates differ significantly from each other. Also the scatter estimates differ significantly from each other. Note also that the off-diagonal elements of both scatter estimates are small.

Location and Scatter Functionals Under Symmetry Assumptions

We now consider the behavior of scatter and location functionals under some symmetry assumptions.

Theorem

Under the assumption of central symmetry, all location functionals are equal to the center of symmetry.

Proof

Let x denote a p-variate random vector with a cumulative distribution function F_x. Let $\theta \in \mathbb{R}^p$ and assume that $x - \theta \sim -(x - \theta)$. Let T be an affine equivariant location functional and assume that $T(F_x)$ exists as finite quantity.

Since T is affine equivariance and since x is symmetric about θ, we have that

$$T(F_x) - \theta = T(F_{x-\theta}) = T(F_{-(x-\theta)}) = T(F_{-x+\theta}) = -T(F_x) + \theta.$$

Thus

$$2T(F_x) = 2\theta$$

and it follows that

$$T(F_x) = \theta.$$

Since T was an arbitrarily chosen location functional, this completes the proof.

Theorem

Under the assumption of multivariate elliptical distribution, all scatter functionals are proportional.

Proof (1/3)

Let x denote a p-variate random vector with a cumulative distribution function F_x. Assume that

$$x = \Omega z + \mu,$$

where $\mu \in \mathbb{R}^p$, $\Omega \in \mathbb{R}^{p \times p}$, Ω is full rank, and $z \sim Oz$ for all orthogonal $O \in \mathbb{R}^{p \times p}$. Let S be an affine equivariant scatter functional and assume that $S(F_x)$ exists as finite quantity.

Proof (2/3)

Since $z \sim Oz$ for all orthogonal $O \in \mathbb{R}^{p \times p}$, it holds that $z \sim PJz$ for all permutation matrices $P \in \mathbb{R}^{p \times p}$ and for all sign change matrices $J \in \mathbb{R}^{p \times p}$. Now it follows from affine equivariance of S that

$$S(F_{PJz}) = PJS(F_z)(PJ)^T$$

and from the property $PJz \sim z$ that

$$S(F_{PJz}) = S(F_z).$$

Thus

$$S(F_z) = PJS(F_z)(PJ)^T.$$

As $S(F_z) = PJS(F_z)(PJ)^T$ holds for all permutation matrices $P \in \mathbb{R}^{p \times p}$ and for all sign change matrices $J \in \mathbb{R}^{p \times p}$, we have that

$$(S(F_z))_{ij} = -(S(F_z))_{ji}, \; i \neq j$$

and

$$(S(F_z))_{ii} = (S(F_z))_{jj}.$$

Thus

$$S(F_z) \propto I.$$

Proof (3/3)

It now follows from above and from affine equivariance of S that

$$S(F_x) = S(F_{\Omega z+\mu}) = \Omega S(F_z)\Omega^T = \Omega c \cdot I\Omega^T = c\Omega\Omega^T,$$

where c is a constant that may depend on S.

Since S was an arbitrarily chosen scatter functional, this completes the proof.

Note that in general different location functionals do not measure the same population quantities. That is true also for scatter functionals — different scatter functional do not necessarily measure the same population quantities!

Next Week

Next week we will talk about principal component analysis (PCA).

References

References I

- K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).
- H. Oja, Multivariate Nonparametric Methods With R, Springer-Verlag, New York, 2010.

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 2: Principal Component Analysis

Lecturer: Pauliina Ilmonen
Slides: Ilmonen

Contents

PCA transformation

PCA-transformation

Principal Component Analysis (PCA) looks for few linear combinations of p variables, losing in the process as little information as possible. More precisely, PCA transformation is an orthogonal linear transformation that transforms a p-variate random vector to a new coordinate system such that, the obtained new variables are uncorrelated, and the greatest possible variance lies on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

PCA-transformation

Let x denote a p-variate random vector with finite mean $E[x] = \mu$, and finite covariance matrix $E[(x-\mu)(x-\mu)^T] = \Sigma$. The Principal Component Transformation is the transformation

$$x \to y = \Gamma^T (x - \mu),$$

where $\Gamma \in \mathbb{R}^{p \times p}$ is orthogonal, $\Gamma^T \Sigma \Gamma = \Lambda = diag(\lambda_1, \cdots, \lambda_p)$ is diagonal and $\lambda_1 \geq \cdots \geq \lambda_p$.

The ith component of y is called the ith principal component of x.

Theoretical Properties

Lecturer:
Pauliina Ilmonen
Slides: Ilmonen

Theorem

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ denote the eigenvalues of Σ, and let y_i denote the ith principal component of x. Then

1. $E[y_i] = 0$,
2. $var(y_i) = E[y_i^2] = \lambda_i$,
3. $cov(y_i, y_j) = E[y_i y_j] = 0, i \neq j$,
4. $var(y_1) \geq \cdots \geq var(y_p) \geq 0$.

Proof.

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ. Let $y = \Gamma^T(x-\mu)$, where $\Gamma \in \mathbb{R}^{p\times p}$ is orthogonal, $\Gamma^T\Sigma\Gamma = \Lambda = diag(\lambda_1, \cdots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p$. Let γ_i denote the ith column vector of Γ. Now

1.

$$E[y_i] = E[\gamma_i^T(x-\mu)] = E[\gamma_i^T x] - E[\gamma_i^T\mu]$$
$$= \gamma_i^T E[x] - \gamma_i^T\mu = \gamma_i^T\mu - \gamma_i^T\mu = 0,$$

and

2., 3., 4.

$$E[(y-E[y])(y-E[y])^T] = E[yy^T] = E[\Gamma^T(x-\mu)(\Gamma^T(x-\mu))^T]$$
$$= \Gamma^T E[(x-\mu)((x-\mu))^T]\Gamma = \Gamma^T\Sigma\Gamma = \Lambda.$$

□

Maximizing Variance

Theorem

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ, and let y_1 denote the first principal component of x. Assume that $a \in \mathbb{R}^p$, $a^T a = 1$. Then $var(y_1) \geq var(a^T x)$.

Proof.

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ. Let $y = \Gamma^T(x-\mu)$, where $\Gamma \in \mathbb{R}^{p\times p}$ is orthogonal, $\Gamma^T\Sigma\Gamma = \Lambda = diag(\lambda_1, \cdots, \lambda_p)$ is diagonal and $\lambda_1 \geq \cdots \geq \lambda_p$. Let γ_i denote the ith column of Γ. Assume that $a \in \mathbb{R}^p$, $a^T a = 1$.

Since the set $\{\gamma_1, \ldots, \gamma_p\}$ is an orthonormal basis of $\mathbb{R}^p$, the vector a can be given as $a = c_1\gamma_1 + \cdots + c_p\gamma_p$. Now, since $\gamma_i^T\gamma_i = 1$, and $\gamma_i^T\gamma_j = 0$ if $j \neq i$, we have that

$$var(a^T x) = a^T\Sigma a = \sum_{j=1}^{p} c_j\gamma_j^T \Big(\sum_{i=1}^{p}\lambda_i\gamma_i\gamma_i^T\Big)\sum_{k=1}^{p} c_k\gamma_k = \sum_{i=1}^{p}\lambda_i c_i^2,$$

and since a satisfies $a^T a = 1$, we have that $\sum_{i=1}^{p} c_i^2 = 1$. Thus, since λ_1 is the largest eigenvalue, the variance $var(a^T x)$ is maximized when $c_1 = 1$, and $c_i = 0, i \neq 1$, and consequently $a = \gamma_1$. This completes the proof. $\square$

Maximizing Variance

Theorem

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ, and let y_k denote the kth principal component of x. Let $b \in \mathbb{R}^p$, $b^T b = 1$. Assume that $b^T x$ is uncorrelated with the first $k-1$ principal components of x. Then $var(y_k) \geq var(b^T x)$.

Proof. This is homework! (The proof is very similar to the previous proof. Note that if $b^T x$ is uncorrelated with the first $k-1$ principal components of x, then b can be given as linear combination of the vectors $\gamma_k, \ldots, \gamma_p$.)

Total Variance

The sum of the first k eigenvalues divided by the sum of all eigenvalues

$$\frac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}$$

represents the proportion of total variance explained by the first k principal components. (Total variation is here understood as the trace of Σ.)

Note that if $y = \Gamma^T(x - \mu)$, then

$$x = \mu + \Gamma y = \mu + \sum_{i=1}^{p} y_i \gamma_i \approx \mu + \sum_{i=1}^{k} y_i \gamma_i.$$

How many components to choose?

Some rules of thumb:

Choose as many components as is needed in order to explain at least 90% (or 80% or 95 %) of the total variance.

Leave out the components that correspond to "small" eigenvalues. (More about this in class.)

Sample Version

Sample version of PCA is obtained by replacing the covariance matrix and the mean vector by their sample estimates. Each p-variate data point is transformed using the sample mean vector and the eigenvector matrix of the sample covariance matrix.

Sample PCA

Let X denote a $n \times p$ data matrix of n independent and identically distributed p-variate observations $x_1, x_2, ..., x_n$ from some continuous distribution with finite mean vector μ, and finite covariance matrix Σ. Let $\bar{x}$ denote the sample mean vector and let G denote the eigenvector matrix of the sample covariance matrix $\hat{\Sigma}$, where the column vectors of G are the eigenvectors of $\hat{\Sigma}$ such that the first vector corresponds to the largest eigenvalue, the second column vector corresponds to the second largest eigenvalue, and so on.

The sample PCA transformation is now given by

$$Y = (X - 1_n \bar{x}^T)G.$$

(Note that now $y_r = G^T(x_r - \bar{x})$.)

Sample PCA, Scores

Consider the transformation given in the previous slide. Now y_{ri} represents the score of the ith principal component on the rth individual.

Sample PCA, Total Variance

Let $\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_p$ denote the eigenvalues of the sample covariance matrix $\hat{\Sigma}$. Now

$$\hat{\lambda}_i = \frac{1}{n} \sum_{r=1}^{n} y_{ri}^2.$$

Thus the contribution of the individual r on the variance $\hat{\lambda}_i$ is given by

$$\frac{\frac{1}{n} y_{ri}^2}{\hat{\lambda}_i}.$$

Sample PCA, Quality of Representation

The quality of the representation of the individual r by the principal axis i is measured by the squared cosines of the angle between the (centered) vectors.

$$cos_r^2(\alpha) = \frac{y_{ri}^2}{\sum_{j=1}^{p}((X - 1_n\bar{x}^T)_{rj})^2}.$$

If the value is close to 1, the quality of the representation is good.

Applications

Applications

- Dimension reduction
- Outlier detection
- Clustering
- Dimension reduction in regression analysis
- ...

Example

Simulated Example

In this example, we simulated a sample from bivariate normal distribution with mean $(3, 2)^T$, and covariance matrix

$$B = \begin{bmatrix} 1.50 & 0.70 \\ 0.70 & 7.00 \end{bmatrix}.$$

PCA transformation was performed. After PCA, the greatest variation is seen in the first axis.

Example

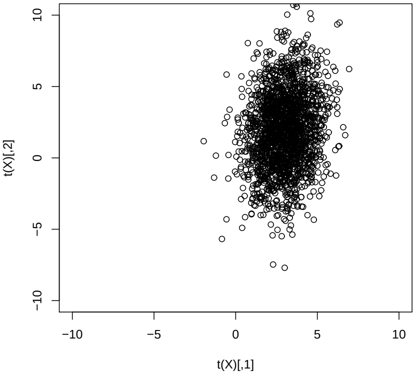


Figure: Bivariate normal distribution.

Example

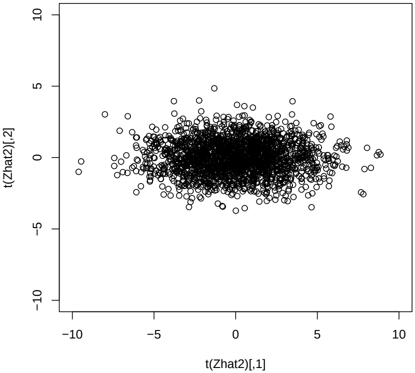


Figure: Bivariate normal distribution after PCA.

Real Data Example

Data Example - Growth

To see how PCA works in practice, let's take a look at a real data example. The data set used in this example is part of a larger sample of height measurements that were collected retrospectively from health centers and schools for construction of the Finnish growth charts. The used data set comprised 525 boys and 571 girls, fullterm, healthy singletons, followed until approximately age 19, with measurements from three to 44 occasions.

Data Example

The original observations were used to estimate each individual growth curve from birth to age 19 by fitting splines. The individuals that did not have enough measurements for fitting the splines were excluded. After that, the remaining observations consisted of 829 (481 boys and 348 girls) estimated height curves. The measurements (based on estimated curves) at ages 8, 9, 10, 11, 12, 13, 14, 15, 16, 17 and 18 years were used in the analysis. Thus PCA was applied to a 11-dimensional sample with 829 observations.

Data Example

PCA was first used for dimension reduction. The first principal component explained 77 %, the second 17 % and the third 4 % of the variance of the data. Thus the first, second and third principal component together already explained 98 % of the variance, and dimension was reduced to three.

Three Principal Components

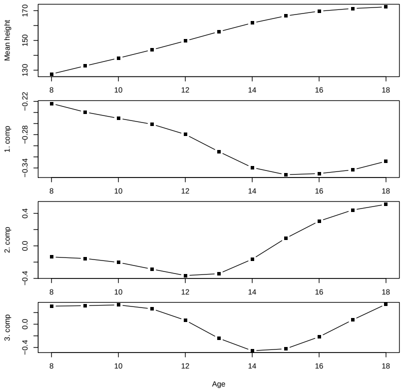


Figure: Mean curve of the estimated data points and the three first principal component curves (the three first column vectors of Γ). The first principal component curve puts emphasis on overall growth (shape of the curve is similar to the mean curve), the second on late growth, and the third on growth around age 14.

To see how the method works on the individual level, the estimated height growth curves of one randomly chosen boy and one randomly chosen girl were presented as sums of their principal component curves. The estimated growth curve of one randomly chosen boy in terms of principal components is presented in Figure 4 and the estimated growth curve of one randomly chosen girl in terms of principal components is presented in Figure 5. The method seems to work very well also on individual level. In these examples only two principal are needed for being very close to the curve based on splines.

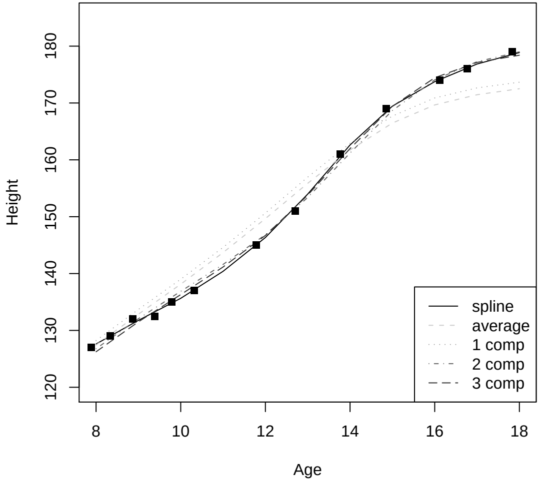


Figure: Estimated growth curve of one randomly chosen boy.

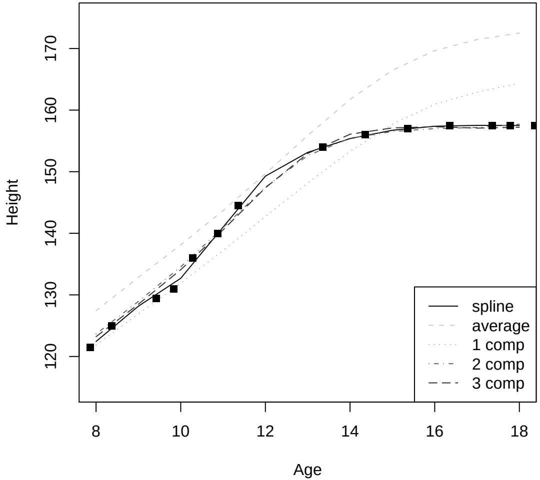


Figure: Estimated growth curve of one randomly chosen girl.

Scatter plot after PCA was considered to see if PCA works in separating genders.

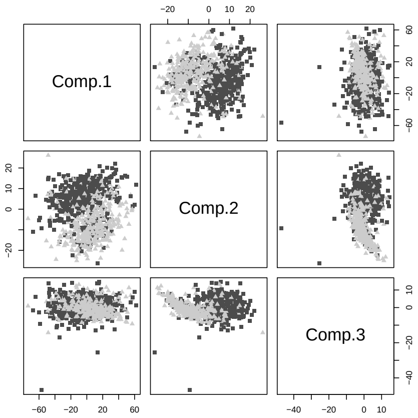


Figure: Scatter plot after PCA. Dark grey squares are used for the boys and light grey triangles for the girls. PCA does not work perfectly in separating the two groups, but one can still see clear differences between the groups. Boys grow later than girls! (Notice the outlying points.)

Words of Warning

Some Words of Warning

- Principal Components are not in general independent
- PCA is a very nonrobust method.
- Traditional PCA is not suitable for qualitative variables.
- PCA transformation is invariant under orthogonal transformations up to heterogeneous sign changes, but it is not affine invariant. In fact, PCA transformation is highly sensitive for scaling of the variables.

More about these issues next week...

Next Week

Next week we will continue talking about principal component analysis.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 3: Principal Component Analysis - part II

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

PCA Using Correlation Matrix

PCA Using Correlation Matrix

As was pointed out last week, PCA is highly sensitive for scaling of the variables. One can address this problem by standardizing the variables first. The data can be standardized by subtracting the sample mean $\bar{x}$, and then dividing each variable by the corresponding square root of the sample variance $\hat{\sigma}_{ii}$. PCA is then applied to this preprocessed data. Note that for standardized variables, the covariance matrix Σ turns into a correlation matrix.

PCA Using Correlation Matrix

If PCA is performed standardizing the variables first, it naturally becomes scale-invariant.

If variables do not have the same natural units, it is better to standardize the data first. For example, if the variables considered are weight, height, age, and IQ, it is a good idea to think about standardizing the data first. But if the variables do share the same units and if there are no large differences between the variances, then one can apply standard PCA.

PCA Using Correlation Matrix

One may address the problem of scale-sensitivity by standardizing the data first. However, this standardization does not make PCA fully invariant under all linear transformations.

Correlation Structure in PCA

Correlation Structure

Theorem

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ. Let σ_{ii} denote the ith diagonal element of Σ. Let $y = \Gamma^T(x - \mu)$, where $\Gamma \in \mathbb{R}^{p \times p}$ is orthogonal, $\Gamma^T \Sigma \Gamma = \Lambda = diag(\lambda_1, \cdots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p$. Let γ_j denote the jth column vector of Γ and let γ_{ij} denote the ith element of it (i.e. γ_{ij} denotes the ij element of Γ). Then

$$corr(x_i y_j) = \rho_{ij} = \frac{\gamma_{ij}\lambda_j}{\sqrt{\sigma_{ii}\lambda_j}}.$$

Correlation Structure

Proof.

Let x denote a p-variate random vector with finite mean vector μ, and finite covariance matrix Σ. Let σ_{ii} denote the ith diagonal element of Σ. Let $y = \Gamma^T(x - \mu)$, where $\Gamma \in \mathbb{R}^{p \times p}$ is orthogonal, $\Gamma^T \Sigma \Gamma = \Lambda = diag(\lambda_1, \cdots, \lambda_p)$ and $\lambda_1 \geq \cdots \geq \lambda_p$. Let γ_j denote the jth column vector of Γ and let γ_{ij} denote the ith element of it (i.e. γ_{ij} denotes the ij element of Γ). Now

$$E[(x - \mu)y^T] = E[(x - \mu)(\Gamma^T((x - \mu)))^T]$$

$$= E[((x - \mu))((x - \mu))^T \Gamma] = \Sigma\Gamma = \Gamma\Lambda.$$

Therefore the covariance between x_i and y_j is $\gamma_{ij}\lambda_j$. Since x_i and y_j have variances σ_{ii} and λ_j, respectively, the correlation between x_i and y_j is given by

$$\rho_{ij} = \frac{\gamma_{ij}\lambda_j}{\sqrt{\sigma_{ii}\lambda_j}}.$$

□

Correlation Structure

It can be said that "the proportion of the variation" of x_i explained by y_j is ρ_{ij}^2. Since the elements of y are uncorrelated, any set S of components explain a proportion

$$\rho_{iS}^2 = \sum_{j \in S} \rho_{ij}^2.$$

Note that when Σ is a correlation matrix, the variance $\sigma_{ii} = 1$ and thus $\rho_{ij} = \gamma_{ij}\sqrt{\lambda_j}$.

Multivariate Linear Regression

Multivariate Linear Regression

Regression analysis is used to predict the value of one or more responses from a set of predictors. Predictors can be continuous or categorical or a mixture of both.

Multivariate Regression Model

Let z be a p-variate random vector of dependent variables such that

$$z = B^T v + u,$$

where v is a q-variate fixed vector of predictors, B is a $q \times p$ matrix of regression parameters, and u is a p-variate vector of random errors with mean 0, and common covariance matrix C. The first element of v is assumed to be 1 (to allow a mean effect).

Multivariate Linear Regression

Assume that we have a size n sample from the multivariate regression model. Then Z is a $n \times p$ data matrix such that

$$Z = VB + U,$$

where V is a known $n \times q$ matrix, B is a $q \times p$ matrix, and U is a $n \times p$ matrix of unobserved random disturbances. The elements of the first column of V are all assumed to be 1, and the rows of U are assumed to be uncorrelated.

Estimation

Assume that Z is a $n \times p$ data matrix such that

$$Z = VB + U,$$

where V is a known $n \times q$ matrix, B is a $q \times p$ matrix, and the $n \times p$ error matrix U is independent of V. The elements of the first column of V are all assumed to be 1. Assume that the rows of the error matrix U are independent and identically distributed with the mean vector $\mu = 0$ and the covariance matrix C. Assume that the inverse of $V^T V$ exists.

Let

$$P = I - V(V^T V)^{-1} V^T.$$

Now, the generalized least squares estimators of B and C can be given as

$$\hat{B} = (V^T V)^{-1} V^T Z$$

and

$$\hat{C} = \frac{1}{n} Z^T P Z.$$

Estimation

The estimate $\hat{B}$ can be used in estimating/predicting the values of the matrix Z,

$$\hat{Z} = V\hat{B}.$$

The estimate of the error matrix is obtained by taking the difference between Z and $\hat{Z}$

$$\hat{U} = Z - V\hat{B}.$$

Trace Correlation and Determinant Correlation

Assume that the matrix Z is centered so that the columns of Z have zero mean. Define now

$$D = (Z^T Z)^{-1} \hat{U}^T \hat{U}.$$

The matrix $\hat{U}^T \hat{U}$ ranges between zero, when all the variation of Z is explained by the regression model, and $Z^T Z$, when no part of the variation in Z is explained by V. Therefore $I - D$ varies between the identity matrix and the zero matrix. It can be shown that all the eigenvalues of $I - D$ lie between 1 and 0.

Trace Correlation and Determinant Correlation

It would be desirable that a measure of multivariate correlation would range between zero and one. This property is satisfied by two often used coefficients, the trace correlation r_T and the determinant correlation r_D,

$$r_T^2 = \frac{1}{p} tr(I - D),$$

and

$$r_D^2 = det(I - D).$$

Note that the coefficient r_D is zero if at least one of the eigenvalues of $I - D$ is zero, and r_T is zero if and only if all the eigenvalues of $I - D$ are zero.

Some Comments

- We assumed that the inverse of $V^T V$ exists. If it does not (or if some of the columns of V are nearly collinear), consider using smaller number of variables.
- One should not use the regression model for predicting outside of the range of the Z values. Behavior of extreme points may be different!
- Traditional L_2 regression is very sensitive to outlying observations.

PCA in Regression Analysis

PCA in Regression Analysis

Linear regression analysis is unstable in the presence of multicollinearity, or near multicollinearity, of the predictors. In this situation, PCA can be used to preprocess the data. Instead of performing regression analysis using the original variables, one can perform it using new variables obtained from PCA

PCA in Regression Analysis

Linear regression analysis is unstable in the presence of highly linearly dependent predictors. This problem is often solved simply by disregarding some of the predictors. Alternatively, PCA can be used to preprocess the data. Instead of performing regression analysis using the original variables, one can perform it using new variables obtained from PCA.

PCA in Regression Analysis

In general, when PCA is used, the principal components with the largest variance are chosen in order to explain as much of the total variation of x as possible. In regression settings, the choice of the components is somewhat different. In the context of regression, it is sensible to choose the components having the largest correlation with the most interesting dependent variables, because the purpose is to use the components in explaining the dependent variables. Fortunately, there is often a tendency in data for the components with largest variances to best explain the dependent variables.

PCA in Regression Analysis

If the original regression equation is given by

$$z = B^T v + u,$$

then also

$$z = A^T w + u,$$

where $w = \Gamma^T z$, Γ^T is the principal component transformation matrix, and $A = \Gamma^T B$. For the corresponding sample version it also holds that if

$$Z = VB + U,$$

then

$$Z = WA + U,$$

where $W = VG$, and $A = G^T B$.

One can now reduce dimension by deleting some of the columns of W.

Next Week

Next week we will talk about robust principal component analysis.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 4: Measures of Robustness, Robust Principal Component Analysis

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Measures of Robustness

Robust Statistical Methods

In statistics, robust methods are methods that perform well – or do not perform too poorly – in the presence of outlying observations.

Robust Statistical Methods, Example

Mean vs median...

Measures of Robustness

Let x denote a random variable or a random vector with a cumulative distribution function F_x, and let $X = \{x_1, x_2, ..., x_n\}$, where $x_1, x_2, ..., x_n$ are n independent and identically distributed observations from the distribution F_x. Consider functional $Q(F_x)$ (or $Q(F_n)$). We wish to measure robustness of that functional.

Influence Function

Influence Function

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Influence function measures the effect on functional Q when the underlying distribution deviates slightly from the assumed one.

$$IF(y, Q, F_x) = \lim_{0<\varepsilon\to 0} \frac{Q((1-\varepsilon)F_x + \varepsilon\delta_y) - Q(F_x)}{\varepsilon},$$

where δ_y is the cumulative distribution function having all its probability mass at y i.e.

$$\delta_y(t) = \begin{cases} 0, & t < y, \\ 1, & t \geq y. \end{cases}$$

Influence Function

Influence function measures the effect of point-mass contamination, and thus it is considered as a measure of local robustness.

A functional with bounded influence function (with respect to for example L_2 norm) is considered as robust and desirable.

Example, Population Mean (Expected Value)

$$IF(y,\mu,F_x) = \lim_{0<\varepsilon\to 0} \frac{\mu((1-\varepsilon)F_x+\varepsilon\delta_y)-\mu(F_x)}{\varepsilon}$$
$$= \lim_{0<\varepsilon\to 0} \frac{E[(1-\varepsilon)x+\varepsilon y]-E[x]}{\varepsilon}$$
$$= \lim_{0<\varepsilon\to 0} \frac{E[x-\varepsilon x+\varepsilon y]-E[x]}{\varepsilon}$$
$$= \lim_{0<\varepsilon\to 0} \frac{E[x]-\varepsilon E[x]+\varepsilon E[y]-E[x]}{\varepsilon}$$
$$= \lim_{0<\varepsilon\to 0} \frac{-\varepsilon E[x]+\varepsilon y}{\varepsilon}$$
$$= \lim_{0<\varepsilon\to 0} -E[x]+y = -E[x]+y = y-E[x] = y-\mu(F_x).$$

This is not bounded with respect to y.

Empirical Influence Function

Empirical Influence Function

Empirical influence function (also called the sensitivity curve) is a measure of the dependence of the estimator on the value of one of the points in the sample.

The empirical influence function can be seen as an estimate of the theoretical influence function.

Empirical Influence Function

Let $X = \{x_1, x_2, ..., x_n\}$, and let $X_y = \{x_1, x_2, ..., x_n, y\}$. Now

$$IF_E(y, Q, F_n) = \frac{Q((1 - \frac{1}{n+1})F_n + \frac{1}{n+1}\delta_y) - Q(F_n)}{\frac{1}{n+1}}$$

$$= (n+1)(Q((1 - \frac{1}{n+1})F_n + \frac{1}{n+1}\delta_y) - Q(F_n))$$

$$= (n+1)(Q(X_y) - Q(X)).$$

Example, Sample Mean

$$IF_E(y, \hat{\mu}, F_n) = (n+1)(\hat{\mu}(X_y) - \hat{\mu}(X)$$

$$= (n+1)(\frac{1}{n+1}(\sum_{i=1}^{n} x_i + y) - \frac{1}{n}\sum_{i=1}^{n} x_i)$$

$$= \sum_{i=1}^{n} x_i + y - \frac{n+1}{n}\sum_{i=1}^{n} x_i$$

$$= y - (\frac{n+1}{n} - 1)\sum_{i=1}^{n} x_i$$

$$= y - (\frac{n+1-n}{n})\sum_{i=1}^{n} x_i$$

$$= y - \frac{1}{n}\sum_{i=1}^{n} x_i = y - \hat{\mu}(X).$$

This is not bounded with respect to y. Note that the empirical influence function estimates the theoretical influence function.

Breakdown Point

Breakdown Point

Another very often used measure of robustness is the breakdown point. Whereas influence function measures local robustness, the breakdown point can be seen as a measure of global robustness.

Breakdown Point

Let $X_n = \{x_1, x_2, ..., x_n\}$, where $x_1, x_2, ..., x_n$ are n independent and identically distributed observations from the distribution F_x. Assume that $m < n$ and replace $x_1, x_2, ..., x_m$ with $x_1^*, x_2^*, ..., x_m^*$. Let $X_n^* = \{x_1^*, x_2^*, ..., x_m^*, x_{m+1}, ..., x_n\}$.

Now, the maximum bias

$$maxBias(m, X_n, Q) = \sup_{x_1^*, x_2^*, \ldots, x_m^*} d(Q(X_n), Q(X_n^*)),$$

where $d(\cdot, \cdot)$ denotes some distance function (for example the Euclidean distance).

The finite sample breakdown point is now given by

$$BP(Q, n) = \min_m \{\frac{m}{n} \mid maxBias(m, X_n, Q) = \infty\},$$

and the (asymptotic) breakdown point

$$BP(Q) = \lim_{n \to \infty} BP(Q, n).$$

Breakdown Point

A functional with large breakdown point is considered as robust. If $BP(Q) = \frac{1}{2}$, then Q is very robust (according to its breakdown point), and if $BP(Q) = 0$, then Q is very nonrobust. When the value is in between $\frac{1}{2}$ and 0, then it is a matter of taste ;-).

Example, Sample Mean

Let $X_n = \{x_1, x_2, ..., x_n\}$ a sample of be independent and identically distributed observations from some distribution F_x. Let $\hat{\mu}(X_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ and let $\hat{\mu}(X_n^*) = \frac{1}{n}(\sum_{i=2}^{n} x_i + x_1^*)$. Let $d(\hat{\mu}(X_n), \hat{\mu}(X_n^*))$ be the Euclidean distance between $\hat{\mu}(X_n)$ and $\hat{\mu}(X_n^*)$. If now $x_1^* \to \infty$, then also $\hat{\mu}(X_n^*) \to \infty$ and consequently

$$maxBias(1, X_n, \hat{\mu}) = \sup_{x_1^*} d(\hat{\mu}(X_n), \hat{\mu}(X_n^*)) = \infty.$$

Contaminating just one data point is enough to make the Euclidean distance arbitrarily large. Thus the finite sample breakdown point

$$BP(\hat{\mu}, n) = \frac{1}{n}$$

and the (asymptotic) breakdown point of the sample mean is

$$BP(\hat{\mu}) = \lim_{n\to\infty} BP(\hat{\mu}, n) = \lim_{n\to\infty} \frac{1}{n} = 0.$$

Example, Sample Median (1/4)

Let $X_n = \{x_1, x_2, ..., x_n\}$ be a sample of independent and identically distributed observations from some distribution F_x. Let $Med(X_n)$ be the sample median calculated from the original sample and let $Med(X_n^*)$ be the sample median calculated from the contaminated sample $X_n^* = \{x_1^*, x_2^*, ..., x_m^*, x_{m+1}, ..., x_n\}$. Let $d(Med(X_n), Med(X_n^*))$ be the Euclidean distance between $Med(X_n)$ and $Med(X_n^*)$.

Example, Sample Median (2/4)

Assume first that n is even. When n is even, the sample median is the average of the two middle values of the ordered observations. Now, one has to contaminate at least half of the observations in order to make the sample median and consequently the Euclidean distance arbitrarily large. Thus, for even n, the number of contaminated observations m has to be at least $n/2$ for

$$maxBias(m, X_n, Med) = \sup_{x_1^*, x_2^*, \ldots, x_m^*} d(Med(X_n), Med(X_n^*)) = \infty$$

to hold, and the finite sample breakdown point is then

$$BP(Med, n) = \min_m \{\frac{m}{n} \mid maxBias(m, X_n, Med) = \infty\} = \frac{n/2}{n} = \frac{1}{2}.$$

Example, Sample Median (3/4)

Assume now that n is odd. When n is odd, the sample median is the middle value of the ordered observations. Now, one has to contaminate at least $(n+1)/2$ observations in order to make the sample median and consequently the Euclidean distance arbitrarily large. Thus, for odd n, the number of contaminated observations m has to be at least $(n+1)/2$ for

$$maxBias(m, X_n, Med) = \sup_{x_1^*, x_2^*, \ldots, x_m^*} d(Med(X_n), Med(X_n^*)) = \infty$$

to hold, and the finite sample breakdown point is then

$$BP(Med, n) = \min_m \{\frac{m}{n} \mid maxBias(m, X_n, Med) = \infty\} = \frac{(n+1)/2}{n} = \frac{n+1}{2n}.$$

Example, Sample Median (4/4)

If n is even, the (asymptotic) breakdown point

$$BP(Med) = \lim_{n\to\infty} BP(Med, n) = \lim_{n\to\infty} \frac{1}{2} = \frac{1}{2}.$$

If n is odd, the (asymptotic) breakdown point

$$BP(Med) = \lim_{n\to\infty} BP(Med, n) = \lim_{n\to\infty} \frac{n+1}{2n}$$

$$= \lim_{n\to\infty} \left(\frac{n}{2n} + \frac{1}{2n}\right) = \lim_{n\to\infty} \left(\frac{1}{2} + \frac{1}{2n}\right) = \frac{1}{2}.$$

Thus, the (asymptotic) breakdown point of sample median is $1/2$.

Breakdown Point, Some Remarks

- The applied distance does not have to be Euclidean.
- Sometimes $maxBias(m, X_n, Q) = \infty$ is not seen as the only "breaking down" case. For example Scatter $= 0$ can be seen as breaking down too.
- For matrices, breaking down is sometimes considered to be equal to the largest eigenvalue approaching ∞.
- It does not make much sense to try construct estimators that have breakdown point larger than $1/2$.

Robust PCA

Robust PCA

If the data can be assumed to arise from elliptical distribution, then principal component analysis can be robustified by replacing the sample covariance matrix with some robust scatter estimate. The reason for that is that, under elliptical distribution, all scatter estimates do estimate the same population quantity (up to the scale). Note that in general (without ellipticity assumption) this does not hold!

Minimum Covariance Determinant (MCD) Method

The determinant (volume) of a covariance matrix, can be seen as a measure of total variation of the data, and it is then called the generalized variance. Data points that are far away from the data cloud increase the volume of the covariance matrix.

Minimum Covariance Determinant (MCD) Method

Minimum Covariance Determinant (MCD) method is a well-known method for robustifying the estimation of the covariance matrix, and the mean vector, under the assumption of multivariate ellipticity.

MCD method is based on considering all subsets containing $p\%$ (usually 50%) of the original observations, and estimating the covariance matrix, and the mean vector, on the data of the subset associated with the smallest covariance matrix determinant. This is equivalent to finding the sub-sample with the smallest multivariate spread. The MCD sample covariance matrix, and the MCD sample mean vector, are then defined as the sample covariance matrix (up to the scale), and the sample mean vector, computed over this sub-sample.

Minimum Covariance Determinant (MCD) Method

Note that, as $\det(AB) = \det(A)\det(B)$ for all square matrices and as the point mass probability of continuous distributions is 0, MCD should be affine equivariant under continuous distributions. However, the fast versions of the algorithm are not necessarily affine equivariant. Some error might occur due to "smart" sub-sampling.

Robust PCA

Under the ellipticity assumption, PCA can be performed using the MCD scatter estimate instead of the traditional sample covariance matrix. MCD estimates are very robust, and thus as a consequence, robust PCA is obtained.

Robust PCA

Note that MCD is not the only possible robust scatter estimate - there exists several robust scatter estimates that all estimate the same population quantity (up to the scale) under the assumption of multivariate ellipticity.

Words of Warning

Some Words of Warning

- It is possible that a functional Q has bounded influence function, but its breakdown point is 0!
- Robust PCA, based on some robust scatter matrix, can be performed under the assumption of multivariate ellipticity. If the ellipticity assumption does not hold, instead of estimating the PCA transformation matrix Γ, one may be estimating some other population quantity.

Next Week

Next week we will talk about bivariate correspondence analysis (CA).

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

P. J. Rousseeuw, Multivariate estimation with high breakdown point, *Mathematical Statistics and Applications* **8** (W. Grossmann, G. Pug, I. Vincze, W. Wertz, eds.), p. 283–297, 1985.

P. J. Rousseeuw, K. Van Driessen, A fast algorithm for the minimum covariance determinant estimator, *Technometrics* **41**, p. 212–223, 1999.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 5: Bivariate Correspondence Analysis

Pauliina Ilmonen

Contents

Correspondence Analysis

Correspondence Analysis

Correspondence analysis is a PCA-type method appropriate for analyzing categorical variables. The aim in bivariate correspondence analysis is to describe dependencies (correspondences) in a two-way contingency table.

Example, Education and Salary

In this lecture, we consider an example where we examine dependencies of categorical variables education and salary.

Frequency Tables

Contingency Tables

We consider a sample of size n described by two qualitative variables, x with categories $A_1, \ldots, A_J$ and y with categories $B_1, \ldots, B_K$. The number of individuals having the modality (category) A_j for the variable x and the modality B_k for the variable y is denoted by n_{jk}. Now the number of individuals having the modality A_j for the variable x is given by

$$n_{j.} = \sum_{k=1}^{K} n_{jk},$$

the number of individuals having the modality B_k for the variable y is given by

$$n_{.k} = \sum_{j=1}^{J} n_{jk},$$

and

$$n = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}.$$

Contingency Tables

The data is often displayed as a two-way contingency table.

	B_1	B_2	$\cdots$	B_K	
A_1	n_{11}	n_{12}	$\cdots$	n_{1K}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2K}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	n_{J1}	n_{J2}	$\cdots$	n_{JK}	$n_{J.}$
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.K}$	n

Table: Contingency table

Example, Education and Salary

We consider size 1000 sample of two categorical variables. Variable x Education is divided to categories A_1 Primary School, A_2 High School, and A_3 University, and variable y Salary is divided to categories B_1 low, B_2 average, and B_3 high.

Example, Education and Salary

We display the Education and Salary data as a two-way contingency table.

	B_1	B_2	B_3	
A_1	150	40	10	200
A_2	190	350	60	600
A_3	10	110	80	200
	350	500	150	1000

Table: Contingency table

- In this sample of 1000 observations, there are 150 individuals that have Primary School education and low salary.
- In this sample of 1000 observations, there are 10 individuals that have Primary School education and high salary.
- In this sample of 1000 observations, there are 110 individuals that have University education and average salary.
- ...

Contingency Tables

The value of the numbers n_{jk} is naturally relative to the total number of observations, n. Thus it is preferable to analyze the contingency table in the form of joint relative frequencies. From the contingency table, it is straightforward to compute the associated relative frequency table (F) where the elements of the contingency table are divided by the number of individuals n leading to $f_{jk} = \frac{n_{jk}}{n}$. The marginal relative frequencies are computed as

$$f_{j.} = \sum_{k=1}^{K} f_{jk}$$

and

$$f_{.k} = \sum_{j=1}^{J} f_{jk}.$$

Contingency Tables

	B_1	B_2	$\cdots$	B_K	
A_1	f_{11}	f_{12}	$\cdots$	f_{1K}	$f_{1.}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2K}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	f_{J1}	f_{J2}	$\cdots$	f_{JK}	$f_{J.}$
	$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.K}$	1

Table: Table of relative frequencies

Example, Education and Salary

	B_1	B_2	B_3	
A_1	0.15	0.04	0.01	0.20
A_2	0.19	0.35	0.06	0.60
A_3	0.01	0.11	0.08	0.20
	0.35	0.50	0.15	1

Table: Table of relative frequencies

- In this sample 15% of individuals have Primary School education and low salary.
- In this sample, 1% of individuals have Primary School education and high salary.
- In this sample, 11% of individuals have University education and average salary.
- ...

Frequencies

The frequency f_{jk} is the estimate of

$$p_{jk} = P(x \in A_j, y \in B_k),$$

and $f_{j.}$ and $f_{.k}$ are the estimates of

$$p_{j.} = P(x \in A_j),$$

and

$$p_{.k} = P(y \in B_k),$$

respectively.

Row Profiles

Tables of Conditional Frequencies

The proportion of individuals that belong to category B_k for the variable y among the individuals that have the modality A_j for the variable x form the so called table of row profiles. The conditional frequencies for fixed j and all k are

$$f_{k|j} = \frac{n_{jk}}{n_{j.}} = \frac{n_{jk}/n}{n_{j.}/n} = \frac{f_{jk}}{f_{j.}}.$$

The frequency $f_{k|j}$ is the estimate of

$$p_{k|j} = P(y \in B_k | x \in A_j).$$

Row Profiles

	B_1	B_2	$\cdots$	B_K	
A_1	$\frac{f_{11}}{f_{1.}}$	$\frac{f_{12}}{f_{1.}}$	$\cdots$	$\frac{f_{1K}}{f_{1.}}$	1
A_2	$\frac{f_{21}}{f_{2.}}$	$\frac{f_{22}}{f_{2.}}$	$\cdots$	$\frac{f_{2K}}{f_{2.}}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	$\frac{f_{J1}}{f_{J.}}$	$\frac{f_{J2}}{f_{J.}}$	$\cdots$	$\frac{f_{JK}}{f_{J.}}$	1

Table: Row profiles

Example, Education and Salary

	B_1	B_2	B_3	
A_1	0.75	0.20	0.05	1
A_2	0.32	0.58	0.10	1
A_3	0.05	0.55	0.40	1

Table: Row profiles

- In this sample 75% of the individuals that have Primary School education, have low salary.
- In this sample, 5% of the individuals that have Primary School education, have high salary.
- In this sample, 55% of the individuals that have University education, have average salary.
- ...

Column Profiles

Tables of Conditional Frequencies

The proportion of individuals that belong to category A_j for the variable x among the individuals that have the modality B_k for the variable y form the table of column profiles. The conditional frequencies for fixed k and all j are

$$f_{j|k} = \frac{n_{jk}}{n_{.k}} = \frac{n_{jk}/n}{n_{.k}/n} = \frac{f_{jk}}{f_{.k}}.$$

The frequency $f_{j|k}$ is the estimate of

$$p_{j|k} = P(x \in A_j | y \in B_k).$$

Column Profiles

	B_1	B_2	$\cdots$	B_K
A_1	$\frac{f_{11}}{f_{.1}}$	$\frac{f_{12}}{f_{.2}}$	$\cdots$	$\frac{f_{1K}}{f_{.K}}$
A_2	$\frac{f_{21}}{f_{.1}}$	$\frac{f_{22}}{f_{.2}}$	$\cdots$	$\frac{f_{2K}}{f_{.K}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_J	$\frac{f_{J1}}{f_{.1}}$	$\frac{f_{J2}}{f_{.2}}$	$\cdots$	$\frac{f_{JK}}{f_{.K}}$
	1	1	$\cdots$	1

Table: Column profiles

Example, Education and Salary

	B_1	B_2	B_3
A_1	0.43	0.08	0.07
A_2	0.54	0.70	0.40
A_3	0.03	0.22	0.53
	1	1	1

Table: Column profiles

- In this sample 43% of the individuals that have low salary, have Primary School education.
- In this sample, 7% of the individuals that have high salary, have Primary School education.
- In this sample, 22% of the individuals that have average salary, have University education.
- ...

Dependence and Independence

Independence

The variables x and y are independent if and only if for all j, k it holds that

$$P(x \in A_j, y \in B_k) = P(x \in A_j)P(y \in B_k),$$

$$P(x \in A_j | y \in B_k) = P(x \in A_j),$$

and

$$P(y \in B_k | x \in A_j) = P(y \in B_k).$$

These equalities can be estimated by

$$f_{jk} \approx f_{j.} f_{.k},$$

$$f_{j|k} = \frac{f_{jk}}{f_{.k}} \approx f_{j.},$$

and

$$f_{k|j} = \frac{f_{jk}}{f_{j.}} \approx f_{.k},$$

respectively.

Theoretical Relative Frequencies Under Independence

We can now define the theoretical relative frequencies and theoretical frequencies under the assumption of independence as follows:

$$f^*_{jk} = f_{j.} f_{.k}$$

and

$$n^*_{jk} = \frac{n_{j.} n_{.k}}{n} = f^*_{jk} n.$$

Example, Education and Salary

	B_1	B_2	B_3	
A_1	150	40	10	200
A_2	190	350	60	600
A_3	10	110	80	200
	350	500	150	1000

Table: Observed frequencies

	B_1	B_2	B_3	
A_1	70	100	30	200
A_2	210	300	90	600
A_3	70	100	30	200
	350	500	150	1000

Table: Theoretical frequencies under independence

Example, Education and Salary

	B_1	B_2	B_3	
A_1	0.15	0.04	0.01	0.20
A_2	0.19	0.35	0.06	0.60
A_3	0.01	0.11	0.08	0.20
	0.35	0.50	0.15	1

Table: Observed relative frequencies

	B_1	B_2	B_3	
A_1	0.07	0.10	0.03	0.20
A_2	0.21	0.30	0.09	0.60
A_3	0.07	0.10	0.03	0.20
	0.35	0.50	0.15	1

Table: Theoretical relative frequencies under independence

Attraction Repulsion Matrix

Attraction Repulsion Matrix

The attraction repulsion indices

$$d_{jk} = \frac{n_{jk}}{n_{jk}^*} = \frac{f_{jk}}{f_{jk}^*} = \frac{f_{jk}}{f_{j.}f_{.k}}$$

can be used to measure dependencies between categorical variables. The attraction repulsion matrix D is a matrix whose elements are the attraction repulsion indices. The element ij of the matrix D is d_{jk}.

Attraction Repulsion Matrix

Note that

$$d_{jk} > 1 \;\Leftrightarrow\; f_{jk} > f_{j.}f_{.k} \;\Leftrightarrow$$
$$f_{j|k} > f_{j.} \text{ and } f_{k|j} > f_{k.}$$

and

$$d_{jk} < 1 \;\Leftrightarrow\; f_{jk} < f_{j.}f_{.k} \;\Leftrightarrow$$
$$f_{j|k} < f_{j.} \text{ and } f_{k|j} < f_{k.}$$

If $d_{jk} > 1$, then the modalities (categories) A_j and B_k are said to be attracted to each other. If $d_{jk} < 1$, then the modalities A_j and B_k are said to repulse each other.

Salary Example

	B_1	B_2	B_3
A_1	2.14	0.40	0.33
A_2	0.90	1.16	0.67
A_3	0.14	1.10	2.67

Table: Attraction repulsion indices

- High salary is more frequent for people with University education.
- High salary is less frequent for people with a Primary School education.
- Low salary is less frequent for people with University education.
- ...

Chi-square Test Statistic

Independence Testing

If the variables (for example salary level and education level) were independent of each other, it would not make sense to assess dependencies between the categories. One can start the analysis by independence testing to see whether there is statistically significant dependency between the variables.

Independence

The independence between variables x and y can be tested using chi-square statistic. The null hypothesis of the test is

$$H_o : p_{jk} = p_{j.}p_{.k}, \text{ for all } j, k$$

and the test statistic is given by

$$\chi^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}.$$

Independence

In the test statistics above, the np_{jk}, under the null, are estimated by n^*_{jk}. When the sample size n is large, the test statistic has, under the null hypothesis, approximately chi-square distribution with $(K-1)(J-1)$ degrees of freedom. Thus the null hypothesis (independence between variables x and y) is rejected at the level α if

$$\chi^2 > \chi^2_{(K-1)(J-1),1-\alpha}.$$

Decomposition of the Chi-square Statistic

Let $Z \in \mathbb{R}^{J \times K}$, where

$$Z_{jk} = \frac{f_{jk} - f_{j.}f_{.k}}{\sqrt{f_{j.}f_{.k}}}.$$

Thus, the matrix Z gives shifted and scaled relative frequencies of the variables. The variables are shifted and scaled such that the elements

$$Z_{jk} = \frac{f_{jk} - f_{j.}f_{.k}}{\sqrt{f_{j.}f_{.k}}} = \frac{f_{jk} - f^*_{jk}}{\sqrt{f^*_{jk}}} = \frac{n_{jk} - n^*_{jk}}{\sqrt{n^*_{jk}}}$$

are the terms that are squared and summed in the chi-square statistic that is used for testing the independence of the variables.

Decomposition of the Chi-square Statistic

A large positive value Z_{jk} indicates a large contribution to the chi-square statistic. This indicates a positive association between row j and column k. (More observations than expected under independence.) A large negative value Z_{jk} also indicates a large contribution to the chi-square statistic, but this indicates a negative association between row j and column k. (Less observations than expected under independence.) Values near zero indicate no contribution to the test statistic. (The number of observations is equal to the expected number under independence.)

Let

$$V = Z^T Z$$

and let

$$W = ZZ^T.$$

Now the chi-square statistic

$$\chi^2 = n(trace(V)) = n(trace(W)).$$

Next Week

Next week we will continue discussion about correspondence analysis.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

L. Simar, An Introduction to Multivariate Data Analysis, Université Catholique de Louvain Press, 2008.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 6: Bivariate Correspondence Analysis - part II

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Correspondence Analysis

Correspondence Analysis (CA)

Correspondence analysis is a PCA-type method appropriate for analyzing categorical variables. The aim in bivariate correspondence analysis is to describe dependencies between the variables and to visualize approximate attraction repulsion indices in lower dimensions. We consider a sample of size n described by two qualitative variables, x with categories $A_1, \ldots, A_J$ and y with categories $B_1, \ldots, B_K$. We use the same notations as last week and start by looking at chi-square distances between the row (or column) profiles of the variables.

Chi-square Distances

Chi-square Distance

When the data is in the form of frequency distribution, the distance between the rows (or columns) can be measured using weighted Euclidean distances. The so called chi-square distance between two rows j_1 and j_2 is given by

$$d(j_1, j_2) = \sum_{k=1}^{K} \frac{1}{f_{.k}} \left(\frac{f_{j_1 k}}{f_{j_1 .}} - \frac{f_{j_2 k}}{f_{j_2 .}}\right)^2.$$

Euclidean distance gives the same weight to each column. The chi-square distance gives the same relative importance to each column proportionally to the marginal relative row frequency. The division of each squared term by the marginal relative column frequency is variance standardizing and compensates for the larger variance in high frequencies and the smaller variance in low frequencies. If no such standardization were performed, the differences between larger proportions would tend to be large and thus dominate the distance calculation, while the differences between the smaller proportions would tend to be swamped.

Chi-square Distance

The chi-square distances between two row profiles can be given as

$$d(j_1, j_2) = \sum_{k=1}^{K} \frac{1}{f_{.k}} \left(\frac{f_{j_1 k}}{f_{j_1 .}} - \frac{f_{j_2 k}}{f_{j_2 .}}\right)^2$$

$$= \sum_{k=1}^{K} \left(\frac{f_{j_1 k}}{f_{j_1 .}\sqrt{f_{.k}}} - \frac{f_{j_2 k}}{f_{j_2 .}\sqrt{f_{.k}}}\right)^2.$$

Thus, if the row profiles are scaled, the usual Euclidean metric can be used on the new scaled data.

Example, Chi-square Distance

	B_1	B_2	
A_1	0.10	0.20	0.30
A_2	0.20	0.40	0.60
A_3	0.01	0.09	0.10
	0.31	0.69	1

Table: Relative frequencies

- The chi-square distances between the first and the second row profile is $\frac{1}{0.31}(\frac{0.1}{0.3} - \frac{0.2}{0.6})^2 + \frac{1}{0.69}(\frac{0.2}{0.3} - \frac{0.4}{0.6})^2 = 0$.
- The chi-square distances between the second and the third row profile is $\frac{1}{0.31}(\frac{0.2}{0.6} - \frac{0.01}{0.1})^2 + \frac{1}{0.69}(\frac{0.4}{0.6} - \frac{0.09}{0.1})^2$
- Note that the chi-square distances between the second and the third row profile is equal to the chi-square distances between the first and the third row profile.

Chi-square Distance

The distance between two columns k_1 and k_2 is given by

$$d(k_1, k_2) = \sum_{j=1}^{J} \frac{1}{f_{j.}} \left(\frac{f_{jk_1}}{f_{.k_1}} - \frac{f_{jk_2}}{f_{.k_2}}\right)^2.$$

$$= \sum_{j=1}^{J} \left(\frac{f_{jk_1}}{f_{.k_1}\sqrt{f_{j.}}} - \frac{f_{jk_2}}{f_{.k_2}\sqrt{f_{j.}}}\right)^2.$$

Correspondence Analysis, Row Profiles

Correspondence Analysis, Row Profiles

Recall that traditional principal component analysis is based on maximizing Euclidean distances. As discussed, in the context of frequency distributions, the proper distance between the variables is the chi-square distance. Thus, in correspondence analysis, a PCA type approach is applied to modified data. Instead of the original relative frequencies f_{jk}, we work on scaled relative frequencies

$$\frac{f_{jk}}{f_{j.}\sqrt{f_{.k}}}.$$

The scaling here is the scaling used in calculating the chi-square distances between the rows. Correspondence analysis is based on maximizing chi-square distances.

Note that the relative row frequency weighted sum

$$\sum_{j=1}^{J} f_{j.} \frac{f_{jk}}{f_{j.}\sqrt{f_{.k}}} = \sqrt{f_{.k}}.$$

Correspondence Analysis, Row Profiles

Let $R \in \mathbb{R}^{J \times K}$, where

$$R_{jk} = \frac{f_{jk}}{f_{j.}\sqrt{f_{.k}}} - \sqrt{f_{.k}}.$$

Let R_j denote the jth row of R and let

$$V = \sum_{j=1}^{J} f_{j.} R_j^T R_j.$$

The matrix R now contains the scaled and centered relative frequencies and the matrix V is a relative row frequency weighted covariance matrix of the rows of R. The data is centered using the relative row frequency weighted mean and the observations are scaled by relative row frequencies. (In traditional covariance matrix the scale is $\frac{1}{n}$.)

Solution

The maximization problem in correspondence analysis is a problem of maximization under constraint, and similarly as in PCA, the solution is given by the eigenvalues and the eigenvectors of the matrix V.

Solution

In correspondence analysis on the row profiles, one finds orthonormal vectors (directions) u_i such that projection $P_i(\cdot)$ onto u_i maximizes the weighted sum of the Euclidean distances,

$$\sum_{j=1}^{J} f_{j\cdot} d^2(0, P_i(R_j)),$$

under the constraint that u_i is orthogonal to all $u_l, \ 1 \leq l < i$.

The vectors u_i are the eigenvectors of the matrix V. In constructing the matrices R and V, the row profiles are scaled and shifted to obtain a maximization problem that involves Euclidean distances as optimization involving chi-square distances directly would be technically difficult.

Remark

Recall the matrix $Z \in \mathbb{R}^{J \times K}$, where

$$Z_{jk} = \frac{f_{jk} - f_{j.} f_{.k}}{\sqrt{f_{j.} f_{.k}}}$$

that is connected to the chi-square independence test. One can show that the matrix

$$V = \sum_{j=1}^{J} f_{j.} R_j^T R_j = Z^T Z.$$

Score Vectors

Let λ_i denote the ith largest eigenvalue of the matrix V and let u_i denote the corresponding unit length eigenvector. Let $u_{i,k}$ denote the kth element of u_i. The score of the row profile j (associated with modality A_j) on the ith CA component is given by

$$\phi_{i,j} = \sum_{k=1}^{K} u_{i,k} R_{jk}.$$

The score vector ϕ_i is centered such that

$$\sum_{j=1}^{J} f_{j.}\phi_{i,j} = 0,$$

and the variance of ϕ_i is λ_i.

Contribution of the Modalities

The contribution of the modality A_j on construction of the axis u_i is given by

$$\frac{f_{j.}(\phi_{i,j})^2}{\lambda_i}.$$

Quality of the Representation

The quality of the representation of the centered row profile R_j by the CA axis i is measured by the squared cosine of angle between the vector OR_j and u_i :

$$cos^2(\alpha) = \left(\frac{< OR_j, u_i >}{||OR_j|| \cdot ||u_i||}\right)^2 = \frac{(\phi_{i,j})^2}{||OR_j||^2}.$$

If the value is close to 1, the quality of the representation is good.

Note that the formula above does not contain the weight f_j, and thus one modality can be:

- Close to the axis u_i and therefore be well represented (well explained).
- Due to a low weight f_j, it can have a low contribution to the axis.

Correspondence Analysis, Column Profiles

Correspondence Analysis, Column Profiles

Correspondence analysis on the column profiles is conducted exactly as correspondence analysis on the row profiles.

Correspondence Analysis, Column Profiles

Let $C \in \mathbb{R}^{J \times K}$, where

$$C_{jk} = \frac{f_{jk}}{f_{.k}\sqrt{f_{j.}}} - \sqrt{f_{j.}}$$

The matrix C contains scaled and shifted column profiles. Let C_k denote the kth column of C and let

$$W = \sum_{k=1}^{K} f_{.k} C_k C_k^T.$$

The matrix C now contains the scaled and centered relative frequencies and the matrix W is a relative column frequency weighted covariance matrix of the rows of C.

Solution

In correspondence analysis on the column profiles, one finds orthonormal vectors (directions) v_h such that projection $P_h(\cdot)$ onto v_h maximizes the weighted sum of Euclidean distances,

$$\sum_{k=1}^{K} f_{.k} d^2(0, P_h(C_k)),$$

under the constraint that v_h is orthogonal to all v_l, $1 \leq l < h$. The solution is given by the eigenvalues and the eigenvectors of the matrix $W = ZZ^T$.

Score Vectors

Let λ_h denote the hth largest eigenvalue of the matrix W and let v_h denote the corresponding unit length eigenvector. Let $v_{h,k}$ denote the kth element of v_h. The score of the column profile k (associated with modality B_k) on the hth CA component is given by

$$\psi_{h,k} = \sum_{j=1}^{J} v_{h,j} C_{jk}.$$

The score vector ψ_h is centered such that

$$\sum_{k=1}^{K} f_{.k} \psi_{h,k} = 0,$$

and the variance of ψ_h is λ_h.

Contribution of the Modalities

The contribution of the modality B_k on construction of the axis v_h is given by

$$\frac{f_{.k}(\psi_{h,k})^2}{\lambda_h}.$$

Quality of the Representation

The quality of the representation of the centered column profile C_k by the CA axis h is measured by the squared cosine of angle between the vector OC_k and v_h.

$$cos^2(\beta) = \left(\frac{< OC_k, v_h >}{||OC_k|| \cdot ||v_h||} \right)^2 = \frac{(\psi_{h,k})^2}{||OC_k||^2}.$$

If the value is close to 1, the quality of the representation is good.

Association Between the Profiles

Association Between the Profiles

It can be shown that the matrices V and W have the same nonzero eigenvalues. Moreover, the eigenvectors u_i can be given in terms of v_i and vice versa:

$$u_i = \frac{1}{\sqrt{\lambda_i}} Z^T v_i$$

and

$$v_i = \frac{1}{\sqrt{\lambda_i}} Z u_i.$$

Association Between the Profiles

Let $H = rank(V) = rank(W)$. The coolest thing in correspondence analysis is that the attraction-repulsion indices d_{jk} can be given in terms of ϕ and ψ as follows

$$d_{jk} = 1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,j} \psi_{h,k}.$$

Association Between the Profiles

The scores are often standardized defining

$$\hat{\psi}_{h,k} = \frac{1}{\sqrt{\lambda_h}} \psi_{h,k}$$

and

$$\hat{\phi}_{h,j} = \frac{1}{\sqrt{\lambda_1}} \phi_{h,j}.$$

Then

$$d_{jk} = 1 + \sqrt{\lambda_1} \sum_{h=1}^{H} \hat{\phi}_{h,j} \hat{\psi}_{h,k}.$$

The attraction-repulsion index d_{jk} is now larger than 1 if and only if the smallest angle between $(\hat{\phi}_{1,j}, ..., \hat{\phi}_{H,j})$ and $(\hat{\psi}_{1,k}, ..., \hat{\psi}_{H,k})$ is less than 90°.

If the row profile j and the column profile k are well represented by the first two CA components, then the attraction-repulsion index

$$d_{jk} \approx 1 + \sqrt{\lambda_1} \sum_{h=1}^{2} \hat{\phi}_{h,j} \hat{\psi}_{h,k}.$$

We can therefore say that the modalities A_j and B_k are attracted to each if the angle between $(\hat{\phi}_{1,j}, \hat{\phi}_{2,j})$ and $(\hat{\psi}_{1,k}, \hat{\psi}_{2,k})$ is less than 90° and they repulse each other if the angle between $(\hat{\phi}_{1,j}, \hat{\phi}_{2,j})$ and $(\hat{\psi}_{1,k}, \hat{\psi}_{2,k})$ is larger than 90°. In this case, one can simply observe the angle from the (double) biplot of the first two components of $\hat{\phi}$ and $\hat{\psi}$.

Example of Correspondence Analysis

Correspondence analysis using the data presented in lecture five. Variable x Education is divided to categories A_1 Primary School, A_2 High School, and A_3 University, and variable y Salary is divided to categories B_1 low, B_2 average, and B_3 high.

	B_1	B_2	B_3	
A_1	150	40	10	200
A_2	190	350	60	600
A_3	10	110	80	200
	350	500	150	1000

Table: Contingency table

Example of Correspondence Analysis

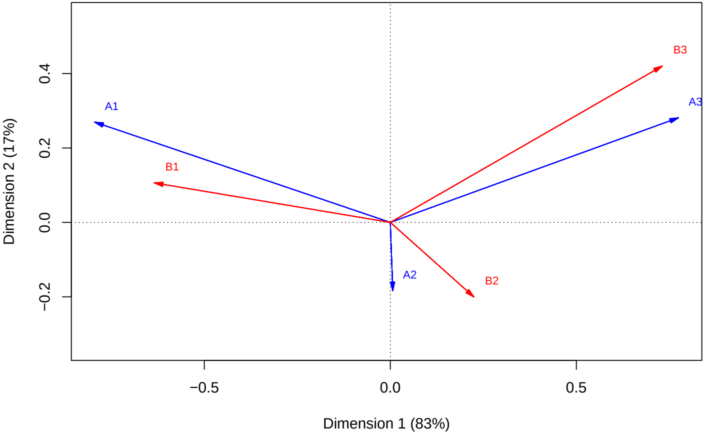


Figure: Salary and education (A1=Primary School education, A2=High School education, A3=University level education, B1=low salary, B2=average salary, B3=high salary)

Next Week

Next week we will talk about multiple correspondence analysis (MCA).

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

L. Simar, An Introduction to Multivariate Data Analysis, Université Catholique de Louvain Press, 2008.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 7: Multiple Correspondence Analysis

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Multiple Correspondence Analysis

Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) is an extension of bivariate correspondence analysis to more than 2 variables.

Frequency Tables

Contingency Tables

We consider a sample of size n described by P qualitative variables $Y_1, ..., Y_P$. The variable Y_p has K_p modalities (categories), and $\sum_{p=1}^{P} K_p$ is the total number of the categories. The number of individuals having the modality l of the variable Y_p is denoted by n_{pl}. We set a variable $x_{ipl} = 1$ if individual i has modality l of Y_p, and we set $x_{ipl} = 0$ otherwise. Now

$$\sum_{l=1}^{K_p} n_{pl} = n,$$

and

$$\sum_{p=1}^{P} \sum_{l=1}^{K_p} n_{pl} = nP.$$

Table of Dummy Variables

The table of K_p dummy variables associated with variable Y_p.

	1	2	$\cdots$	K_p	
1	x_{1p1}	x_{1p2}	$\cdots$	x_{1pK_p}	1
2	x_{2p1}	x_{2p2}	$\cdots$	x_{2pK_p}	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{np1}	x_{np2}	$\cdots$	x_{npK_p}	1
	n_{p1}	n_{p2}	$\cdots$	n_{pK_p}	n

Table: Table of dummy variables

Complete Disjunctive Table

Now we introduce the $n \times K$ table/matrix $X = [X_1, \dots, X_P]$, called the complete disjunctive table.

	X_1			$\cdots$	X_P			
	X_{11}	$\cdots$	X_{1K_1}	$\cdots$	X_{P1}	$\cdots$	X_{PK_P}	$\sum_{p=1}^{P} \sum_{l=1}^{K_p} x_{ipl}$
1	x_{111}	$\cdots$	x_{11K_1}	$\cdots$	x_{1P1}	$\cdots$	x_{1PK_P}	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i11}	$\cdots$	x_{i1K_1}	$\cdots$	x_{iP1}	$\cdots$	x_{iPK_P}	P
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n11}	$\cdots$	x_{n1K_1}	$\cdots$	x_{nP1}	$\cdots$	x_{nPK_P}	P
$\sum_{i=1}^{n} x_{ipl}$	n_{11}	$\cdots$	n_{1K_1}	$\cdots$	n_{P1}	$\cdots$	n_{PK_P}	nP

Table: Complete disjunctive table

Example: Party Snacks

A group of kids were asked to select party snacks. Each kid chose one cookie, one milk shake and one salty snack. Here we have a sample of 4 individuals and 3 variables — $n = 4, P = 3$.

- Variable X_1 cookie has two options (modalities/categories) — chocolate chip cookie (1) and oat cookie (2).
- Variable X_2 milk shake has three options — vanilla (1), strawberry (2), and chocolate (3).
- Variable X_3 salty snack has two options — pop corn (1), and potato chips (2).

Now $K = K_1 + K_2 + K_3 = 2 + 3 + 2 = 7$.

Example: Party Snacks

We display the party snack data as a complete disjunctive table.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	$\sum_{p=1}^{7}\sum_{l=1}^{K_p} x_{ipl}$
1	0	1	1	0	0	1	0	3
2	0	1	1	0	0	0	1	3
3	1	0	0	0	1	1	0	3
4	0	1	0	1	0	0	1	3
$\sum_{i=1}^{n} x_{ipl}$	1	3	2	1	1	2	2	12

Table: Complete disjunctive table

- The first kid chose an oat cookie, vanilla milk shake and pop corn.
- The third kid chose a chocolate chip cookie, chocolate milk shake and pop corn.

Multiple Correspondence Analysis

Bivariate correspondence analysis is now applied to the complete disjunctive table!

Relative Frequency Tables

From the complete disjunctive table, one can compute the associated relative frequency table (F), where the elements of the complete disjunctive table are divided by the total sum nP leading to

$$f_{ipl} = \frac{x_{ipl}}{nP} \quad (i = 1, \ldots, n; p = 1, \ldots, P; l = 1, \ldots, K_p).$$

We have P variables and n individuals and $f_{ipl} = \frac{1}{nP}$ or $f_{ipl} = 0$. Thus the marginal relative frequencies are computed as

$$f_{i..} = \sum_{p=1}^{P} \sum_{l=1}^{K_p} \frac{x_{ipl}}{nP} = P \frac{1}{nP} = \frac{1}{n}$$

and

$$f_{.pl} = \sum_{i=1}^{n} \frac{x_{ipl}}{nP} = \frac{n_{pl}}{nP}.$$

Example: Party Snacks

We display the party snacks data as a relative frequency table.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	$f_{i..}$
1	0	$\frac{1}{12}$	$\frac{1}{12}$	0	0	$\frac{1}{12}$	0	$\frac{1}{4}$
2	0	$\frac{1}{12}$	$\frac{1}{12}$	0	0	0	$\frac{1}{12}$	$\frac{1}{4}$
3	$\frac{1}{12}$	0	0	0	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{4}$
4	0	$\frac{1}{12}$	0	$\frac{1}{12}$	0	0	$\frac{1}{12}$	$\frac{1}{4}$
$f_{.pl}$	$\frac{1}{12}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{2}{12}$	1

Table: Relative frequency table

Row Profiles

Row Profiles

The idea behind MCA, like in bivariate correspondence analysis, is to apply a PCA type approach on one hand to the row profiles, and on the other hand to the column profiles of the relative frequencies table F. The coordinate pl of the row profile $l_i(1 \times K)$ associated with individual i is given as

$$(l_i)_{pl} = \frac{f_{ipl}}{f_{i..}} = \frac{\frac{x_{ipl}}{nP}}{\frac{1}{n}} = \frac{x_{ipl}}{nP}\frac{n}{1} = \frac{x_{ipl}}{P}, \qquad i = 1, \ldots, n.$$

As

$$\sum_{i=1}^{n} \frac{1}{n}(l_i)_{pl} = \sum_{i=1}^{n} \frac{1}{n}\frac{x_{ipl}}{P} = \frac{n_{pl}}{nP},$$

the n row profiles weighted by the marginal relative frequencies $(1/n)$ compose a point cloud in $\mathbb{R}^K$ with a center given by the relative marginal profile

$$G_l = (\frac{n_{11}}{nP}, \ldots, \frac{n_{1K_1}}{nP}, \ldots, \frac{n_{P1}}{nP}, \ldots, \frac{n_{PK_P}}{nP}).$$

Example: Party Snacks

The row profiles of the party snacks data is given below.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	
1	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	$\frac{1}{3}$	0	1
2	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	1
3	$\frac{1}{3}$	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$	0	1
4	0	$\frac{1}{3}$	0	$\frac{1}{3}$	0	0	$\frac{1}{3}$	1

Table: Row profiles

Row Profiles

Intuitively, the distance between two individuals is small if they have many modalities in common, and the distance between the individual i and the center increases as the modalities taking by the individual i becomes rare ($x_{ipl} = 1$ for n_{pl} small).

Row Profiles

More formally, the chi-square distances between two row profiles l_{i_1} and l_{i_2} can be given as

$$d^2(l_{i_1}, l_{i_2}) = \sum_{p=1}^{P} \sum_{l=1}^{K_P} \frac{1}{f_{.pl}}((l_{i_1})_{pl} - (l_{i_2})_{pl})^2$$

$$= \frac{n}{P} \sum_{p=1}^{P} \sum_{l=1}^{K_P} \frac{1}{n_{pl}}(x_{i_1 pl} - x_{i_2 pl})^2.$$

Example: Party Snacks

The distance between the first kid and the second kid in the party snacks data is

$$\left(\frac{4}{3}(1(0-0)^2+\frac{1}{3}(1-1)^2+\frac{1}{2}(1-1)^2+1(0-0)^2+1(0-0)^2+\frac{1}{2}(1-0)^2+\frac{1}{2}(0-1)^2)\right)$$

$$=\frac{4}{3}\approx 1.33.$$

Column Profiles

Column Profiles

The coordinate i of the column profile c_{pl} $(n \times 1)$ associated with the modality l of Y_p is given as

$$(c_{pl})_i = \frac{f_{ipl}}{f_{.pl}} = \frac{\frac{x_{ipl}}{nP}}{\frac{n_{pl}}{nP}} = \frac{x_{ipl}}{nP}\frac{nP}{n_{pl}} = \frac{x_{ipl}}{n_{pl}}, \qquad p = 1, \ldots, P; l = 1, \ldots, K_p.$$

As

$$\sum_{p=1}^{P}\sum_{l=1}^{K_p} f_{.pl}(c_{pl})_i = \sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{n_{pl}}{nP}\frac{x_{ipl}}{n_{pl}} = \sum_{p=1}^{P}\sum_{l=1}^{K_p} \frac{x_{ipl}}{nP} = \frac{P}{nP} = \frac{1}{n},$$

the K column profiles weighted by the marginal relative frequencies $(\frac{n_{pl}}{nP})$ compose a point cloud in $\mathbb{R}^K$ with the center given by the relative marginal profile $G_c = (\frac{1}{n}, \ldots, \frac{1}{n})$.

Example: Party Snacks

The column profiles of the party snacks data are given below.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}
1	0	$\frac{1}{3}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0
2	0	$\frac{1}{3}$	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$
3	1	0	0	0	1	$\frac{1}{2}$	0
4	0	$\frac{1}{3}$	0	1	0	0	$\frac{1}{2}$
	1	1	1	1	1	1	1

Table: Column profiles

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Column Profiles

Intuitively, the χ^2 distance between two modalities is small if the same individuals take these two modalities together, and the distance between the modality l of Y_p and the center increases as the modality becomes more rare (n_{pl} small).

Column Profiles

More formally, the chi-square distances between two column profiles $c_{p_1 l_1}$ and $c_{p_2 l_2}$ can be given as

$$d^2(c_{p_1 l_1}, c_{p_2 l_2}) = \sum_{i=1}^{n} \frac{1}{f_{i..}}((c_{p_1 l_1})_i - (c_{p_2 l_2})_i)^2$$

$$= n \sum_{i=1}^{n} (\frac{x_{ip_1 l_1}}{n_{p_1 l_1}} - \frac{x_{ip_2 l_2}}{n_{p_2 l_2}})^2.$$

Example: Party Snacks

The distance between modality 1 of Y_1 (chocolate chip cookie) and modality 2 of Y_2 (strawberry milk shake) is

$$4((0-0)^2+(0-0)^2+(1-0)^2+(0-1)^2)=8$$

Attraction Repulsion Indices

Attraction Repulsion Indices

Let $n_{p_1 l_1, p_2 l_2}$ be the number of individuals having the modality l_1 of the variable Y_{p_1} and the modality l_2 of the variable Y_{p_2}. Now the attraction repulsion index $d_{p_1 l_1, p_2 l_2}$ between the modality l_1 of the variable Y_{p_1} and the modality l_2 of the variable Y_{p_2} is given by

$$d_{p_1 l_1, p_2 l_2} = \frac{n_{p_1 l_1, p_2 l_2}/n}{n_{p_1 l_1}/n n_{p_2 l_2}/n} = \frac{n_{p_1 l_1, p_2 l_2}}{\frac{n_{p_1 l_1} n_{p_2 l_2}}{n}}.$$

Attraction Repulsion Indices

It is clear that if the attraction repulsion index is larger than one, the individuals are more inclined to take both modalities simultaneously than under the hypothesis of independence. And vice-versa, if the attraction repulsion index is smaller than one, the individuals are less inclined to take both modalities simultaneously than under the hypothesis of independence.

Attraction Repulsion Indices

The attraction repulsion index $d_{i,pl}$ between the individual i and the modality l of the variable Y_p is defined as follows.

$$d_{i,pl} = \frac{f_{ipl}}{f_{i..}f_{.pl}} = \frac{x_{ipl}}{n_{pl}/n}.$$

Now, clearly

$$d_{i,pl} = 0,$$

if $x_{ipl} = 0$ and

$$d_{i,pl} = \frac{n}{n_{pl}},$$

if $x_{ipl} = 1$. Thus, if the individual i does not have the modality l of the variable Y_p, then the attraction repulsion index $d_{i,pl}$ is equal to 0, and if the individual i does have the modality l of Y_p, then the attraction repulsion index $d_{i,pl}$ increases as the l of Y_p becomes rare.

Multiple Correspondence Analysis

Multiple Correspondence Analysis

To maximize chi-square distances and to obtain a representation in lower dimension, PCA type transformation is applied on the two data clouds: the row profiles and the column profiles. A transformation of the profiles is necessary to center the variables, and to be able to base the maximization problem on euclidian distances instead of χ^2 distances directly:

$$(l_i^\circ)_{pl} = \frac{(l_i)_{pl}}{\sqrt{f_{.pl}}} - \sqrt{f_{.pl}} \text{ and } (c_{pl}^\circ)_i = \frac{(c_{pl})_i}{\sqrt{f_{i..}}} - \sqrt{f_{i..}}.$$

Solution

The solution of the problem of maximization associated with the transformed row and column profiles is given respectively by the eigenvalues and the eigenvectors of the matrices $V(K \times K)$ and $W(n \times n)$ where

$V = T^T T$ and $W = TT^T$ where the elements of T are given by $\frac{x_{ipl} - n_{pl}/n}{\sqrt{Pn_{pl}}}$.

Note that here also, the matrix V is a relative row frequency weighted covariance matrix of the scaled and shifted row profiles and the matrix W is a relative column frequency weighted covariance matrix of the scaled and shifted column profiles.

Multiple Correspondence Analysis

The MCA components for the individuals are derived from the eigenvectors of the matrix V, and the MCA components for the modalities from the eigenvectors of the matrix W.

Let $H = rank(V) = rank(W)$. The scores of the individuals are given as

$$\phi_{h,i} = \sum_{k=1}^{K} u_{h,k}(l_i^{\circ})_k \qquad h = 1, \ldots, H,$$

where $u_{h,k}$ is the kth element of the eigenvector associated with the hth largest eigenvalues of V.

The scores for the modalities are given as

$$\psi_{h,pl} = \sum_{i=1}^{n} v_{h,i}(c_{pl}^{\circ})_i \qquad h = 1, \ldots, H.$$

Contribution of the Modalities

Contribution of the modality l of Y_p on the variance of the new variable ψ_h is given by

$$C(pl, h) = \frac{f_{.pl}\psi^2_{h,pl}}{\lambda_h} = \frac{n_{pl}\psi^2_{h,pl}}{nP\lambda_h}.$$

Global contribution of the variable Y_p is given by

$$C(p, h) = \sum_{l=1}^{K_p} C(pl, h).$$

Graphical Presentation

Comparison of the Modalities

The attraction repulsion index

$$d_{p_1 l_1, p_2 l_2} = 1 + \sum_{h=1}^{H} \psi_{h, p_1 l_1} \psi_{h, p_2 l_2}.$$

The graphical output of MCA is the approximation of the previous formula using few dimensions. Suppose that the modalities are well represented in two dimensions. Then we can plot the two first MCA components and interpret the proximity between the points on the first principal plan with the following approximation

$$d_{p_1 l_1, p_2 l_2} \approx 1 + \sum_{h=1}^{2} \psi_{h, p_1 l_1} \psi_{h, p_2 l_2}.$$

Comparison of the Individuals

The proximity between two individuals i_1 and i_2 is defined as

$$d_{i_1,i_2} = 1 + \sum_{h=1}^{H} \phi_{h,i_1} \phi_{h,i_2}.$$

Two individuals are close if they have in general the same modalities.
Now d_{i_1,i_2} can be approximated by

$$d_{i_1,i_2} \approx 1 + \sum_{h=1}^{2} \phi_{h,i_1} \phi_{h,i_2}.$$

Simultaneous Comparison

The attraction repulsion index

$$d_{i,pl} = 1 + \sum_{h=1}^{H} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i} \psi_{h,pl},$$

and thus again

$$d_{i,pl} \approx 1 + \sum_{h=1}^{2} \frac{1}{\sqrt{\lambda_h}} \phi_{h,i} \psi_{h,pl}.$$

Simultaneous Comparison

The scores are often standardized defining

$$\hat{\phi}_{1,j} = \frac{1}{\sqrt{\lambda_1}} \phi_{1,j}$$

and

$$\hat{\phi}_{2,j} = \frac{1}{\sqrt{\lambda_2}} \phi_{2,j}.$$

Then

$$d_{i,pl} \approx 1 + \sum_{h=1}^{2} \hat{\phi}_{h,i} \psi_{h,pl},$$

and the final graphical representation can be given simultaneously as a double biplot.

Example

Example of MCA: Extended Party Snack Data

Disclaimer: This example data set is randomly generated. Please do not draw real life conclusions from it.

	X_{11}	X_{12}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	$\sum_{p=1}^{7}\sum_{l=1}^{K_p} x_{ipl}$
1	0	1	1	0	0	1	0	3
2	0	1	1	0	0	0	1	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
25	1	0	0	0	1	0	1	3
$\sum_{i=1}^{n} x_{ipl}$	16	9	9	6	10	14	11	

Table: Complete disjunctive table

Example of MCA

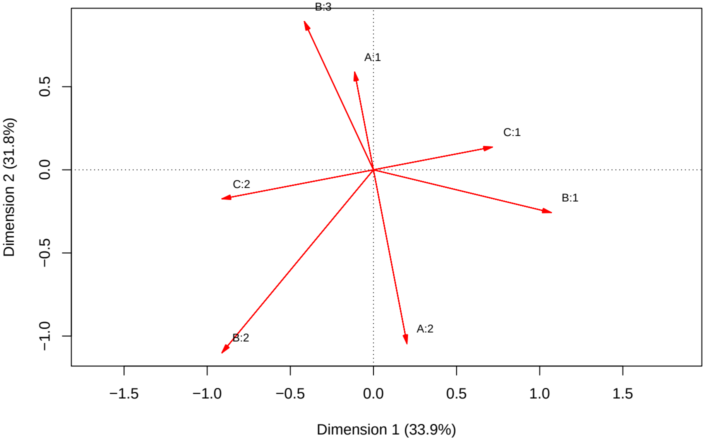


Figure: Result of MCA (A1=chocolate chip cookie, A2=oat cookie, B1=vanilla milk shake, B2=strawberry milk shake, B3=chocolate milk shake, C1=pop corn, C2=potato chips.) It seems that kids that like chocolate chip cookies like chocolate milk shake as well.

Some Remarks

Some Remarks

When performing MCA, it is better to take into account variables that have more or less the same number of modalities. (The number of modalities has an effect on the analysis.) It is also advised to avoid having very rare modalities. (Rare modalities have a big impact on analysis, and that makes MCA quite nonrobust method.) One can preprocess the data by grouping modalities if necessary.

Next Week

Next week we will talk about canonical correlation analysis.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

L. Simar, An Introduction to Multivariate Data Analysis, Université Catholique de Louvain Press, 2008.

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 8: Canonical Correlation Analysis

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Canonical Correlation Analysis

Canonical Correlation Analysis

Canonical correlation analysis involves partition of variables into two vectors x and y. The aim is to find linear combinations $\alpha^T x$ and $\beta^T y$ that have the largest possible correlation.

Canonical Correlation Analysis

Let x be a p-variate random vector and let y be a q-variate random vector. The object in canonical correlation analysis is to find linear combinations

$$u_k = \alpha_k^T x$$

and

$$v_k = \beta_k^T y$$

that maximizes the correlation $|corr(u_k, v_k)|$ between u_k and v_k subject to

$$var(u_k) = var(v_k) = 1,$$

and

$$corr(u_k, u_t) = 0, corr(v_k, v_t) = 0, \; t < k.$$

Correlation

Quick reminder:

$$corr(w_1, w_2) = \frac{E[(w_1 - \mu_{w_1})(w_2 - \mu_{w_2})]}{\sigma_{w_1}\sigma_{w_2}}.$$

Canonical Correlation Analysis

The vectors α_k and β_k are called the kth canonical vectors and

$$\rho_k = |corr(u_k, v_k)|$$

are called canonical correlations.

Canonical Correlation Analysis

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Whereas principal component analysis considers interrelationships within a set of variables, canonical correlation analysis considers relationships between two groups of variables.

Canonical Correlation Analysis, Examples

- Exercise — health.
- Open book exams — closed book exams.
- Job satisfaction — performance.

Canonical Correlation Analysis, Regression Analysis

Canonical correlation analysis can be seen as an extension of multivariate regression analysis. However, note that in canonical correlation analysis there is no assumption of causal asymmetry - x and y are treated symmetrically!

Canonical Correlation Analysis, Solution

Let $z = (x^T, y^T)^T$, and let

$$cov(z) = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Define

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

and

$$M_2 = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Canonical Correlation Analysis, Solution

Now, the canonical vectors α_k are the eigenvectors of M_1 (α_k corresponds to the kth largest eigenvalue), the canonical vectors β_k are the eigenvectors of M_2, and ρ_k^2 are the eigenvalues of the matrix M_1 (and of M_2 as well). The proof of this solution can be found from pages 283-284 of [1].

Canonical Correlation Analysis, Solution

Note that the eigenvectors α_k and β_k do not have length$= 1$!
Requirements

$$var(u_k) = var(\alpha_k^T x) = 1$$

and

$$var(v_k) = var(\beta_k^T y) = 1$$

define the lengths of the eigenvectors.

Canonical Correlation Analysis, Solution

If the covariance matrices Σ_{11} and Σ_{22} are not full rank, similar results may be obtained using generalized inverses. One may also consider dimension reduction as a first step.

Canonical Correlation Analysis, Sample Version

Sample estimates $\hat{\alpha}_k$, $\hat{\beta}_k$ and $\hat{\rho}_k$ of α_k, β_k and ρ_k, respectively, are obtained by using sample covariance matrices calculated from the samples $x_1, x_2, ..., x_n$, $y_1, y_2, ..., y_n$ and $z_1, z_2, ..., z_n$.

Testing Independence

Testing Independence

Assume that $z = (x^T, y^T)^T \sim N_{p+q}(\mu, \Sigma)$. Consider testing

H_0 : x and y are independent,

against

H_1 : x and y are not independent.

Testing Independence

Let $m = min\{p, q\}$, and let

$$T = -(n - \frac{1}{2}(p + q + 3)) \ln(\prod_{k=1}^{m}(1 - \hat{\rho}_k^2)).$$

Now, under H_0, and under the assumption of multivariate normality, the test statistic T is asymptotically distributed as $\chi^2(pq)$.

Testing Partial Independence

Assume that $z = (x^T, y^T)^T \sim N_{p+q}(\mu, \Sigma)$. Consider testing

H_0 : Only s of the canonical correlation coefficients are nonzero,

against

H_1 : The number of nonzero canonical correlation coefficients is larger than s.

Testing Partial Independence

Let $m = min\{p, q\}$, and let

$$T_s = -(n - \frac{1}{2}(p + q + 3)) \ln(\prod_{k=s+1}^{m} (1 - \hat{\rho}_k^2)).$$

Now, under H_0, and under the assumption of multivariate normality, the test statistic T is asymptotically distributed as $\chi^2((p - s)(q - s))$.

Independence Testing

If the normality assumption $z = (x^T, y^T)^T \sim N_{p+q}(\mu, \Sigma)$ does not hold, the p-values of the above mentioned test statistics can be approximated using permutations.

Scoring and Predicting

Scoring and Predicting

Let X and Y denote the $n \times p$ and $n \times q$ data matrices for n individuals, and let $\hat{\alpha}_k$ and $\hat{\beta}_k$ denote the kth (sample) canonical vectors. Then the $n \times 1$ vectors

$$\eta_k = X\hat{\alpha}_k$$

and

$$\phi_k = Y\hat{\beta}_k$$

denote the scores of the n individuals on the kth canonical correlation variables.

Scoring and Predicting

If the x and y variables are interpreted as the "predictor" and "predicted" variables, respectively, then the η_k score vector can be used to predict the ϕ_k score vector by using least square regression:

$$(\tilde{\phi_k})_i = \hat{\rho}_k((\eta_k)_i - \hat{\alpha}_k^T \bar{x}) + \hat{\beta}_k^T \bar{y}.$$

The canonical correlation $\hat{\rho}_k$ estimates the proportion of the variance of ϕ_k that is explained by the regression on x.

Example

Example: closed book exams — open book exams.

Example

Marks in open-book (O) and closed-book (C) exams:

i	Mechanics (C)	Vectors (C)	Algebra (O)	Analysis (O)	Statistics (O)
1	77	82	67	67	81
2	63	78	80	70	81
3	75	73	71	66	81
⋮	⋮	⋮	⋮	⋮	⋮
100	46	52	53	41	40

Source: K. V. Mardia, J. T. Tent, J. M. Bibby, Multivariate analysis, Academic Press, London, 2003 (reprint of 1979).

Example

Means:

Variable	Mean
x_1	38.9545
x_2	50.5909
y_1	50.6023
y_2	46.6818
y_3	42.3068

Example

Covariance matrix

$$
\begin{array}{c}
\qquad\quad \Sigma_{11} \qquad\qquad\qquad\qquad \Sigma_{12} \\
\Sigma = \left[\begin{array}{cc|ccc}
302.3 & 125.8 & 100.4 & 105.1 & 116.1 \\
 & 170.9 & 84.2 & 93.6 & 97.9 \\
\hline
 & & 111.6 & 110.8 & 120.5 \\
 & & & 217.9 & 153.8 \\
 & & & & 294.4
\end{array}\right] \\
\qquad\quad \Sigma_{21} \qquad\qquad\qquad\qquad \Sigma_{22}
\end{array}
$$

Example

Calculate the eigenvectors

$$M_1 = \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \Rightarrow \hat{\alpha}_k$$

and

$$M_2 = \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \Rightarrow \hat{\beta}_k.$$

Here

$$\hat{\alpha}_1 = \begin{bmatrix} 0.0260 \\ 0.0518 \end{bmatrix}$$

and

$$\hat{\beta}_1 = \begin{bmatrix} 0.0824 \\ 0.0081 \\ 0.0035 \end{bmatrix}.$$

Example

$$u_1 = 0.0260x_1 + 0.0518x_2$$

and

$$v_1 = 0.0824y_1 + 0.0081y_2 + 0.0035y_3.$$

The highest correlation occurs between an average of x_1 and x_2 weighted on x_2 and an average of $y1$, $y2$ and y_3, heavily weighted on y_1
The canonical correlations

$$\rho_1 = 0.6630$$

and

$$\rho_2 = 0.0412.$$

Example

Predicting

$$\left(\tilde{\phi}_k\right)_i = \hat{\rho}_k\left((\eta_k)_i - \hat{\alpha}_k^T \bar{x}\right) + \hat{\beta}_k^T \bar{y}.$$

Here

$$\left(\tilde{\phi}_1\right)_i = 0.6630\left((\eta_1)_i - (0.0260 * 38.9545 + 0.0518 * 50.5909)\right)$$
$$+(0.0824 \cdot 50.6023 + 0.0081 \cdot 46.6818 + 0.0035 \cdot 42.3068)$$

$$\approx 0.6630(\eta_1)_i + 2.2905$$

$$\approx 0.6630(0.0260(x_1)_i + 0.0518(x_2)_i) + 2.2905$$

$$\approx 0.0172(x_1)_i + 0.0343(x_2)_i + 2.2905.$$

Note that this almost predicts y_1.

Words of Warning

Some Words of Warning

- The procedure maximizes the correlation between the linear combination of variables — it can be more than difficult to interpret the results.
- Correlation does not automatically imply causality.

Next Week

Next week we will talk about discriminant analysis and classification.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 9: Discriminant Analysis and Classification

Lecturer: Pauliina Ilmonen
Slides: Ilmonen/Kantala

Contents

Discriminant Analysis

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Discriminant Analysis

The aim in discriminant analysis is to find a way to separate two or more classes of objects or events. That is then used in classification of new observations.

Discriminant Analysis

Consider $g,\ g > 1$, categories (populations or groups). The object in discriminant analysis is to find a rule for allocating an individual to one of these g groups based on his measurements. For example, the population might consist of different diseases and the measurement is the symptoms of a patient. Thus one is trying to find a rule that helps in diagnosing new patients' diseases based on their symptoms.

Discriminant Analysis, Normal Variables

Lecturer:
Pauliina Ilmonen
Slides:
Ilmonen/Kantala

Discriminant Analysis, Normal Variables

Let $n \times p$ matrix

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_g \end{bmatrix},$$

where each X_i, $i \in 1, ..., g$, is an $n_i \times p$ data matrix corresponding to group/population i coming from normal distribution $N(\mu_i, \Sigma_i)$. We here assume that the covariance matrices Σ_i are always of full rank.

Discriminant Analysis, Normal Variables

The probability density function of $N(\mu, \Sigma)$ distributed variables (with full rank covariance matrix) can be given as

$$(2\pi)^{-p/2} det(\Sigma)^{-1/2} exp(-1/2((x-\mu)^T \Sigma^{-1}(x-\mu)))$$

and the parameters μ and Σ can be estimated consistently by the sample mean vector and the sample covariance matrix, respectively.

Discriminant Analysis, Normal Variables

Under the assumption of normal distributions, an observation x can be allocated to one of the g groups on the basis of estimated probability density functions. Let $S_i = cov(X_i)$, and let $\bar{x}_i = mean(X_i)$. The observation x is allocated to group j, if

$$ln(det(S_j)) + (x - \bar{x}_j)^T S_j^{-1} (x - \bar{x}_j) < ln(det(S_i)) + (x - \bar{x}_i)^T S_i^{-1} (x - \bar{x}_i), \text{ for all } i \neq j.$$

Discriminant Analysis, Normal Variables

If the g groups are assumed to come from normal distributions with equal covariance matrices, then a consistent estimate of the common covariance matrix Σ is given by

$$S = \frac{1}{n-g}\sum_{i=1}^{g}(n_i - 1)S_i.$$

An observation x is allocated to group j, if

$$(x-\bar{x}_j)^T S^{-1}(x-\bar{x}_j) < (x-\bar{x}_i)^T S^{-1}(x-\bar{x}_i), \text{ for all } i \neq j.$$

Fisher's Linear Discriminant Function

Fisher's Linear Discriminant Function

Let $n \times p$ matrix

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_g \end{bmatrix},$$

where each X_i, $i \in 1, ..., g$, is an $n_i \times p$ data matrix corresponding to group/population i.

Fisher's Linear Discriminant Function

Let

$$W = \sum_{i=1}^{g} (n_i - 1) S_i,$$

where $S_i = cov(X_i)$, and let

$$B = \sum_{i=1}^{g} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^T.$$

The matrix W measures within group dispersions and the matrix B measures dispersion between groups.

Fisher's Linear Discriminant Function

Fisher's linear discriminant function is the linear function $a^T x$, where a is the maximizer of

$$\frac{a^T Ba}{a^T Wa}.$$

Thus Fisher's linear discriminant function is a linear function that maximizes the ratio of between groups dispersion and within group dispersions.

The solution is obtained by setting a to be equal to the eigenvector of $W^{-1}B$ that corresponds to the largest eigenvalue.

Fisher's Linear Discriminant Function

Once linear discriminant function has been calculated, an observation x can be allocated to one of the g groups on the basis of its discriminant score $a^T x$. The observation x is allocated to the population whose mean score is closest to the $a^T x$. That is, x is allocated to group j, if

$$|a^T x - a^T \bar{x}_j| < |a^T x - a^T \bar{x}_i|, \text{ for all } i \neq j.$$

Fisher's Linear Discriminant Function

Fisher's linear discriminant function is most important in the special case of $g = 2$ groups. Then the matrix B has rank 1, and it can be written as

$$B = \frac{n_1 n_2}{n} dd^T,$$

where $d = \bar{x}_1 - \bar{x}_2$. Thus, $W^{-1}B$ has only one non-zero eigenvalue and that equals to

$$tr(W^{-1}B) = \frac{n_1 n_2}{n} d^T W^{-1} d.$$

The corresponding eigenvector is

$$a = W^{-1} d.$$

Fisher's LDA, Example 1

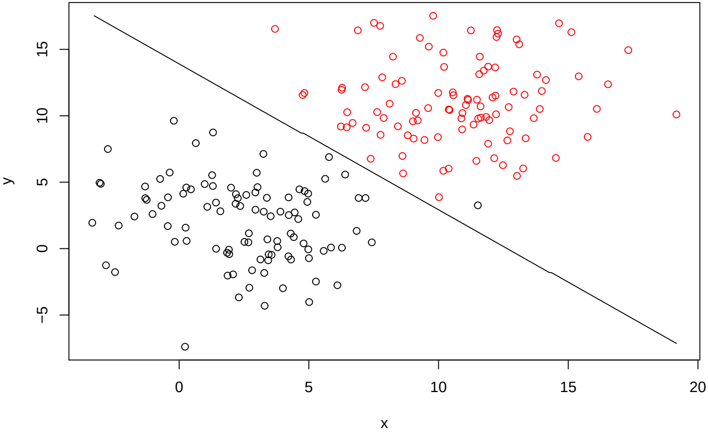


Figure: Fisher's linear discriminant analysis under normality (two groups).

Fisher's LDA, Example 2

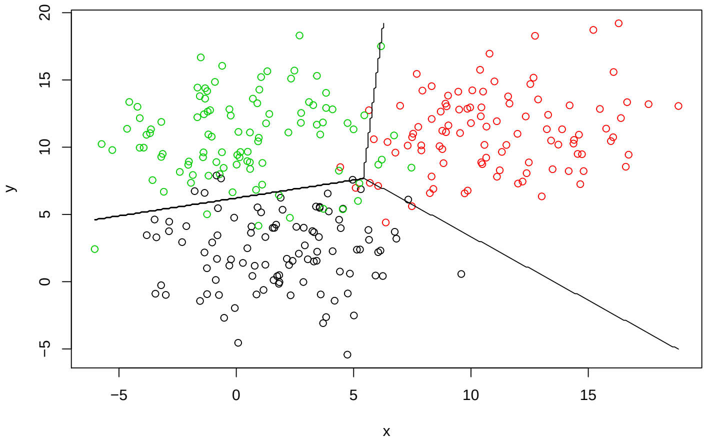


Figure: Pairwise Fisher's linear discriminant analysis under normality (three groups).

Statistical Depth

Statistical Depth

Let $S_n = \{x_1, ..., x_n\}$ denote a set of p variate observations from distribution F_x. Statistical depth $D(y, S_n)$ measures centrality of any p variate y with respect to S_n. The value of $D(y, S_n)$ is always between 0 and 1 and the larger the value of $D(y, S_n)$ is, the more central y is with respect to S_n.

Mahalanobis Depth

Let $S_n = \{x_1, ..., x_n\}$ denote a set of p variate observations from distribution F_x. The Mahalanobis depth $D_M(y, S_n)$ is defined as follows.

$$D_M(y, S_n) = \frac{1}{1 + d^2},$$

with

$$d = \sqrt{(y - \bar{x})^T C^{-1} (y - \bar{x})},$$

where $\bar{x}$ is the sample mean vector and C the sample covariance matrix calculated from the sample S_n.
Similar depth functions may be constructed by replacing the sample mean vector with some other location vector and the sample covariance matrix by some other scatter matrix.

Mahalanobis Depth, population version

Let x denote a p variate random variable with cumulative distribution function F_x. The population Mahalanobis depth $D_M(y, F_x)$ is defined as follows.

$$D_M(y, F_x) = \frac{1}{1 + d^2},$$

with

$$d = \sqrt{(y - \mu)^T \Sigma^{-1} (y - \mu)},$$

where $\mu = \mu(F_x)$ is the mean vector and $\Sigma = \Sigma(F_x)$ is the covariance matrix of the random variable x.

Half Space Depth

Let $S_n = \{x_1, ..., x_n\}$ denote a set of p variate observations from distribution F_x. The half space depth $D_H(y, S_n)$ is defined as follows.

$$D_H(y, S_n) = \min_{u \in U} \frac{1}{n} |\{x_i \in S_n \mid u^T(x_i - y) \geq 0\}|,$$

where U denotes the unit sphere in $\mathbb{R}^p$.

Half Space Depth, Example

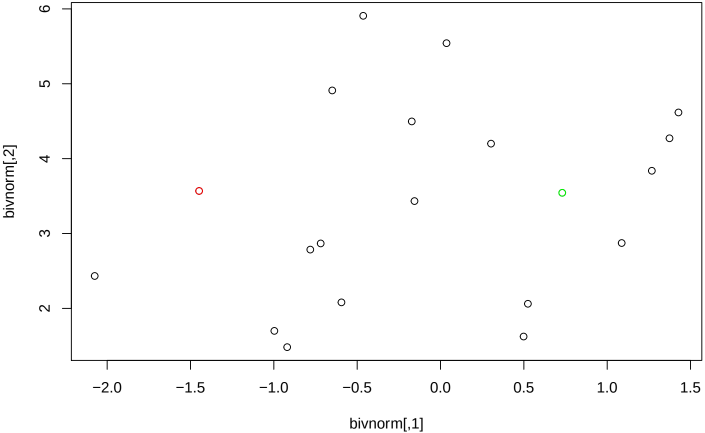


Figure: Bivariate normal distribution. The half space depth value of the red point is $2/20 = 0.1$. The half space depth value of the green point is $5/20 = 0.25$.

Half Space Depth, Population Version

Let x denote a p variate random variable with cumulative distribution function F_x. The population half space depth $D_H(y, F_x)$ is defined as follows.

$$D_H(y, F_x) = \inf_{u \in U} P(u^T(x - y) \geq 0),$$

where U denotes the unit sphere in $\mathbb{R}^p$.

Depth Functions

Mahalanobis depth and half space depth are just two examples of statistical depth functions. There are several other depth functions that have been presented in the literature.
Let x denote a p variate random variable with cumulative distribution function F_x. In general, depth functions should fulfill the following properties (Zuo and Serfling):

- Affine invariance: For any p vector b and any $p \times p$ matrix A, $D(y, F_x) = D(Ay + b, F_{Ax+b})$.
- Maximality at center: If there exist a unique point of symmetry θ such that $\theta + x$ is distributed as $\theta - x$, then $D(\theta, F_x) = \sup_y D(y, F_x)$.
- Monotonicity with respect to the deepest point: If there exist a deepest point α, then for any p vector v $D(\alpha + tv, F_x)$ is monotonically decreasing function of $t > 0$.
- Vanishing at infinity:$D(y, F_x) \to 0$, as $||y|| \to \infty$.

Classification Based on Statistical Depth

Classification Based on Statistical Depth

Consider two samples $S_n = \{x_1, ..., x_n\}$ and $T_m = \{z_1, ..., z_m\}$ from distributions F_x and F_z, respectively. A new observation y can now be allocated as coming from F_x or F_z by using a depth function. If $D(y, S_n) \geq D(y, T_m)$, the observation y is allocated as coming from F_x, and otherwise it is allocated as coming from F_z.

The procedure generalizes naturally to several distributions. The observation is allocated as coming from the distribution F_w that corresponds to the largest depth value for y.

Misclassification Rates

Misclassification Rates

In discriminant analysis, it is desirable to find such classification rules that reduce misclassification as much as possible. In practice one can also take into account the costs of misclassification. For example, it can be worse not to detect an illness than to classify a healthy individual as ill.

Misclassification Rates

Calculating exact misclassification rates can be difficult or even impossible when exact underlying distributions are not known.

Misclassification Rates

Misclassification rates are often estimated by calculating sample misclassification rates. After defining a classification rule, the data is classified according to that rule, and sample misclassification rate is obtained. Note that estimated misclassification rates obtained this way grossly underestimate the true misclassification rates - even when sample sizes n_i are large. The problem comes from the fact that the same sample is used to construct the rule and also to test the quality of the classification

Misclassification Rates, Training Sample

Misclassification rates can also be estimated by dividing the original sample into two parts. A training sample (for example 80% of the observations) is used to construct the rule. The rest of the sample is used in approximating the misclassification rate. However, this approach requires large sample sizes and the evaluated classification rule is not the same rule as the one that would be obtained using the entire original sample.

Other Approaches

Other Approaches

- Classification based "closest neighbors" or on local depths.
- Random forest classification.
- Context related classification.
- ...

Next Week

Next week we will talk about clustering.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

- R. Y. Liu, J. M. Parelius, K. Singh, Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference (with discussion), The Annals of Statistics, 27, 783–858, 1999.
- Y. Zuo, R. Serfling, General notions of statistical depth function, The Annals of Statistics, 28, 461–482, 2000.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 10: Clustering

Pauliina Ilmonen

Contents

Clustering

Clustering

Let $x_1, x_2, ..., x_n$ be measurements of p variables on n objects that are believed to be heterogeneous. The aim in cluster analysis is to group these objects into k homogeneous classes. The number of classes, k, is also often unknown (but usually assumed to be a lot smaller than n).

In multisample problem one has m samples and the aim is to group the m samples into k homogeneous classes.

Clustering

Clustering methods rely on two (separate) issues:

- The choice of a distance or dissimilarity measure between objects.
- The choice of a group building algorithm.

Clustering

Cluster analysis is a difficult problem in a general framework.

An intuitively appealing approach:

1. Define all the possible partitions of the n $p-$variate data points into k classes, $k = 1, 2, ..., n$.
2. For each obtained partition, compute the value of a chosen criterion.
3. Select the partition that optimizes the criterion.

Problem: The number of combinations that have to be computed (even when n is small) is huge! For example, for $n = 12$, the number of possible partitions is over 4 millions.

Agglomerative Hierarchical Algorithms

Agglomerative Hierarchical Algorithms

Agglomerative hierarchical algorithms are methods that start from n classes and go step by step to $n-1, n-2, \ldots$ nested classes.

Agglomerative Hierarchical Algorithms

1. Start from the finest partition: n clusters, each containing one data point x_i
2. Calculate distances $d_{ij} = d(x_i, x_j)$, where d is an appropriate distance between individuals.
3. Find the minimal distance and group together the corresponding individuals.
4. Compute distances between obtained groups using an appropriate linkage function.
5. Find the minimal distance and group together the corresponding closest groups.
6. Repeat steps 4 and 5 until you have one single group.

How Many Clusters to Choose?

In agglomerative hierarchical algorithms, the minimal value in step (3 and) 5 provides the so called aggregation level for each step. (In other words: the aggregation level is the distance between the clusters that were grouped.) The number of clusters is chosen based on the aggregation level. High level indicates grouping of heterogeneous clusters. Thus one can decide to "cut" at a desired level.

Which Linkage Functions to Use?

There are several ways to measure the distance between groups:

- The minimum linkage:

$$d(A,B) = \min_{x_i \in A, x_j \in B} d(x_i, x_j).$$

- The maximum linkage:

$$d(A,B) = \max_{x_i \in A, x_j \in B} d(x_i, x_j).$$

- The average linkage:

$$d(A,B) = \frac{1}{n_A n_B} \sum_{x_i \in A} \sum_{x_j \in B} d(x_i, x_j).$$

- The Ward linkage:

$$d(A,B) = \frac{n_A n_B}{n_A + n_B} d(c_A, c_B),$$

where c_A and c_B are the centers of the clusters A and B, respectively.

Linkage functions

Minimum linkage is simple, but the problem is that quite different groups could be clustered together just for having two close elements (chaining). This approach is still very often used in practice. Also in maximum linkage, the problem is that the decision is based on single points. If one wishes to avoid these problems, the average linkage provides a safer choice.

How to Choose the Distance?

With quantitative data, the euclidian distance

$$d^2(x_i, x_j) = (x_i - x_j)^T (x_i - x_j)$$

is a classical choice.

How to Choose the Distance?

Also principal component metric is quite popular choice. Then

$$d^2(x_i, x_j) = (x_i - x_j)^T D^{-1}(x_i - x_j),$$

where $D = diag(s_1^2, ..., s_p^2)$ and s_t^2 is the variance of the tth component of x.

How to Choose the Distance?

There are plenty of other choices: Manhattan distance, Maximum distance, ...

How to Choose the Distance?

With qualitative data, if one wishes to perform a cluster analysis of the row profiles of a contingency table, one could use the chi-square distances between the row profiles (as in MCA).

How to Choose the Distance?

Also context related distances can be used. For example, if one of the variables is considered being more important than the other variables, then one can put more weight on that.

Hierarchical Clustering Methods

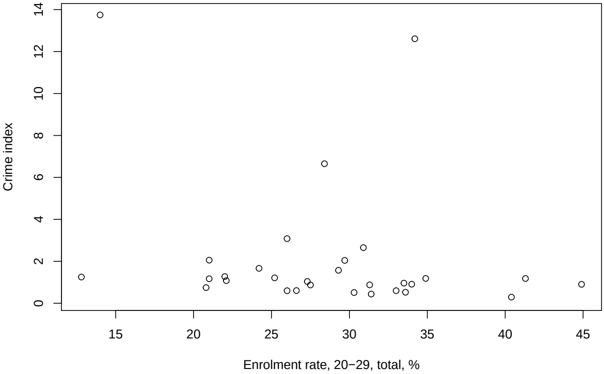
Agglomerative hierarchical clustering algorithms are "bottom up" methods. One may also start from one cluster and split step by step. These "top down" methods are called divisive hierarchical clustering algorithms.

Agglomerative Hierarchical Clustering, Example

The data consists of the education enrolment rate of 20–29-year-olds in the OECD countries (where data was available) and a crime index based on the UN homicide and robbery rates. Agglomerative hierarchical clustering was applied to the data set. The metric used was Euclidean distance and linkage criteria was the average linkage.

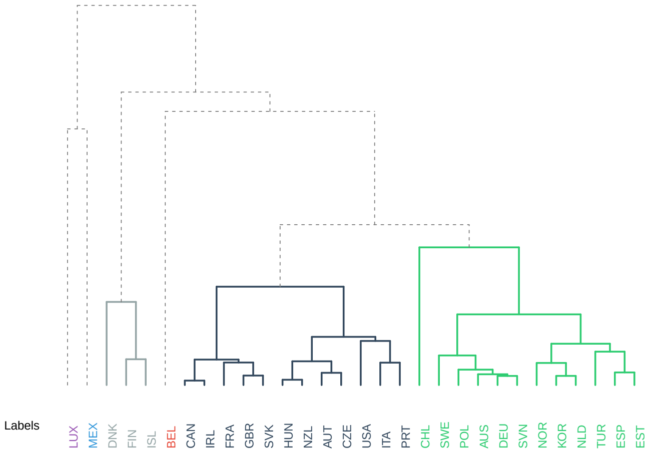
Agglomerative Hierarchical Clustering, Example

Scatter plot of the data:



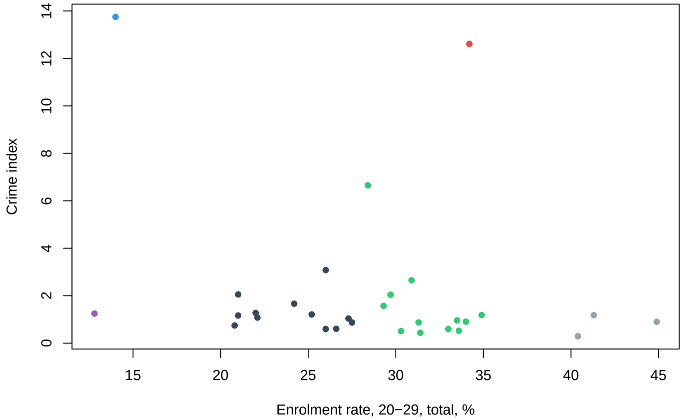
Agglomerative Hierarchical Clustering, Example Continues

The results of the cluster analysis presented as a classification tree:



Agglomerative Hierarchical Clustering, Example Continues

Scatter plot using the cluster colours from the previous slide.



Nonhierarchical Clustering Methods

Agglomerative hierarchical clustering algorithms and divisive hierarchical clustering algorithms are popular, but there exists several other clustering methods. The use of nonhierarchical clustering methods usually requires knowledge of the number of clusters.

Moving Centers Method

Moving Centers Method (k-means clustering)

Moving centers clustering method is based on calculating distances from "centers". The method requires knowing the number of clusters k.

Moving Centers Method (k-means clustering)

1. Choose randomly k data points $c_1, \ldots, c_k$ out of $x_1, \ldots, x_n$.
2. Define k sets $A_1, \ldots, A_k$ such that $A_t = \{x_i \mid d(x_i, c_t) \leq d(x_i, c_j), \text{ for } j \neq t\}$.
3. Calculate new centers $c_1, \ldots, c_k$ (for example sample means) of the sets $A_1, \ldots, A_k$.
4. Repeat steps 2 and 3 until convergence.

Distances

As when applying hierarchical clustering, also when applying moving centers clustering, one has to consider the context and decide what distance is (the most) appropriate. One also has to decide how to define the "center." Usually sample mean is used, but other locations can be used as well.

Moving Centers Method, Problems and Solutions

- Problem: Possible empty clusters in step 2. Solution: Choose one new center randomly.
- Problem: The algorithm always converges, but sometimes to a local optimum and sometimes very slowly. Solution: Choose the initial points wisely.

Moving Centers Method, Choosing the Initial Points Wisely

k-means++ initialization

Words of Warning

Some Words of Warning

- Optimal clustering methods are computationally very very heavy — in general not doable using standard computers and software.
- Different clustering methods can produce different solutions.
- The chosen distances, and methods to calculate distances between sets, may have an effect on the outcome.

Next Week

Next week we will talk about some very recently developed multivariate methods.

References

References I

K. V. Mardia, J. T. Kent, J. M. Bibby, Multivariate Analysis, Academic Press, London, 2003 (reprint of 1979).

References II

- R. V. Hogg, J. W. McKean, A. T. Craig, Introduction to Mathematical Statistics, Pearson Education, Upper Sadle River, 2005.
- R. A. Horn, C. R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.
- R. A. Horn, C. R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

References III

L. Simar, An Introduction to Multivariate Data Analysis, Université Catholique de Louvain Press, 2008.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 11: New Winds

New Winds

We have superb guest lecturers. Our guest lecturers talk about their research related to multivariate statistics.

MS-E2112 Multivariate Statistical Analysis (5cr)
Lecture 12: Summary

Summary

Lecture 12 is a summary lecture. We review some of the course materials and talk about the exam. There are no lecture slides for this lecture