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

Lecturer: Pauliina Ilmonen
Slides: Ilmonen

Contents

PCA transformation

Theoretical Properties

Sample Version

Applications

Example

Real Data Example

Words of Warning

References

PCA transformation

PCA-transformation

PCA transformation
Theoretical Properties
Sample Version
Applications
Example
Real Data Example
Words of Warning
References

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

Principal Components

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.

$$
\begin{aligned}
E[y_i] = E[\gamma_i^T(x - \mu)] = E[\gamma_i^T x] - E[\gamma_i^T \mu] \\
= \gamma_i^T E[x] - \gamma_i^T \mu = \gamma_i^T \mu - \gamma_i^T \mu = 0,
\end{aligned}
$$

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 $a \in \mathbb{R}^p,\ a^Ta = 1$.

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

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

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

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

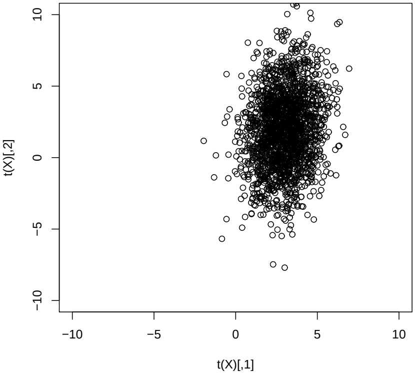

Figure: Bivariate normal distribution.

Example

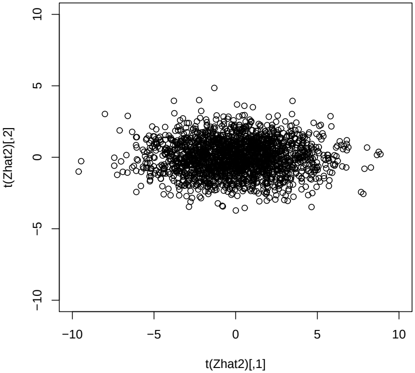

Figure: Bivariate normal distribution after PCA.

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

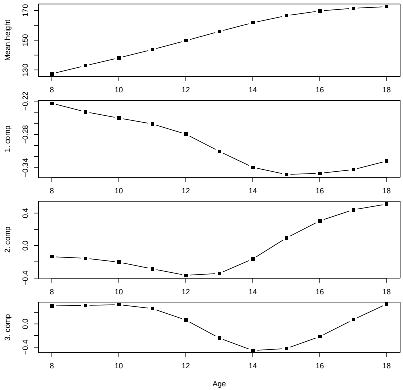

Figure: Mean curve of the estimated data points and the three first principal component curves (the three first column vectors of Γ). The first principal component curve puts emphasis on overall growth (shape of the curve is similar to the mean curve), the second on late growth, and the third on growth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

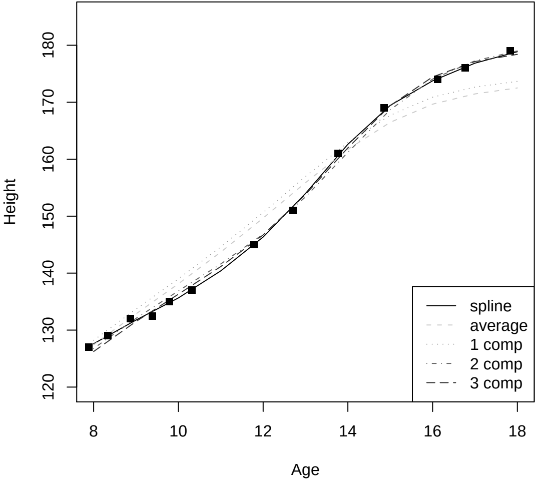

Figure: Estimated growth curve of one randomly chosen boy.

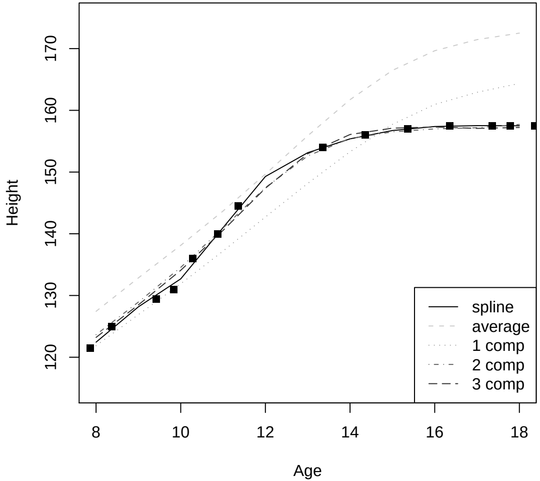

Figure: Estimated growth curve of one randomly chosen girl.

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

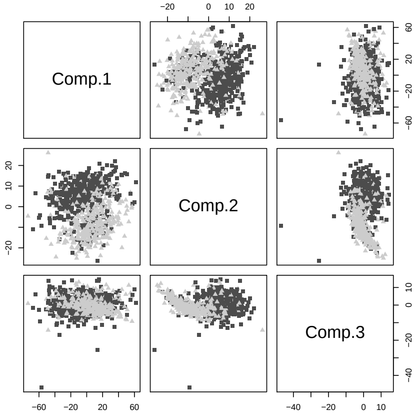

Figure: Scatter plot after PCA. Dark grey squares are used for the boys and light grey triangles for the girls. PCA does not work perfectly in separating the two groups, but one can still see clear differences between the groups. Boys grow later than girls! (Notice the outlying points.)

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.