STATISTICAL LEARNING SYSTEMS

LECTURE 8: UNSUPERVISED LEARNING: FINDING STRUCTURE IN DATA

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

The project is co-financed by the European Union within the framework of European Social Fund

Principal Component Analysis

We are given n p-dimensional data points i.e. a cloud of n points in a p-dimensional space.

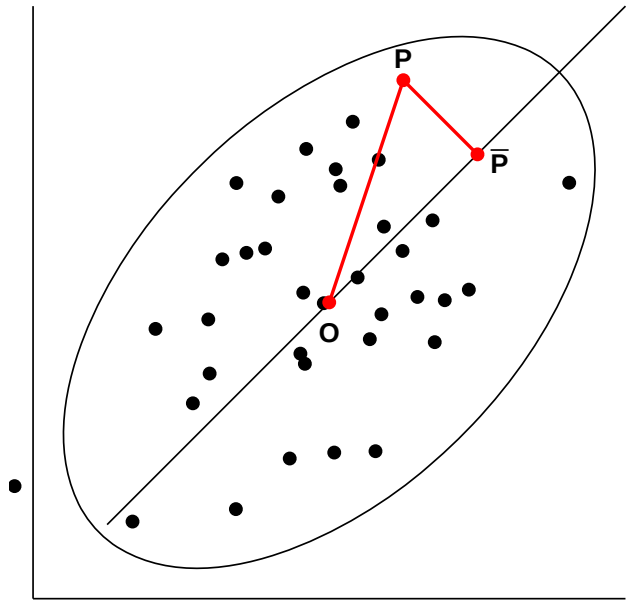
Aim: provide the 'best' r-dimensional representation of this cloud, where $r < p$.

Principal component analysis (PCA) is one of the realizations of this aim with certain adopted meaning of 'best'.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal Component Analysis

Consider $p = 2$ and a two-dimensional cloud of points. Position coordinate center O at the centroid of this points ($\equiv \mathbf{x}_i := \mathbf{x}_i - \bar{\mathbf{x}}$).



We look for one-dimensional representation of this cloud such that the displacement of points is relatively small.

Principal Component Analysis

Observe that Pythagoras' theorem implies ($\bar{P}_i$: projection of P_i)

$$(OP_i)^2 = (O\bar{P}_i)^2 + (P_i\bar{P}_i)^2.$$

and thus

$$\sum_{i=1}^{n}(OP_i)^2 = \sum_{i=1}^{n}(O\bar{P}_i)^2 + \sum_{i=1}^{n}(P_i\bar{P}_i)^2.$$

As the lefthand side does not depend on the direction of the line, we see that
minimization of $\sum_{i=1}^{n}(P_i\bar{P}_i)^2 \equiv$ maximization of $\sum_{i=1}^{n}(O\bar{P}_i)^2$ or equivalently, maximization of

$$\frac{1}{n-1}\sum_{i=1}^{n}(O\bar{P}_i)^2$$

i.e. the variance of the projections on the considered line.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal Component Analysis

This justifies the following strategy of PCA (for general p):

(i) For $\mathbf{x}_1, \ldots \mathbf{x}_n \in R^p$ find direction $\mathbf{a}_1$ such that $||\mathbf{a}_1|| = 1$ and the variance of points projected onto this direction, $\mathbf{a}_1^T \mathbf{x}_1, \ldots \mathbf{a}_1^T \mathbf{x}_n$, is the largest.

(ii) Find direction $\mathbf{a}_2$ such $||\mathbf{a}_2|| = 1$ and $\mathbf{a}_2$ is perpendicular to $\mathbf{a}_1$ such that the the variance of points projected onto this direction, $\mathbf{a}_2^T \mathbf{x}_1, \ldots \mathbf{a}_2^T \mathbf{x}_n$, is the largest among all perpendicular directions.

(iii) continue to choose $\mathbf{a}_1, \ldots, \mathbf{a}_r$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal Component Analysis

Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathbf{T}}$, $i = 1, \ldots, n$ and $\mathbf{a}_i = (a_{i1}, \ldots, a_{ip})^{\mathbf{T}}$.
$y_i = \mathbf{a}_i^T \mathbf{x}$ is called the i^{th} principal component;
$\mathbf{a}_i$ -vector of loadings (direction) of the i^{th} principal component.

The r principal component values for sth sample point are thus given by

$$\begin{aligned} y_{s1} &= a_{11}x_{s1} + a_{12}x_{s2} + \ldots + a_{1p}x_{sp} \\ y_{s2} &= a_{21}x_{s1} + a_{22}x_{s2} + \ldots + a_{2p}x_{sp} \\ &\ldots \\ y_{sr} &= a_{r1}x_{s1} + a_{r2}x_{s2} + \ldots + a_{rp}x_{sp} \end{aligned}$$

$y_{s1}, y_{s2}, \ldots, y_{sr}$ - r principal component scores for the s^{th} individual.

Note: as principal directions are orthogonal, projections of a data set on different directions are uncorrelated.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal Component Analysis

How to find principal directions ?

A simple algebraic solution exists:
Consider the empirical covariance matrix $\mathbf{S}$ corresponding to the cloud of points. Find eigenvalues λ_i of $\mathbf{S}$ and order them: $\lambda_1 \geqslant \lambda_2 \cdots \geqslant \lambda_p \geqslant 0$. Principal directions are given by eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_p$ (of unit length) corresponding to the ordered eigenvalues.

Another point of view: PCA yields a transformation of data matrix $X = (x_{ij})$ with consecutive observations being rows such that

$$\mathbf{Y}_{n\times p} = \mathbf{X}_{n\times p}\mathbf{A}_{p\times p}\,,$$

where columns of $\mathbf{A}$ are eigenvectors of covariance matrix $\mathbf{S}$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal component analysis

PCA, a proof (we assume for convenience that the data are centered, and hence $\mathbf{S} = \mathbf{X}'\mathbf{X}/n = (1/n)\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x_i}'$):

Write the Lagrange function and equate its derivative to zero:

$$\frac{d}{d\mathbf{a}}\{\mathbf{a}'\mathbf{S}\mathbf{a} - \lambda\mathbf{a}'\mathbf{a}\} = 0.$$

Thus

$$\mathbf{S}\mathbf{a} = \lambda\mathbf{a}$$

what, in fact, ends the proof.

Remark: If the data are not centered, i.e. $\sum_i \mathbf{x}_i \neq 0$, we can center them replacing $\mathbf{X}$ by $(\mathbf{I} - \mathbf{M})\mathbf{X}$, where $\mathbf{M} = \mathbf{1}\mathbf{1}'/n$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal component analysis

$$\mathbf{S} = \mathbf{V}\mathbf{\Delta}\mathbf{V}'$$

with

$$\mathbf{V} = [\mathbf{a}_{(1)} \ldots \mathbf{a}_{(p)}]_{(p,p)},$$

but we usually employ the singular value decomposition of $\mathbf{X}$,

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$$

to get

$$[\mathbf{y}_{(1)} \ldots \mathbf{y}_{(p)}]_{(n,p)} = \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{D}.$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal Component Analysis

How to choose the number of principal directions r ?

Fact: The variance of projections of $\mathbf{x}_1, \ldots, \mathbf{x}_n$ on the hyperplane spanned by $\mathbf{a}_1, \mathbf{a}_2 \ldots, \mathbf{a}_r$ is equal to $\lambda_1 + \lambda_2 \cdots + \lambda_r$.

First choice of r:

$$P_k = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Take as r

$$\min r : P_r \geqslant \text{given threshold } \gamma \ (\text{usually } \gamma \approx 0.7 - 0.9)$$

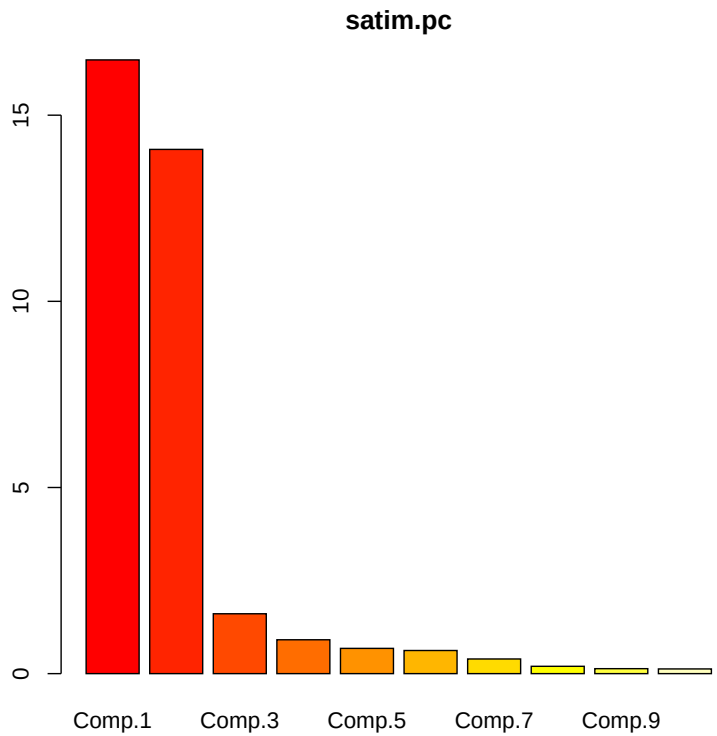
Interpretation : The principal directions chosen explain at least $100\gamma\%$ of the variability present in the data.

Second choice of r:

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal Component Analysis

Consider a scree-plot i.e. the plot of λ_i's against index i and choose as r the minimal index i_0 such that the plot levels off for $i > i_0$.



This does not always work: it can happen that we have many components with comparable and small variabilities which jointly have non-negligible impact on the variability of data.

Applications

A plot of the first two or three component scores frequently yields a new insight into the structure of the data. Some ideas:

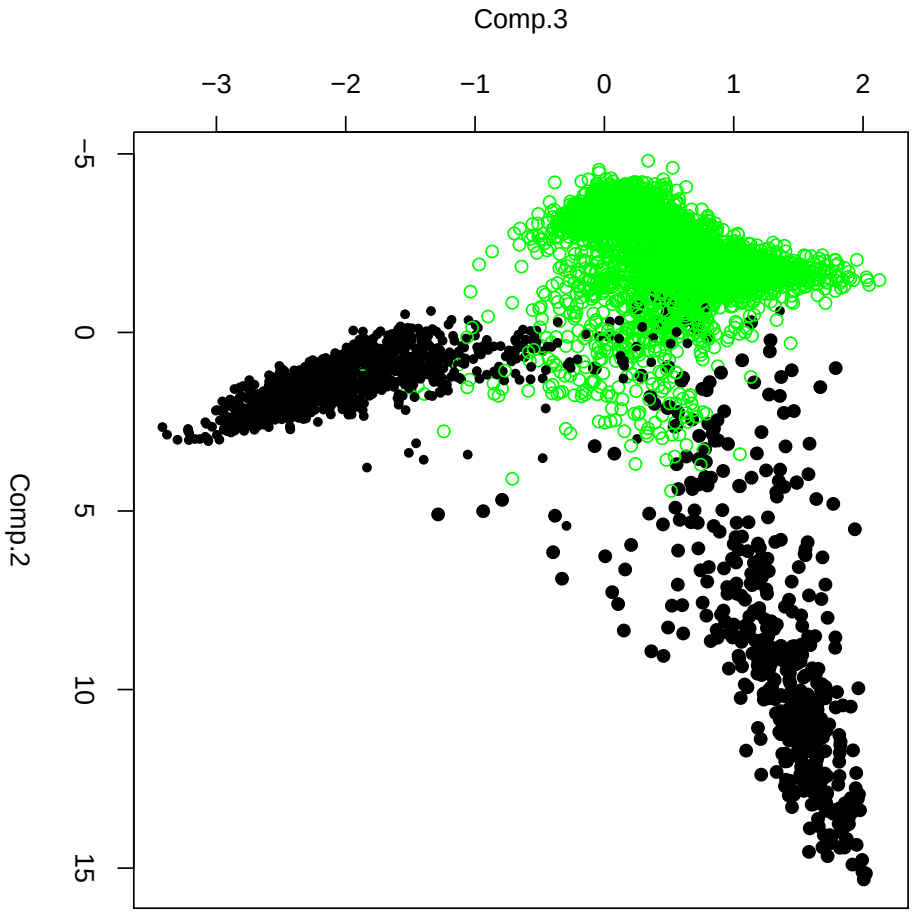
- useful to get some idea about possible clusters, outliers etc.

- PCA is frequently used as a feature extraction method. We work with the first few principal components instead of original variables. This is used e.g. in PCA regression when response is regressed on the first r principal components of predictors. This should be used with caution: principal components do not use the response. It is possible that a lesser principal component is actually very important in predicting the response.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal component analysis

In classification problems with large number of attributes it frequently pays off to perform LDA or QDA on several first principal components of $\mathbf{x}$. Scatterplot of the 2nd and 3rd principal components for sattelite image data:



Principal component analysis - towards nonlinear PCA

Assume for convenience that the data are centered. The task of PCA can be stated as that of seeking in R^p of such a subspace of dimension k, spanned by an orthonormal basis of vectors $\mathbf{a}_i$, $i = 1, \ldots, k$ that the value of

$$E\left(\|\mathbf{x} - \sum_{i=1}^{k} (\mathbf{a}_i^{\mathrm{T}} \mathbf{x}) \mathbf{a}_i \|^2 \right)$$

is minimized. Note that the projection of any $\mathbf{x} \in R^p$ on the subspace spanned by the $\mathbf{a}_i$ has the form

$$\sum_{i=1}^{k} (\mathbf{a}_i^{\mathrm{T}} \mathbf{x}) \mathbf{a}_i .$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal component analysis - towards nonlinear PCA

One can replace in given criterion the $\mathbf{a}_i^{\mathrm{T}}\mathbf{x}$ by some nonlinear functions of the inner products $g_i(\mathbf{a}_i^{\mathrm{T}}\mathbf{x})$, $i = 1, \ldots, k$, to obtain a nonlinear setup for PCA. This task bears some similarity to Independent Component Analysis and we shall turn later only to that latter problem.

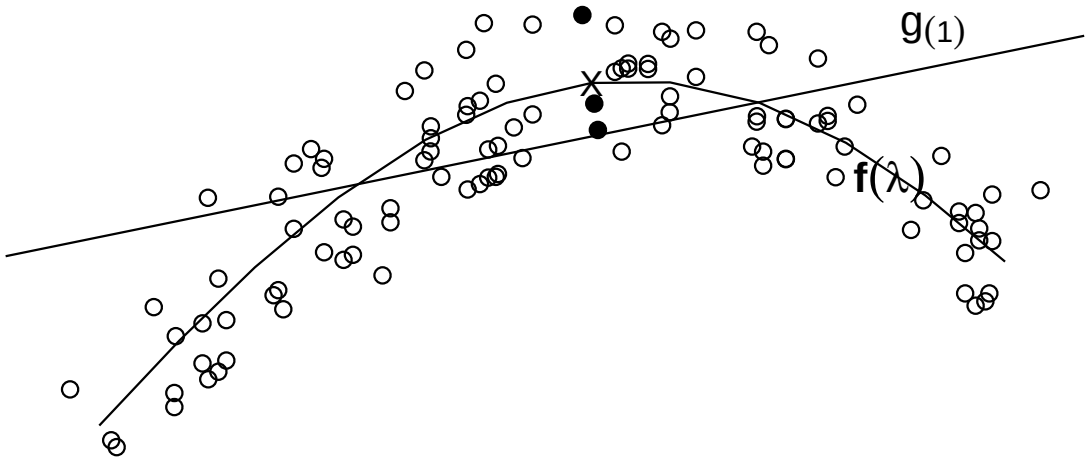
Now, we shall sketch a slightly different problem, namely that of constructing principal curves and surfaces. Principal curve $\mathbf{f}$ in R^p, parameterized by a real valued parameter λ is defined as

$$E[\mathbf{x}|\lambda(\mathbf{x}) = \lambda] = \mathbf{f}(\lambda);$$

here $\lambda(\mathbf{x})$ is a projection of $\mathbf{x}$ on $\mathbf{f}$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Principal component analysis



Kernel principal components

Transform data into a feature space by some $\phi : R^p \longrightarrow R^M$. Assume for convenience that the data are centered in the feature space (or first center the transformed data replacing matrix $\boldsymbol{\phi} = [\phi(\mathbf{x}_1)', \ldots, \phi(\mathbf{x}_n)']'$ by $(\mathbf{I} - \mathbf{M})\boldsymbol{\phi}$ and, by some abuse of notation, keep the symbol $\boldsymbol{\phi}$ for the centered data).

Now, take the covariance matrix in the feature space,

$$\mathbf{C} = \frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i)\phi(\mathbf{x}_i)',$$

and perform its spectral decomposition

$$\mathbf{C}\mathbf{v}_k = \lambda_k \mathbf{v}_k,$$

$k = 1, \ldots, M$. Write

$$\frac{1}{n} \sum_{i=1}^{n} \phi(\mathbf{x}_i)\{\phi(\mathbf{x}_i)'\mathbf{v}_k\} = \lambda_k \mathbf{v}_k.$$

Hence, if only $\lambda_k > 0$, we obtain

Kernel principal components

$$\mathbf{v}_k = \sum_{i=1}^{n} a_{ki}\phi(\mathbf{x}_i).$$

Upon substitution,

$$\frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)'\sum_{m=1}^{n}a_{km}\phi(\mathbf{x}_m) = \lambda_k\sum_{i=1}^{n}a_{ki}\phi(\mathbf{x}_i)$$

and, premultiplying both sides by $\phi(\mathbf{x}_\ell)'$, we get

$$\frac{1}{n}\sum_{i=1}^{n}k(\mathbf{x}_\ell,\mathbf{x}_i)\sum_{m=1}^{n}a_{km}k(\mathbf{x}_i,\mathbf{x}_m) = \lambda_k\sum_{i=1}^{n}a_{ki}k(\mathbf{x}_\ell,\mathbf{x}_i),$$

where

$$k(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)'\phi(\mathbf{x}_j).$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Kernel principal components

In matrix notation, for $l = 1, \ldots, n$, we thus have

$$\mathbf{K}^2 \mathbf{a}_k = \lambda_k n \mathbf{K} \mathbf{a}_k,$$

where $\mathbf{a}_k$ is the n-vector with elements a_{ki}. Finally,

$$\mathbf{K} \mathbf{a}_k = \lambda_k n \mathbf{a}_k,$$

the eigenproblem for an $n \times n$ matrix $\mathbf{K}$. For any $\mathbf{x}$ from the original space, its projection onto the eigenvector $\mathbf{v}_k$ in the feature space is given by

$$\phi(\mathbf{x})^{\mathrm{T}} \mathbf{v}_k = \sum_{i=1}^{n} a_{ki} k(\mathbf{x}, \mathbf{x}_i).$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Sparse principal component analysis

During the lecture, we shall discuss the paper by Hui Zou, Trevor Hastie and Robert Tibshirani (2004).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Factor analysis

Let $\mathbf{x} = [x^{(1)}, \ldots, x^{(p)}]' \in R^p$ be an observation and let $\mathbf{z} = = [z^{(1)}, \ldots, z^{(k)}]' \in R^k$, $k < p$, be a k-vector of factors (latent variables of a kind).

Assume that $\mathbf{x}$ has expected value $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$ (in practice, both are unknown and have to be estimated from data).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Factor analysis

Let

$$\mathbf{x} = \mathbf{m} + \mathbf{\Delta z} + \mathbf{e},$$

where $\mathbf{m} = [m^{(1)}, \ldots, m^{(p)}]'$ is the mean vector of $\mathbf{x}$, $\mathbf{\Delta}_{(p,k)}$ is a matrix of unknown coefficients δ_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, k$, $\mathbf{e} = [e^{(1)}, \ldots, e^{(p)}]'$ is a random vector,

$$E(\mathbf{z}) = \mathbf{0}, \quad \mathrm{Cov}(\mathbf{z}) = \mathbf{I},$$

where $\mathrm{Cov}(\mathbf{z})$ denotes covariance matrix of $\mathbf{z}$,

$$E(\mathbf{e}) = \mathbf{0}, \quad \mathrm{Cov}(\mathbf{e}) \equiv \mathbf{\Psi} = \mathrm{diag}(\psi_{11}, \ldots, \psi_{pp}),$$

$$\mathrm{Cov}(\mathbf{z}, \mathbf{e}) = \mathbf{0}$$

and $\mathrm{Cov}(\mathbf{z}, \mathbf{e})$ is a covariance matrix of $\mathbf{z}$ and $\mathbf{e}$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Factor analysis

It follows that each centered component of $\mathbf{x}$, $x^{(i)} - m^{(i)}$, is a linear combination of k uncorrelated common factors and a random variable (often referred to as specific variate) $e^{(i)}$ which is uncorrelated with the factors:

$$x^{(i)} = m^{(i)} + \sum_{j=1}^{k} \delta_{ij} z^{(j)} + e^{(i)},$$

$i = 1, \ldots, p$. The coefficients or weights δ_{ij} are referred to as factor loadings.

It also follows from the above that

$$\mathbf{\Sigma} = \mathbf{\Delta}\mathbf{\Delta}' + \mathbf{\Psi}.$$

Our objective is to determine k and the elements of $\mathbf{\Delta}$ and $\mathbf{\Psi}$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Factor analysis

Note, however, that for any nonsingular orthogonal transformation of $\mathbf{z}$, we get

$$\mathbf{x} = \mathbf{m} + (\mathbf{\Delta G})(\mathbf{G}'\mathbf{z}) + \mathbf{e},$$

where $\mathbf{G}_{(k,k)}$ is any orthogonal matrix, and

$$\mathbf{\Sigma} = (\mathbf{\Delta G})(\mathbf{\Delta G})' + \mathbf{\Psi} = \mathbf{\Delta\Delta}' + \mathbf{\Psi}.$$

Shortly put, matrix $\mathbf{\Delta}$ cannot be determined in a unique way. We therefore impose additional conditions on the elements of $\mathbf{\Delta}$ and $\mathbf{\Psi}$ or on $\mathbf{\Delta}$ solely. The simplest such condition requires that

$$\mathbf{\Delta}^{\mathrm{T}}\mathbf{\Delta}$$

be diagonal.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Factor analysis

Having identified the factor space in this way, we are at liberty to rotate the axes into any position that provides better interpretation of of the rotated weights (for the given application).

Regarding estimation process, observe that we have $\frac{1}{2}p(p+1)$ items of information (elements of $\mathbf{S}$) from which to estimate pk factor loadings and p specific variances. Given the constraint that $\mathbf{\Delta}^{\mathrm{T}}\mathbf{\Delta}$ is diagonal, we thus need for estimability of parameters that

$$\frac{1}{2}p(p+1) \geqslant p(k+1) - \frac{1}{2}k(k-1)$$

i.e. that $(p-k)^2 \geqslant p+k$.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Factor analysis (FA)

It easy to see that, at least formally, PCA and FA are very close one to another. Indeed, we have

$$\mathbf{x} = \mathbf{\Delta z} + \mathbf{e}$$

and

$$\mathbf{y} = \mathbf{\Gamma}'\mathbf{x}.$$

Moreover, since $\mathbf{\Gamma}$ is orthogonal, we in fact have

$$\mathbf{x} = \mathbf{\Gamma y}.$$

Despite the apparent similarity, the two problems are qualitatively very different (we shall discuss this issue during the lecture).

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund