MA012 Statistika II

11. Analýza hlavních komponent (PCA)

Ondřej Pokora (pokora@math.muni.cz)

Ústav matematiky a statistiky, Přírodovědecká fakulta, Masarykova univerzita, Brno

(podzim 2015)

Analýza hlavních komponent

Analýza hlavních komponent (Principal component analysis, PCA) je statistická metoda pro snížení dimenze (počtu) proměnných, které bývají často korelované. Připomeňme např. velmi častý problém multikolinearity ve vícenásobné lineární regresi.

PCA využívá ortogonalizace báze vektorového prostoru původních náhodných veličin. Nové bázové náhodné veličiny, nazývané hlavní komponenty, jsou nekorelované. Současně je volba nové báze optimální v tom, že pro vysvětlení zadaného podíly celkové variability dat stačí uvažovat nižší počet hlavních komponent, než byl počet (dimenze) původních náhodných veličin. Jednotlivé hlavní komponenty jsou postupně hledány tak, aby vysvětlily co největší část celkové variability dat pomocí co nejžšího počtu nových náhodnách veličin.

Nevýhou PCA může být nemožnost interpretace hlavních komponent, např. z důvodu nekompatibility použitých jednotek. V praxi toto často řešíme uvažováním bezrozměrných veličin.

Pozorování ve standardní bázi

Data n pozorování k veličin máme v matici $\boldsymbol{X}$ rozměru $n \times k$, $n > k$. Můžeme to chápat jako n pozorování v k-rozměrném prostoru. Uvažujeme-li standardní bázi jednotkových vektorů v $\mathbb{R}^k$,

$$\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k,$$

jednotlivé řádky matice $\boldsymbol{X}$ jsou souřadnice jednotlivých pozorování v této standardní bázi, která je ortonormální.

Předpokládejme nyní, že máme obecnou ortonormální bázi

$$\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in \mathbb{R}^k, \qquad \boldsymbol{u}_i' \boldsymbol{u}_j = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}.$$

Vytvořme matici $\boldsymbol{U}$ tak, že tyto bázové vektory umístíme do sloupců. Taková matice $\boldsymbol{U}$ je potom ortogonální, tzn. platí

$$\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_k, \qquad \boldsymbol{U}^{-1} = \boldsymbol{U}'.$$

Pro každé $\boldsymbol{x} \in \mathbb{R}^k$ pak platí

$$\boldsymbol{x} = a_1\boldsymbol{u}_1 + \cdots + a_k\boldsymbol{u}_k = \boldsymbol{U}\boldsymbol{a}, \qquad \boldsymbol{a} = \boldsymbol{U}'\boldsymbol{x},$$

kde $\boldsymbol{a} = (a_1, \ldots, a_k)'$ jsou souřadnice $\boldsymbol{x}$ v bázi sloupců matice $\boldsymbol{U}$.

Volba optimální báze

Bázi $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$ chceme zvolit tak, aby pro $\boldsymbol{x} \in \mathbb{R}^k$ platilo

$$\boldsymbol{x} \approx \boldsymbol{x}^*,$$

kde $\boldsymbol{x}^*$ je lineární kombinací pouze prvních $r < k$ bázových vektorů,

$$\boldsymbol{x}^* = a_1\boldsymbol{u}_1 + \cdots + a_r\boldsymbol{u}_r = \boldsymbol{U}^*\boldsymbol{a}^*,$$

s redukovanou maticí $\boldsymbol{U}^*$ tvořenou prvními r sloupci matice $\boldsymbol{U}$. Máme tedy

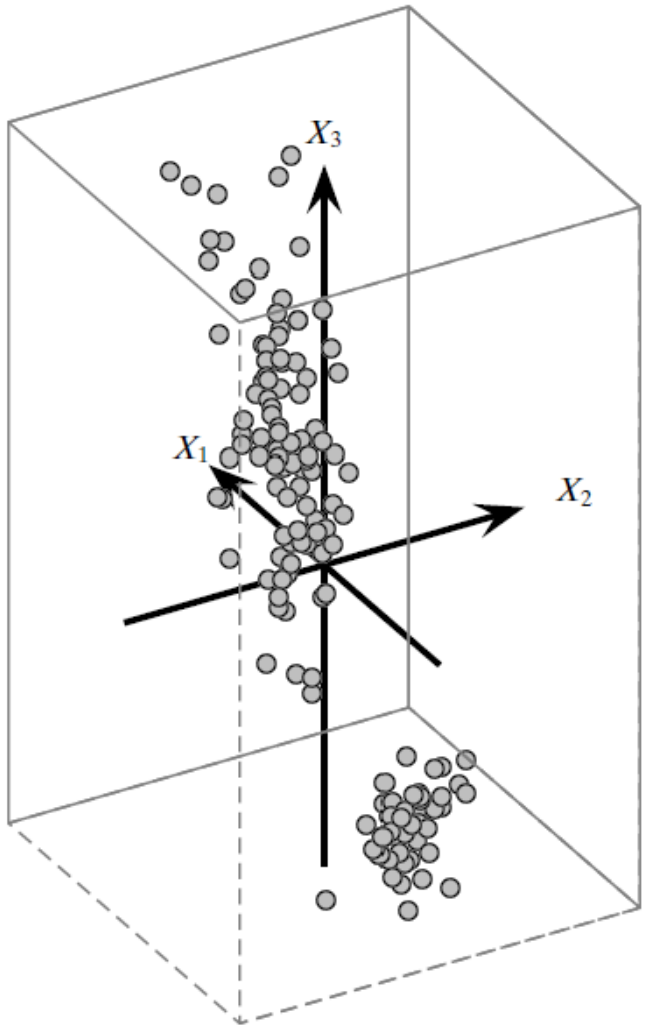
$$\boldsymbol{x}^* = \boldsymbol{U}^*\boldsymbol{a}^* = \underbrace{\boldsymbol{U}^*(\boldsymbol{U}^*)'}_{\boldsymbol{P}}\boldsymbol{x}, \quad \text{neboť} \quad \boldsymbol{a}^* = (\boldsymbol{U}^*)'\boldsymbol{x}.$$

Aproximace $\boldsymbol{x}^*$ je tedy projekcí $\boldsymbol{x}$ do podprostoru generovaného sloupci redukované matice $\boldsymbol{U}^*$, a to pomocí projekční matice $\boldsymbol{P}$.

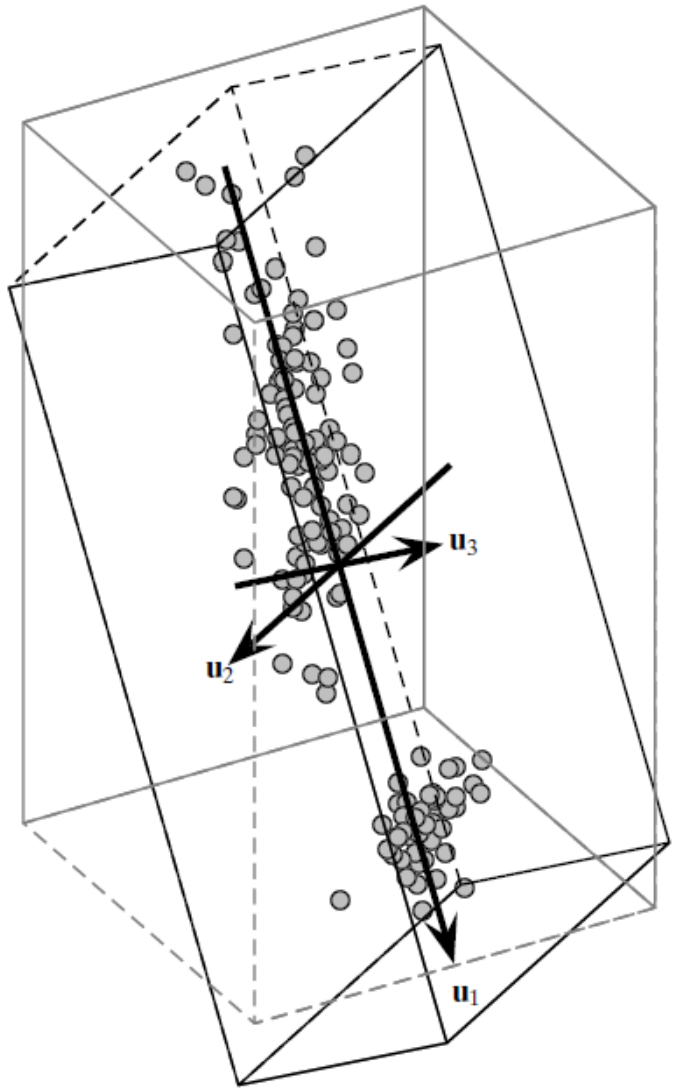
Dopustíme se přitom chyby

$$\boldsymbol{\varepsilon} = \boldsymbol{x} - \boldsymbol{x}^* = a_{r+1}\boldsymbol{u}_{r+1} + \cdots + a_k\boldsymbol{u}_k.$$

Různé volby bází



(a) Original Basis



(b) Optimal Basis

Analýza hlavních komponent

Od začátku: chceme nalézt první vektor $\boldsymbol{u}$ optimální ortonormální báze, pomocí něhož dokážeme $\boldsymbol{x}$ aproximovat pomocí $\boldsymbol{x}^*$ nejlépe z pohledu vysvětleného rozptylu.

Dále předpokládáme, že data v matici $\boldsymbol{X}$ jsou centrována; v opačném případě je centrujeme, tzn. od každého sloupce odečteme jeho průměr.

Projekce i-tého pozorování $\boldsymbol{x}_i$, do směru bázového vektoru $\boldsymbol{u}$ je tvaru

$$\boldsymbol{x}_i^* = a_i \boldsymbol{u}, \qquad i = 1, \ldots, n.$$

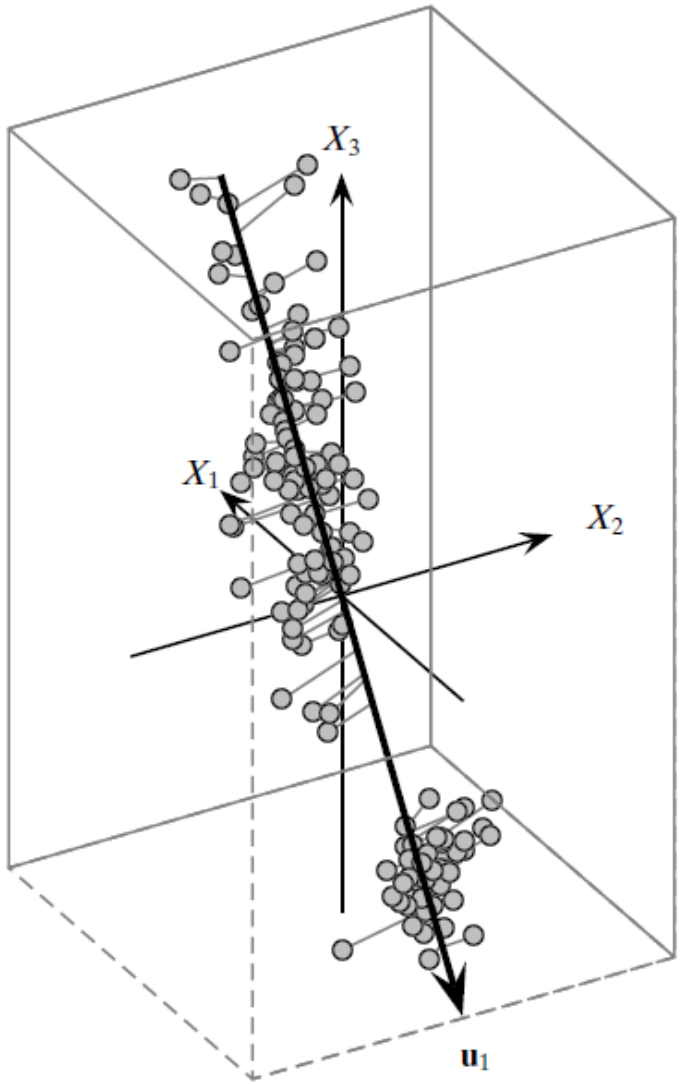
Rozptyl projektovaných pozorování $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_n^*$ je rovný

$$\sigma_{\boldsymbol{u}}^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i - \underbrace{\mu_i}_{=0})^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2 = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{u}'\boldsymbol{x}_i)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{u}'\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{u} = \boldsymbol{u}'\underbrace{\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)}_{\boldsymbol{\Sigma}}\boldsymbol{u} = \boldsymbol{u}'\boldsymbol{\Sigma}\boldsymbol{u},$$

kde $\boldsymbol{\Sigma}$ je kovarianční matice $(k \times k)$ centrovaných veličin.

Volba optimálního směru



Obrázky: Zaki & Meira: Data mining and analysis, 2014.

Analýza hlavních komponent

Optimální vektor $\boldsymbol{u}$, $\boldsymbol{u}'\boldsymbol{u} = 1$, hledáme tak, abychom maximalizovali rozptyl $\sigma^2_{\boldsymbol{u}}$ pozorování projektovaných do směru $\boldsymbol{u}$. Řešíme tedy úlohu na vázaný extrém

$$\sigma^2_{\boldsymbol{u}} = \boldsymbol{u}'\boldsymbol{\Sigma}\boldsymbol{u} \to \max, \qquad \text{za podmínky} \qquad \boldsymbol{u}'\boldsymbol{u} = 1.$$

Použijeme Lagrangeovu metodu. Lagrangeova funkce L s Lagrangeovým multiplikátorem λ je

$$L(\boldsymbol{u}) = \boldsymbol{u}'\boldsymbol{\Sigma}\boldsymbol{u} - \lambda(\boldsymbol{u}'\boldsymbol{u} - 1),$$

pro niž nalezneme pomocí stacionárního bodu $\frac{\partial L(\boldsymbol{u})}{\partial \boldsymbol{u}} = \boldsymbol{0}$ řešení zadané vztahem

$$\boldsymbol{\Sigma}\boldsymbol{u} = \lambda\boldsymbol{u}.$$

To znamená, že λ je vlastní číslo kovarianční matice $\boldsymbol{\Sigma}$ a $\boldsymbol{u}$ je jemu odpovídající vlastní vektor.

Přitom hledaný rozptyl je rovný $\sigma^2_{\boldsymbol{u}} = \lambda$. Má-li být $\sigma^2_{\boldsymbol{u}}$ maximální, je třeba vzít za λ největší vlastní číslo kovariační matice $\boldsymbol{\Sigma}$ a za $\boldsymbol{u}$ jemu odpovídající vlastní vektor, který označujeme jako první hlavní komponentu $\boldsymbol{u}_1$.

Analýza hlavních komponent

Na uvedeném principu je založena metoda hlavních komponent. Postupně hledáme vektory ortonormální báze, které s co nejmenším dimenzí generovaného podprostoru vysvětlují co nejvíce variability dat.

Ve druhém kroku tedy hledáme druhou hlavní komponentu $\boldsymbol{u}_2$,

$$\boldsymbol{u}_2'\boldsymbol{u}_2 = 1, \qquad \boldsymbol{u}_1'\boldsymbol{u}_2 = 0,$$

tak, aby se maximalizovala variabilita v datech projektovaných do podprostoru generovaného vektory $\boldsymbol{u}_1$ a $\boldsymbol{u}_2$.

Analýza hlavních komponent

Kovarianční matice $\boldsymbol{\Sigma}$ datové matice $\boldsymbol{X}$ rozměru $n \times k$ je symetrická a pozitivně semidefinitní čtvercová matice řádu k, tzn. její vlastní čísla jsou vždy nezáporná, a počet kladných vlastních čísel odpovídá hodnosti matice $\boldsymbol{\Sigma}$. Uspořádejme tato vlastní čísla do nerostoucí posloupnosti

$$\lambda_1 \geq \cdots \geq \lambda_k$$

a označme k nim příslušející vlastní vektory $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k$.

Věta 1

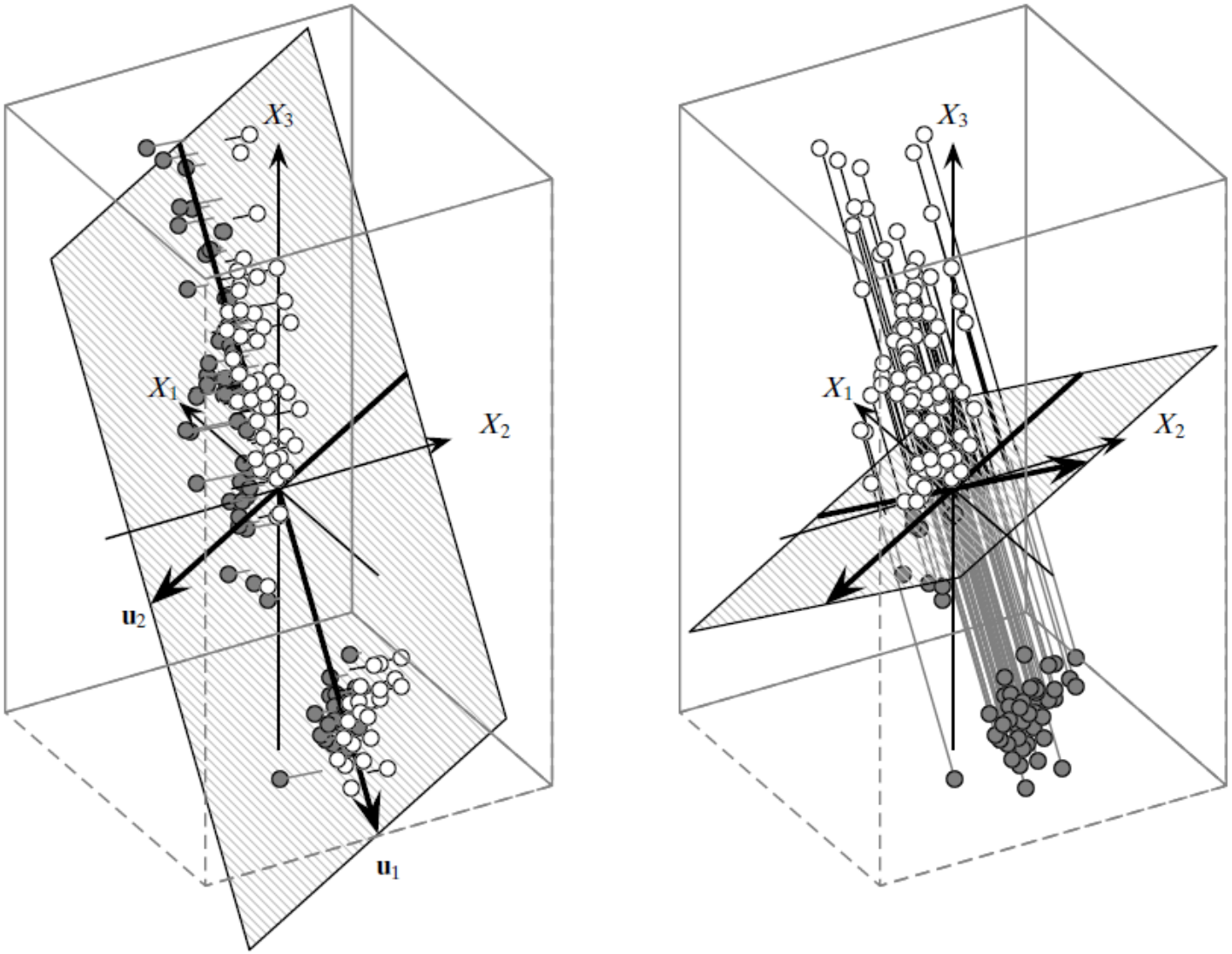
Platí $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$,

kde sloupce matice $\boldsymbol{U}$ *jsou vlastní vektory a* $\boldsymbol{\lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$.

Celková variabilita dat v matici $\boldsymbol{X}$ *je rovna*

$$\sum_{i=j}^{k} \lambda_j = \operatorname{Tr} \boldsymbol{\Sigma}.$$

Variabilita v různých bázích



(a) Optimal basis

(b) Nonoptimal basis

Algoritmus PCA I

1. Pro datovou matici $\boldsymbol{Y}$ typu $n \times k$ spočítáme výběrové průměry pro každou veličinu (tj. sloupcové průměry)

$$\boldsymbol{\mu}' = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i' = \left(\frac{1}{n}\sum_{i=1}^{n} x_{i1}, \ \ldots, \ \frac{1}{n}\sum_{i=1}^{n} x_{ik} \right).$$

2. Vytvoříme matici cetrovaných veličin rozměru $n \times k$,

$$\boldsymbol{X} = (1, \ldots, 1)'\boldsymbol{\mu}',$$

   složenou z n řádků $\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n'$.

3. Označíme $\boldsymbol{\Sigma}$ kovariační matici matice $\boldsymbol{X}$,

$$\boldsymbol{\Sigma} = \mathrm{cov}\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'.$$

Algoritmus PCA II

4. Spočítáme vlastní čísla kovarianční matice $\boldsymbol{\Sigma}$, seřadíme je do nerostoucí posloupnosti a odpovídající vlastní vektory uspořádáme (ve stejném pořadí) do sloupců matice $\boldsymbol{U}$,

$$\lambda_1 \geq \cdots \geq \lambda_k, \qquad \boldsymbol{U} = (\boldsymbol{u}_1 \,|\, \cdots \,|\, \boldsymbol{u}_k)\,.$$

5. Nalezneme vhodnou dimenzi $r < k$, např. s podmínkou na minimální hodnotu podílu rozptylu

$$\frac{\sum_{j=1}^{r} \lambda_j}{\sum_{j=1}^{k} \lambda_j} \geq \alpha$$

pro stanovenou hodnotu $\alpha \in (0; 1)$, často se uvádí $\alpha = 0{,}8$.

6. Vytvoříme matici redukované báze hlavních komponent $\boldsymbol{U}_r$ tak, že z matice $\boldsymbol{U}$ vezmeme pouze prvních r sloupců.

7. Datová matice v bázi hlavních komponent $\boldsymbol{A}$ je rozměru $n \times r$ a je tvořena řádky $\boldsymbol{a}'_i$, kde

$$\boldsymbol{a}'_i = \boldsymbol{U}_r \boldsymbol{x}'_i, \qquad i = 1, \ldots, n.$$

Příklad

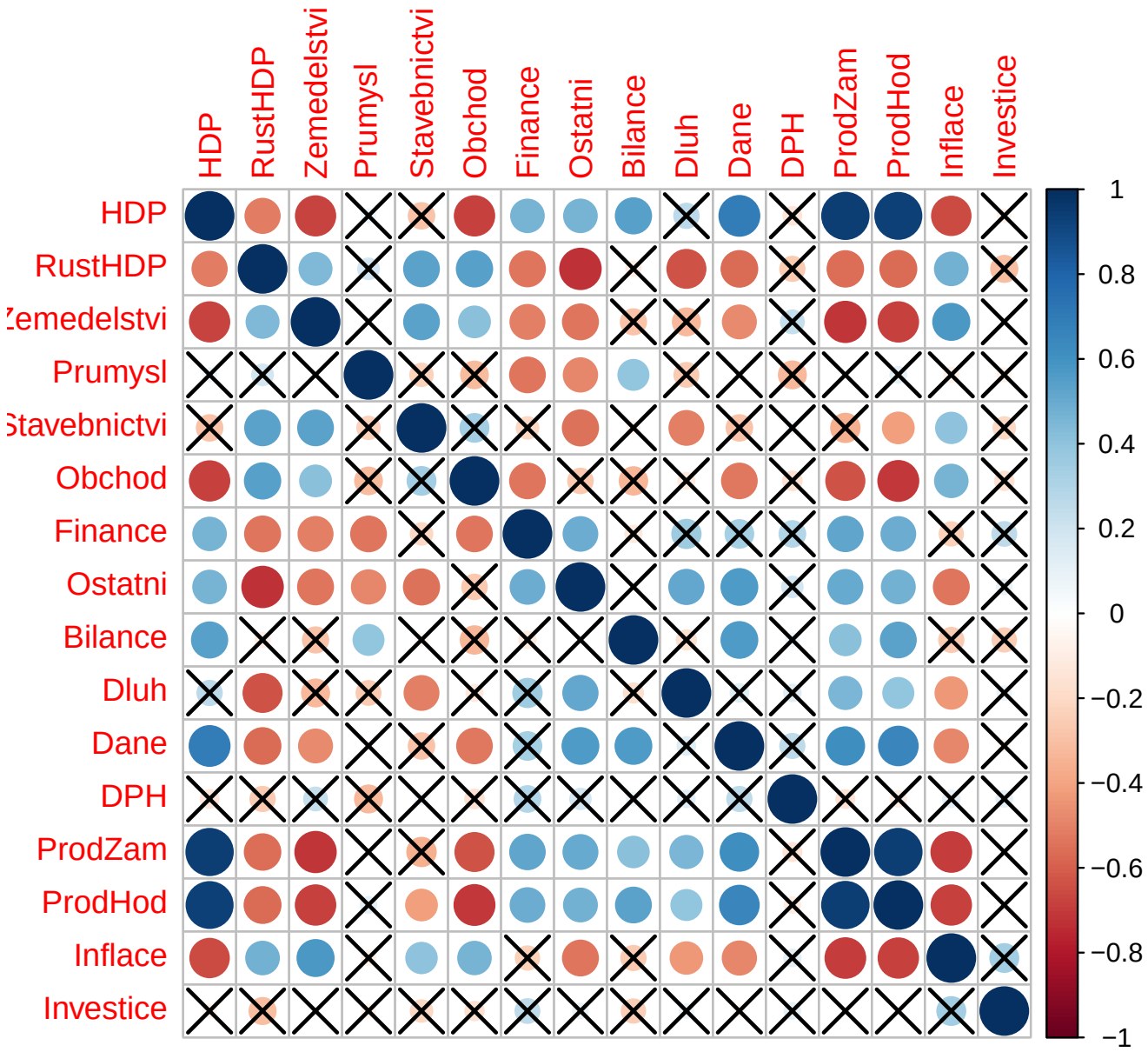
V *R* je klasická PCA naprogramována ve funkci `prcomp`, která pro hledání vlastních čísel a vektorů využívá klasický spektrální rozklad kovarianční matice $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$.

V praxi se však obvykle používá tzv. *singular value decomposition (SVD)* datové matice $\boldsymbol{X} = \boldsymbol{L}\boldsymbol{\Delta}\boldsymbol{R}'$, která je výpočetně výhodnější. V *R* tento postup využívá funkce `princomp`.

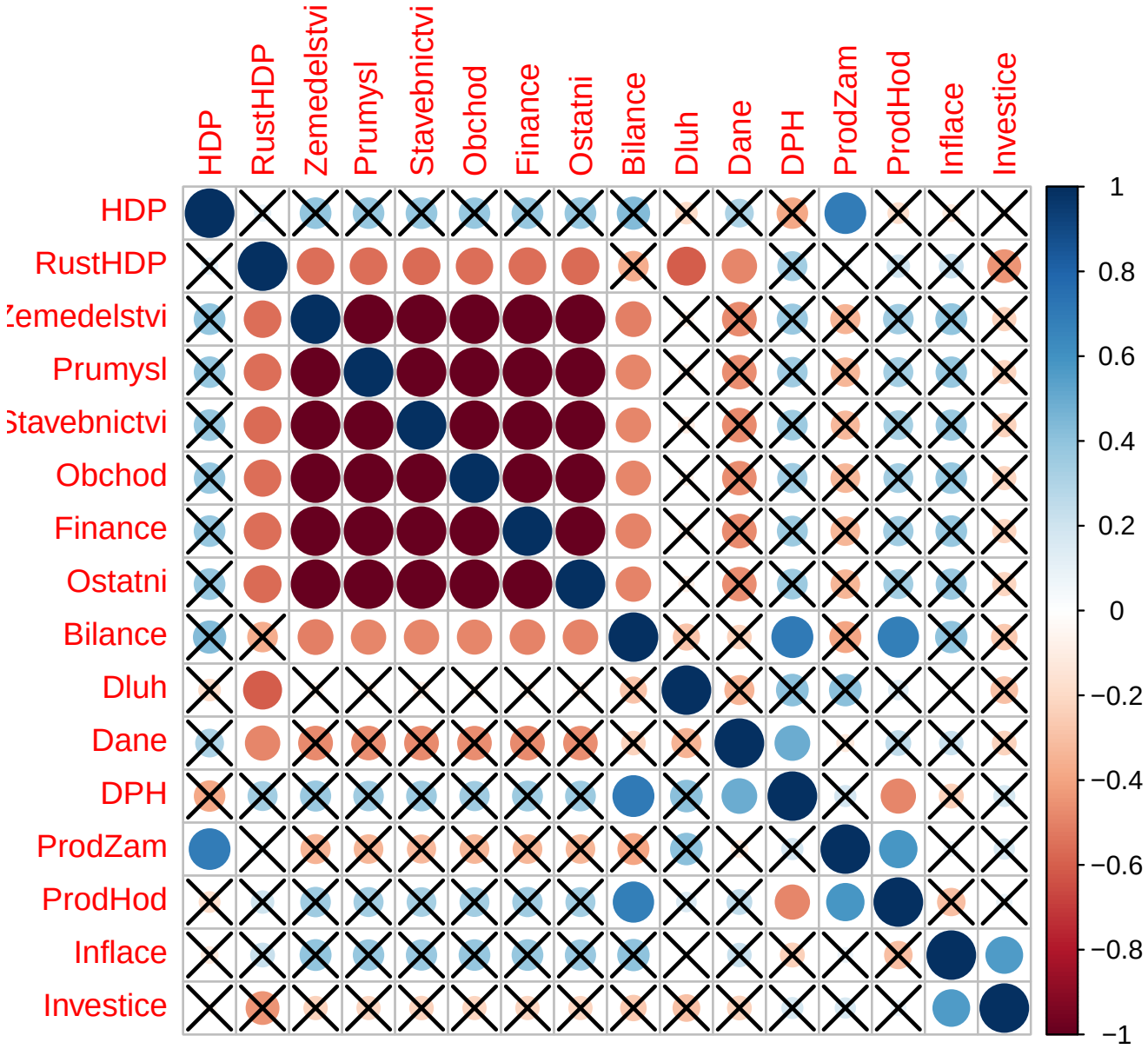
Pro grafické reprezetace výsledků PCA pak používáme mj. `screeplot` a `biplot`.

Příklad 1

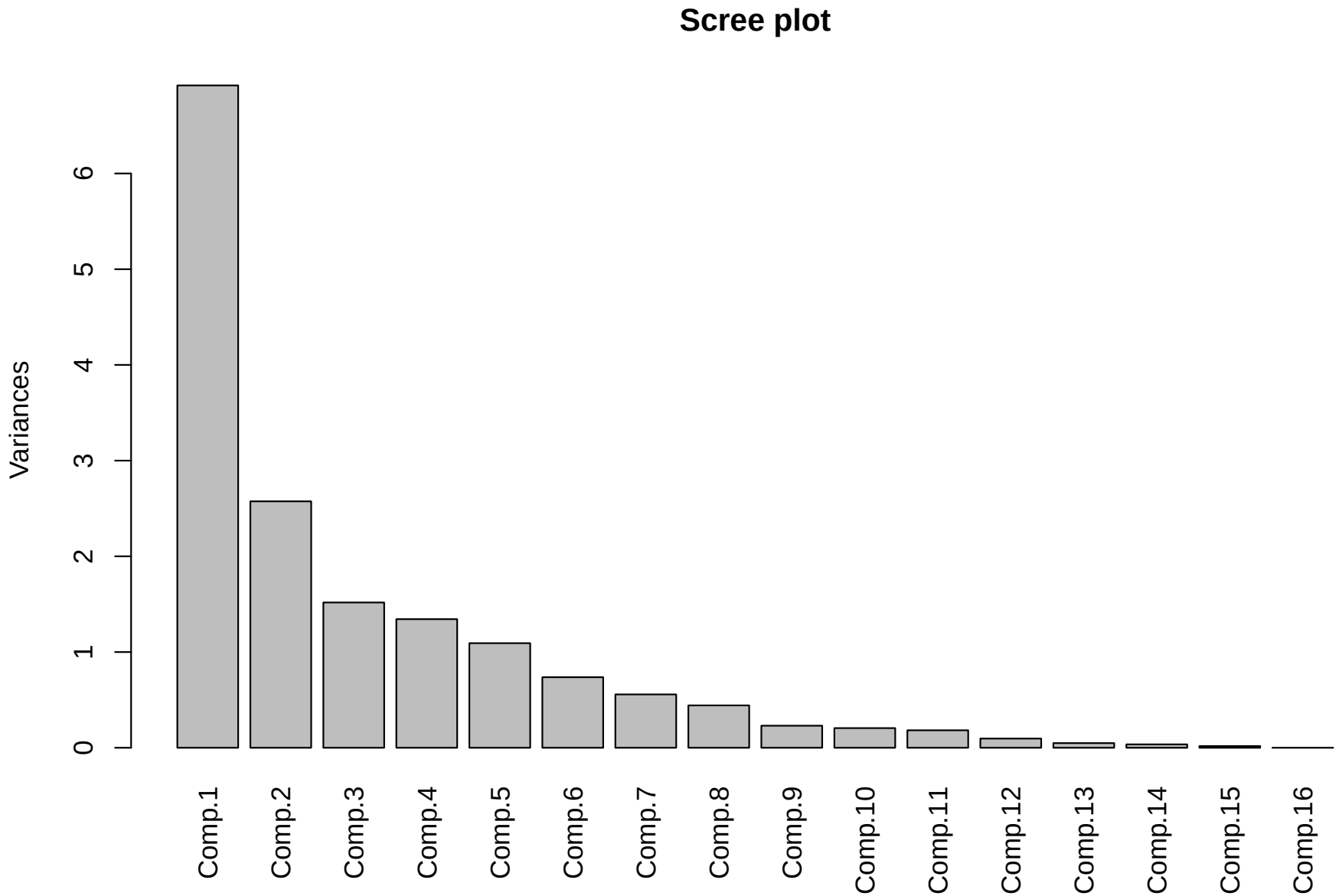
V souboru `ukazatele.csv` *je 16 vybraných ekonomických a finančních ukazatelů 29 evropských zemí z roku 2007, které publikoval EUROSTAT v roce 2009. Proveďte analýzu hlavních komponent a identifikujte hlavní komponenty ukazatelů.*


HDP
RustHDP
Zemedelstvi
Prumysl
Stavebnictvi
Obchod
Finance
Ostatni
Bilance
Dluh
Dane
DPH
ProdZam
ProdHod
Inflace
Investice
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

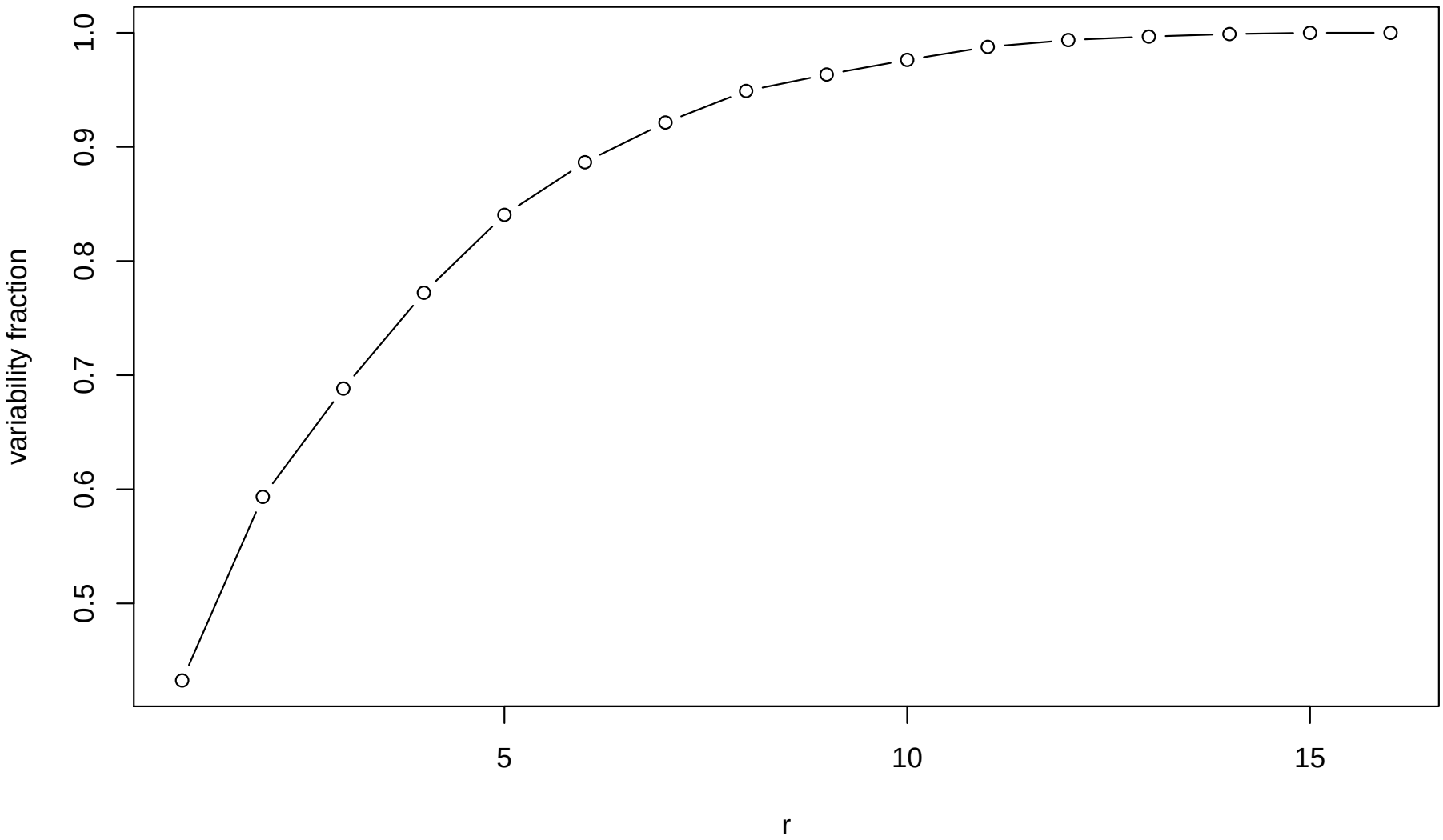
Korelogram



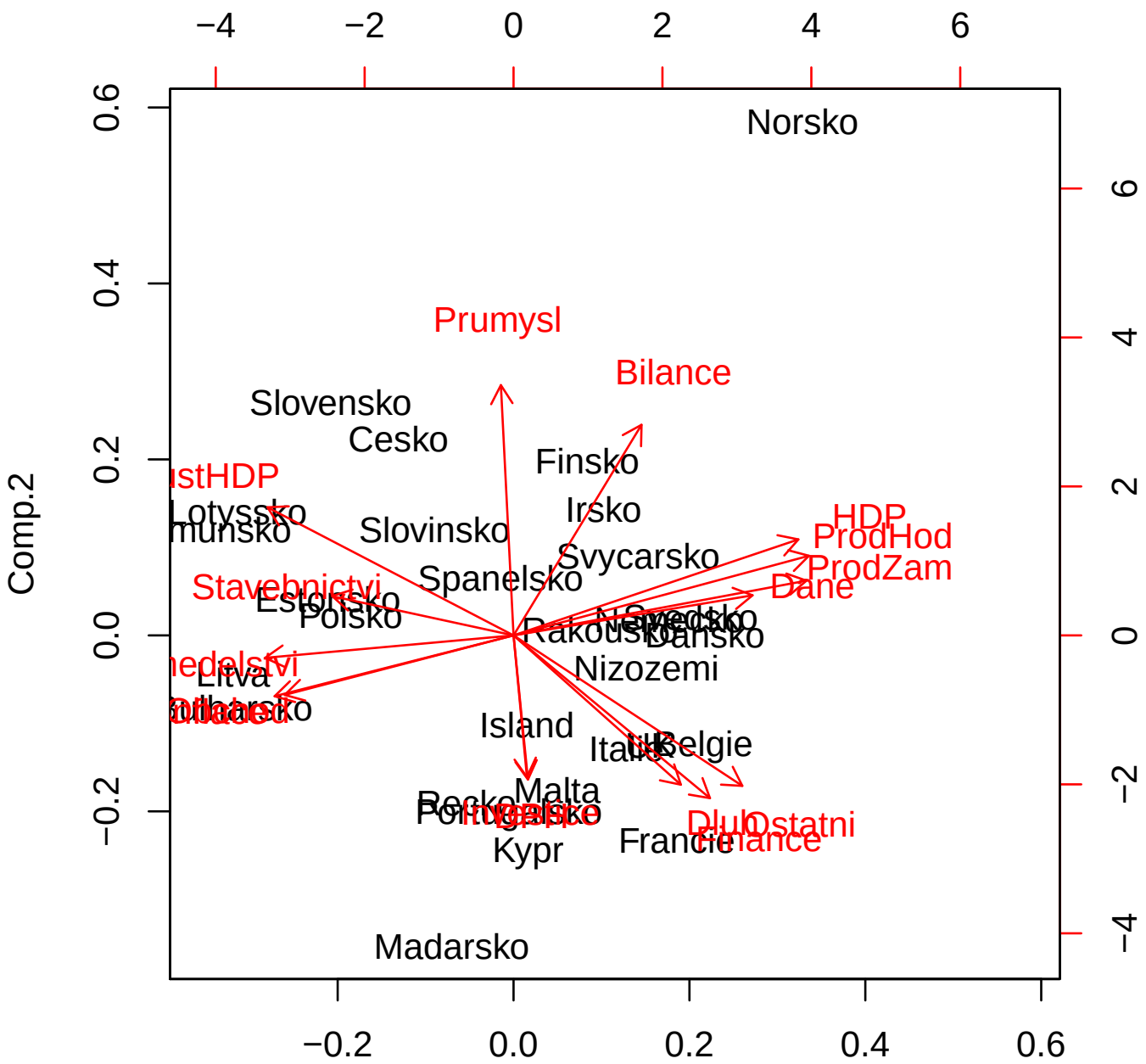
Korelogram parciálních korelací



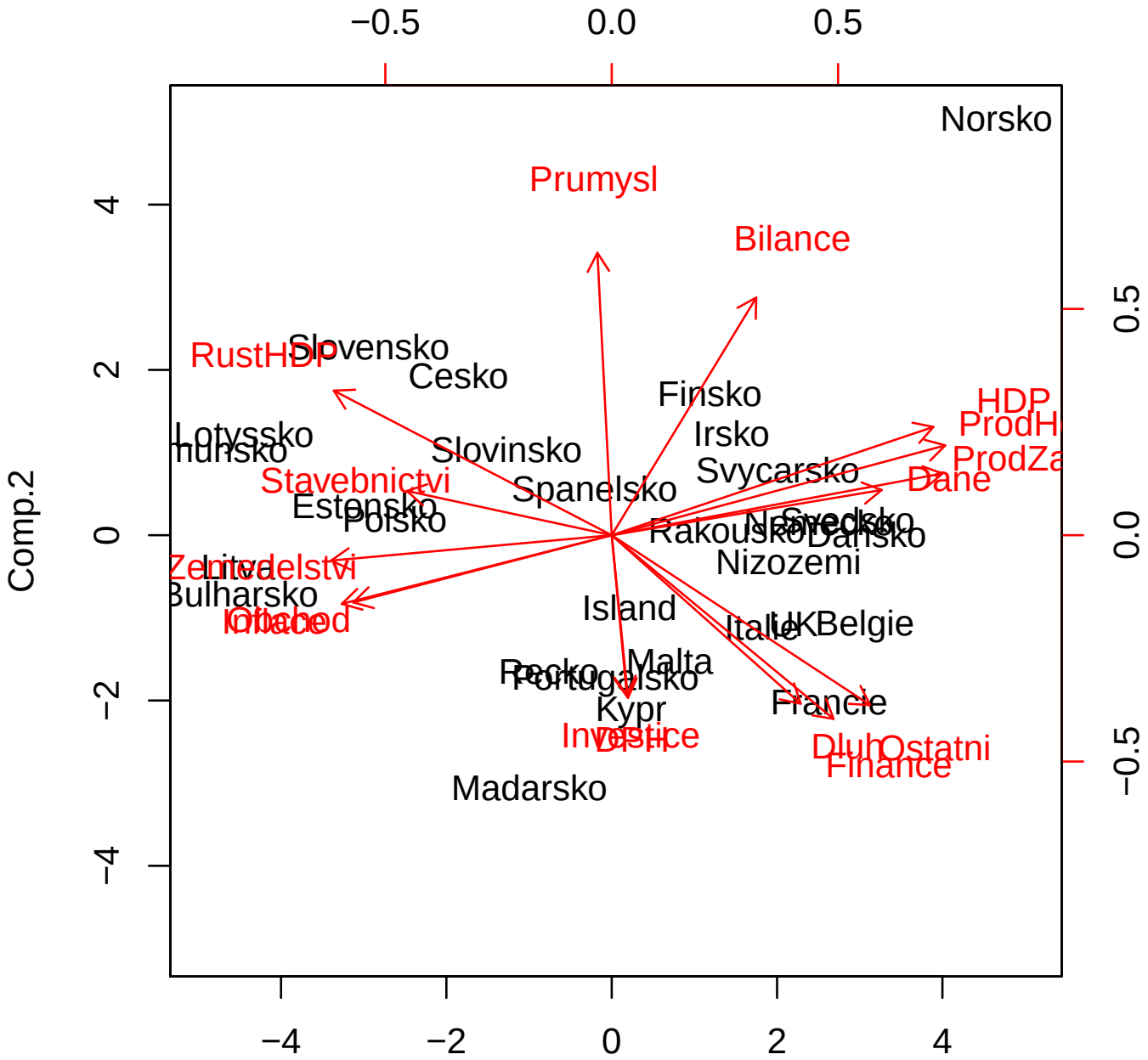
Tzv. suťový graf



Graf podílu vysvětlené variability


Norsko
Prumysl
Bilance
Slovensko
Cesko
Finsko
Irsko
Slovinsko
Svycarsko
Spanelsko
Estonsko
Polsko
Litva
Island
Belgie
Malta
Kypr
Francie
Madarsko
HDP
ProdHod
ProdZam
Dane
Nizozemi
Comp.1
Comp.2

Tzv. biplot



Tzv. biplot (v jiné škále)