

Vícerozměrné statistické metody

Ordinační analýzy – přehled metod

Jiří Jarkovský, Simona Littnerová

Vícerozměrné statistické metody

Analýza hlavních komponent jako příklad výpočtu redukce dimenzionality pomocí ordinační analýzy

Analýza hlavních komponent

- Analýza hlavních komponent je typickou metodou ze skupiny ordinačních analýz
- Pracuje s asociací proměnných popisujících objekty a snaží se na základě jejich korelací/kovariancí stanovit dimenze zahrnující větší podíl variability než připadá na původní proměnné
- Předpoklady jsou obdobné jako při výpočtu korelací a kovariancí:
 - nepřítomnost odlehlých hodnot (s výjimkou situace kdy analýzu provádíme za účelem identifikace odlehlých hodnot)
 - nepřítomnost více skupin objektů (s výjimkou situace kdy analýzu provádíme za účelem detekce přirozeně existujících shluků spjatých s největší variabilitou souboru)
- Datový soubor musí mít více objektů než proměnných, pro získání stabilních výsledků se doporučuje alespoň 10x tolik objektů než proměnných, ideální je 40-60x více objektů než proměnných
- Cíle analýzy
 - Popis a vizualizace vztahů mezi proměnnými
 - Výběr neredundantních proměnných pro další analýzy
 - Vytvoření zástupných faktorových os pro použití v dalších analýzách
 - Identifikace shluků v datech spjatých s variabilitou dat
 - Identifikace vícerozměrně odlehlých objektů

Výpočet faktorových os

- Výpočetně vychází analýza hlavních komponent z korelační/kovarianční asociační matice (a obdobně i další ordinační analýzy, pouze pomocí jiných asociačních metrik)
- Vlastní výpočet je pak realizován prostřednictvím výpočtu vlastních čísel a vlastních vektorů této matice
- Vlastní vektory a vlastní čísla
 - Existují pro čtvercové matice
 - Vyžadují aby hodnota matice odpovídala jejímu řádu, tedy pouze pro matice v nichž neexistuje lineární závislost. Tento fakt komplikuje (nebo znemožňuje) výpočet při přítomnosti zcela redundantních (lineárně závislých) proměnných
 - Vlastní čísla matice jsou ve vazbě na variabilitu vyčerpanou vytvářenými faktorovými osami
 - Vlastní vektory definují směr nových faktorových os v prostoru původních proměnných
 - Existuje několik možných vyjádření vlastních čísel a vlastních vektorů, proto je před interpretací výstupů nezbytné znát algoritmus použitý v SW

Vlastní čísla a vlastní vektory

Výpočet vlastních čísel pro matici A

$$|A - \lambda_i I| = 0$$

$$\left| \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\left| \begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0 \quad \Rightarrow$$

$$\begin{vmatrix} 2 - \lambda & 2 \\ 2 & 5 - \lambda \end{vmatrix} = 0$$

$$(2 - \lambda)(5 - \lambda) - 4 = 0$$

$$\lambda^2 - 7\lambda + 6 = 0$$

$$\lambda_1 = 6$$

$$\lambda_2 = 1$$

Výpočet vlastního vektoru l_1 ,
pro l_2 je výpočet obdobný

$$\lambda_1 = 6$$

$$\left(\begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix} - 6 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = 0$$

$$\begin{bmatrix} -4 & 2 \\ 2 & -1 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \end{bmatrix} = 0$$

$$-4u_{11} + 2u_{21} = 0$$

$$2u_{11} - 1u_{21} = 0$$

$$u_{11} = 1$$

$$-4 + 2u_{21} = 0$$

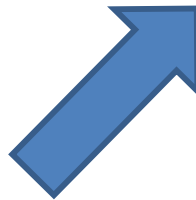
$$u_{21} = 2$$

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Příklad výpočtu

Primární data

1	2	3	4	5
SEPALLEN	SEPALWID	PETALLEN	PETALWID	IRISTYPE
5.0	3.3	1.4	0.2	SETOSA
6.4	2.8	5.6	2.2	VIRGINIC
6.5	2.8	4.6	1.5	VERSICO
6.7	3.1	5.6	2.4	VIRGINIC
6.3	2.8	5.1	1.5	VIRGINIC
4.6	3.4	1.4	0.3	SETOSA
6.9	3.1	5.1	2.3	VIRGINIC
6.2	2.2	4.5	1.5	VERSICO
5.9	3.2	4.8	1.8	VERSICO
4.6	3.6	1.0	0.2	SETOSA
6.1	3.0	4.6	1.4	VERSICO
6.0	2.7	5.1	1.6	VERSICO
6.5	3.0	5.2	2.0	VIRGINIC
5.6	2.5	3.9	1.1	VERSICO
6.5	3.0	5.5	1.8	VIRGINIC
5.8	2.7	5.1	1.9	VIRGINIC
6.8	3.2	5.9	2.3	VIRGINIC
5.1	3.3	1.7	0.5	SETOSA
5.7	2.8	4.5	1.3	VERSICO
6.2	3.4	5.4	2.3	VIRGINIC
7.7	3.8	6.7	2.2	VIRGINIC
6.3	3.3	4.7	1.6	VERSICO
6.7	3.3	5.7	2.5	VIRGINIC
7.6	3.0	6.6	2.1	VIRGINIC
4.9	2.5	4.5	1.7	VIRGINIC
5.5	3.5	1.3	0.2	SETOSA
6.7	3.0	5.2	2.3	VIRGINIC
7.0	3.2	4.7	1.4	VERSICO
6.4	3.2	4.5	1.5	VERSICO
6.1	2.8	4.0	1.3	VERSICO
4.8	3.1	1.6	0.2	SETOSA
5.9	3.0	5.1	1.8	VIRGINIC
5.5	2.4	3.8	1.1	VERSICO
6.3	2.5	5.0	1.9	VIRGINIC
6.4	3.2	5.3	2.3	VIRGINIC
5.2	3.4	1.4	0.2	SETOSA
4.9	3.6	1.4	0.1	SETOSA
5.4	3.0	4.5	1.5	VERSICO
7.9	3.8	6.4	2.0	VIRGINIC
4.4	3.2	1.3	0.2	SETOSA
6.7	3.3	5.7	2.1	VIRGINIC
5.0	3.5	1.6	0.6	SETOSA
5.8	2.6	4.0	1.2	VERSICO



Korelační matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.000	-0.118	0.872	0.818
SEPALWID	-0.118	1.000	-0.428	-0.366
PETALLEN	0.872	-0.428	1.000	0.963
PETALWID	0.818	-0.366	0.963	1.000

Kovarianční matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	0.686	-0.042	1.274	0.516
SEPALWID	-0.042	0.190	-0.330	-0.122
PETALLEN	1.274	-0.330	3.116	1.296
PETALWID	0.516	-0.122	1.296	0.581

Kovarianční nebo korelační matice?

Korelační matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.000	-0.118	0.872	0.818
SEPALWID	-0.118	1.000	-0.428	-0.366
PETALLEN	0.872	-0.428	1.000	0.963
PETALWID	0.818	-0.366	0.963	1.000

- Jednoznačně v případě nesrovnatelných jednotek (např. věk vs. krevní tlak)
- Korelace je vlastně kovariance standardizovaná na variabilitu dat, tedy kovariance na standardizovaných datech = korelace
- Diagonála obsahuje hodnotu 1
 - Úplná korelace proměnné sama se sebou
 - Standardizovaný rozptyl
- Ostatní buňky obsahují vzájemné korelace proměnných

Kovarianční matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	0.686	-0.042	1.274	0.516
SEPALWID	-0.042	0.190	-0.330	-0.122
PETALLEN	1.274	-0.330	3.116	1.296
PETALWID	0.516	-0.122	1.296	0.581

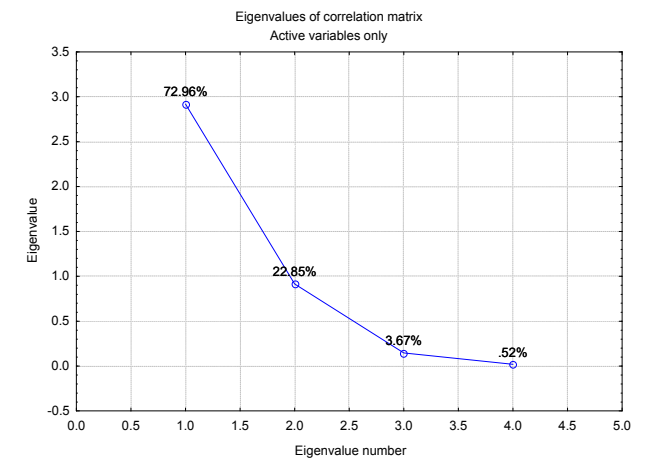
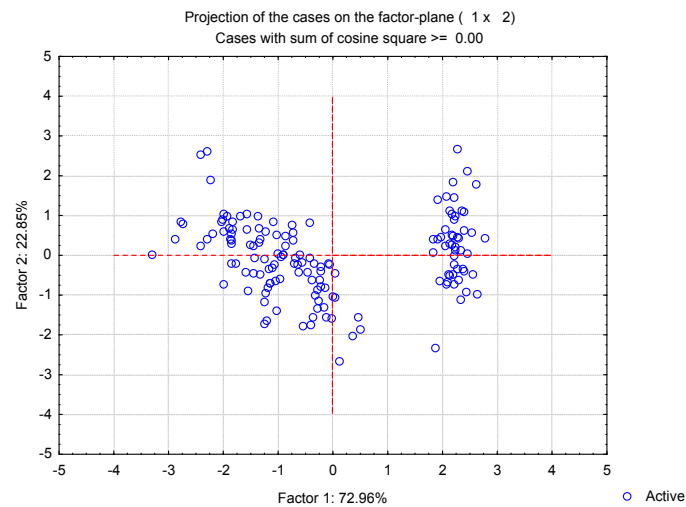
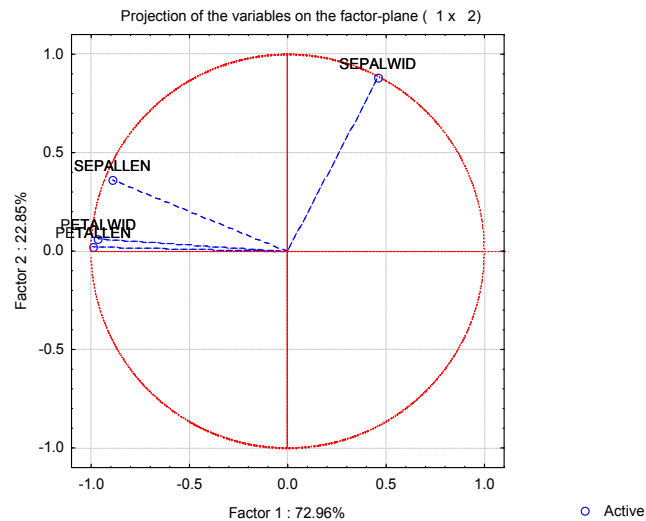
- Lze použít v případě proměnných o stejných jednotkách a podobném významu (např. rozměry objektu)
- Má smysl v případě, že chceme zohlednit absolutní hodnoty a rozsah proměnných
- Diagonála obsahuje hodnotu rozptylu proměnných
- Ostatní buňky obsahují kovarianci (= sdílený rozptyl) proměnných

Výstupy PCA

- Vlastní čísla (eigenvalues)
- Vlastní vektory (eigenvectors)
- Communalities
- Souřadnice objektů
- Scree plot
- Biplot

Eigenvectors of correlation matrix (Irisdat)				
Active variables only				
Variable	Factor 1	Factor 2	Factor 3	Factor 4
SEPALLEN	-0.521066	0.377418	0.719566	-0.261286
SEPALWID	0.269347	0.923296	-0.244382	0.123510
PETALLEN	-0.580413	0.024492	-0.142126	0.801449
PETALWID	-0.564857	0.066942	-0.634273	-0.523597

Eigenvalues of correlation matrix, and related statistics				
Active variables only				
Value number	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	2.918498	72.96245	2.918498	72.9624
2	0.914030	22.85076	3.832528	95.8132
3	0.146757	3.66892	3.979285	99.4821
4	0.020715	0.51787	4.000000	100.0000



Vlastní čísla (Eigenvalues)

Korelační matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.000	-0.118	0.872	0.818
SEPALWID	-0.118	1.000	-0.428	-0.366
PETALLEN	0.872	-0.428	1.000	0.963
PETALWID	0.818	-0.366	0.963	1.000



	Eigenvalue	% Rozptylu	Kumulativní eigenvalue	Kumulativní % rozptylu
1	2.918	73.0	2.918	73.0
2	0.914	22.9	3.833	95.8
3	0.147	3.7	3.979	99.5
4	0.021	0.5	4.000	100.0

- Spjaty s vytvářenými faktorovými osami
- Suma eigenvalues = počet proměnných (suma standardizovaných rozptylů)
- Hodnota eigenvalue je ve vztahu k variabilitě vztahu proměnných vyčerpané příslušnou faktorovou osou
- Hodnota eigenvalue = kolikrát více vyčerpává faktorová osa variability než by na ni připadalo rovnoměrným rozdělením (eigenvalue=1)

Kovarianční matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	0.686	-0.042	1.274	0.516
SEPALWID	-0.042	0.190	-0.330	-0.122
PETALLEN	1.274	-0.330	3.116	1.296
PETALWID	0.516	-0.122	1.296	0.581

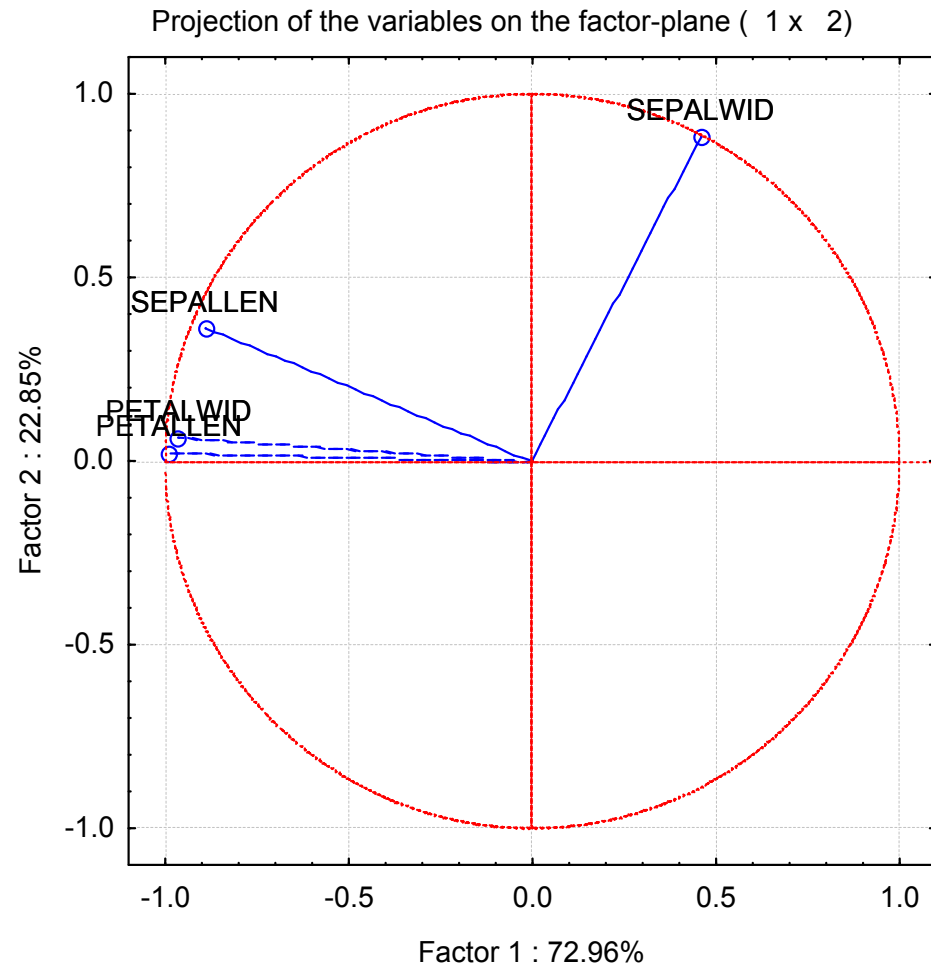


	Eigenvalue	% Rozptylu	Kumulativní eigenvalue	Kumulativní % rozptylu
1	4.228	92.5	4.228	92.5
2	0.243	5.3	4.471	97.8
3	0.078	1.7	4.549	99.5
4	0.024	0.5	4.573	100.0

- Spjaty s vytvářenými faktorovými osami
- Suma eigenvalues = suma rozptylu
- Velikost eigenvalue je ve vztahu k variabilitě vyčerpané příslušnou faktorovou osou
- Hodnota eigenvalue/průměrné eigenvalue = kolikrát více vyčerpává faktorová osa variability než by na ni připadalo rovnoměrným rozdělením

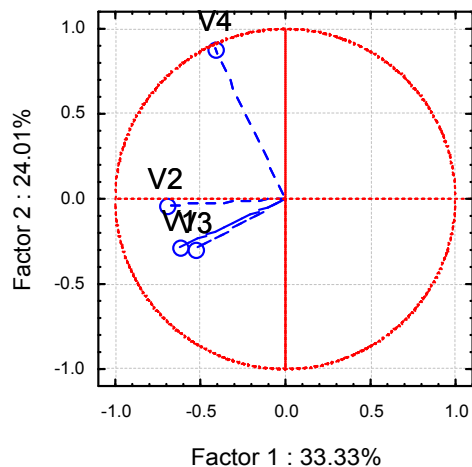
Interpretace vyčerpané variability faktorovými osami

- Variabilita vyčerpaná faktorovými osami je vztažena pouze k použitým proměnným
- Nevypovídá nic o proměnných nezahrnutých do analýzy !!!!
- Orientačně odpovídá počtu (nebo rozptylu) proměnných navázaných na příslušnou osu
- Souvisí i s počtem proměnných v analýze, čím více proměnných, tím spíše bude variabilita vyčerpaná první osou nižší (platí samozřejmě pouze v případě, že nejsou přidávány silně redundantní proměnné)
- V případě silně redundantních proměnných tyto redundantní proměnné zvyšují variabilitu vyčerpanou na příslušné faktorové ose, s níž jsou spjaty



Vyčerpaná variabilita a redundance proměnných

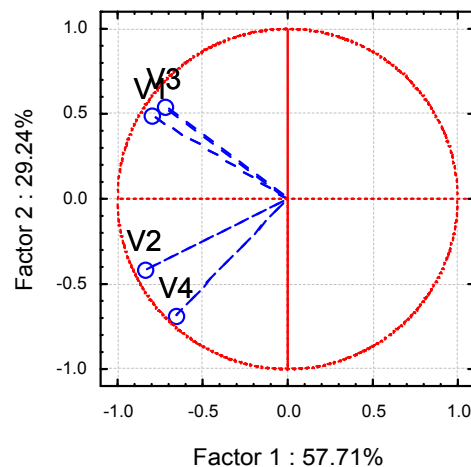
Příklad 1



	V1	V2	V3	V4
V1	1.00	0.19	0.10	0.05
V2	0.19	1.00	0.13	0.11
V3	0.10	0.13	1.00	0.05
V4	0.05	0.11	0.05	1.00

- Slabé korelace mezi proměnnými
- Vyčerpaná variabilita na první ose jen mírně převyšuje 1/4

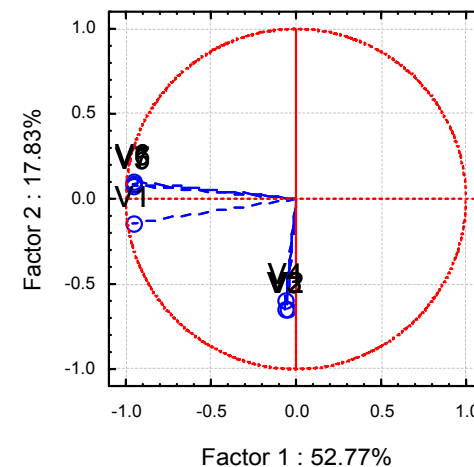
Příklad 2



	V1	V2	V3	V4
V1	1.00	0.52	0.71	0.14
V2	0.52	1.00	0.30	0.72
V3	0.71	0.30	1.00	0.20
V4	0.14	0.72	0.20	1.00

- Silné korelace mezi proměnnými
- Vyčerpaná variabilita na první ose představuje více než polovinu celkové variability

Příklad 3



	V1	V2	V3	V4	V5	V6	V7
V1	1.00	0.19	0.12	0.12	0.90	0.89	0.89
V2	0.19	1.00	0.12	0.09	-0.01	-0.01	-0.03
V3	0.12	0.12	1.00	0.12	0.02	0.02	0.02
V4	0.12	0.09	0.12	1.00	0.02	-0.01	0.03
V5	0.90	-0.01	0.02	0.02	1.00	0.90	0.90
V6	0.89	-0.01	0.02	-0.01	0.90	1.00	0.90
V7	0.89	-0.03	0.02	0.03	0.90	0.90	1.00

- K příkladu 1 přidány proměnné redundantní k V1
- Výsledek PCA se kompletně mění, první osa vyčerpává přes polovinu variability díky redundantním proměnným

Vlastní vektory

Korelační matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.000	-0.118	0.872	0.818
SEPALWID	-0.118	1.000	-0.428	-0.366
PETALLEN	0.872	-0.428	1.000	0.963
PETALWID	0.818	-0.366	0.963	1.000



Standardizace na délku 1

	Factor 1	Factor 2	Factor 3	Factor 4
SEPALLEN	-0.521	0.377	0.720	-0.261
SEPALWID	0.269	0.923	-0.244	0.124
PETALLEN	-0.580	0.024	-0.142	0.801
PETALWID	-0.565	0.067	-0.634	-0.524

Kovarianční matice

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	0.686	-0.042	1.274	0.516
SEPALWID	-0.042	0.190	-0.330	-0.122
PETALLEN	1.274	-0.330	3.116	1.296
PETALWID	0.516	-0.122	1.296	0.581



	Factor 1	Factor 2	Factor 3	Factor 4
SEPALLEN	-0.361	0.657	0.582	-0.315
SEPALWID	0.085	0.730	-0.598	0.320
PETALLEN	-0.857	-0.173	-0.076	0.480
PETALWID	-0.358	-0.075	-0.546	-0.754

Standardizace na délku druhé odmocniny eigenvalue (směrodatná odchylka)

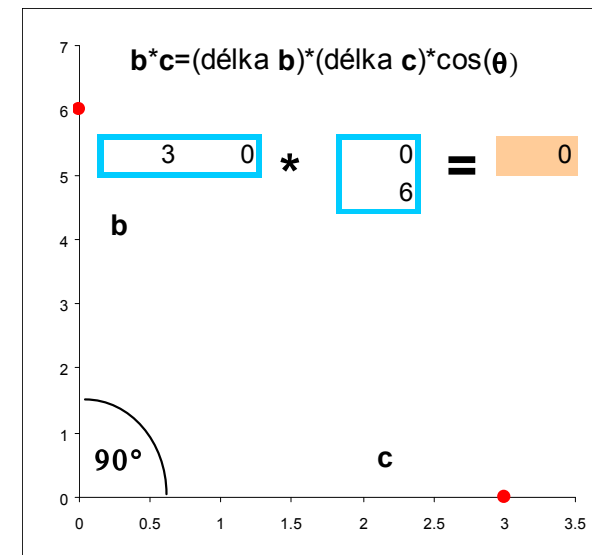
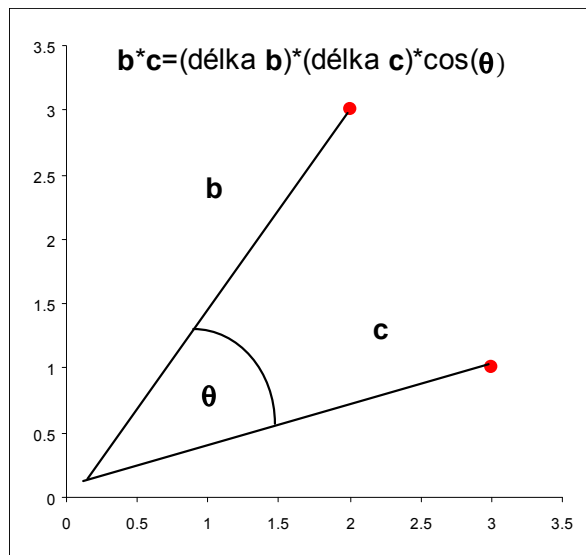
	Factor 1	Factor 2	Factor 3	Factor 4
SEPALLEN	-0.890	0.361	0.276	-0.038
SEPALWID	0.460	0.883	-0.094	0.018
PETALLEN	-0.992	0.023	-0.054	0.115
PETALWID	-0.965	0.064	-0.243	-0.075

	Factor 1	Factor 2	Factor 3	Factor 4
SEPALLEN	-0.743	0.323	0.163	-0.049
SEPALWID	0.174	0.360	-0.167	0.049
PETALLEN	-1.762	-0.085	-0.021	0.074
PETALWID	-0.737	-0.037	-0.153	-0.116

- Vlastní vektory popisují směr kterým v prostoru původních proměnných směřují faktorové osy
- Eigenvektory mohou být různým způsobem standardizovány a vizualizovány; interpretace výstupů (tzv. biplotů) se liší podle použité standardizace

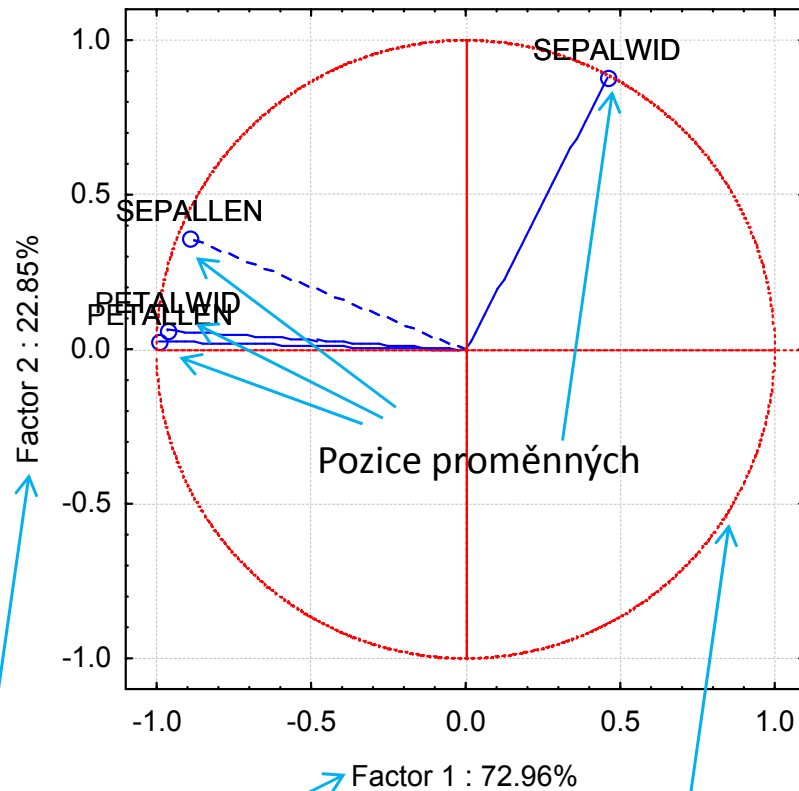
Vlastnosti vlastních vektorů

- Vlastní vektory jsou navzájem ortogonální (nezávislé, svírající úhel 90°)
- Z hlediska interpretace definují nezávislé proměnné, tedy nesoucí **zcela unikátní informaci** o objektech
- Definují směr nových faktorových os v prostoru původních proměnných a umožňují počítat pozici objektů na nových faktorových osách
- **Geometrie součinu vektorů** - Součin vektorů lze spočítat jako součin jejich délek násobený cosinem úhlu, který svírají. Pokud 2 vektory svírají pravý úhel je jejich součin 0 a nazývají se **orthogonální vektory**. Matice, jejíž sloupcové vektory navzájem svírají pravý úhel se nazývá **orthogonální matice**.



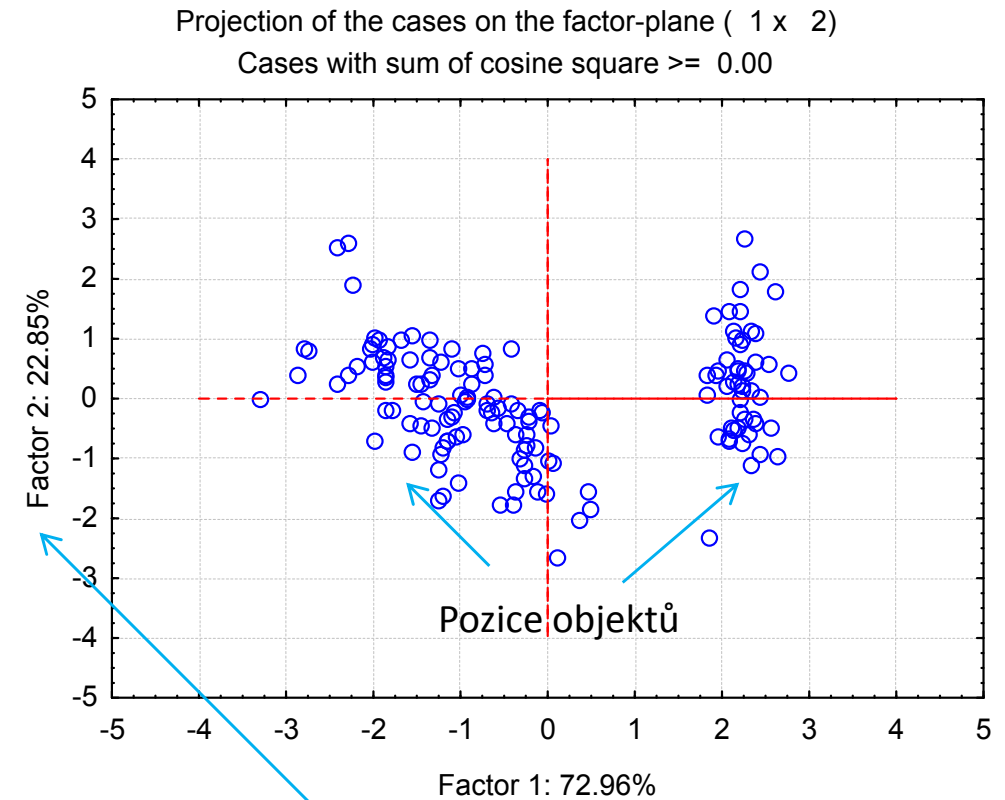
Biplot

- Biplot – současná vizualizace pozice proměnných a objektů
- Několik typů biplotů s různou interpretací
- Pro zjednodušení interpretace je možné hodnoty na osách násobit konstantou



Variabilita vyčerpaná faktorovými osami

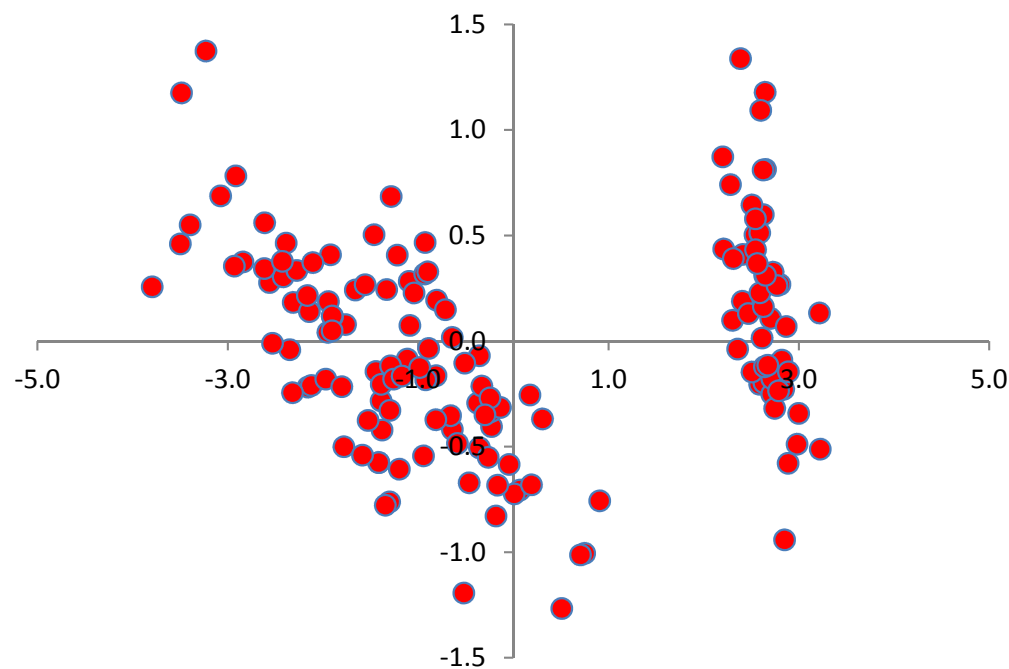
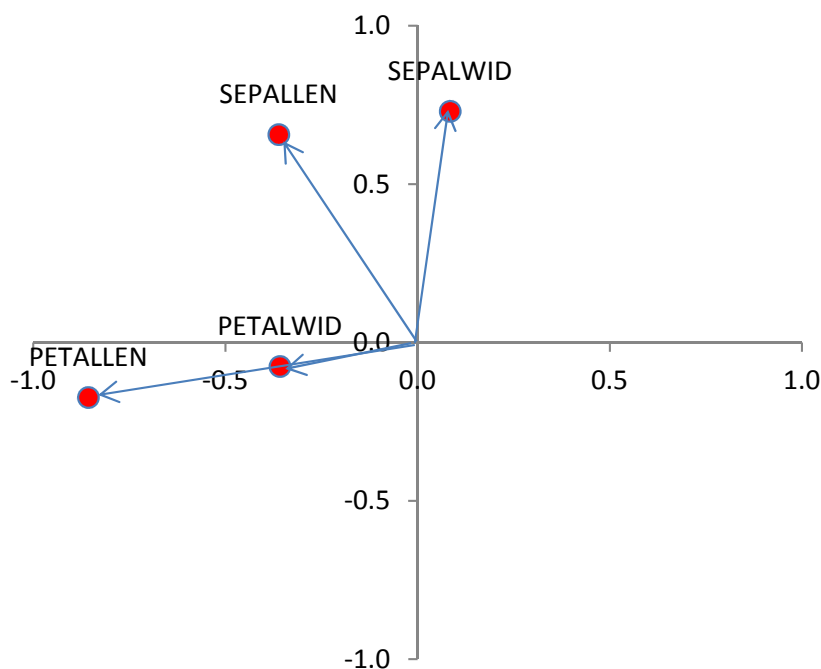
Jednotková kružnice -
Hranice příspěvku k
definici faktorové osy



Variabilita vyčerpaná faktorovými osami

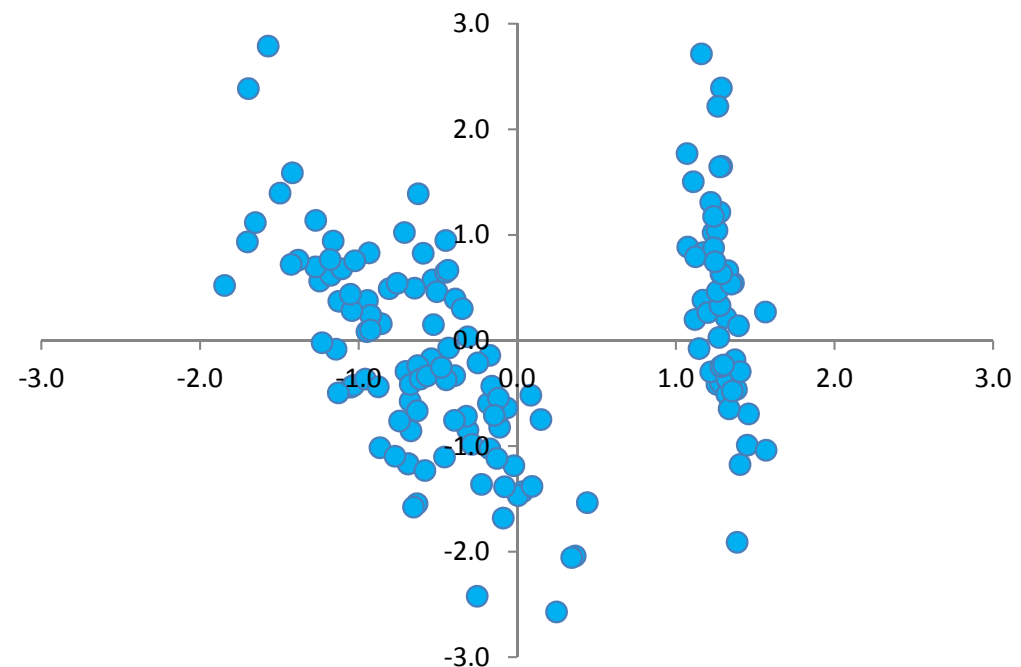
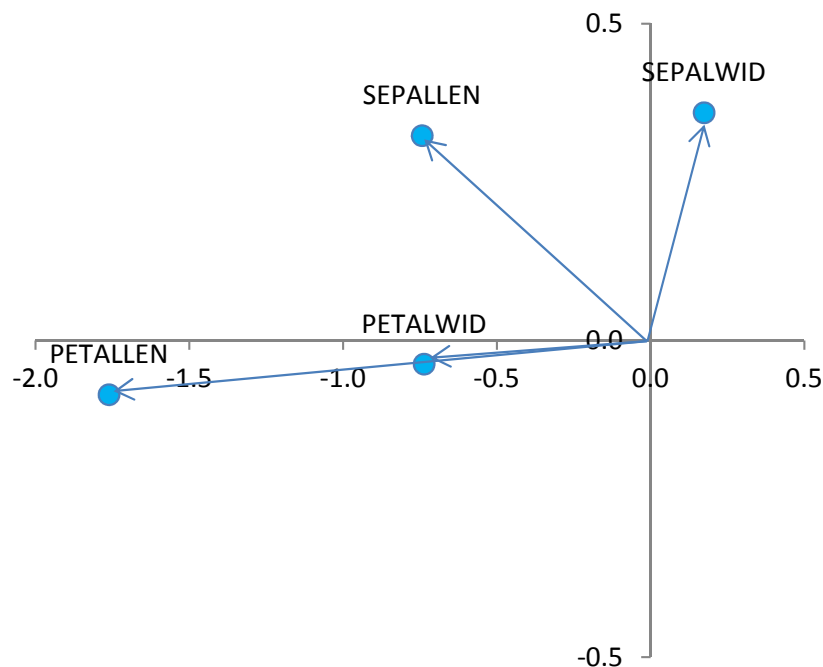
Standardizace eigenvektorů a její interpretace I

- Standardizace délky eigenvektorů na jednotkovou délku
 - Při vizualizaci vede na tzv. Biplot vzdáleností (distance biplot)
 - Pozice objektů na faktorových osách mají rozptyl= příslušné eigenvalue
 - Interpretace biplotu
 - Umožňuje interpretovat euklidovské vzdálenosti objektů v prostoru PCA (jsou aproximací euklidovských vzdáleností v původním prostoru)
 - Projekce objektu v pravém uhlu na původní proměnnou aproximuje pozici objektu na této původní proměnné
 - Délka projekce jednotlivých původních proměnných v prostoru faktorových os popisuje jejich příspěvek k definici daného faktorového prostoru
 - Úhly mezi původními proměnnými ve faktorovém prostoru nemají žádnou interpretaci



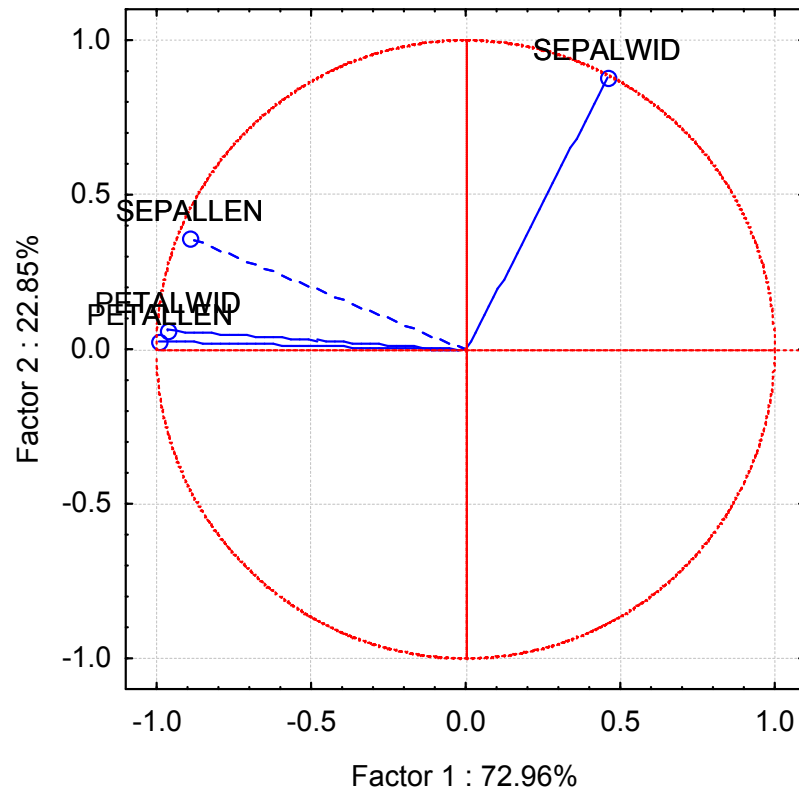
Standardizace eigenvektorů a její interpretace II

- Standardizace délky eigenvektorů na druhou odmocninu z eigenvalue
 - Při vizualizaci vede na tzv. Biplot korelací (correlation biplot)
 - Pozice objektů na faktorových osách mají jednotkový rozptyl
 - Interpretace biplotu
 - euklidovské vzdálenosti objektů v prostoru PCA nejsou aproximací euklidovských vzdáleností v původním prostoru
 - Projekce objektu v pravém uhlu na původní proměnnou aproximuje pozici objektu na této původní proměnné
 - Délka projekce jednotlivých původních proměnných v prostoru faktorových os popisuje jejich směrodatnou odchylku
 - Úhly mezi původními proměnnými ve faktorovém prostoru souvisí s jejich korelací
 - Není vhodný pokud má smysl interpretovat vzdálenosti (vzájemné vztahy) mezi objekty

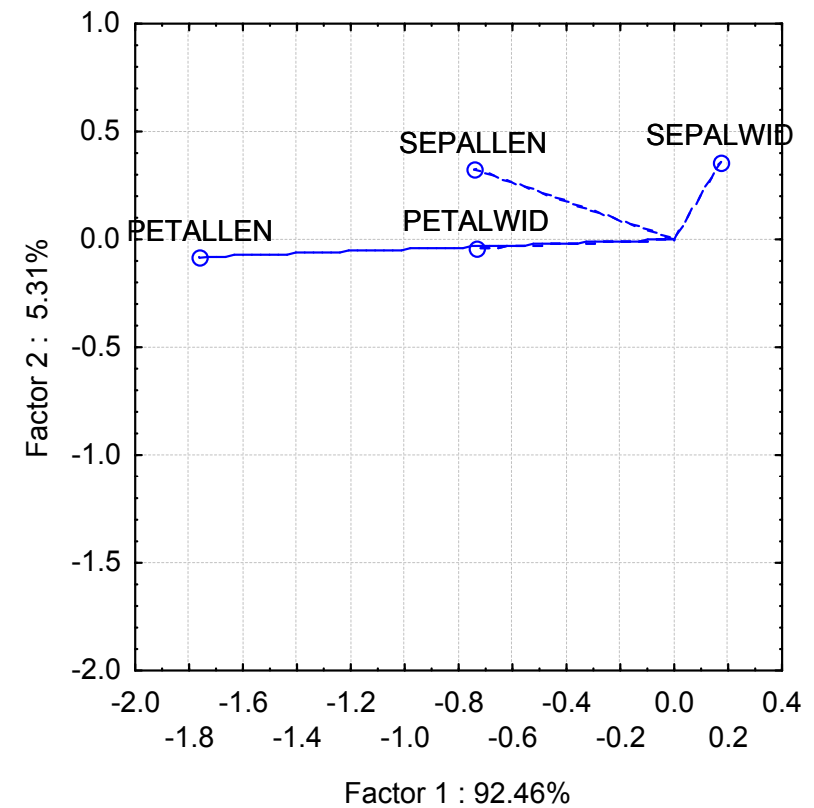


Correlation biplot

Korelační matice

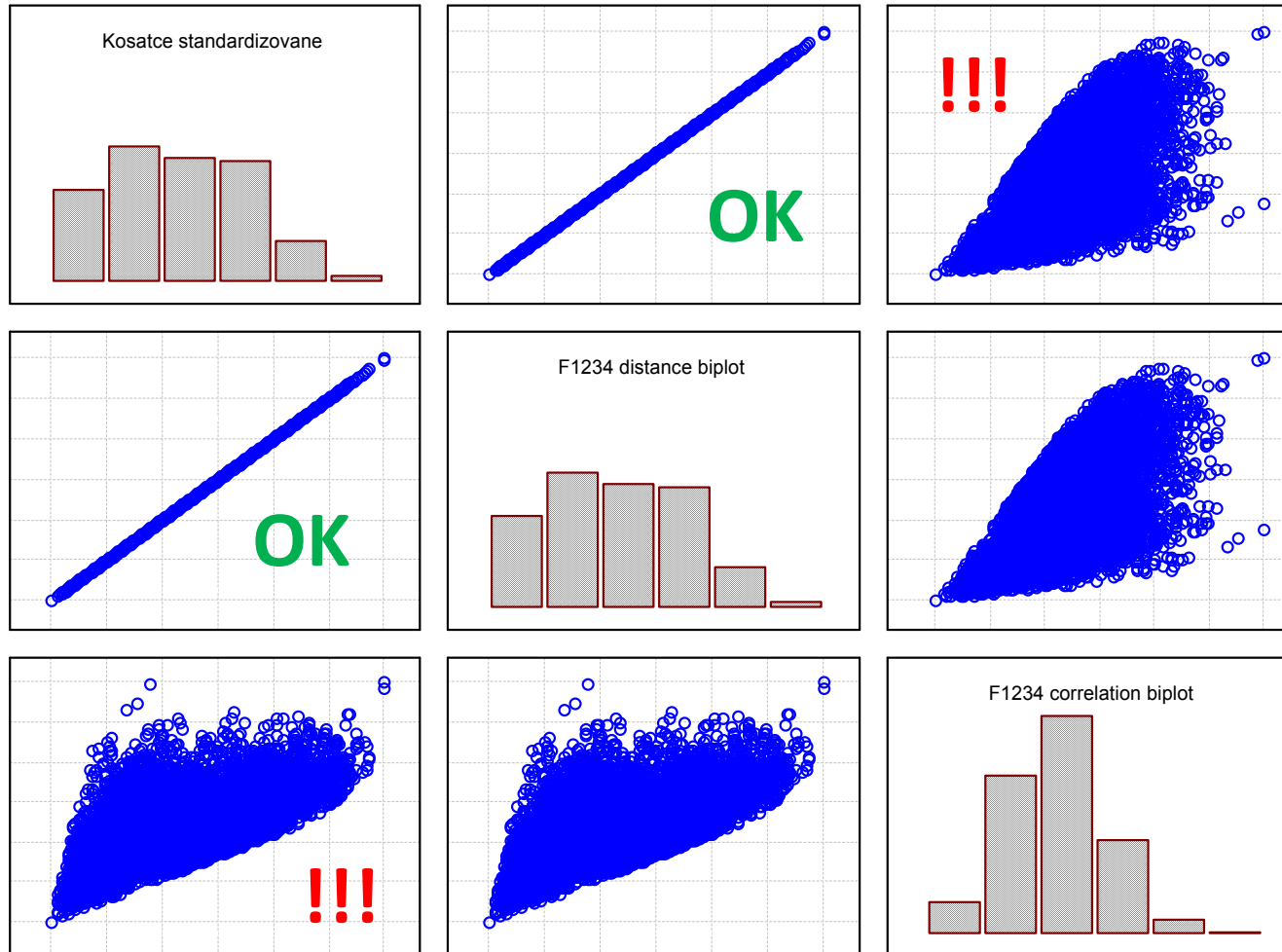


Kovarianční matice



Zachování vzdáleností objektů v původním prostoru vzhledem k různým typům biplotu

- Pouze distance biplot zachovává vzdálenostní vztahy mezi objekty, v případě korelačního biplotu není možná interpretace těchto vzdáleností



Standardizace eigenvektorů a její vliv na projekci původních proměnných: shrnutí

	Kovarianční matice		Korelační matice	
Původní proměnná (centrovaná)	Standardizace eigenvektoru			
	$\sqrt{\lambda_k}$	1	$\sqrt{\lambda_k}$	1
Celková délka	s_j	1	1	1
Úhly proměnných v redukovaném prostoru	Projekce kovariancí (korelací)	90° rotace systému os	Projekce korelací	90° rotace systému os
Hranice příspěvku k definici faktorové osy	$s_j \sqrt{d/p}$	$\sqrt{d/p}$	$\sqrt{d/p}$	$\sqrt{d/p}$
Projekce na faktorovou osu k	$u_{jk} \sqrt{\lambda_k}$ Kovariance s k	u_{jk} Proporcionální kovarianci s k	$u_{jk} \sqrt{\lambda_k}$ Korelace s k	u_{jk} Proporcionální korelaci s k
Korelace s faktorovou osou k	$\frac{u_{jk} \sqrt{\lambda_k}}{s_j}$	$\frac{u_{jk} \sqrt{\lambda_k}}{s_j}$	$u_{jk} \sqrt{\lambda_k}$	$u_{jk} \sqrt{\lambda_k}$

λ_k Eigenvalue faktorové osy k

d Počet původních proměnných

u_{jk} Hodnota eigenvektoru faktorové osy k pro původní proměnnou j

s_j Směrodatná odchylka původní proměnné j

p Počet faktorových os

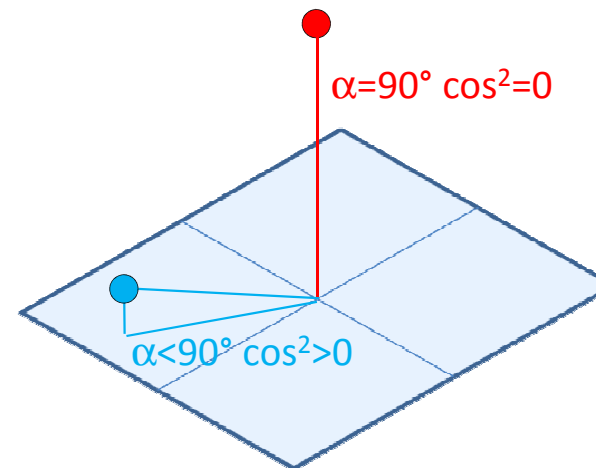
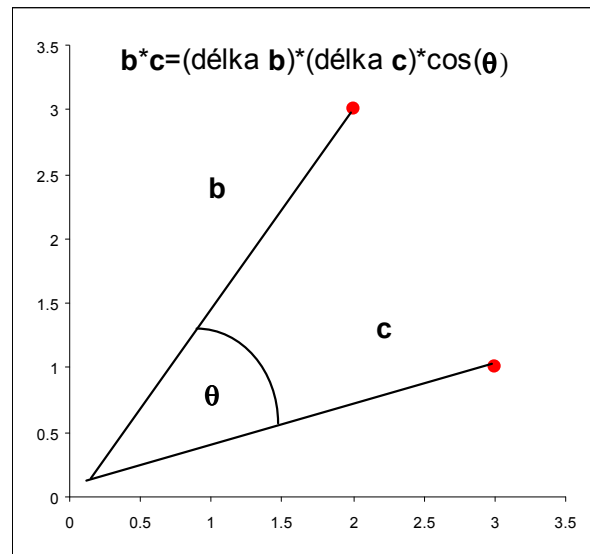
Communalities

- Jde o podíl variability sdílené s jinými proměnnými, zde s postupně se zvyšujícím počtem faktorových os

	From 1	From 2	From 3	From 4
SEPALLEN	0.792	0.923	0.999	1.000
SEPALWID	0.212	0.991	1.000	1.000
PETALLEN	0.983	0.984	0.987	1.000
PETALWID	0.931	0.935	0.994	1.000

Cosinus²

- Souvisí s geometrickým významem cosinu při násobení vektorů, kdy $\cos=0$ znamená ortogonální vztah vektorů
- V PCA se používá jako filtr pro zobrazení objektů v biplotu, kdy objekty s $\cos^2 \sim 0$ jsou umístěny kolmo k rovině definované vybranými faktorovými osami a tedy nejsou v tomto pohledu interpretovatelné

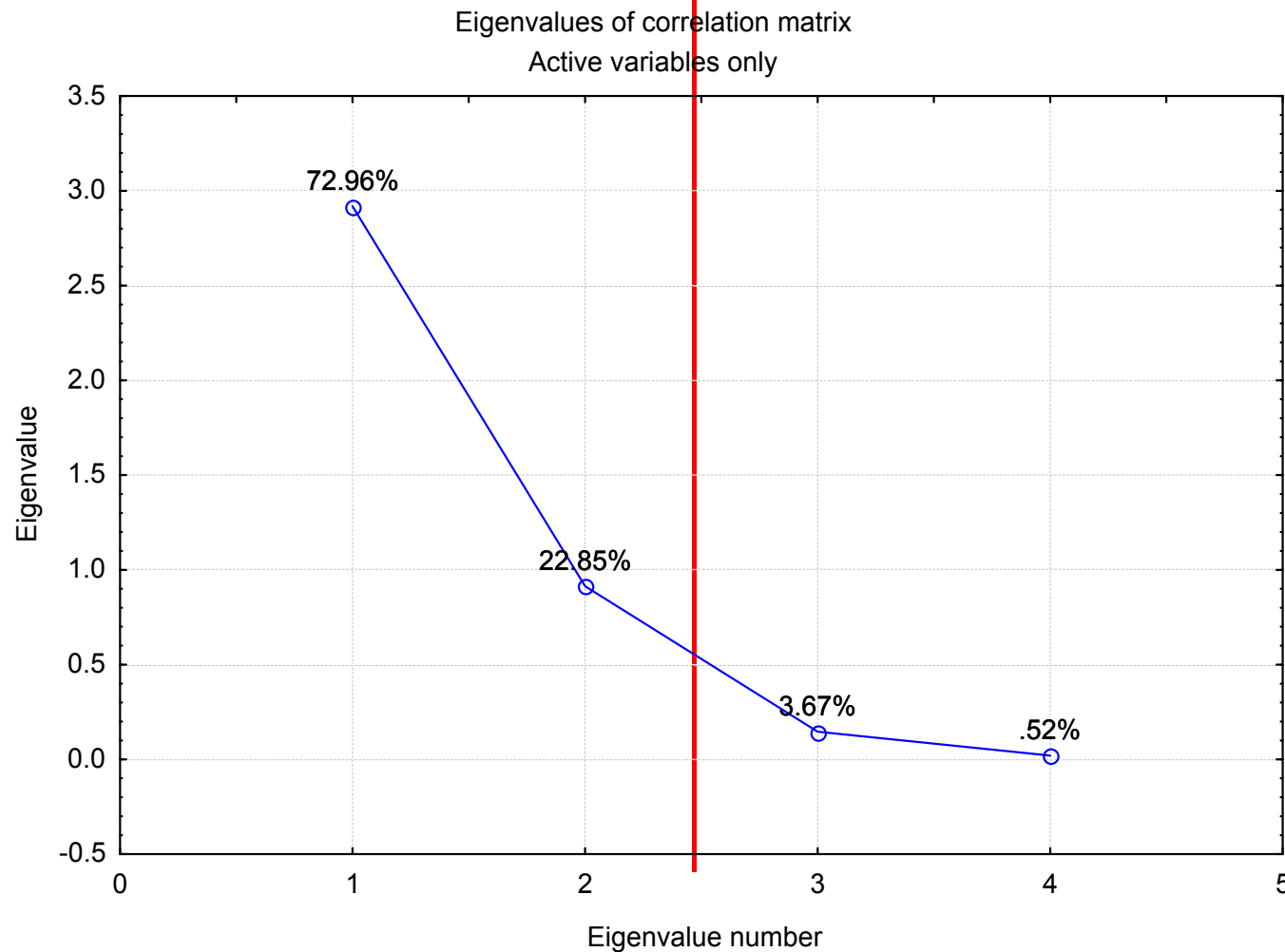


Identifikace optimálního počtu faktorových os pro další analýzu

- Jedním z cílů ordinační analýzy je výběr menšího počtu dimenzí pro další analýzu
- Řada pravidel pro výběr optimálního počtu dimenzí, optimální je samozřejmě skončit s výběrem dvou, maximálně tří dimenzí (s výjimkou speciálních aplikací typu analýzy obrazů MRI, kde je úspěchem redukce z milionu dimenzi na desítky)
- Kaiser Guttmanovo kritérium:
 - Pro další analýzu jsou vybrány osy s vlastním číslem >1 (korelace) nebo větším než je průměrné eigenvalue (kovariance)
 - Logika je vybírat osy, které přispívají k vysvětlení variability dat více než připadá rovnoměrným rozdělením variability
- Scree plot
 - Grafický nástroj hledající zlom ve vztahu počtu os a vyčerpané variability
- Sheppard diagram
 - Grafická analýza vztahu mezi vzdálenostmi objektů v původním prostoru a redukovaném prostoru o daném počtu dimenzí

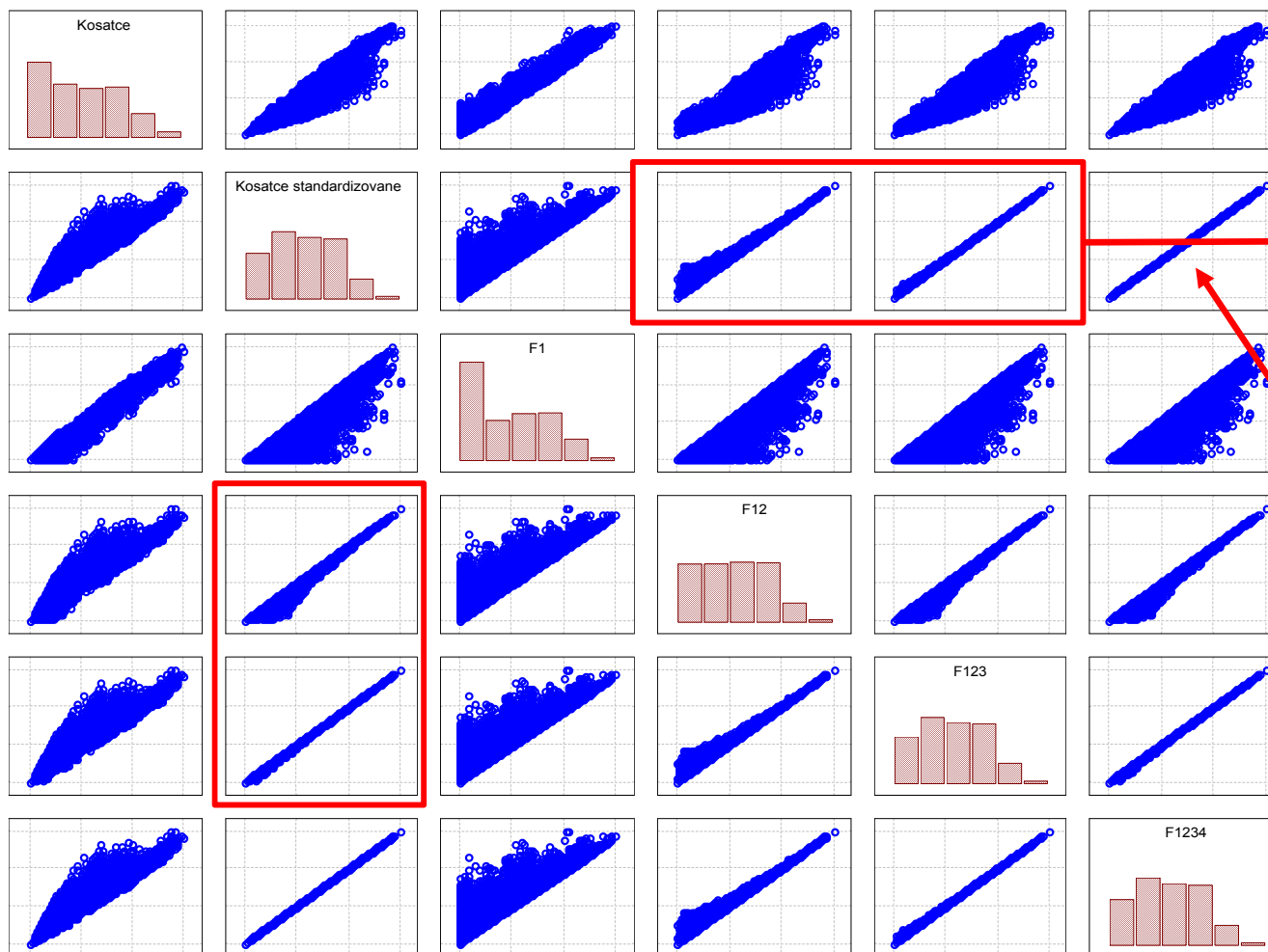
Scree plot

Zlom ve vztahu mezi počtem eigenvalue a jimy vyčepanou variabilitou – pro další analýzu použity první dvě faktorové osy



Sheppard diagram

- Vztahuje vzdálenosti v prostoru původních proměnných ke vzdálenostem v prostoru vytvořeném PCA
- Je třeba brát ohled na typ PCA (korelace vs. kovariance)
- Obecná metoda určení optimálního počtu dimenzí v ordinační analýze (třeba respektovat použitou asociační metriku)



Za optimální z hlediska zachování vzdáleností objektů lze považovat dvě nebo tři dimenze

Při použití všech dimenzí jsou vzdálenosti perfektně zachovány

Shrnutí

- Analýza hlavních komponent je základním nástrojem pro analýzu variability spojitých proměnných a jejich vztahů
- Kromě spojitých proměnných mohou být vstupem i binární proměnné (popřípadě kategoriální data ve formě tzv. dummies), ale je třeba mít na paměti jednak omezení vyplývající z double zero problému, jednak omezení týkající se poměru počtu proměnných a objektů
- Při výpočtu je nezbytné mít na paměti omezení výpočtu vyplývající z předpokladů analýzy korelací a kovariancí
- Analýza hlavních komponent může být počítána za různým účelem, tomu je třeba přizpůsobit výběr použitého algoritmu a výběr výstupů pro další interpretaci
- Při interpretaci výstupů analýzy hlavních komponent je třeba zvažovat
 - Použitý algoritmus a jeho implementace v použitém SW
 - Typ výstupu PCA a omezení jeho interpretace (standardizace eigenvektorů, typy biplotů apod.)
 - Praktická interpretace výstupů a vliv artefaktů dat (redundantní proměnné, několik metod měření jednoho parametru apod.)

Vícerozměrné statistické metody

Faktorová analýza

Faktorová analýza

- Faktorová analýza se snaží vysvětlit strukturu dat pomocí tzv. společných faktorů vysvětlujících sadu původních proměnných
- Čím se principiálně liší od analýzy hlavních komponent?
 - Analýza hlavních komponent – vysvětlení maxima variability v datech
 - Faktorová analýza – vysvětlení maxima kovariance mezi popisnými proměnnými
- Čím se prakticky liší od analýzy hlavních komponent?
 - Hlavním praktickým rozdílem je rotace proměnných tak aby se vytvořené faktorové osy daly dobře interpretovat
 - Výhodou je lepší interpretace vztahu původních proměnných
 - Nevýhodou je prostor pro subjektivní názor analytika daný výběrem rotace
- Typy faktorové analýzy
 - Vysvětlující (Explanatory) – snaží se identifikovat minimální počet faktorů pro vysvětlení dat
 - Potvrzující (Confirmatory) – testuje hypotézy ohledně skryté struktury v datech

Společné faktory a základní možné rotace

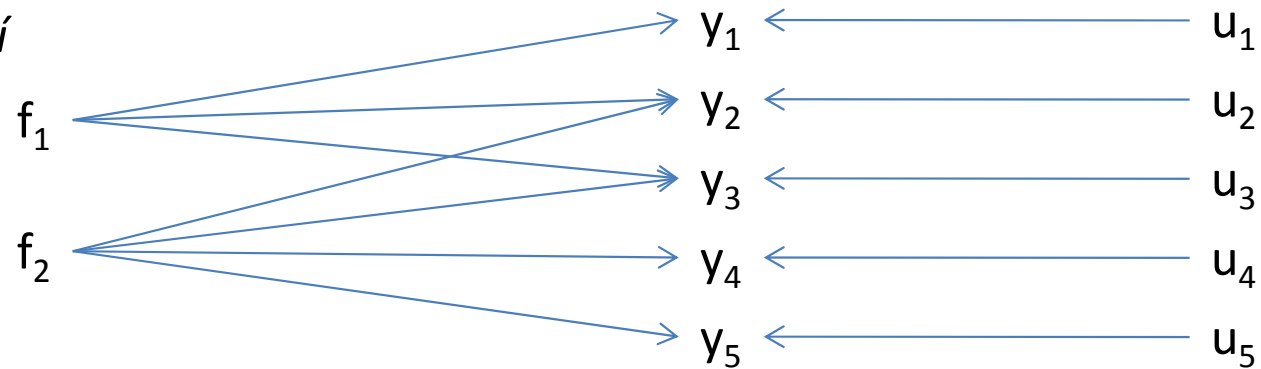
Společný faktor

Pozorovaná proměnná

Unikátní faktor

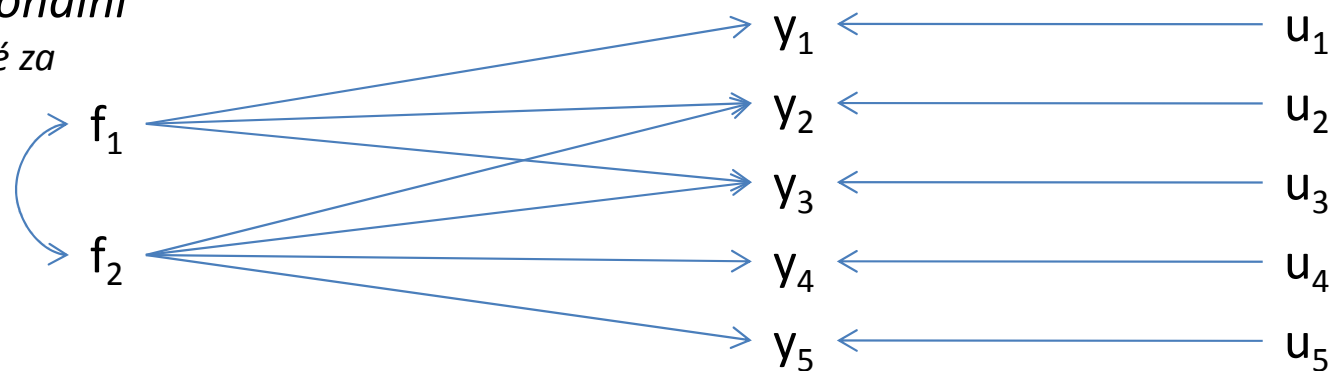
Rotace ortogonální

- Nezávislé faktory



Rotace neortogonální

- Faktory jsou závislé za účelem zvýšení interpretovatelnosti



Faktorová analýza – postup výpočtu

1. Extrakce prvotních faktorů z kovarianční matice (analogie eigenvektorů v PCA)

- Oproti PCA pracuje pouze s částí variability každé proměnné (tzv. communality), která je sdílena společnými faktory
- Několik možných algoritmů – principiálně factoring, metoda nejmenších čtverců, maximum likelihood apod.
- Výsledkem je komplexní struktura faktorů (obdobná PCA), kde řada faktorů má významné loadings (~ vztah) k původním proměnným, počet takových faktorů je tzv. komplexita faktorů.

2. V druhém kroku je rotací dosaženo zjednodušení struktury faktorů, tj. vztah mezi společnými faktory a původními proměnnými je zjednodušen (každá původní proměnná má hlavní vztah s jedním faktorem nebo malým počtem faktorů)

- Dva hlavní typy rotace
 - Ortogonální – faktory nemohou být korelovány, jsou tedy zcela nezávislé
 - Neortogonální - faktory mohou být korelovány, nejsou tedy zcela nezávislé; vzhledem ke korelacím obtížnější interpretace

Faktorová analýza - rotace

- Ortogonální rotace
 - Quartimax – minimalizuje sumu čtverců loadings původních proměnných na faktorových osách, tedy zjednodušuje řádky matice loadings (=každá původní proměnná má největší loadings na jedné faktorové ose)
 - Varimax – zjednodušuje sloupce matice loadings
 - Equimax – zjednodušuje řádky i sloupce matice loadings
 - Biquartimax – varianta equimax
- Neortogonální rotace
 - Oblimax
 - Quartimin
 - Oblimin
 - Covarimin
 - Biquartimin
 - Atd.

Vícerozměrné statistické metody

Korespondenční analýza

Korespondenční analýza

- Vstupní data:
 - Tabulka obsahující souhrny proměnných (počty, průměry) za skupiny respondentů
- Výstupy analýzy
 - Vztahy všech původních faktorů a/nebo skupin respondentů v jednoduchém xy grafu
- Kritické problémy analýzy
 - Skupiny s malým počtem hodnot mohou být zatíženy značným šumem a náhodnou chybou
 - Obtížná interpretace velkého množství malých skupin respondentů
- Výpočet probíhá prostřednictvím singular value decomposition na matici chi-square vzdáleností (tedy na matici příspěvků buněk tabulky k celkovému chi-square obdobně jako v klasickém testu dobré shody na kontingenční tabulce)

Analýza kontingenčních tabule jako princip výpočtu vícerozměrných analýz

- Abundance taxonů (nebo počet jakýchkoliv objektů) na lokalitách lze brát jako kontingenční tabulku a mírou vztahu mezi řádky (lokality) a sloupci (taxony) je velikost chi-kvadrátu

$$\chi^2_{(1)} = \frac{\left[\begin{array}{c} \text{pozorovaná} \\ \text{četnost} \end{array} - \begin{array}{c} \text{očekávaná} \\ \text{četnost} \end{array} \right]^2}{\text{očekávaná četnost}}$$

Počítáno pro každou buňku tabulky

	☠	😊
A	10	0
B	0	10

Pozorovaná tabulka

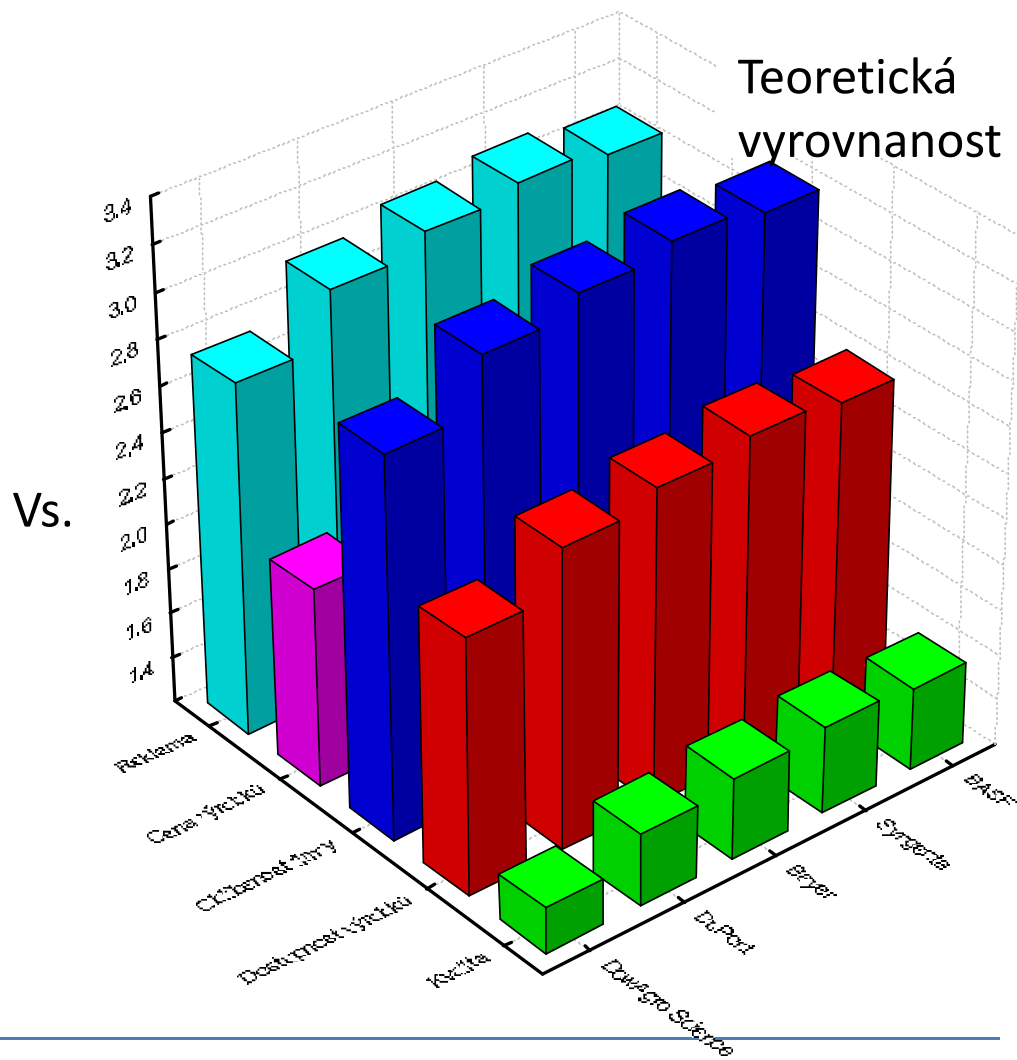
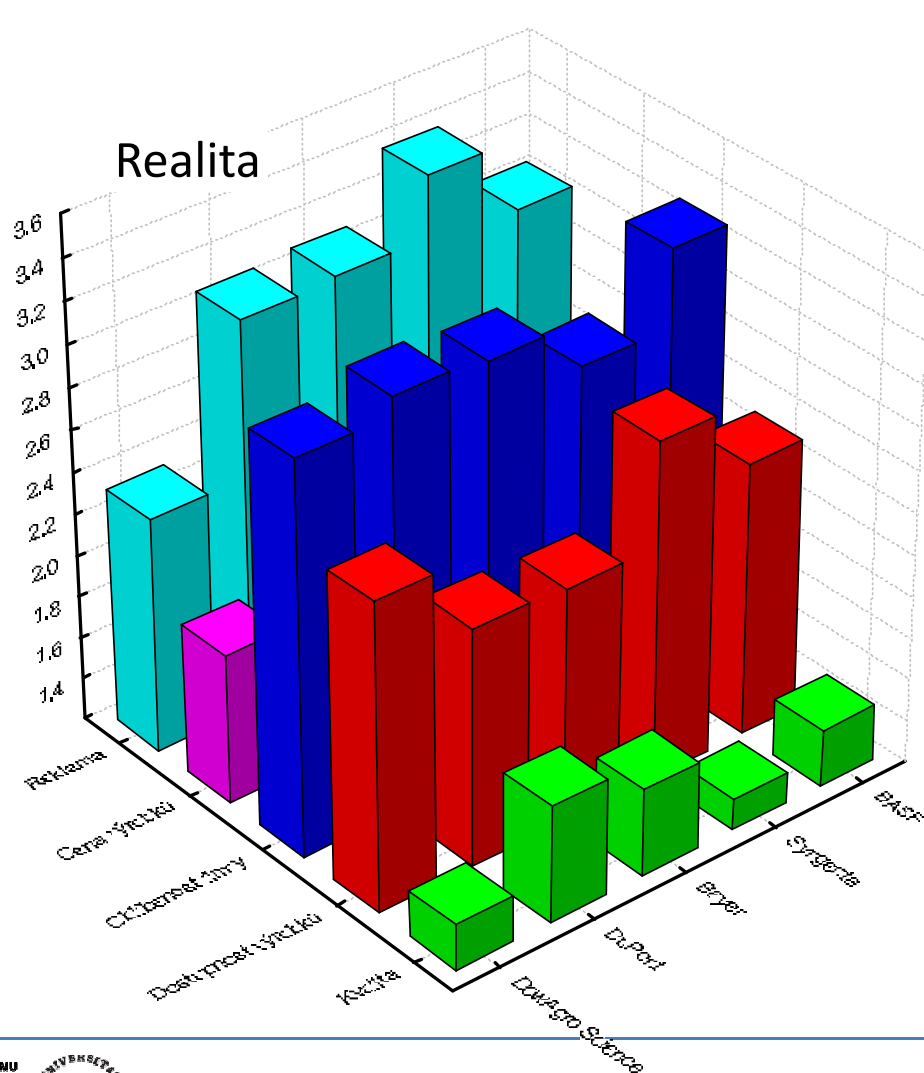
	☠	😊
A	5	5
B	5	5

Očekávaná tabulka

Hodnota chi-kvadrátu definuje míru odchylky dané buňky (v našem kontextu vztahu taxon-lokalita) od situace, kdy mezi řádky a sloupci (taxon-lokalita) není žádný vztah

Princip korespondenční analýzy

- Korespondenční analýza hledá, které kombinace řádků a sloupců hodnocené tabulky nejvíce přispívají k její variabilitě



Výstupy korespondenční analýzy

STATISTICA - [Data: mark_pruzkum* (5v by 5c)]

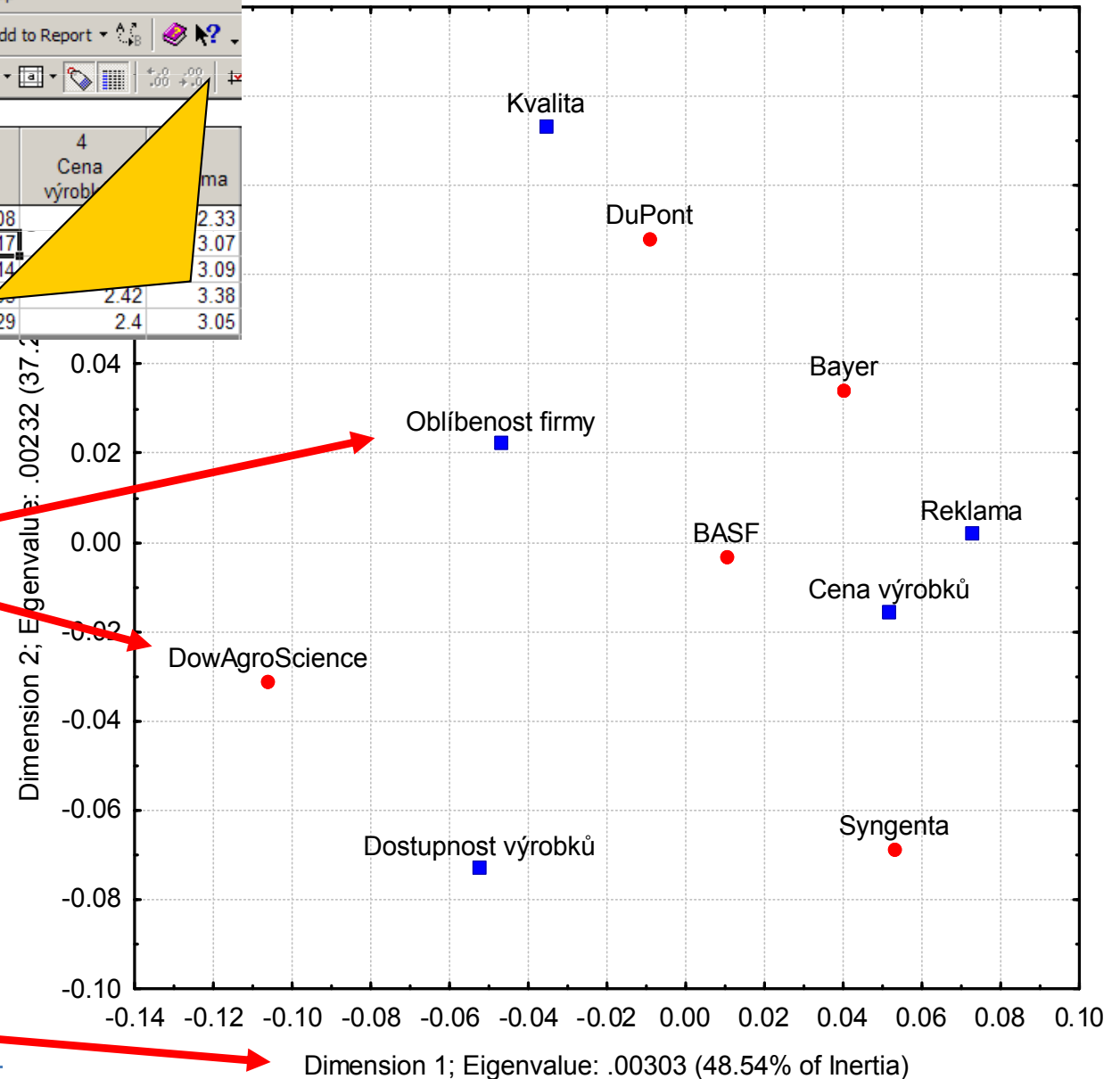
File Edit View Insert Format Statistics Graphs Tools Data Window Help

Arial 10 B I U

	1 Kvalita	2 Dostupnost výrobků	3 Oblíbenost firmy	4 Cena výrobků	5 Reklama
DowAgro Science	1.42	2.67	3.08	2.42	2.33
DuPont	1.76	2.34	3.17	2.42	3.07
Bayer	1.62	2.32	3.14	2.42	3.09
Syngenta	1.35	2.81	2.99	2.42	3.38
BASF	1.47	2.51	3.29	2.4	3.05

Vzájemná pozice faktorů a skupin respondentů: vzájemnou pozici lze interpretovat

Variabilita vyčerpaná danou faktorovou osou



Vícerozměrné statistické metody

Multidimensional scaling (Nemetrické vícerozměrné škálování)

Multidimensional scaling

- Jde o iterační algoritmus řešící převod libovolné asociační matice do Euklidovského prostoru (různé SW tak mohou dosahovat mírně odlišné výsledky)
- Cílem je dosáhnout řešení, které při nejmenším počtu vytvořených os zachovává pořadí vzdáleností objektů v původní asociační matici
- Vstupem analýzy je libovolná asociační matice (včetně nemetrických koeficientů)
- Výstupem je zadaný počet „faktorových os“

Multidimensional scaling: Příklad

- Data vzdáleností evropských měst - > rekonstrukce mapy

STATISTICA - [Data: mesta_vzdalenosti (21v by 24c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help

Arial 10 B I U

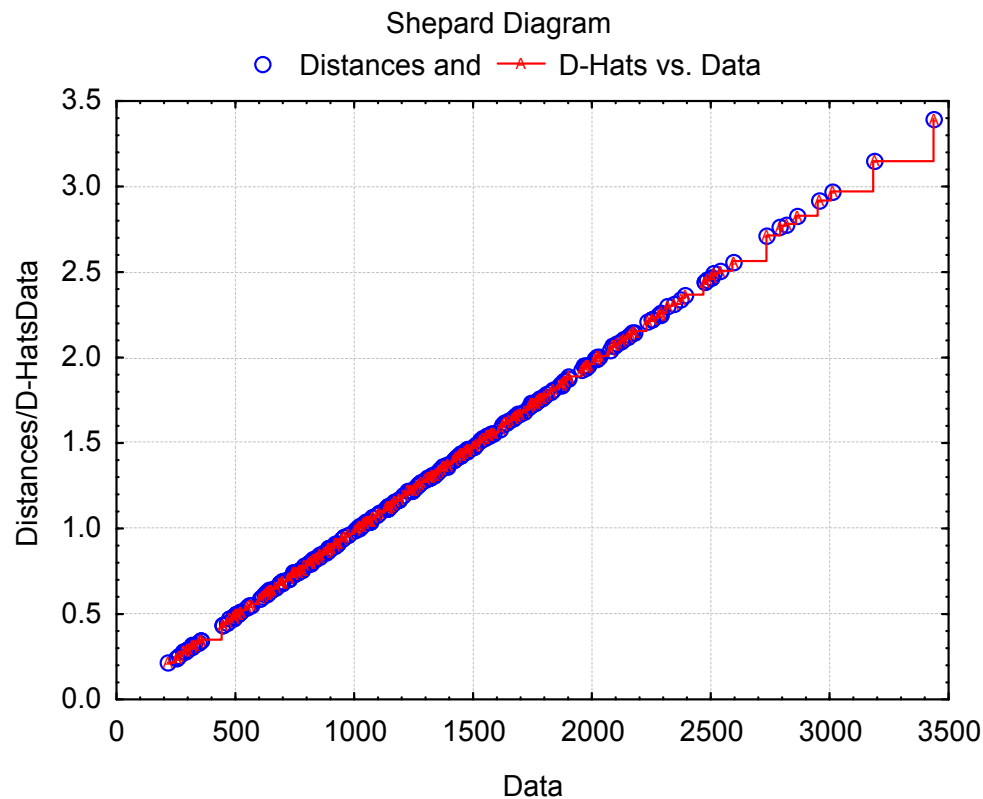
C:\Users\Jarkovsky\Desktop\FSTA\mesta_vzdalenosti.xlsx : Sheet1

	1	2	3	4	5	6	7	8	9	10
	Barcelon	Bělehrad	Berlín	Brusel	Bukurešť	Budapešť	Kodaň	Dublin	Hamburg	Istanbul
Barcelona	0	1528	1497	1062	1968	1498	1757	1469	1471	2230
Bělehrad	1528	0	999	1372	447	316	1327	2145	1229	809
Berlín	1497	999	0	651	1293	689	354	1315	254	1735
Brusel	1062	1372	651	0	1769	1131	766	773	489	2178
Bukurešť	1968	447	1293	1769	0	639	1571	2534	1544	445
Budapešť	1498	316	689	1131	639	0	1011	1894	927	1064
Kodaň	1757	1327	354	766	1571	1011	0	1238	287	2017
Dublin	1469	2145	1315	773	2534	1894	1238	0	1073	2950
Hamburg	1471	1229	254	489	1544	927	287	1073	0	1983
Istanbul	2230	809	1735	2178	445	1064	2017	2950	1983	0
Kiev	2391	976	1204	1836	744	894	1326	2513	1440	1052
Londýn	1137	1688	929	318	2088	1450	955	462	720	2496
Madrid	504	2026	1867	1314	2469	1975	2071	1449	1785	2734
Miláno	725	885	840	696	1331	788	1157	1413	900	1669
Moskva	3006	1710	1607	2253	1497	1565	1558	2792	1779	1753
Mnichov	1054	773	501	601	1186	563	838	1374	610	1582
Paříž	831	1445	876	261	1869	1247	1025	776	744	2253
Praha	1353	738	280	721	1076	443	633	1465	492	1507
Řím	856	721	1181	1171	1137	811	1529	1882	1307	1373
Saint Petersburg	2813	1797	1319	1903	1740	1556	1143	2314	1414	2099
Sofia	1745	329	1318	1697	296	629	1635	2471	1554	502
Stockholm	2276	1620	810	1280	1742	1316	521	1626	809	2171
Vídeň	1347	489	523	914	855	216	868	1680	742	1273
Varšava	1862	826	516	1159	946	545	667	1823	750	1386

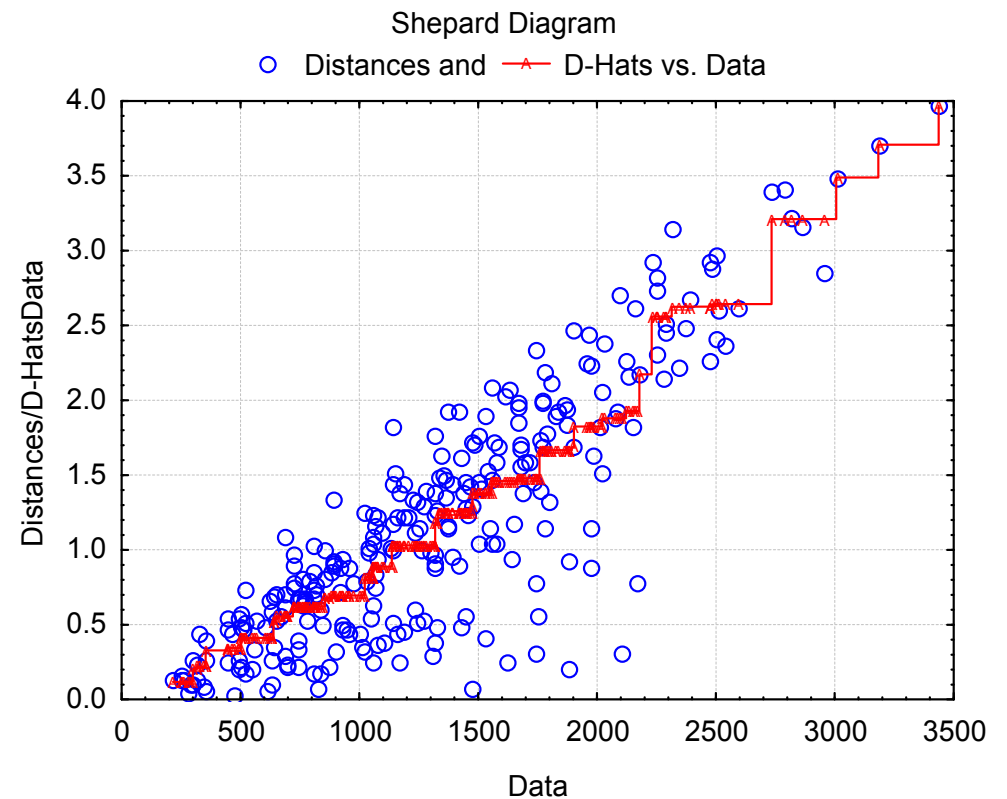
Multidimensional scaling: Příklad

- Kvalita dodržení pořadí vzdáleností v datech při daném počtu os je kontrolována Shepardovým diagramem

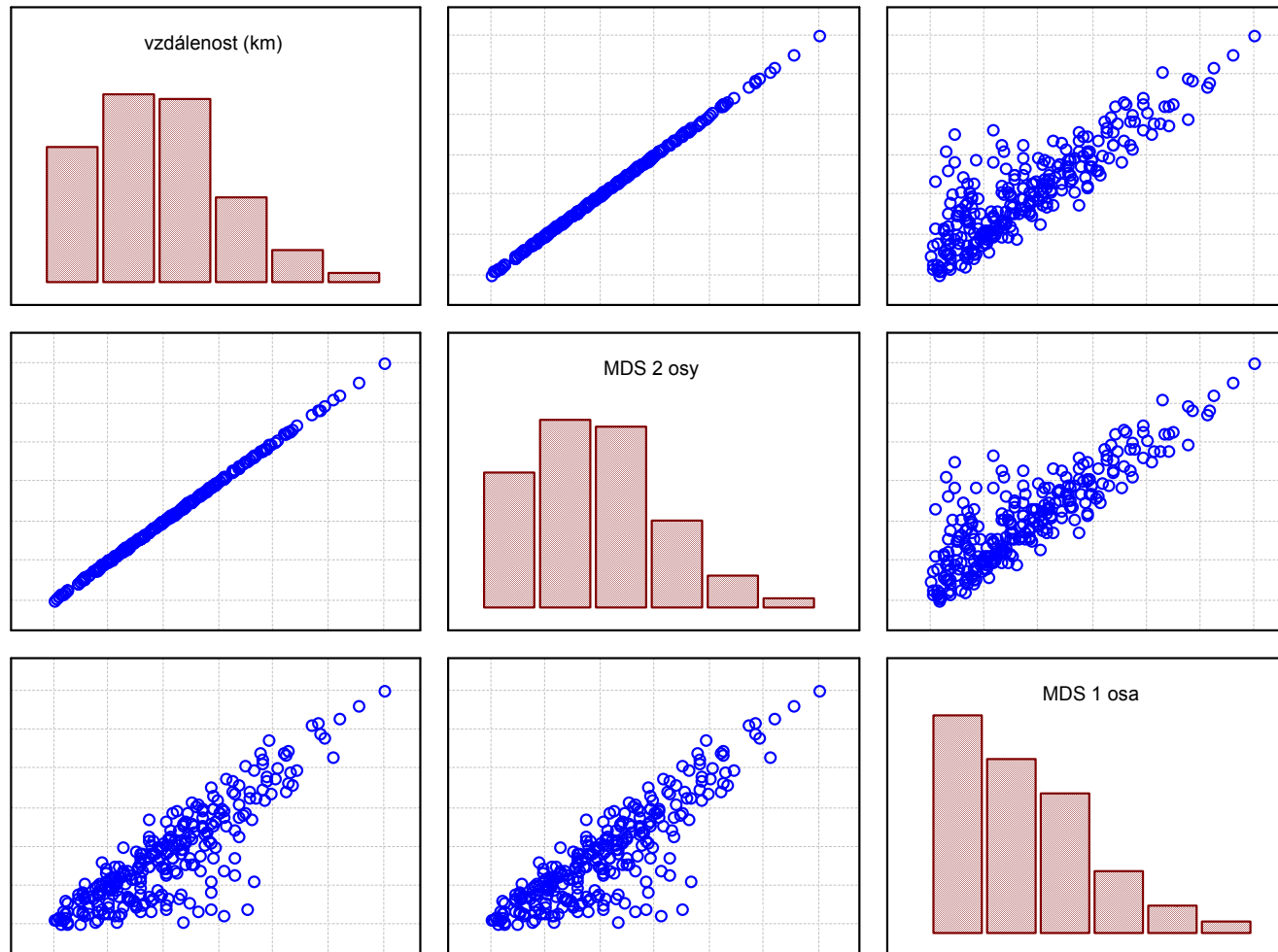
2 osy



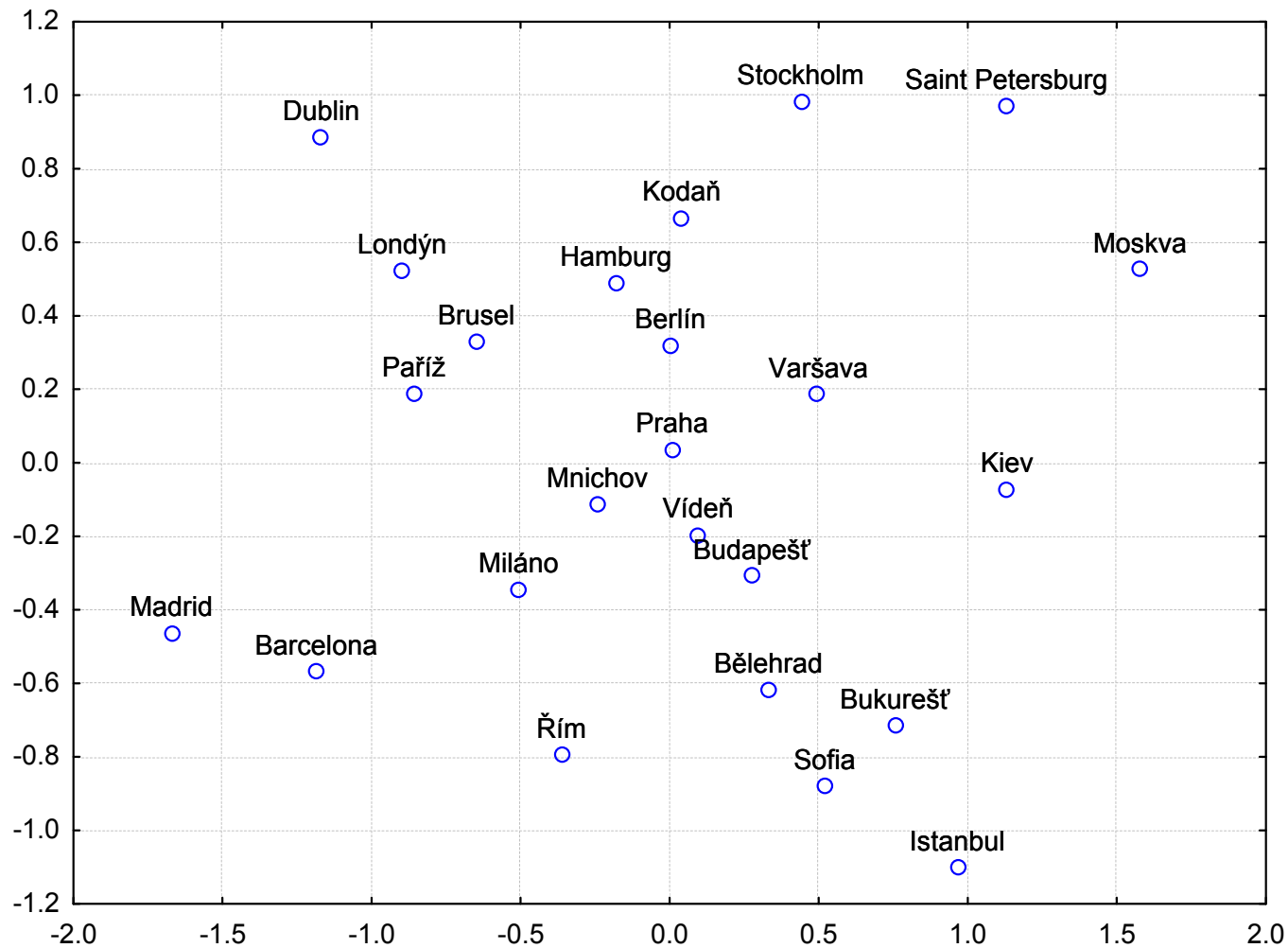
1 osa



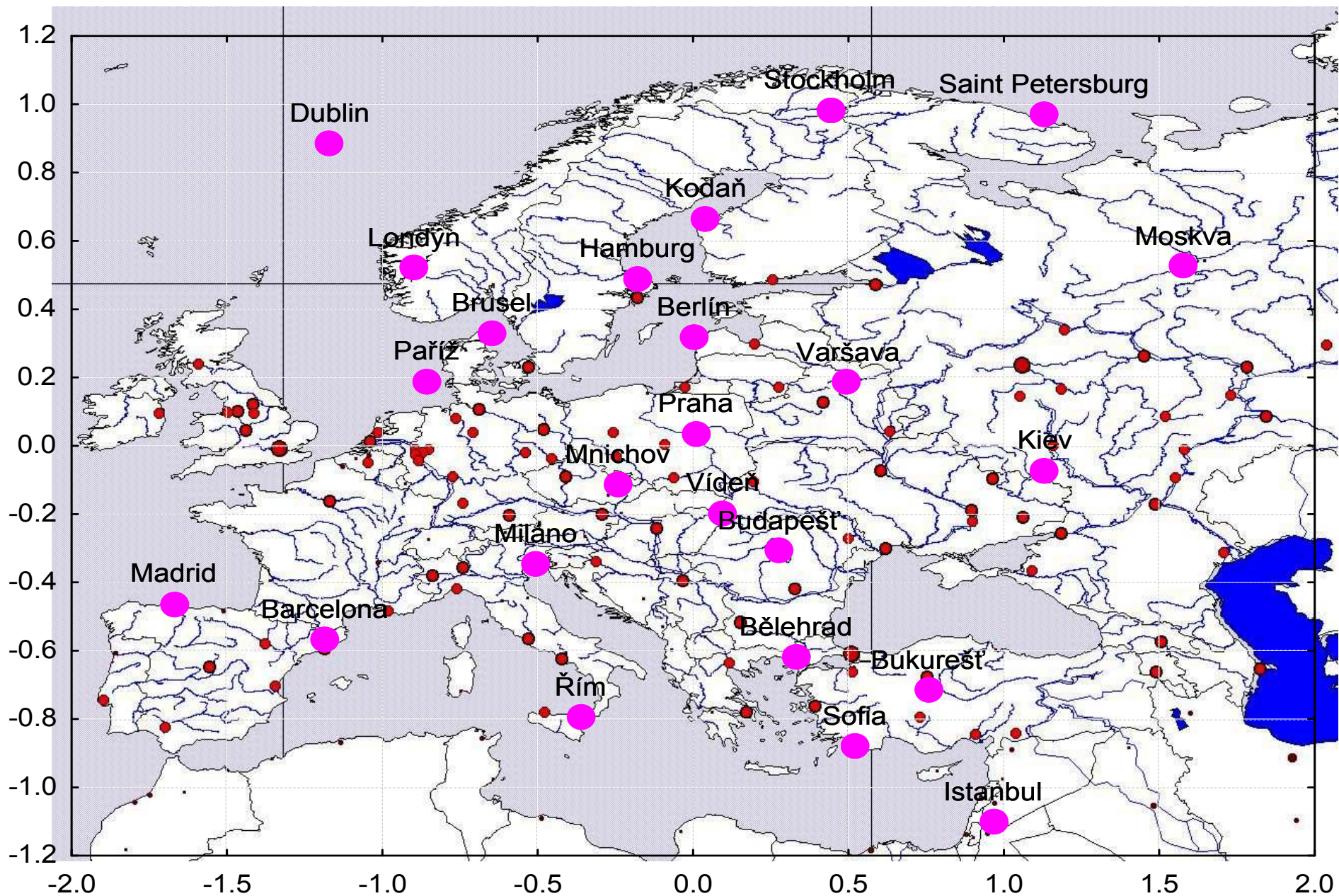
Vzdálenosti v původních datech a vytvořených faktorových osách



Reprezentace výstupu



Reprezentace výstupu



Ordinační analýzy: shrnutí

- Analýza hlavních komponent, faktorová analýza, korespondenční analýza a multidimensional scaling se snaží zjednodušit vícerozměrnou strukturu dat výpočtem souhrnných os
- Metody se liší v logice tvorby těchto os
 - Maximální variabilita (analýza hlavních komponent, korespondenční analýza)
 - Maximální interpretovatelnost os (faktorová analýza)
 - Převod asociační matice do Euklidovského prostoru (multidimensional scaling)