UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

BAKALÁŘSKÁ PRÁCE

Biplot a jeho aplikace

Vedoucí bakalářské práce:
RNDr. Karel Hron, Ph.D.
Rok odevzdání: 2010

Vypracovala:
Alžběta Kalivodová
Aplikovaná statistika, III. ročník

Prohlášení

Prohlašuji, že jsem vytvořila tuto diplomovou práci samostatně pod vedením RNDr. Karla Hrona, Ph.D. a že jsem v seznamu použité literatury uvedla všechny zdroje použité při zpracování práce.

V Olomouci dne 30. března 2010

Poděkování

Ráda bych na tomto místě poděkovala vedoucímu bakalářské práce RNDr. Karlu Hronovi, Ph.D. za obětavou spolupráci i za čas, který mi věnoval při konzultacích. Dále bych ráda poděkovala všem svým blízkým, že se mnou měli trpělivost, a také svému počítači.

Obsah

Úvod

Úkolem této práce je popsat vlastnosti biplotu jako v současnosti hojně užívaného grafického nástroje mnohorozměrné statistické analýzy. Biplot se totiž často užívá i při statistické analýze speciálních typů dat, například tzv. kompozičních dat, nesoucích pouze relativní informaci.

V první kapitole připomeneme některé poznatky z teorie matic, včetně tzv. singulárního rozkladu matice. Tyto znalosti se nám budou hodit při dalších výpočtech. Dále uvedeme vybrané číselné charakteristiky náhodného vektoru a Mahalanobisovu vzdálenost. V následující kapitole se seznámíme s metodou hlavních komponent, a to jak v její teoretické, tak i ve výběrové podobě. Tato mnohorozměrná statistická metoda je totiž základem pro tvorbu samotného biplotu - hlavního tématu této práce. Závěrečnou, neméně důležitou, částí budou příklady z oblasti aplikací.

Toto téma jsem si vybrala hned z několika důvodů. Statistika je mým hlavním oborem a biplot se mi zdá zajímavým, a podle mě nedoceněným, nástrojem explorativní analýzy dat. Protože je biplot poměrně novou grafickou technikou, cítila jsem též potřebu napomoci jeho rozšíření do obecného povědomí. Myslím si totiž, že výsledný výstup je dobře čitelný i pro matematického laika.

1 Poznatky z teorie matic

1.1 Základní pojmy

Nejprve se seznámíme s některými základními vlastnostmi matic, které budeme dále potřebovat při samotné konstrukci a odvozování biplotu. Zejména si připomeneme, co je to singulární rozklad matice a osvětlíme některé jeho vlastnosti. Při tvorbě této kapitoly byly informace čerpány zejména z [9], [11], [14].

Definice 1.1.

Nechť $\mathbf{u} = (u_1, \ldots, u_p)^T$ *a* $\mathbf{v} = (v_1, \ldots, v_p)^T$ *jsou p-složkové reálné sloupcové vektory, tedy* $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$.

Řekneme, že $\mathbf{u}$ *a* $\mathbf{v}$ *jsou ortogonální, jestliže platí* $\mathbf{u}^T\mathbf{v} = 0$. *Vektory* $\mathbf{u}$, $\mathbf{v}$ *jsou ortonormální, jestliže* $\mathbf{u}^T\mathbf{v} = 0$ *a zároveň pro jejich euklidovskou normu platí* $||\mathbf{u}|| = ||\mathbf{v}|| = \sqrt{v_1^2 + \ldots + v_p^2} = \sqrt{\sum_{i=1}^{p} v_i^2} = 1$.

Definice 1.2. *Reálnou čtvercovou matici* $\mathbf{A}$ *nazýváme ortogonální, jestliže platí* $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$, *kde* $\mathbf{I}$ *je jednotková matice příslušného řádu, tj. jestliže jsou její sloupce vzájemně ortonormální vektory.*

Definice 1.3.

1. Čtvercovou matici $\mathbf{A}$ *stupně n nazýváme pozitivně definitní, je-li symetrická a platí-li pro každý nenulový vektor* $\mathbf{x}$ *nerovnost* $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$. *Značíme* $\mathbf{A} > 0$.

2. Čtvercovou matici $\mathbf{A}$ *stupně n nazýváme pozitivně semidefinitní, je-li symetrická a platí-li pro libovolný vektor* $\mathbf{x}$ *nerovnost* $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$. *Značíme* $\mathbf{A} \geq 0$.

1.2 Singulární rozklad matice

Mějme dánu reálnou matici $\mathbf{X}$ o rozměrech $n \times p$, zapisujeme $\mathbf{X}_{n,p}$, a dále $\mathbf{U}_{n,n}$, $\mathbf{D}_{n,p}$ a $\mathbf{V}_{p,p}$. Matici $\mathbf{X}$ lze rozložit na součin

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T, \tag{1.1}$$

kde $\mathbf{U}$ a $\mathbf{V}$ jsou ortogonální matice. Tento vztah včetně značení byl převzat z [4] (str. 64, vztah 6.1).

Poznamenejme, že pro matice $\mathbf{U}$ a $\mathbf{V}$ tedy platí

$$\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = \mathbf{I} \quad \text{a} \quad \mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathbf{I}. \tag{1.2}$$

Dále $\mathbf{D}$ je matice nezáporných, tzv. singulárních, hodnot. Ty se nacházejí na hlavní diagonále a jsou uspořádány sestupně. Tedy

$$d_{11} \geq d_{22} \geq \cdots \geq d_{kk} \geq 0 \quad \text{kde} \quad k = \min(n, p); \tag{1.3}$$

přitom prvky mimo hlavní diagonálu jsou rovny nule.
Maticově obdržíme (v případě $n > p$)

$$\mathbf{D} = \begin{pmatrix} d_{11} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d_{pp} \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}. \tag{1.4}$$

Můžeme říci, že matice $\mathbf{X}$ o rozměrech $n \times p$ má maximálně $k = \min(n, p)$ různých singulárních hodnot.

Sloupce matice $\mathbf{U}$ se nazývají skóry (scores), odpovídající sloupce matice $\mathbf{V}$ se nazývají zátěže (loadings). Sloupce příslušných matic budeme značit $\mathbf{u}_i$, resp. $\mathbf{v}_j$, $i = 1, \ldots, n$; $j = 1, \ldots, p$. Singulární hodnoty mají potom vypovídající funkci o jejich vztahu. Čím blíže k nule je číslo d_{ii} pro $i = 1, 2, \ldots, k$, tím mají odpovídající skóry a zátěže v celkovém rozkladu menší vliv. Pěkně je toto vidět, zapíšeme-li rozklad $\mathbf{X}$ pomocí diagonálních prvků matice $\mathbf{D}$ a sloupců matic $\mathbf{U}, \mathbf{V}$,

$$\mathbf{X}_{n,p} = \mathbf{u}_1 d_{11} \mathbf{v}_1 + \mathbf{u}_2 d_{22} \mathbf{v}_2 + \cdots + \mathbf{u}_k d_{kk} \mathbf{v}_k = \sum_{i=1}^{k} d_{ii} \mathbf{u}_i \mathbf{v}_i^T. \tag{1.5}$$

Kladné hodnoty d_{ii} pro $i = 1, 2, \ldots, k$ budeme nazývat cenné hodnoty matice $\mathbf{X}$. Přitom platí

$$\mathbf{X}\mathbf{v}_i = d_{ii}\mathbf{u}_i \qquad \text{a} \qquad \mathbf{X}^T\mathbf{u}_i = d_{ii}\mathbf{v}_i, \quad i = 1, \ldots, k. \tag{1.6}$$

Singulární hodnoty se přitom mění se změnou prvků matice $\mathbf{X}$. Například, pokud vynásobíme všechny prvky této matice dvěma, normované velikosti prvků $\mathbf{U}$ a $\mathbf{V}$ se nezmění, ale singulární hodnoty budou dvakrát větší. Singulární hodnoty d_{ii} jsou vlastně odmocněná vlastní čísla čtvercových matic $\mathbf{X}\mathbf{X}^T$ a $\mathbf{X}^T\mathbf{X}$, jak si ukážeme za chvíli.

Součin skóru a příslušné singulární hodnoty se nazývá hlavní komponenta (principal component). Rozklad čtvercových matic $\mathbf{X}$ $\mathbf{X}^T$ a $\mathbf{X}^T\mathbf{X}$ tedy potom vede k tzv. metodě hlavních komponent. Tou se budeme podrobně zabývat ve třetí kapitole.

Zmíníme se o rozkladu matic $\mathbf{X}\mathbf{X}^T$ a $\mathbf{X}^T\mathbf{X}$ podrobněji:

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{V}\mathbf{D}^T\mathbf{U}^T = \mathbf{U}(\mathbf{D}\mathbf{D}^T)\mathbf{U}^T, \tag{1.7}$$

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{V}^T = \mathbf{V}(\mathbf{D}^T\mathbf{D})\mathbf{V}^T, \tag{1.8}$$

tedy sloupce matice $\mathbf{U}$ jsou vlastní vektory matice $\mathbf{X}\mathbf{X}^T$ a sloupce matice $\mathbf{V}$ jsou vlastní vektory matice $\mathbf{X}^T\mathbf{X}$. Jiný zápis, ze kterého je toto zřetelnější:

$$\mathbf{X}\mathbf{X}^T\mathbf{u}_i = d_{ii}^2\mathbf{u}_i \quad \text{pro} \quad i = 1, \ldots, k = \min(n, p), \tag{1.9}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{v}_i = d_{ii}^2\mathbf{v}_i \quad \text{pro} \quad i = 1, \ldots, k = \min(n, p). \tag{1.10}$$

Poznamenejme, že ve statistice matice $\mathbf{X}$ často představuje tzv. datovou matici. Její řádky jsou tvořeny n objekty, na nichž jsme změřili hodnoty p statistických znaků.

2 Číselné charakteristiky náhodného vektoru

2.1 Teoretické a výběrové charakteristiky náhodného vektoru

Než postoupíme dále, uvedeme si základní číselné charakteristiky náhodného vektoru. Na konci kapitoly ještě připomeneme pojem Mahalanobisovy vzdálenosti. Hlavním zdrojem při tvorbě této kapitoly byly knihy [2], [3], [8] a internetové stránky [10].

Definice 2.1. *Nechť je dán náhodný vektor* $\mathbf{X} = (X_1, \ldots, X_p)^T$ *na pravděpodobnostním prostoru* $(\Omega, \mathcal{A}, \mathsf{P})$ *a nechť existují střední hodnoty* $\mathsf{E}(X_1), \ldots, \mathsf{E}(X_p)$ *jeho složek. Pak se vektor* $\mathsf{E}(\mathbf{X}) = (\mathsf{E}(X_1), \ldots, \mathsf{E}(X_p))^T$ *nazývá střední hodnota náhodného vektoru* $\mathbf{X}$.

Tuto definici můžeme také interpretovat tak, že střední hodnota náhodného vektoru $\mathbf{X} = (X_1, \ldots, X_p)^T$ je vektor středních hodnot jeho složek (náhodných veličin $X_1, \ldots, X_p$).

Poznamenejme, že $\mathsf{E}(\mathbf{X} - \mathsf{E}(\mathbf{X})) = 0$. Tato zřejmá vlastnost je teoretickým podkladem pro tzv. centrování dat v popisné statistice, kdy od každého sloupce (hodnot znaku) v datové matici odečteme průměr hodnot znaku.

Obdobně definujeme střední hodnotu matice $\underline{\mathbf{X}}_{n,p}$, jejíž prvky jsou náhodné veličiny,

$$\mathsf{E}(\underline{\mathbf{X}}) = \mathsf{E}\begin{pmatrix} X_{11} & \cdots & X_{1p} \\ \cdots & \cdots & \cdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix} = \begin{pmatrix} \mathsf{E}(X_{11}) & \cdots & \mathsf{E}(X_{1p}) \\ \cdots & \cdots & \cdots \\ \mathsf{E}(X_{n1}) & \cdots & \mathsf{E}(X_{np}) \end{pmatrix}. \tag{2.1}$$

Definice 2.2. *Nechť má náhodný vektor* $\mathbf{X}$ *konečné druhé momenty. Potom můžeme definovat kovarianci složek (náhodných veličin)* X_i *a* X_j *jako*

$$\operatorname{cov}(X_i, X_j) = \mathsf{E}(X_i - \mathsf{E}(X_i))(X_j - \mathsf{E}(X_j)), \ i, j = 1, \ldots, p.$$

Pro $i = j$ je kovariance zřejmě rovna rozptylu, tedy $\mathrm{cov}(X_i, X_i) = \mathrm{var}(X_i)$.

Definice 2.3. *Čtvercová matice řádu* n $\mathrm{var}(\mathbf{X}) = \mathrm{cov}(X_i, X_j)_{i,j=1}^n$ *se nazývá varianční matice.*

Věta 2.1. *Varianční matice je symetrická a pozitivně semidefinitní.*

Důkaz: Symetrie je zřetelná z maticového zápisu

$$\mathrm{var}(\mathbf{X}) = \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_2, X_1) & \mathrm{var}(X_2) & \cdots & \mathrm{cov}(X_2, X_n) \\ \cdots & \cdots & \ddots & \cdots \\ \mathrm{cov}(X_n, X_1) & \mathrm{cov}(X_n, X_2) & \cdots & \mathrm{var}(X_n) \end{pmatrix};$$

víme totiž, že $\mathrm{cov}(X_i, X_j) = \mathrm{cov}(X_j, X_i),\ i, j = 1, \ldots, n$.

Pro důkaz pozitivní semidefinitnosti zvolme libovolný vektor $\mathbf{c} = (c_1, \ldots, c_n)^T$. Rozptyl každé náhodné veličiny je nenulový, nenulový je také rozptyl veličiny $\mathbf{c}^T\mathbf{X}$. Použijeme vlastnosti varianční matice lineární transformace číselným vektorem $\mathbf{a} \in \mathbb{R}^m$ a číselnou maticí $\mathbf{B}_{m,n}$, $\mathrm{var}(\mathbf{a}+\mathbf{B}\mathbf{X}) = \mathbf{B}\mathrm{var}(\mathbf{X})\mathbf{B}^T$, tedy $\mathrm{var}(\mathbf{c}^T\mathbf{X}) = \mathbf{c}^T\mathrm{var}(\mathbf{X})\mathbf{c} \geq 0$. Důkaz této věty je převzat z [2], str. 39, důkaz věty 3.3. □

Varianční matice může být definována i "maticově" jako

$$\mathrm{var}(\mathbf{X}) = \mathsf{E}\left[(\mathbf{X} - \mathsf{E}(\mathbf{X}))(\mathbf{X} - \mathsf{E}(\mathbf{X}))^T\right] = \mathsf{E}(\mathbf{X}\mathbf{X}^T) - (\mathsf{E}(\mathbf{X}))(\mathsf{E}(\mathbf{X}))^T. \qquad (2.2)$$

Situace v matematické statistice je opačná než v teorii pravděpodobnosti, kde společně s náhodným vektorem známe i jeho číselné charakteristiky. Zde vycházíme z p-rozměrného náhodného výběru $\mathbf{X}_1, \ldots, \mathbf{X}_n$ z rozdělení vektoru $\mathbf{X}$ a pomocí vhodných statistik se snažíme co nejlépe odhadnout skutečnou hodnotu $\mathsf{E}(\mathbf{X})$, respektive $\mathrm{var}(\mathbf{X})$. Tedy pracujeme s p proměnnými zjištěnými u n náhodně vybraných objektů. Příslušné teoretické charakteristiky označme $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, tedy $\boldsymbol{\mu} = \mathsf{E}(\mathbf{X}), \boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X})$.

Definice 2.4. *Vektor aritmetických průměrů jednotlivých složek náhodných vektorů* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *z p-rozměrného náhodného výběru z rozdělení vektoru* $\mathbf{X}$*,*

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i,$$

se nazývá výběrová střední hodnota.

Dále si zavedeme tzv. Wishartovu matici, která je v teorii mnohorozměrné statistické analýzy velmi oblíbená. Tato matice je čtvercová řádu p, symetrická a má tvar

$$\mathbf{W} = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T. \tag{2.3}$$

Odhadem varianční matice $\mathrm{var}(\mathbf{X})$ je výběrová varianční matice $\mathbf{S}$.

Definice 2.5. *Nechť máme dán náhodný výběr* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *z rozdělení vektoru* $\mathbf{X}$*. Výběrovou varianční maticí nazveme matici* $\mathbf{S}$

$$\mathbf{S} = \frac{1}{n-1}\mathbf{W} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = \frac{1}{n-1} \left(\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T - n\bar{\mathbf{X}}\bar{\mathbf{X}}^T \right).$$

Věta 2.2. *Nechť máme dán náhodný výběr* $\mathbf{X}_1, \ldots, \mathbf{X}_n$ *z rozdělení, které má střední hodnotu* $\boldsymbol{\mu}$ *a varianční matici* $\boldsymbol{\Sigma}$*. Potom platí*

$\mathsf{E}(\bar{\mathbf{X}}) = \boldsymbol{\mu}, \qquad \mathrm{var}(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}, \qquad \mathsf{E}(\mathbf{S}) = \boldsymbol{\Sigma}.$

Důkaz: Důkaz je uveden v [2], str. 68, důkaz věty 5.2. □

2.2 Mahalanobisova vzdálenost

Mahalanobisova vzdálenost byla zavedena roku 1936 indickým matematikem P.C. Mahalanobisem. Je užívána především ve své výběrové podobě, kdy její realizace vyjadřuje vzdálenost pozorování $\mathbf{X}_i$ od centra distribuce datového souboru, vyjádřeného pomocí výběrové střední hodnoty $\bar{\mathbf{X}}$, vzhledem ke kovarianční struktuře, dané výběrovou varianční maticí $\mathbf{S}$.

Definice 2.6. *Mějme* $\mathbf{X} = (X_1, \ldots, X_p)^T$ *náhodný vektor, který má střední hodnotu* $\boldsymbol{\mu}$ *a varianční matici* $\boldsymbol{\Sigma}$*. Mahalanobisova vzdálenost je definována vztahem*

$$D_M(\mathbf{X}) = \sqrt{(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})}.$$

Mahalanobisova vzdálenost může být také definována jako vzdálenost dvou různých náhodných vektorů $\mathbf{X}$ a $\mathbf{Y}$, které mají stejné rozdělení s varianční maticí $\boldsymbol{\Sigma}$:

$$D_M(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \mathbf{Y})}. \tag{2.4}$$

Jak již bylo řečeno na začátku, nejpoužívanější je Mahalanobisova vzdálenost ve své výběrové formě

$$D_M(\mathbf{X}_i) = \sqrt{(\mathbf{X}_i - \bar{\mathbf{X}})^T \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}})}. \tag{2.5}$$

3 Metoda hlavních komponent

3.1 Úvod

Jak si ukážeme v další kapitole, metoda hlavních komponent je základem pro konstrukci biplotu. Jejím tvůrcem je Karl Pearson (1901). Cílem této metody je zredukovat dimenzi mnohorozměrných dat tak, aby se stala jednoduchými a dobře čitelnými, ale abychom touto redukcí ztratili co nejméně informace. Data se zobrazují skrze hlavní komponenty, což jsou skryté veličiny, které vysvětlují jejich variabilitu a vzájemnou závislost. Ve své teoretické podobě jsou hlavní komponenty vlastně lineární kombinace původních složek náhodného vektoru. Při této metodě nejsou data nijak členěna, ale posuzujeme je jako rovnocenné. Při tvorbě hlavních komponent vycházíme ze singulárního rozkladu, který jsme si popsali v první kapitole. Při zpracování této kapitoly bylo čerpáno zejména z [1], [4], [7].

Mějme $\mathbf{X} = (X_1, \ldots, X_p)^T$ náhodný vektor s rozdělením, které má střední hodnotu $\mathsf{E}(\mathbf{X}) = \boldsymbol{\mu}$ a pozitivně semidefinitní varianční matici

$$\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{X}) = \mathsf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T]. \tag{3.1}$$

Dále je dána matice $\mathbf{G} = (\mathbf{g}_1, \ldots, \mathbf{g}_p)$, kde $\mathbf{g}_1, \ldots, \mathbf{g}_p$ jsou ortonormální vlastní vektory matice $\boldsymbol{\Sigma}$. Platí pro ně tedy

$$\mathbf{g}_i^T \mathbf{g}_j = 0 \quad \text{pro} \quad i \neq j \quad \text{a} \quad \mathbf{g}_i^T \mathbf{g}_i = 1 \quad \text{pro} \quad i, j = 1, \ldots, p, \tag{3.2}$$

tedy matice $\mathbf{G}$ je ortogonální.

Hlavní komponenty jsou vyjádřeny pomocí náhodného vektoru $\mathbf{Z}$,

$$\mathbf{Z} = \mathbf{G}^T(\mathbf{X} - \boldsymbol{\mu}) \tag{3.3}$$

nebo též jednotlivě jako náhodné veličiny

$$Z_i = \mathbf{g}_i^T(\mathbf{X} - \boldsymbol{\mu}) \quad \text{pro} \quad i = 1, \ldots, p. \tag{3.4}$$

Rozptyl i-té hlavní komponenty je potom

$$\text{var}(Z_i) = \mathsf{E}\left[\mathbf{g}_i^T(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\mathbf{g}_i\right] = \mathbf{g}_i^T\boldsymbol{\Sigma}\mathbf{g}_i \quad \text{pro} \quad i = 1, \ldots, p. \tag{3.5}$$

Libovolné dvě hlavní komponenty jsou nekorelované, protože jsou příslušné vlastní vektory $\mathbf{g}_i$ ortonormální. Takto lze tedy vytvořit p komponent; z hlediska zmenšování dimenze dat (která je naším hlavním cílem) je ale lepší mít komponent méně.

Označme d_i vlastní čísla matice $\boldsymbol{\Sigma}$ seřazená sestupně $d_1 \geq d_2 \geq \ldots \geq d_r \geq 0$. Je-li $r < p$, pak zbývající vlastní čísla (je jich $p - r$) jsou nulová. Pokud je $\boldsymbol{\Sigma}$ pozitivně definitní, jsou všechny vlastní hodnoty kladná reálná čísla.

Hlavním kritériem konstrukce veličin Z_i je požadavek jejich maximálního rozptylu v daném směru. Hlavní komponenty tedy obdržíme maximalizací funkce $\mathbf{g}_i^T\boldsymbol{\Sigma}\mathbf{g}_i$ za podmínky $\mathbf{g}_i^T\mathbf{g}_i = 1$. Získáme funkci s Lagrangeovými multiplikátory

$$\phi_i = \mathbf{g}_i^T\boldsymbol{\Sigma}\mathbf{g}_i - d_i(\mathbf{g}_i^T\mathbf{g}_i - 1), \quad i = 1, \ldots, r. \tag{3.6}$$

Parciální derivací podle $\mathbf{g}_i$ obdržíme:

$$\frac{\partial \phi_i}{\partial \mathbf{g}_i} = 2\boldsymbol{\Sigma}\mathbf{g}_i - 2d_i\mathbf{g}_i = 0, \tag{3.7}$$

$$(\boldsymbol{\Sigma} - d_i\mathbf{I})\mathbf{g}_i = 0, \tag{3.8}$$

maticově potom

$$\boldsymbol{\Sigma}\mathbf{G} = \mathbf{G}\mathbf{D}, \tag{3.9}$$

kde $\mathbf{D} = \mathrm{Diag}(d_1, \ldots, d_r)$, tedy $\mathbf{D}$ je matice, která má na diagonále čísla d_i, $i = 1, \ldots, r$, a mimo diagonálu 0. Varianční matici $\boldsymbol{\Sigma}$ můžeme vyjádřit jako

$$\boldsymbol{\Sigma} = \mathbf{G}\mathbf{D}\mathbf{G}^T \tag{3.10}$$

nebo můžeme vyjádřit $\mathbf{D}$,

$$\mathbf{G}^T\boldsymbol{\Sigma}\mathbf{G} = \mathbf{D}. \tag{3.11}$$

Střední hodnota hlavních komponent je nulová. Vycházíme z:

$$\mathsf{E}(\mathbf{Z}) = \mathbf{G}^T\big[\mathsf{E}(\mathbf{X} - \boldsymbol{\mu})\big] = \mathbf{G}^T\big[\mathsf{E}(\mathbf{X}) - \mathsf{E}(\boldsymbol{\mu})\big] = \mathbf{G}^T(\boldsymbol{\mu} - \boldsymbol{\mu}) = 0. \tag{3.12}$$

Varianční matice má tvar

$$\mathrm{var}(\mathbf{Z}) = \mathbf{G}^T\mathrm{var}(\mathbf{X} - \boldsymbol{\mu})\mathbf{G} = \mathbf{G}^T\boldsymbol{\Sigma}\mathbf{G} = \mathbf{D}. \tag{3.13}$$

Prvky g_{ij} matice $\mathbf{G}$ vyjadřují vliv veličiny X_i na Z_j, $i, j = 1, \ldots, p$, $\mathbf{G}$ se nazývá matice zátěží (loading matrix).

Trochu jiný náhled na konstrukci hlavních komponent spočívá v tom, že hledáme takový vektor reálných čísel $\mathbf{c} = (c_1, \ldots, c_p)^T$, který splňuje podmínku $\mathbf{c}^T\mathbf{c} = 1$ a pro který má veličina $\mathbf{c}^T\mathbf{X}$ největší rozptyl. $\mathbf{X}$ je centrovaný náhodný vektor. Protože $\mathrm{var}(\mathbf{c}^T\mathbf{X}) = \mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$, maximalizujeme vlastně výraz $\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$. Tato maximální hodnota je d_1 a platí, pokud $\mathbf{c} = \mathbf{g}_1$. Získali jsme první hlavní komponentu $Z_1 = \mathbf{g}_1^T\mathbf{X}$. Dále budeme hledat znovu vektor $\mathbf{c} \in \mathbb{R}^p$ za daných podmínek, tentokrát ale přibývá ještě jedna podmínka, a to, že musí být nekorelovaný s veličinou Z_1. Tato podmínka nekorelovanosti je dosažena pokud $\mathbf{c}^T\mathbf{g}_1 = 0$. Takto dostaneme $\mathbf{c} = \mathbf{g}_2$ a druhou hlavní komponentu $Z_2 = \mathbf{g}_2^T\mathbf{X}$. Ukazuje se, že tento proces pokračuje obdobně i dále, tedy do doby, než najdeme všechny hlavní komponenty

$$Z_i = \mathbf{g}_i^T\mathbf{X}, \quad i = 1, \ldots, r. \tag{3.14}$$

Poznamenejme ovšem již nyní, že při tvorbě biplotu používáme pouze první dvě hlavní komponenty a první dva sloupce matice $\mathbf{G}$.

V praxi se objevuje problém při tvorbě hlavních komponent veličin $\mathbf{X} = (X_1, \ldots, X_p)^T$, které jsou dány v různých jednotkách. Změnou měřítka se totiž mohou podstatně změnit hodnoty hlavních komponent. Proto často nevycházíme z původních veličin, ale provedeme transformaci znormováním složek náhodného vektoru. Takto tedy pracujeme s náhodným vektorem $\mathbf{Y} = (Y_1, \ldots, Y_p)^T$, který vznikne odečtením středních hodnot od původních veličin a podělením příslušnou směrodatnou odchylkou,

$$Y_i = \frac{X_i - \mathsf{E}(X_i)}{\sqrt{\mathrm{var}(X_i)}}, \quad i = 1, \ldots, p. \tag{3.15}$$

Jak již bylo řečeno, v praxi se snažíme popsat mnohorozměrnou strukturu náhodného vektoru $\mathbf{X}$ pomocí několika málo komponent Z_i. Hlavním kritériem je přitom procentuální podíl celkové variability vektoru $\mathbf{X}$, tj. $\sum_{i=1}^{p} \mathrm{var}(X_i)$, který se pomocí veličin Z_i podaří vysvětlit. Požadovaná hodnota tohoto podílu přitom závisí na konkrétní situaci a na dimenzi p.

3.2 Hlavní komponenty ve výběru

Z praktického hlediska je pro nás ovšem důležitá zejména výběrová obdoba metody hlavních komponent. Nechť máme dán náhodný výběr z rozdělení vektoru $\mathbf{X}$, tj. nezávislé a stejně rozdělené náhodné vektory $\mathbf{X}_1, \ldots, \mathbf{X}_n$, který uspořádáme do datové matice $\mathbf{X}$ o rozměrech $n \times p$. *Dále až do konce práce tedy budeme zápisem* $\mathbf{X}$ *rozumět tuto matici.* Nechť je dána výběrová střední hodnota $\bar{\mathbf{X}}$ a výběrová varianční matice $\mathbf{S}$, definované v předchozí kapitole.

Jako analogii k (3.3) můžeme výběrové hlavní komponenty (resp. jednotlivé výběrové komponenty - sloupce matice $\mathbf{Z}_{n,p}$) vypočítat ze vztahů

$$\mathbf{Z} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}^T)\hat{\mathbf{G}}, \tag{3.16}$$

$$\mathbf{Z}_j = (\mathbf{X} - \mathbf{1}\bar{\mathbf{X}}^T)\hat{\mathbf{g}}_j \quad \text{pro} \quad j = 1, \ldots, p. \tag{3.17}$$

Přitom $\mathbf{1}$ je vektor n jedniček a $\hat{\mathbf{G}} = (\hat{\mathbf{g}}_1, \ldots, \hat{\mathbf{g}}_p)$ je matice, jejíž sloupce jsou jednotlivé vlastní vektory matice $\mathbf{S}$.

Podobně jako v teoretickém případě (3.11) i zde

$$\hat{\mathbf{G}}^T\mathbf{S}\hat{\mathbf{G}} = \hat{\mathbf{D}}, \tag{3.18}$$

kde $\hat{\mathbf{D}} = Diag(\hat{g}_1, \ldots, \hat{g}_p)$ je diagonální matice vlastních čísel (seřazených sestupně) matice $\mathbf{S}$.

Poznamenejme přitom, že matice $\mathbf{Z}$ má stejné rozměry jako datová matice $\mathbf{X}$ a její hodnoty $Z_{ij}, i = 1, \ldots, n,\ j = 1, \ldots, p$, můžeme označit jako skóry. Realizací náhodného výběru poté obdržíme uvedené charakteristiky v jejich známé, číselné podobě.

Ani při výběrové metodě hlavních komponent nezapomínáme na škálování jednotlivých statistických znaků (teoreticky viz vztah (3.15)).

4 Konstrukce biplotu

V této kapitole se již budeme věnovat hlavní náplni této práce - biplotu. Biplot se dá zjednodušeně vysvětlit jako dvoudimenzionální zobrazení objektů a proměnných v jednom grafu. Slovo biplot pochází z angličtiny a "bi" na začátku značí právě dvě dimenze. Autorem této statistické metody je K. R. Gabriel, který ji poprvé popsal roku 1971 v článku [5]. Hlavním zdrojem pro tuto kapitolu bylo opět skriptum [4].

Biplot je tedy grafické zobrazení statistického souboru o rozsahu n odpovídající p statistickým znakům $X_1, \ldots, X_p$, vyjádřeným pomocí datové matice $\mathbf{X}$ (předpokládejme přitom, bez újmy na obecnosti, že pracujeme již s centrovanými daty). Biplot má několik druhů, zde představíme ten, který vychází z metody hlavních komponent. O ostatních se můžeme dočíst například v knize autorů Gower a Hand [6].

Zobrazení v rovinném grafu je z hlediska interpretace dat výhodné a přehledné. Při tomto zobrazení musíme ale předpokládat, že datová matice $\mathbf{X}$ má alespoň dva sloupce. Pro matici s větším počtem sloupců pak využijeme informaci z prvních dvou hlavních komponent.

Nejprve je pro konstrukci biplotu důležité vyjádření matice $\mathbf{X}$ pomocí matic $\mathbf{U}$, $\mathbf{D}$ a $\mathbf{V}$ (viz (1.1) a (1.5)). Dále ze vztahů (1.7) a (1.8), můžeme konstatovat, že n sloupců $\mathbf{U}$ představuje ortonormální vlastní vektory matice $\mathbf{X}\mathbf{X}^T$ a p sloupců $\mathbf{V}$ jsou ortonormální vlastní vektory matice $\mathbf{X}^T\mathbf{X}$.

Mějme tedy matici $\mathbf{X}$ o rozměrech $n \times p$ s hodností $k < \min(n, p)$. Princip konstrukce biplotu je založen na nahrazení matice $\mathbf{X}$ pomocí její aproximace $\mathbf{X}_{(2)}$ s hodností rovnou dvěma, která se jeví optimální z hlediska minimalizace součtu čtverců odchylek jejích prvků od příslušných prvků matice $\mathbf{X}$. Ve vyjádření matice $\mathbf{X}_{(2)}$ přitom použijeme pouze první dva sloupce matice $\mathbf{U}$ a první dva sloupce matice $\mathbf{V}$ ze singulárního rozkladu. Maticově lze tuto skutečnost zapsat jako (upozorňujeme přitom na předefinování matic $\mathbf{U}$, $\mathbf{D}$ a $\mathbf{V}$)

$$\mathbf{X} \approx \mathbf{X}_{(2)} = \mathbf{U}\mathbf{D}\mathbf{V}^T = (\mathbf{u}_1, \mathbf{u}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \end{pmatrix}. \tag{4.1}$$

Přitom je zřejmé, že $\mathbf{X}_{(2)}$ je opět rozměrů $n \times p$. Můžeme ji rozdělit takto:

$$\mathbf{X}_{(2)} = \mathbf{G}\mathbf{H}^T, \tag{4.2}$$

kde

$$\mathbf{G} = (\mathbf{u}_1, \mathbf{u}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{1-c}, \tag{4.3}$$

$$\mathbf{H} = (\mathbf{v}_1, \mathbf{v}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c} \tag{4.4}$$

pro $0 \leq c \leq 1$. Připomeneme si tzv. cenné hodnoty $\mathbf{X}$ zmíněné v kapitole 1.2. Po volbě čísla c tedy máme rozděleny první dvě cenné hodnoty mezi matice $\mathbf{G}$ a $\mathbf{H}$ a můžeme takto získat již zmíněné různé druhy biplotů. Biplot je potom tvořen právě řádky matic $\mathbf{G}$ a $\mathbf{H}$ o rozměrech $n \times 2$ a $p \times 2$.

Pro $c = 1$ potom matice $\mathbf{G}$ a $\mathbf{H}$ vychází

$$\mathbf{G} = \begin{pmatrix} \mathbf{g}_1^T \\ \vdots \\ \mathbf{g}_n^T \end{pmatrix} = \sqrt{n-1}(\mathbf{u}_1, \mathbf{u}_2), \tag{4.5}$$

$$\mathbf{H} = \begin{pmatrix} \mathbf{h}_1^T \\ \vdots \\ \mathbf{h}_p^T \end{pmatrix} = \frac{1}{\sqrt{n-1}}(\mathbf{v}_1, \mathbf{v}_2) \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}. \tag{4.6}$$

Pro volbu $c = 1$ tedy matice $\mathbf{G}$ a $\mathbf{H}$ (až na konstantu $\sqrt{n-1}$) představují skóry a zátěže prvních dvou hlavních komponent, jak bylo zmíněno v kapitole 1.2.

Jak si uvedeme ještě později při samotné grafické intepretaci biplotu, řádky matice $\mathbf{G}$ v grafu představují body a řádky matice $\mathbf{H}$ vrcholy šipek, které vycházejí

z uspořádané dvojice průměrů sloupců matice $\mathbf{G}$. V praxi ovšem často pracujeme s centrovanými daty, to znamená, že od hodnot sloupců datové matice $\mathbf{X}$ odečítáme aritmetické průměry těchto sloupců. V tomto případě je střed, ze kterého vycházejí šipky, umístěn v bodě $[0, 0]$.

Nyní si ukážeme některé vlastnosti biplotu (za předpokladu práce s centrovanými daty).

Při součinu řádků matic $\mathbf{G}$ a $\mathbf{H}$ nám vychází

$$\mathbf{g}_{i.}^T \mathbf{h}_{j.} = \sqrt{n-1}\mathbf{u}_{i.}^T \frac{1}{\sqrt{n-1}} (\mathbf{v}_{j.}^T \mathbf{D})^T = \mathbf{u}_{i.}^T \mathbf{D} \mathbf{v}_{j.} \approx x_{ij}. \tag{4.7}$$

Druhé mocniny délek vektorů $\mathbf{h}_i$ aproximují rozptyl statistických znaků X_i, protože

$$\begin{aligned} \mathbf{H}\mathbf{H}^T &\overset{1.}{=} \big(\frac{1}{\sqrt{n-1}}\mathbf{V}\mathbf{D}\big)\big(\frac{1}{\sqrt{n-1}}\mathbf{D}\mathbf{V}^T\big) \overset{2.}{=} \frac{1}{n-1}\mathbf{V}\mathbf{D}^2\mathbf{V}^T \\ &\overset{3.}{=} \frac{1}{n-1}(\mathbf{V}\mathbf{D}\mathbf{U}^T)(\mathbf{U}\mathbf{D}\mathbf{V}^T) \overset{4.}{=} \frac{1}{n-1}\mathbf{X}_{(2)}^T\mathbf{X}_{(2)} \approx \frac{1}{n-1}\mathbf{X}^T\mathbf{X} \overset{5.}{=} \mathbf{S} \end{aligned} \tag{4.8}$$

a diagonální prvky matice $\mathbf{H}\mathbf{H}^T$ jsou rovny $\mathbf{h}_i^T\mathbf{h}_i = \|\mathbf{h}_i\|^2$.

Nyní si jednotlivé kroky podrobně zdůvodníme:

1. Je to pouze jiný zápis vzorce (4.6), využíváme zde vlastnosti transpozice součinu matic.

2. Využíváme toho, že $\mathbf{D}$ je diagonální matice nezáporných singulárních hodnot. Můžeme tedy psát $\mathbf{D}\mathbf{D}^T = \mathbf{D}^2$.

3. Vychází z rovnosti (1.8) a z toho, že $\mathbf{U}^T\mathbf{U} = \mathbf{I}$.

4. Využíváme rovností (1.1) a (4.1) a dále vlastností transpozice násobených matic.

5. Využíváme vztah pro výpočet výběrové varianční matice. Zde se vyskytuje $\bar{\mathbf{X}}$ (výběrová střední hodnota řádků matice $\mathbf{X}$), která je ale rovna nulovému vektoru, protože data jsou centrovaná. Dále využíváme vztahu $\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T = \mathbf{X}\mathbf{X}^T$.

Kosinus úhlů $\mathbf{h}_i$ a $\mathbf{h}_j, i \neq j$, aproximuje korelační koeficient mezi X_i, X_j:

$$\cos(\mathbf{h}_i, \mathbf{h}_j) = \frac{\mathbf{h}_i^T \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|} \approx r_{ij}. \tag{4.9}$$

Jak již bylo řečeno, biplot nám slouží ke zjednodušení grafického zobrazení více než dvou statistických znaků. Přitom je takto metoda hlavních komponent využívána v tom smyslu, že biplot je vlastně zobrazení skórů a zátěží prvních dvou hlavních komponent datové matice $\mathbf{X}$ při volbě $c = 1$ (ve vztazích (4.3) a (4.4)).

Rozdíl mezi metodou hlavních komponent a biplotem je v normování. Když si graficky zobrazíme matice skórů a zátěží z metody hlavních komponent, nebudou normované. Toto normování nám naopak zajistí biplot, přitom normovacími konstantami jsou singulární hodnoty. Matice $\mathbf{G}$, $\mathbf{H}$ zmíněné v této kapitole takto vlastně představují výběrové skóry a zátěže (v tomto pořadí) prvních dvou hlavních komponent.

Jestliže v grafu reprezentují body jednotlivé objekty (pomocí skórů), šipky reprezentují jednotlivé statistické znaky. Délky jednotlivých šipek jsou přibližně rovny rozptylům příslušných statistických znaků, viz (4.8). Obecně řečeno, čím je šipka delší (znak má větší rozptyl), tím je vliv příslušného znaku na uspořádání dat větší. Kosinus úhlu mezi dvěma šipkami zobrazuje hodnotu korelačního koeficientu daných znaků. Jednoduše řečeno, čím je úhel mezi šipkami menší, tím je lineární vztah odpovídajících si statistických znaků těsnější. Další aspekty interpretace biplotu si ukážeme již přímo na konkrétních příkladech v následující kapitole.

Na konec celého procesu je dobré zjistit procentuální podíl celkové variability souboru (součet rozptylů jednotlivých statistických znaků) vysvětlené pomocí prvních dvou hlavních komponent Z_1, Z_2, tedy přesnost aproximace původní mnohorozměrné struktury dat (někdy se místo rozptylu uvažuje směrodatná odchylka). Za dobrý výsledek v tomto ohledu budeme nejčastěji považovat více

jak 75 %. Toto číslo je ale velmi subjektivní a je dáno zkušenostmi a rozměrem původních dat. Budeme se řídit pravidlem, že čím více statistických znaků (složek) bude obsahovat původní soubor, tím menší procentuální podíl vysvětlené variability budeme považovat za vyhovující.

5 Příklady

Na následujících příkladech si ukážeme využití biplotu v praxi. První příklad je převzatý ze skripta [4], v dalších jsou využívány datové soubory z internetových stránek Českého statistického úřadu a knihovny datových tabulek, která je umístěna na internetových stránkách `http://lib.stat.cmu.edu/DASL` (konkrétně [17], [18]). K výpočtům a grafickému vyjádření je využíván statistický software R (`www.r-project.org`)[12].

5.1 Studenti

Třída 88 studentů prvního ročníku vysoké školy technického směru je testována z pěti předmětů. Každý student může získat v každém z testů maximálně 100 bodů. Výsledky jsou zaznamenány a seřazeny v tabulce (Příloha A). Zkratky jednotlivých předmětů jsou tyto: ME = mechanika, AG = analytická geometrie, LA = lineární algebra, AN = matematická analýza, ES = elementární statistika. Úkolem je zjistit, jaký je vzájemný vztah mezi jednotlivými studenty a také mezi jednotlivými předměty.

Jak již bylo řečeno na začátku, k výpočtům a grafickým zobrazením je používán statistický software R. Nejdříve nastavíme příslušnou zdrojovou knihovnu v daném počítači. K tomu použijeme příkaz `setwd()`. Zdrojová data jsou uložena v tabulce v textovém souboru `Stud.txt`. Poté zadáme data do softwaru jako matici o 88 řádcích (studenti) a 5 sloupcích (předměty):

```
>X=matrix(scan("Stud.txt"),ncol=5,byrow=T)
```

Dále označíme jednotlivé sloupce jmény:

```
>colnames(X)=c("ME","AG","LA","AN","ES")
```

Následuje příkaz `summary(princomp(X))`, který ukazuje, kolik procent celkové variability statistického souboru (po vynásobení 100) je vysvětleno pomocí hlavních komponent, nově vytvořených statistických znaků. V našich příkladech se díváme

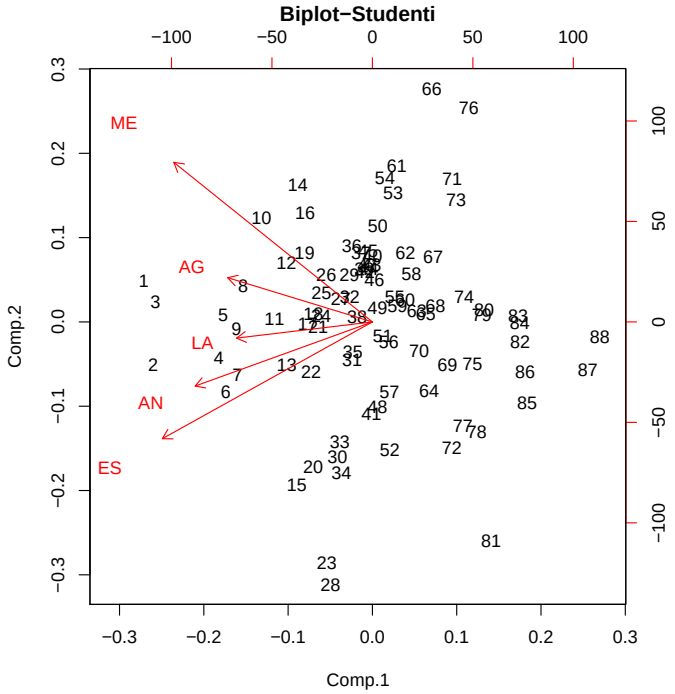
na první dva sloupce, které odpovídají prvním dvěma (nejvýznamnějším) hlavním komponentám. V tomto případě máme tabulku:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	26.061142	14.1355705	10.12760414	9.14706148	5.63807655
Proportion of Variance	0.619115	0.1821424	0.09349705	0.07626893	0.02897653
Cumulative Proportion	0.619115	0.8012575	0.89475453	0.97102347	1.00000000

První řádek tabulky nám ukazuje směrodatné odchylky jednotlivých znaků $Z_1, \ldots, Z_5$. Druhý řádek vyjadřuje, že znak Z_1 (první hlavní komponenta) vysvětluje 62 % a znak Z_2 (druhá hlavní komponenta) dalších 18 % celkové variability, což dá v součtu 80 %, a tedy velmi dobrý výsledek. Tento součet je vidět v poli třetího řádku, sloupce s nadpisem `Comp.2` (číslo 0.8012575). Biplot tedy bude velmi dobře odrážet skutečnou strukturu mnohorozměrného datového souboru. Poslední příkaz, který byl použit,

`>biplot(princomp(X),main="Biplot-Studenti")`,

vykreslí biplot pro naše data:



Nyní shrneme, jak lze biplot interpretovat. Můžeme říci, že hodnoty znaku Z_1 (`Comp.1` v předchozím grafu) zhruba reprezentují obecnou úspěšnost studentů (při orientaci zleva doprava, tj. od nejlepších k nejhorším, lze porovnat s tabulkou v Příloze A). Když se podíváme na směry šipek, vidíme, že menší úhly mají mezi sebou AG a ME na jedné straně a AN, LA, ES na straně druhé. Tady se potvrzuje předběžný předpoklad, že studenti, kteří mají odpovídající převládající typ myšlení (prostorové, analytické) budou dosahovat lepších výsledků v odpovídajících předmětech (tj. ME, AG, resp. LA, AN, ES). Přitom směry těchto šipek nám též pomáhají určit interpretaci pro znak Z_2 (`Comp.2`), který takto představuje přechod od prostorového myšlení k analytickému (jež odpovídá dosaženým výsledkům v jednotlivých testech v příslušných předmětech). Laicky tedy můžeme říci, že studenti, kteří se nacházejí v dolní části grafu jsou na tom lépe v předmětech LA, AN, ES, zatímco studenti nahoře jsou zběhlejší ve zbylých předmětech. Konkrétně se můžeme podívat na výsledky studenta číslo 28, jehož výsledky jsou postupně 18, 44, 50, 57, 81. Tedy je opravdu lepší ve skupině posledních tří předmětů. Naopak student číslo 66 (výsledky 59, 53, 37, 22, 19) má lepší skóre v prvních dvou předmětech. Nakonec poznamenejme, že délky jednotlivých šipek ukazují, jaké odpovídající znaky (zde ME a ES) mají na uspořádání pozorování v grafu největší vliv.

5.2 Zemědělství

Na dalším příkladu si názorně ukážeme, jak se v biplotu projevují proměnné (statistické znaky) s výrazně vyššími hodnotami (a rozptylem) než mají ostatní proměnné. V příkladu jsou použita reálná data z internetových stránek Českého statistického úřadu [16]. Příslušná tabulka má název Hektarové výnosy sklizně hlavních zemědělských plodin podle krajů v roce 2007 (Příloha B). Proměnnými jsou zde pšenice, ječmen, brambory, řepka, slunečnice a pícniny. Tou, jejíž hodnoty převyšují ostatní, jsou brambory. Jednotlivá pozorování představují kraje České republiky (je jich tedy 14) a v grafu jsou označeny zkratkami státních poznávacích značek automobilů. Jednotlivé zkratky jsou uvedeny v již zmíněné tabulce.

5.2.1 Zemědělství s bramborami

První biplot byl vytvořen pro všechny proměnné. Příkazy v softwaru jsou analogické jako u předchozího příkladu, proto stačí zmínit pouze některé.

Zde je vidět, jak jsou pojmenovány jednotlivé proměnné:

```
>colnames(X)=c("Ps","Je","Br","Re","Sl","Pi")
```

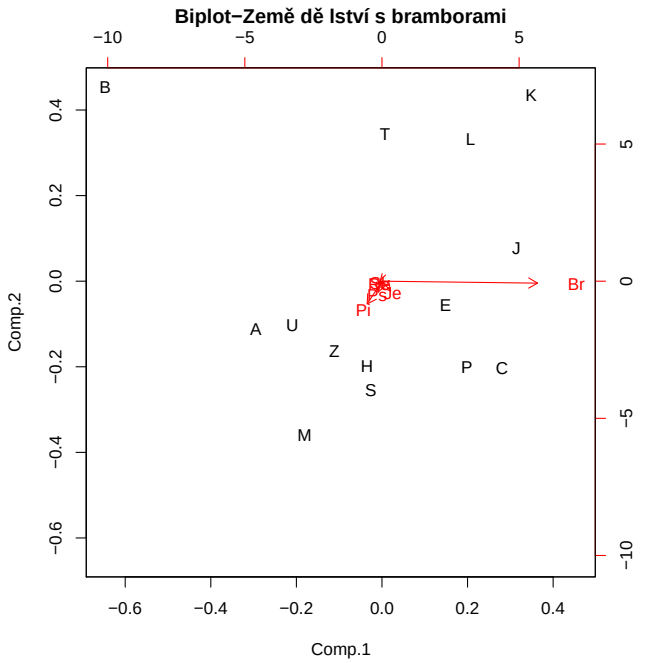
Následuje pojmenování jednotlivých pozorování - krajů České republiky:

```
>rownames(X)=c("A","S","C","P","K","U","L","H","E","J","B","M","Z","T")
```

Dále je důležitá tabulka charakteristik jednotlivých hlavních komponent:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.907905	0.33129175	0.191553760	0.105538424	0.067728392	1.443441e-02
Proportion of Variance	0.957296	0.02886385	0.009649712	0.002929235	0.001206353	5.479378e-05
Cumulative Proportion	0.957296	0.98615991	0.995809618	0.998738853	0.999945206	1.000000

Vidíme tedy, že znak Z_1 vysvětluje skoro 96 % a znak Z_2 pouze 2,9 % celkové variability, což dá v součtu téměř 99 %. Výsledná vysvětlená variabilita se může zdát vynikající. Ovšem, když se podíváme, jaký je nepoměr mezi jednotlivými znaky, už se tento výsledek tak dobrý nejeví. Toto je také patrné z biplotu.



Z grafu je na první pohled patrné, že brambory "ovládly" ostatní proměnné. Jejich vliv je největší. Je to dáno tím, že biplot se snaží vysvětlit co nejvíce rozptylu datového souboru, proto vzal v úvahu především proměnnou brambory - díky největším hodnotám má též zdaleka největší rozptyl. Uspořádání prvků zleva doprava nám ukazuje úspěšnost jednotlivých krajů, co se týče produkce brambor. Pořadí je zde ale opačné než u předchozího příkladu - od nejhoršího po nejlepší, což souvisí s orientací jednotlivých hlavních komponent. Ostatní plodiny zde nemají vliv. Úhly mezi šipkami, a tedy vztahy jednotlivých proměnných, nejsou moc zřetelné, takže je bohužel nemůžeme nijak interpretovat.

5.2.2 Zemědělství bez brambor

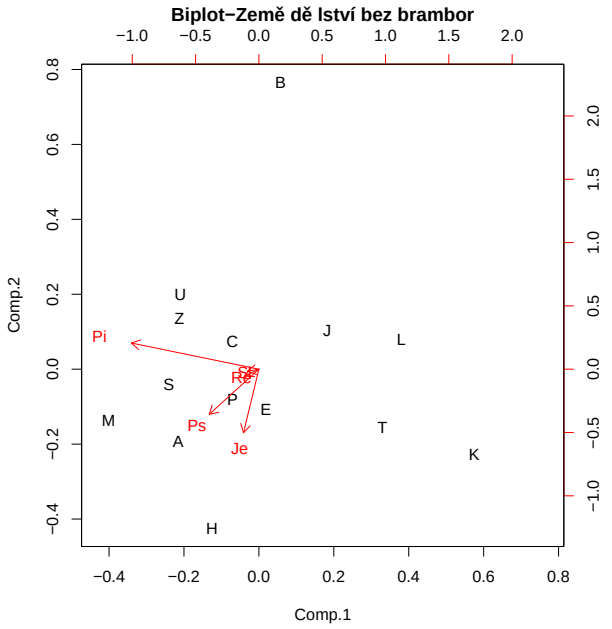
V druhém případě je proměnná brambory vynechána, tím pádem jsme získali pouze 5 proměnných a vliv proměnné brambory je eliminován.

Podíl celkové variability je v tomto případě menší, ale pořád dostačující:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	0.3664893	0.2178510	0.1532782	0.06893191	0.017657245
Proportion of Variance	0.6385857	0.2256398	0.1117011	0.02259108	0.001482321
Cumulative Proportion	0.6385857	0.8642255	0.9759266	0.99851768	1.000000000

Znak Z_1 vysvětluje 64 % a znak Z_2 22 % celkové variability, jejich součet je tedy 86 % . Biplot zřejmě vyjadřuje kovarianční strukturu zbylých znaků lépe než předchozí případ.

Graf, který je uveden na následující straně, je evidentně zřetelnější a čitelnější. Uspořádání prvků zleva doprava nám opět ukazuje úspěšnost jednotlivých krajů. Tentokrát v celkové produkci (od nejlepšího po nejhorší). Jde nám o hektarové výnosy, a tedy o kvalitu půdy. Zde je také krásně vidět geografické rozložení jednotlivých krajů. Jako nejhorší nám totiž vyšel kraj Karlovarský, zato nejlepší výsledek má Olomoucký kraj - úrodná Haná. Nejmenší úhel je mezi pšenicí a ječmenem, tedy jejich vztah je nejsilnější. Toto je opět logické vzhledem k tomu, že jsou obě plodiny obilovinami.



Největší vliv (nejdelší šipku) mají pícniny. Tato skutečnost je zřejmá z tabulky - pícniny mají druhé nejvyšší hodnoty po bramborách. Poslední fakt, který je zde zřetelný, je odlehlost pozorování Jihomoravský kraj. Všimněme si přitom, proč je toto odlehlé pozorování právě nahoře - tedy nejblíže ze všech šipek k šipce pícnin. Hektarový výnos této plodiny převažuje v daném kraji nad ostatními.

I tento biplot je ovšem značně ovlivněn různou variabilitou jednotlivých statistických znaků. Také jsme v tomto případě přišli o informaci z jednoho statistického znaku (brambor), což biplotu dále ubírá na relevantnosti.

5.2.3 Zemědělství škálované

Postup, který byl použit v předchozích případech, se ale obecně nezdá příliš vhodný. Pokud chceme zjistit vztahy mezi jednotlivými proměnnými bez toho, aby byly ovlivňovány jejich vysokými (nebo naopak nízkými) hodnotami a s tím souvisejícími hodnotami rozptylů znaků, použijeme škálování. Přeškálujeme tedy všechny hodnoty jednotlivých znaků tak, aby jejich průměr byl nula a rozptyl byl roven jedné. Takto se sice změní hodnoty proměnných, ale základní struktura datového souboru zůstane zachována.

Vezmeme tedy původní matici hodnot:

```
>X=matrix(scan("Zems.txt"),ncol=6,byrow=T)
```

Pojmenujeme její řádky a sloupce:

```
>colnames(X)=c("Ps","Je","Br","Re","Sl","Pi")
>rownames(X)=c("A","S","C","P","K","U","L","H","E","J","B","M","Z","T")
```

A nakonec data přeškálujeme:

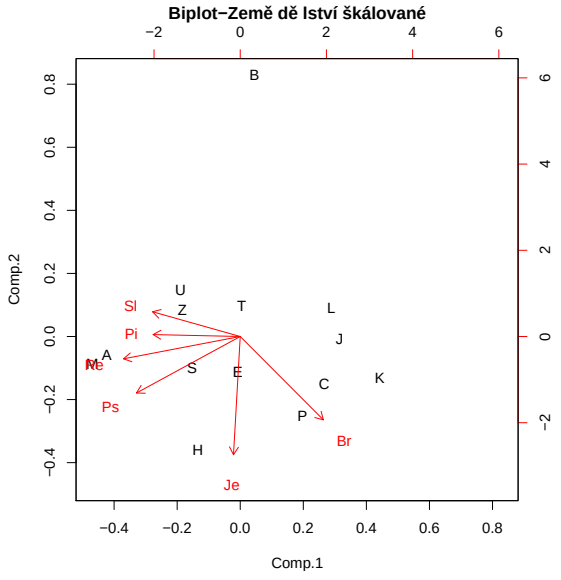
```
>M=scale(X, center = TRUE, scale = TRUE)
```

Přeškálované hodnoty jsou v tabulce v Příloze C. Nyní s novou datovou maticí **M** pracujeme stejně jako v předchozích příkladech s maticí **X**.

Tabulka charakteristik jednotlivých hlavních komponent:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.6738023	1.2312319	0.8448974	0.59072591	0.39008319	0.197253384
Proportion of Variance	0.5028538	0.2720904	0.1281272	0.06263333	0.02731165	0.006983648
Cumulative Proportion	0.5028538	0.7749442	0.9030714	0.96570470	0.99301635	1.000000000

Znak Z_1 vysvětluje 50 % a znak Z_2 27 % celkové variability, tedy součet je 77 % . Tato hodnota je vyhovující.



Tento biplot je zřejmě nejvýstižnější ze všech tří, které máme k dispozici. Uspořádání prvků zleva doprava nám ukazuje úspěšnost jednotlivých krajů v celkové produkci (od nejlepšího po nejhorší). Uspořádání se v některých aspektech liší od grafu Zemědělství bez brambor, vyjma například postavení krajů Olomouckého a Karlovarského. Úhel mezi pšenicí a ječmenem se značně zvětšil (jejich vzájemný vliv se tedy zmenšil), zato mezi pícninami a pšenicí je menší (vztah je silnější). Tentokrát šipky pšenice a ječmene nejdou stejným směrem. To znamená, že produkce těchto dvou plodin není tolik korelovaná. Šipka odpovídající proměnné brambory směřuje na opačnou stanu než většina ostatních - její produkce je negativně korelovaná. Jediná plodina, se kterou má alespoň nějakou pozitivní korelaci, je ječmen. Nejdelší šipku, a tím pádem i největší vliv, mají brambory a ječmen současně, o moc menší není ani vliv řepky a pšenice. Ovšem v tomto grafu mají větší význam pro interpretaci spíše směry šipek než jejich délky. Můžeme pozorovat vztah mezi řepkou a slunečnicí a také se zde opakuje odlehlé pozorování Jihomoravský kraj. Ještě bychom se měli zaměřit na postavení Moravskoslezského kraje. Ten stojí zhruba uprostřed grafu nad spojnicí všech šipek. Toto postavení nám říká, že daný kraj je v produkci jednotlivých plodin zhruba uprostřed. Když se podíváme do tabulky v Příloze C, je to opravdu tak. V produkci slunečnice je sice nejlepší (je totiž v grafu nejblíže její šipce), tuto výhodu ale naopak srazí předposlední místo v pícninách.

5.3 Inteligence a tělesné proporce

Následující příklad byl vybrán jako ukázka, že u biplotu je třeba být při interpretaci výsledků občas velmi obezřetný. Data jsou převzata z knihovny tabulek dat, kterou nalezneme na internetu [17].

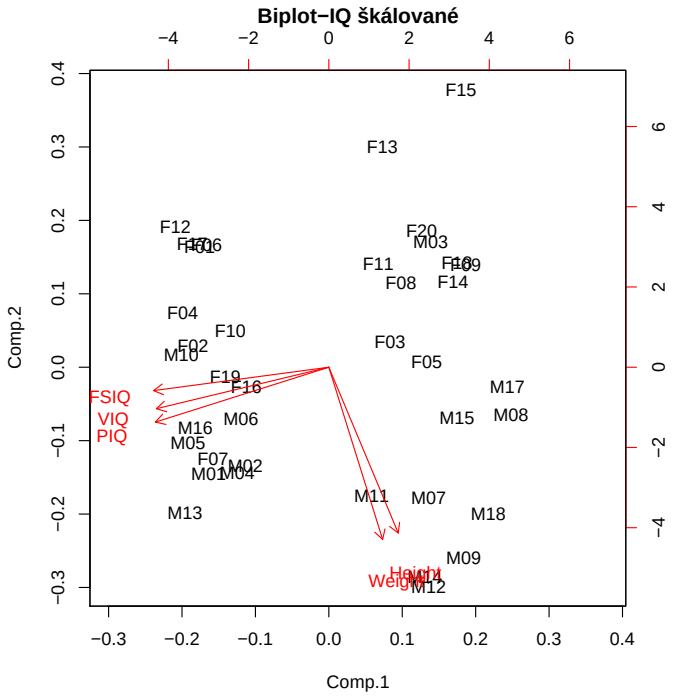
Jistý doktor Willerman dal v roce 1991 dohromady skupinu 40 studentů psychologie z Jihozápadní univerzity ve Velké Británii. Studenti vyplnili čtyři testy (slovní zásoba, podobnosti, Block Design a představivost obrázků) z Wechslerova testu inteligence pro dospělé. Z těchto testů určil doktor jednotlivé inteligence respondentů - celková (FSIQ), verbální (VIQ) a představivostní (PIQ). Dále je

uvedena informace o váze (Weight - v librách) a výšce (Height - v palcích) u testovaných. Protože u dvou osob chyběly údaje, pracujeme pouze s 38 pozorováními. Studenti jsou označeni podle pohlaví a očíslováni pro jednoduchou orientaci. Tabulka údajů je uvedena v Příloze D.

Když se podíváme na data, je nám jasné, že není dobré počítat v původních jednotkách. Proto tabulku nejprve přeškálujeme. Příkazy v softwaru jsou obdobné jako u předchozích příkladů, a proto je zbytečné je zde znovu uvádět. Uvedeme tedy pouze tabulku charakteristik jednotlivých hlavních komponent:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.5842328	1.2627884	0.54709289	0.52014131	0.44060808
Proportion of Variance	0.5155251	0.3275465	0.06148002	0.05557181	0.03987648
Cumulative Proportion	0.5155251	0.8430717	0.90455171	0.96012352	1.000000000

Znak Z_1 vysvětluje necelých 52 % a znak Z_2 skoro 33 % celkové variability, dohromady 84 % . Tato hodnota je opět vyhovující.



Z grafu je na první pohled patrné, že jednotlivé inteligence nesouvisí s výškou a váhou jedince, protože hodnota úhlu mezi těmito skupinkami (inteligence a tělesné proporce) se pohybuje kolem 90 stupňů. Dále vidíme, že jednotlivé druhy inteligence se svým významem neliší, je tedy zřejmé, že tyto tři proměnné jsou provázané. Též výška a váha mají na uspořádání pozorování v biplotu zhruba stejný vliv. Uspořádání prvků zprava doleva nám ve většině případů ukazuje celkovou úspěšnost. V tomto příkladu ale porovnáváme dvě nekorelované skupiny proměnných odlišného typu. Tedy celkovou "úspěšnost" zde nelze jednoznačně určit a především v interpretaci výsledků musíme být velmi opatrní. Je potřeba pozorně sledovat směry jednotlivých šipek. Uspořádání prvků zleva doprava nám seřazuje testované jedince podle hodnoty celkové inteligence. Tedy čím více vlevo se daný responent nachází, tím vyšší má IQ. Naopak uspořádání prvků shora dolů nám ukazuje rozložení výšky a váhy. Nahoře jsou hubení a nízcí lidé, dole jsou těžší a vyšší. Obecně lze říci, že v horní polovině grafu je většina žen, muži jsou naopak dole.

5.4 Cigarety a rakovina v USA

Poslední příklad je opět převzatý z knihovny datových tabulek z internetu [18], dále zde využíváme informace z internetových stránek [13], [15]. Tento příklad použijeme k posouzení, zda můžeme nějaké proměnné vynechat bez výraznějšího ovlivnění celkových výsledků.

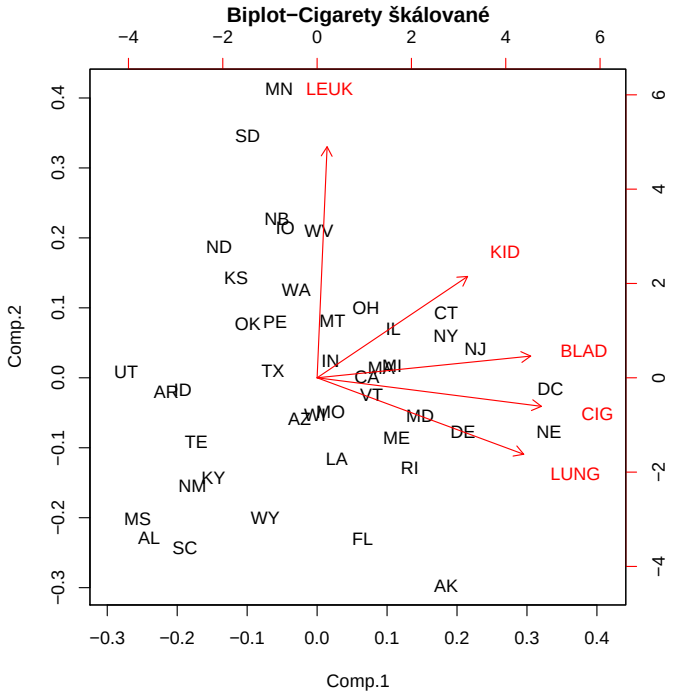
Data (Příloha E) znázorňují počet cigaret prodaných ve vybraných 43 státech USA včetně the District of Columbia v roce 1960, přepočítáno vždy na jednoho obyvatele daného státu. Dále je zde uvedena úmrtnost na 100 obyvatel na různé formy rakoviny. Zkratky jednotlivých proměnných jsou tyto: CIG = Počet vykouřených cigaret na 1 obyvatele, BLAD = Úmrtnost na 100 obyvatel na rakovinu močového měchýře, LUNG = Úmrtnost na rakovinu plic, KID = Úmrtnost na rakovinu ledvin, LEUK = Úmrtnost na leukémii.

Příkazy v softwaru se opět nijak neliší od ostatních. Vzhledem k různým velikostem hodnot proměnných použijeme škálování. Tabulka charakteristik jed-

notlivých hlavních komponent vypadá následovně:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	1.6051368	1.0631717	0.8227164	0.52739123	0.4738793
Proportion of Variance	0.5272764	0.2313242	0.1385206	0.05692198	0.0459568
Cumulative Proportion	0.5272764	0.7586006	0.8971212	0.95404320	1.00000000

Znak Z_1 vysvětluje 53 % a znak Z_2 23 % celkové variability. Součet 76 % tedy opět vyhovuje.



Uspořádání prvků zleva doprava je dáno počtem cigaret a výskytem rakoviny (od nejmenšího výskytu po největší), jako nejlepší nám tedy z tohoto pohledu vychází stát Utah. To je zřejmé i z tabulky - hodnota počtu prodaných cigaret je nejmenší. Uspořádání zdola nahoru ukazuje zasažení dané oblasti leukémií. Dole jsou státy s nejmenším výskytem (Alaska), naopak nahoře jsou státy s vysokým počtem případů této nemoci (Minnesota). Je to zřejmé i z postavení šipky leukémie - ta ukazuje opravdu zdola nahoru. Úhel mezi leukémií a cigaretami

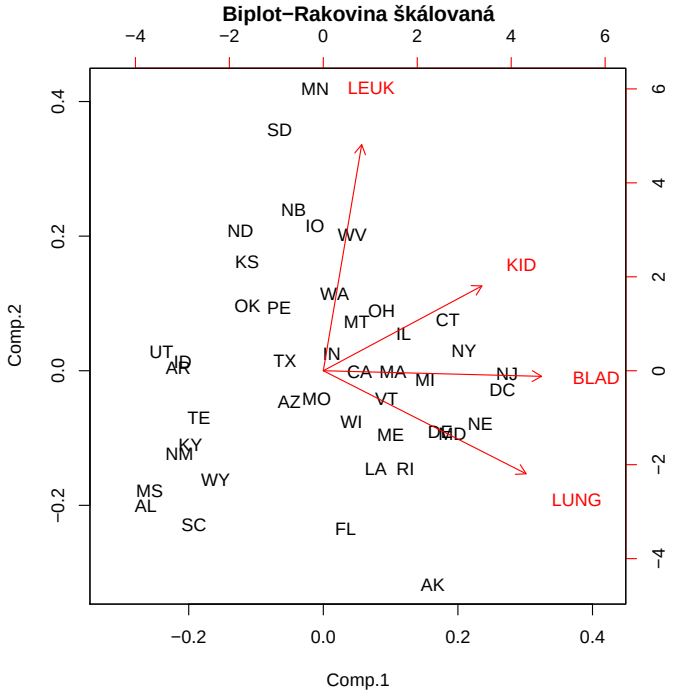
je 90 stupňů, tyto proměnné jsou tedy nekorelované. Tento závěr je podle mého názoru logický, protože kouření nemá vliv na vznik leukémie. Šipka leukémie jde obecně opačným směrem než šipky ostatní. Můžeme to interpretovat tak, že leukémie je zcela specifický druh rakoviny. Úhel mezi cigaretami a rakovinou plic je stejný jako úhel mezi cigaretami a rakovinou močového měchýře. První vztah není nijak zvláštní, protože plíce jsou hlavním orgánem postiženým u kuřáků. Druhý vztah může být však pro někoho na první pohled překvapující, ovšem podle posledních studií je kouření významným rizikovým faktorem pro vznik rakoviny močového měchýře.

Nakonec můžeme najít na tomto grafu ještě jednu zajímavost. Uspořádání prvků částečně odpovídá uspořádání států na mapě. Například stát Florida je jak v grafu, tak na mapě, na okraji. Ještě lépe je to vidět u Aljašky (Alasky), která je v grafu mimo ostatní státy - vpravo dole. Z geografického hlediska stojí také mimo ostatní státy USA, nachází se totiž na sever od území Kanady. Dále státy Minnesota, South Dakota a North Dakota jsou u sebe, Dakoty si jen vyměnily místo (North Dakota je v grafu níž) a poslední příklad New York, Connecticut a New Jersey jsou malé státy na severovýchodě. V této oblasti se nachází také v grafu. Abychom si mohli ověřit dané závěry, v Příloze F je umístěna mapa Spojených států amerických.

Nyní si ukážeme, co se stane, když proměnnou CIG vynecháme. Opět škálujeme a tabulku charakteristik jednotlivých hlavních komponent následně dostaneme v tomto tvaru:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.3630421	1.0535728	0.8207924	0.51719566
Proportion of Variance	0.4752726	0.2839575	0.1723419	0.06842802
Cumulative Proportion	0.4752726	0.7592301	0.9315720	1.00000000

Znak Z_1 vysvětluje 47.5 % celkové variability, tedy méně než v předchozím případu, znak Z_2 28 %. Součet je 76 % (stejný jako s cigaretami).



Šipka leukémie se nyní posunula blíže k ostatním. Tedy vztah mezi ní a ostatními proměnnými se mírně zesílil, ostatní úhly zůstaly zhruba stejné. Uspořádání prvků zleva doprava vyjadřuje opět množství výskytů rakovin (od nejmenšího po největší), ale už bez počtu cigaret. Tentokrát je "nejlepším" státem Mississippi, Utah je až na druhém místě. Objekty se posunuly jen některé a pouze nepatrně, tedy poznatky o geografickém rozložení zůstávají zachovány.

Z daných závěrů můžeme usuzovat, že na tento graf proměnná cigarety nemá téměř žádný vliv, proto je celkem jedno, jestli ji vynecháme. Já osobně bych ji ale asi zachovala. Ukáže nám totiž vliv kouření na vznik jednotlivých druhů rakovin. I když některé závěry nejsou na první pohled zřejmé (například již zmíněná rakovina močového měchýře), zdá se, že jsou správné.

Závěr

Přiznávám, že psaní toho textu pro mě někdy nebylo jednoduché, především se "prokousat" některými částmi teorie. Naopak při části praktické jsem se "vyřádila". Hledání vhodných příkladů sice bylo ze začátku složité, nakonec jsem ale, doufám, našla zajímavé nejen z matematického, ale i interpretačního hlediska. Nejzajímavější pro mě byl příklad o zemědělství; ne, že bych skutečnosti obsažené v tabulce neznala, spíše mě zaujalo, jak se s daty dá pracovat a kolik existuje možných postupů. Možná i proto je tento příklad nejobsáhlejší.

Když jsem se seznamovala s tématem a viděla první příklady, měla jsem pocit, že je interpretace biplotu velmi jednoduchá. Základní pravidla pro "výklad" grafů jsou daná. Ve svých příkladech jsem se ale přesvědčila, že ačkoliv graf vypadá jednoduše, výklad může být někdy obtížnější. To je také vidět v příkladu Inteligence a tělesné proporce. Moje slova z úvodu, že grafická reprezentace může být jednoduchá i pro matematického laika, je tedy přece jen nutné poněkud relativizovat. Přesněji bych tedy tuto myšlenku formulovala tak, že i laik se může v biplotu vyznat a "číst" v něm, ovšem tento musí mít též kvalitní komentář příslušného odborníka.

Příloha A

Výsledky studentů prvního ročníku vysoké školy technického směru.

Student	ME	AG	LA	AN	ES	Student	ME	AG	LA	AN	ES
1	77	82	67	67	81	37	46	56	57	49	32
2	63	78	80	70	81	38	45	42	55	56	40
3	75	73	71	66	81	39	42	60	54	49	33
4	55	72	63	70	68	40	40	63	53	54	25
5	63	63	65	70	63	41	23	55	59	53	44
6	53	61	72	64	73	42	48	48	49	51	37
7	51	67	65	65	68	43	41	63	49	46	34
8	59	70	68	62	56	44	46	52	53	41	40
9	62	60	58	62	70	45	46	61	46	38	41
10	64	72	60	62	45	46	40	57	51	52	31
11	52	64	60	63	54	47	49	49	45	48	39
12	55	67	59	62	44	48	22	58	53	56	41
13	50	50	64	55	63	49	35	60	47	54	33
14	65	63	58	56	37	50	48	56	49	42	32
15	31	55	60	57	73	51	31	57	50	54	34
16	60	64	56	54	40	52	17	53	57	43	51
17	44	69	53	53	53	53	49	57	47	39	26
18	42	69	61	55	45	54	59	50	47	15	46
19	62	46	61	57	45	55	37	56	49	28	45
20	31	49	62	63	62	56	40	43	48	21	61
21	44	61	52	62	46	57	35	35	41	51	50
22	49	41	61	49	64	58	38	44	54	47	24
23	12	58	61	63	67	59	43	43	38	34	49
24	49	53	49	62	47	60	39	46	46	32	43
25	54	49	56	47	53	61	62	44	36	22	42
26	54	53	46	59	44	62	48	38	41	44	33
27	44	56	55	61	36	63	34	42	50	47	29
28	18	44	50	57	81	64	18	51	40	56	30
29	46	52	65	50	35	65	35	36	46	48	29
30	32	45	49	57	64	66	59	53	37	22	19
31	30	69	50	52	45	67	41	41	43	30	33
32	46	49	53	59	37	68	31	52	37	27	40
33	40	27	54	61	61	69	17	51	52	35	31
34	31	42	48	54	68	70	34	30	50	47	36
35	36	59	51	45	51	71	46	40	47	29	17
36	56	40	56	54	35	72	10	46	36	47	39

Student	ME	AG	LA	AN	ES	Student	ME	AG	LA	AN	ES
73	46	37	45	15	30	81	3	9	51	47	40
74	30	34	43	46	18	82	7	51	43	17	22
75	13	51	50	25	31	83	15	40	43	23	18
76	49	50	38	23	9	84	15	38	39	28	17
77	18	32	31	45	40	85	5	30	44	36	18
78	8	42	48	26	40	86	12	30	32	35	21
79	23	38	36	48	15	87	5	26	15	20	20
80	30	24	43	33	25	88	0	40	21	9	14

Příloha B

Hektarové výnosy sklizně hlavních zemědělských plodin podle krajů v roce 2007.

Území	SPZ	Pšenice	Ječmen	Brambory	Řepka	Slunečnice	Pícniny
Hlavní město Praha	A	5.29	3.78	24.26	3.18	2.13	6.34
Středočeský	S	5.02	3.88	26.19	3.09	2.13	6.48
Jihočeský	C	4.75	3.91	28.37	3.00	2.00	6.35
Plzeňský	P	4.74	4.07	27.77	3.03	2.01	6.33
Karlovarský	K	4.73	3.85	28.77	3.01	2.00	5.40
Ústecký	U	4.93	3.69	24.88	3.11	2.13	6.49
Liberecký	L	4.64	3.72	27.79	3.03	2.10	5.73
Královehradecký	H	5.02	4.20	26.09	3.10	2.12	6.27
Pardubický	E	4.93	3.87	27.41	3.07	2.16	6.13
Vysočina	J	4.75	3.73	28.59	3.00	2.04	5.99
Jihomoravský	B	4.54	3.30	21.73	3.02	2.12	6.30
Olomoucký	M	5.23	3.88	25.10	3.14	2.28	6.62
Zlínský	Z	4.86	3.81	25.58	3.09	2.23	6.50
Moravskoslezský	T	4.87	3.78	26.36	3.05	2.29	5.69

Údaje jsou uvedeny v tunách.

Příloha C

Přeškálovaná tabulka Hektarové výnosy sklizně hlavních zemědělských plodin podle krajů v roce 2007.

Území	SPZ	Pšenice	Ječmen	Brambory	Řepka	Slunečnice	Pícniny
Hlavní město Praha	A	1.94053570	-0.19378343	-1.062253797	2.07219013	0.06050784	0.4264185
Středočeský	S	0.66705915	0.29948349	-0.080985503	0.44034040	0.06050784	0.8169700
Jihočeský	C	-0.60641741	0.44746356	1.027390083	-1.19150932	-1.31604545	0.4543150
Plzeňský	P	-0.65358320	1.23669064	0.722332582	-0.64755941	-1.21015674	0.3985220
Karlovarský	K	-0.70074900	0.15150341	1.230761750	-1.01019269	-1.31604545	-2.1958560
Ústecký	U	0.24256696	-0.63772366	-0.747027713	0.80297367	0.06050784	0.8448665
Liberecký	L	-1.12524119	-0.48974359	0.732501165	-0.64755941	-0.25715831	-1.2752703
Královehradecký	H	0.66705915	1.87793763	-0.131828420	0.62165704	-0.04538088	0.2311427
Pardubický	E	0.24256696	0.25015680	0.539298082	0.07770713	0.37817398	-0.1594088
Vysočina	J	-0.60641741	-0.44041689	1.139244500	-1.19150932	-0.89249060	-0.5499603
Jihomoravský	B	-1.59689917	-2.56146465	-2.348579591	-0.82887605	-0.04538088	0.3148323
Olomoucký	M	1.65754091	0.29948349	-0.635173296	1.34692358	1.64883856	1.2075215
Zlínský	Z	-0.08759363	-0.04580336	-0.391127296	0.44034040	1.11939498	0.8727631
Moravskoslezský	T	-0.04042783	-0.19378343	0.005447455	-0.28492614	1.75472727	-1.3868564

Bezrozměrná jednotka.

Příloha D

IQ a tělesné proporce studentů Jihozápadní univerzity.

Gender	FSIQ	VIQ	PIQ	Weight	Height
F01	133	132	124	118	64.5
M01	139	123	150	143	73.3
M02	133	129	128	172	68.8
F02	137	132	134	147	65.0
F03	99	90	110	146	69.0
F04	138	136	131	138	64.5
F05	92	90	98	175	66.0
M03	89	93	84	134	66.3
M04	133	114	147	172	68.8
F06	132	129	124	118	64.5
M05	141	150	128	151	70.0
M06	135	129	124	155	69.0
F07	140	120	147	155	70.5
F08	96	100	90	146	66.0
F09	83	71	96	135	68.0
F10	132	132	120	127	68.5
M07	100	96	102	178	73.5
F11	101	112	84	136	66.3
M08	80	77	86	180	70.0
M09	97	107	84	186	76.5
F12	135	129	134	122	62.0
M10	139	145	128	132	68.0
F13	91	86	102	114	63.0
M11	141	145	131	171	72.0
F14	85	90	84	140	68.0
M12	103	96	110	187	77.0
F15	77	83	72	106	63.0
F16	130	126	124	159	66.5
F17	133	126	132	127	62.5
M13	144	145	137	191	67.0
M14	103	96	110	192	75.5
M15	90	96	86	181	69.0
F18	83	90	81	143	66.5
F19	133	129	128	153	66.5
M16	140	150	124	144	70.5
F20	88	86	94	139	64.5
M17	81	90	74	148	74.0
M18	89	91	89	179	75.5

Příloha E

Počet vykouřených cigaret a výskyt 4 druhů rakoviny ve vybraných státech USA.

Stát	Zkratka	CIG	BLAD	LUNG	KID	LEUK
Alabama	AL	18.20	2.90	17.05	1.59	6.15
Arizona	AZ	25.82	3.52	19.80	2.75	6.61
Arkansas	AR	18.24	2.99	15.98	2.02	6.94
California	CA	28.60	4.46	22.07	2.66	7.06
Connecticut	CT	31.10	5.11	22.83	3.35	7.20
Delaware	DE	33.60	4.78	24.55	3.36	6.45
District of Columbia	DC	40.46	5.60	27.27	3.13	7.08
Florida	FL	28.27	4.46	23.57	2.41	6.07
Idaho	ID	20.10	3.08	13.58	2.46	6.62
Illinois	IL	27.91	4.75	22.80	2.95	7.27
Indiana	IN	26.18	4.09	20.30	2.81	7.00
Iowa	IO	22.12	4.23	16.59	2.90	7.69
Kansas	KS	21.84	2.91	16.84	2.88	7.42
Kentucky	KY	23.44	2.86	17.71	2.13	6.41
Louisiana	LA	21.58	4.65	25.45	2.30	6.71
Maine	ME	28.92	4.79	20.94	3.22	6.24
Maryland	MD	25.91	5.21	26.48	2.85	6.81
Massachusetts	MA	26.92	4.69	22.04	3.03	6.89
Michigan	MI	24.96	5.27	22.72	2.97	6.91
Minnesota	MN	22.06	3.72	14.20	3.54	8.28
Mississippi	MS	16.08	3.06	15.60	1.77	6.08
Missouri	MO	27.56	4.04	20.98	2.55	6.82
Montana	MT	23.75	3.95	19.50	3.43	6.90
Oregon	NB	23.32	3.72	16.70	2.92	7.80
Nebraska	NE	42.40	6.54	23.03	2.85	6.67
New Jersey	NJ	28.64	5.98	25.95	3.12	7.12
New Mexico	NM	21.16	2.90	14.59	2.52	5.95
New York	NY	29.14	5.30	25.02	3.10	7.23
North Dakota	ND	19.96	2.89	12.12	3.62	6.99
Ohio	OH	26.38	4.47	21.89	2.95	7.38
Oklahoma	OK	23.44	2.93	19.45	2.45	7.46
Pennsylvania	PE	23.78	4.89	12.11	2.75	6.83
Rhode Island	RI	29.18	4.99	23.68	2.84	6.35
South Carolina	SC	18.06	3.25	17.45	2.05	5.82
South Dakota	SD	20.94	3.64	14.11	3.11	8.15
Tennessee	TE	20.08	2.94	17.60	2.18	6.59
Texas	TX	22.57	3.21	20.74	2.69	7.02

Stát	Zkratka	CIG	BLAD	LUNG	KID	LEUK
Utah	UT	14.00	3.31	12.01	2.20	6.71
Vermont	VT	25.89	4.63	21.22	3.17	6.56
Washington	WA	21.17	4.04	20.34	2.78	7.48
Wisconsin	WI	21.25	5.14	20.55	2.34	6.73
West Virginia	WV	22.86	4.78	15.53	3.28	7.38
Wyoming	WY	28.04	3.20	15.92	2.66	5.78
Alaska	AK	30.34	3.46	25.88	4.32	4.90

Příloha F

Spojené státy americké

Literatura

[1] Anděl, J., *Matematická statistika*, 1. vydání, Praha, Praha, SNTL + Alfa, 1978.

[2] Anděl, J., *Statistické metody*, 3. vydání, Praha: MATFYZPRESS, 2003.

[3] Anděl, J., *Základy matematické statistiky*, 2. opravené vydání, Praha: MATFYZPRESS, 2007.

[4] Filzmoser, P., *Multivariate Statistik*, TU Wien, 2007.

[5] Gabriel, K. R., *The biplot graphic display of matrices with application to principal component analysis*, Biometrika, 1971.

[6] Gower, J.C., Hand, D.J, *Biplots*, Chapman & Hall, London, UK, 1996.

[7] Hebák, P. a kol., *Vícerozměrné statistické metody [3]*, 1. vydání, Praha: INFORMATORIUM, 2005.

[8] Hebák, P., Hustopecký, J., *Vícerozměrné statistické metody s aplikacemi*, 1. vydání, Praha, SNTL + Alfa, 1987.

[9] Jukl, M., *Lineární algebra - Euklidovské vektorové prostory, Homomorfizmy vektorových prostorů*, 1. vydání, Olomouc: Univerzita Palackého Olomouc, 2006.

[10] Mahalanobis distance [online], dostupné z: *http://en.wikipedia.org/wiki/Mahalanobis_distance* [citováno 7. 5. 2009].

[11] Massart, D.L, Vander Heyden, Y., From tables to visuals: PCA I, PCA II, Vrije Universiteit Brussel, Belgium, článek v časopise [online], dostupné z: *http://chromatographyonline.findanalytichem.com/lcgc/data/ articlestandard/lcgceurope/462004/133038/article.pdf* [citováno 6. 4. 2009].

[12] Nápověda k softwaru R [online], dostupné z: *http://www.r-project.org/* [citováno 20. 10. 2009].

[13] Rakovina močového měchýře [online], dostupné z: *http://theses.cz/id/jawkgj* [citováno 12. 10. 2009].

[14] Singular value decomposition [online], dostupné z: *http://en.wikipedia.org/wiki/Singular_Value_Decomposition* [citováno 16. 4. 2009].

[15] Spojené státy americké [online], dostupné z: *http://cs.wikipedia.org/wiki/Spojené_státy_americké* [citováno 14. 9. 2009].

[16] Statistická ročenka České republiky 2008 - Zěmědělství, tabulka 14-9 Hektarové výnosy sklizně hlavních zemědělských plodin podle krajů v roce 2007 [online], dostupné z: *http://www.czso.cz/csu/2008edicniplan.nsf/kapitola/10n1-08-2008-1400* [citováno 20. 7. 2009].

[17] The data and story library - Brain Size [online], dostupné z: *http://lib.stat.cmu.edu/DASL/Datafiles/Brainsize.html* [citováno 1. 8. 2009].

[18] The data and story library - Smoking and Cancer [online], dostupné z: *http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html* [citováno 18. 8. 2009].