UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULTA

DIPLOMOVÁ PRÁCE

Vícerozměrné škálování

Katedra matematické analýzy a aplikací matematiky
Vedoucí diplomové práce: **doc. RNDr. Karel Hron, Ph.D.**
Vypracovala: **Bc. Nikola Štefelová**
Studijní program: N1101 Matematika
Studijní obor Matematika a její aplikace
Forma studia: prezenční
Rok odevzdání: 2016

BIBLIOGRAFICKÁ IDENTIFIKACE

Autor: Bc. Nikola Štefelová

Název práce: Vícerozměrné škálování

Typ práce: Diplomová práce

Pracoviště: Katedra matematické analýzy a aplikací matematiky

Vedoucí práce: doc. RNDr. Karel Hron, Ph.D.

Rok obhajoby práce: 2016

Abstrakt: Vícerozměrné škálování (MDS) je název pro skupinu statistických metod založených na redukci mnohorozměrného prostoru objektů (pozorování) a průzkumové analýze vztahů mezi nimi. MDS pracuje s různými typy relací mezi objekty, přičemž nejčastěji jde o číselně vyjádřenou vzájemnou nepodobnost. Úkolem MDS je na základě daných nepodobností mezi objekty získat jejich reprezentaci v prostoru nízké dimenze. Základním dělením metod je to na metrické a nemetrické MDS. Cílem diplomové práce je provést ucelený přehled nejznámějších metod tohoto typu a aplikovat je na reálná data.

Klíčová slova: vícerozměrné škálování, metrické, nemetrické, proximitní matice, vzdálenost, nepodobnost, disparita, Stress, korespondenční analýza

Počet stran: 59

Počet příloh: 1

Jazyk: český

BIBLIOGRAPHICAL IDENTIFICATION

Author: Bc. Nikola Štefelová

Title: Multidimensional Scaling

Type of thesis: Master's

Department: Department of Mathematical Analysis and Application of Mathematics

Supervisor: doc. RNDr. Karel Hron, Ph.D.

The year of presentation: 2016

Abstract: Multidimensional scaling (MDS) is the name for a set of statistical methods based on reduction of a hight-dimensional space of objects (observations) and an exploratory analysis of their relationships. MDS works with various types of relations between objects, most often with numerically expressed mutual dissimilarity. The task of MDS is to use the given dissimilarities between the objects to get their representation in a low-dimensional space. The basic division of the methods is that into metric and non-metric MDS. The aim of this thesis is to provide a comprehensive overview of the best-known MDS methods and apply them to real data.

Key words: multidimensional scaling, metric, nonmetric, proximity matrix, distance, dissimilarity, disparity, Stress, correspondence analysis

Number of pages: 59

Number of appendices: 1

Language: Czech

Prohlášení

Prohlašuji, že jsem diplomovou práci zpracovala samostatně pod vedením pana doc. RNDr. Karla Hrona, Ph.D. a všechny použité zdroje jsem uvedla v seznamu literatury.

V Olomouci dne ....

podpis

Obsah

Seznam obrázků

Seznam tabulek

Poděkování

Ráda bych na tomto místě poděkovala svému vedoucímu diplomové práce panu doc. RNDr. Karlu Hronovi, Ph.D. za vstřícný přístup, čas strávený při konzultacích a cenné rady a připomínky.

Úvod

Jedním z hlavních problémů, které řeší statistické metody pracující s mnohorozměrnými daty, je otázka optimálního snížení dimenze dat napomáhajícího k objasnění vztahů mezi pozorovanými objekty. Mezi nejvýznamnější metody tohoto typu patří analýza hlavních komponent a faktorová analýza. Obdobnou úlohu řeší i vícerozměrné škálování (dále budeme používat už jen zkratku MDS, z angl. Multidimensional Scaling). Místo datové matice však pracuje s proximitní maticí vyjadřující nepodobnost mezi objekty.

Úkolem této diplomové práce je podat ucelený přehled nejvýznamnějších metod MDS a aplikovat je na reálná data. Obsah práce je rozdělen do šesti kapitol.

První kapitola nás uvede do problematiky MDS. Popíšeme si situaci a úkol, který řeší, dozvíme se, jak je to s jednoznačností řešení a jak a podle čeho lze metody MDS dělit. Také se budeme zabývat historií této metody a možnostmi jejího využití. Blíže se pak podíváme na proximitní matici, která stojí na počátku každé úlohy řešené pomocí MDS.

V druhé kapitole se zaměříme na skupinu metrických metod. Představíme si klasický model MDS a jeho méně restriktivní variantu. Dále si ukážeme, jak vhodně zvolit dimenzi řešení, a na ilustrativním příkladě si přiblížíme výpočetní algoritmus.

Ve třetí kapitole se budeme věnovat nemetrickému MDS. Vysvětlíme si, co to jsou disparity a jakým postupem je můžeme určit. Poté si objasníme algoritmus vedoucí k nalezení řešení, který je založen na minimalizaci ztrátové funkce *Stress*. Budeme se také zabývat otázkou správné volby dimenze řešení a významem Shepardova diagramu. Získané poznatky budeme demonstrovat na ilustrativním příkladě.

Ve čtvrté kapitole se budeme koncentrovat na to, jak lze v úlohách MDS využít statistický software R. Seznámíme se s funkcemi, jež slouží k řešení metrických i nemetrických úloh, a prozkoumáme jejich parametry.

V páté kapitole budeme metodu MDS používat na reálných datech. Ukážeme si příklad použití metrického i nemetrického MDS.

V poslední (šesté) kapitole se podíváme na aplikaci MDS v jiné mnohorozměrné statistické metodě a to korespondenční analýze. Dozvíme se, jak lze k průzkumu vztahů mezi kategoriálními proměnnými v kontingenčních tabulkách uplatnit duální škálování. Užitečnost této metody předvedeme na příkladě.

V textu práce budeme používat následující značení. Matice budeme značit velkým tučným písmem (tedy například $\mathbf{M}$, případně $\mathbf{M}^{(n\times m)}$ pokud budeme chtít zdůraznit její rozměry). Vektory budeme označovat malým tučným písmem, přičemž $\mathbf{v}$ bude představovat sloupcový vektor.

1 Seznámení se s metodou MDS

Text této kapitoly se opírá o publikace [2], [3], [4] a [5].

Uvažujme situaci, kdy máme n objektů a na každém z nich napozorovaných p proměnných (standardně $n > p$). Na objekty lze pohlížet jako na n bodů v $p-$rozměrném prostoru o souřadnicích $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top,\ i = 1, \ldots, n$. *Matici souřadnic objektů* (datovou matici) $\mathbf{X}^{(n\times p)}$ získáme uspořádáním $\mathbf{X}= (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$.

V úlohách vícerozměrného škálování nemáme typicky na vstupu matici $\mathbf{X}$, ale tzv. *proximitní matici* $\mathbf{\Delta}^{(n\times n)}$, která je teoreticky z jisté matice souřadnic $\mathbf{X}$ (obvykle nám neznámého rozměru p) získána. Prvky matice $\mathbf{\Delta}$ představují číselně vyjádřenou nepodobnost mezi objekty.

Úkolem MDS je na základě daných nepodobností mezi objekty vytvořit jejich prostorovou reprezentaci. V podstatě jde o techniku spočívající v redukci dimenze dat, neboť cílem je najít takovou množinu bodů v prostoru nízké dimenze q, jejichž vzájemné vzdálenosti co nejlépe odpovídají těm v relativní konfiguraci bodů v prostoru vyšší dimenze p. Získanou matici souřadnic budeme značit $\mathbf{Y}^{(n\times q)}$. Pro snadné grafické zobrazení se snažíme volit $q = 2$, eventuálně $q = 3$. Graf s výslednými body nazýváme *mapou objektů*. Mapa objektů nám má pomoci ozřejmit jejich vzájemné vztahy, jež z proximitní matice nemusí být patrné.

Je třeba poznamenat, že MDS dává řešení jednoznačné až na rotaci, reflexi a translaci. Zřejmě se vzdálenosti mezi výslednými body nezmění, pokud konfiguraci bodů otočíme, zrcadlově převrátíme, či posuneme. Tedy, je-li $\mathbf{y}_i,\ i = 1, \ldots, n$, řešení MDS, pak také $\mathbf{z}_i = \mathbf{A}\mathbf{y}_i + \mathbf{b},\ i = 1, \ldots, n$, kde $\mathbf{A}$ je ortogonální matice a $\mathbf{b}$ vektor posunutí, představuje řešení MDS. Problém s lokací konfigurace se obvykle řeší umístěním jejího středu do počátku, tj. $\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i = \mathbf{0}$.

Podle typu dat v proximitní maticí se MDS dělí na *metrické* a *nemetrické*. V metrickém MDS představují nepodobnosti kvantitativní data v podobě vzdáleností, zatímco v nemetrickém MDS mají nepodobnosti charakter kvalitativních dat ordinálního typu.

Kořeny MDS sahají do 30. let 20. století. Mezi vědce, kteří se zasloužili o rozvoj metrických metod, patří především G. Young, A. Householder, W. Torgerson, S. Messick, R. Abelson a L. Tucker. Za zakladatele nemetrických metod považujeme R. Sheparda, na jehož práci navázal J. Kruskal.

MDS lze využít všude tam, kde nás může zajímat otázka, proč jsou určité předměty zkoumání vnímány stejně, či odlišně, tedy např. v psychologii, sociologii, politologii či marketingu.

1.1 Proximitní matice

Vstupem do algoritmu MDS je proximitní matice (matice relací mezi objekty) $\mathbf{\Delta}^{(n\times n)} = (\delta_{ij})$, kde δ_{ij} je číslo vyjadřující *nepodobnost* mezi $i-$tým a $j-$tým objektem. Matici $\mathbf{\Delta}$ též nazýváme *maticí nepodobností*.

Nepodobnost nemusí být jediným určením vztahu mezi dvěma objekty. Někdy může proximitní matice obsahovat *podobnosti* s_{ij}, jež se ovšem obvykle převádí na nepodobnost. Uvažujeme-li například δ_{ij} v intervalu $\langle 0, 1\rangle$, pak nejjednodušší transformací je $\delta_{ij} = 1 - s_{ij}$.

V klasickém modelu metrického MDS považujeme každé číslo z výchozí proximitní matice za *vzdálenost* d_{ij} mezi $i-$tým a $j-$tým objektem. Matici $\mathbf{\Delta}$ pak ztotožňujeme s *maticí vzdáleností* $\mathbf{D}^{(n\times n)} = (d_{ij})$. V MDS se nejčastěji používá vzdálenost euklidovská, ale lze počítat i s jinými typy. Uveďme přehled těch nejužívanějších:

- *euklidovská vzdálenost*: $d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2}$,
- *čtvercová euklidovská vzdálenost*: $d_{ij} = \sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2$,
- *vážená euklidovská vzdálenost*: $d_{ij} = \sqrt{\sum_{k=1}^{p} w_k \left(x_{ik} - x_{jk}\right)^2}$, $\sum_{k=1}^{p} w_k = 1$,
- *Minkowského vzdálenost*: $d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{p} \left|x_{ik} - x_{jk}\right|^{\lambda}}$, $\lambda > 0$,

- *Mahalanobisova vzdálenost*: $d_{ij} = \sqrt{\big(\mathbf{x}_i - \mathbf{x}_j\big)^\top \mathbf{S}^{-1} \big(\mathbf{x}_i - \mathbf{x}_j\big)}$,
$$\mathbf{S} = \tfrac{1}{n-1} \sum_{i=1}^{n} \big(\mathbf{x}_i - \overline{\mathbf{x}}\big)\big(\mathbf{x}_i - \overline{\mathbf{x}}\big)^\top,$$

- *Hemmingova (Manhattanská) vzdálenost*: $d_{ij} = \sum_{k=1}^{p} \big|x_{ik} - x_{jk}\big|$,

- *Čebyševova vzdálenost*: $d_{ij} = \max\limits_{k=1,\ldots,p} \big|x_{ik} - x_{jk}\big|$.

$\boldsymbol{\Delta}$ je symetrickou maticí s nezápornými prvky, jejíž hlavní diagonálu tvoří nuly, tj.

$$\delta_{ij} \geq 0,\ \delta_{ii} = 0,\ \delta_{ij} = \delta_{ji},\ \forall i, j = 1, \ldots, n. \tag{1}$$

Ve skutečnosti nemusí být $\boldsymbol{\Delta}$ vždy symetrická. V takovém případě z ní symetrickou matici vytvoříme, např. úpravami $\delta_{ij} \leftarrow \frac{1}{2}(\delta_{ij} + \delta_{ji})$, pro ta i a j, která nesplňují třetí podmínku z (1).

Proto, abychom na nepodobnosti mohli pohlížet jako na metrické vzdálenosti, vyžadujeme, aby δ_{ij} splňovaly tzv. trojúhelníkovou nerovnost, tj.

$$\delta_{ij} \leq \delta_{ik} + \delta_{kj},\ \forall i, j, k = 1, \ldots, n. \tag{2}$$

Proximitní matici můžeme získat dotazováním, expertními odhady apod., případně ji vypočítat z datové matice, máme-li $\mathbf{X}$ k dispozici.

2 Metrické MDS

Obsah této kapitoly vychází ze zdrojů [2], [3], [4] a [5].

V metrickém MDS jsou nepodobnosti δ_{ij} v proximitní matici $\mathbf{\Delta}^{(n\times n)}$ vyjádřeny pomocí vzdáleností d_{ij}.

Máme-li k dispozici souřadnice vyjadřující umístění objektů v $p-$rozměrném prostoru, pak snadno vypočítáme jejich vzájemné vzdálenosti. Metrické MDS nám umožní vyřešit opačný problém. S využitím daných vzdáleností mezi objekty nalezneme souřadnice odpovídající jejích poloze. Pro zjednodušení grafického zobrazení řešení hledáme tyto souřadnice v prostoru dimenze $q < p$, přičemž je žádoucí, aby $q = 2$, případně $q = 3$.

Řešení se hledá tak, aby euklidovské vzdálenosti mezi body v redukovaném prostoru d^*_{ij} byly nejlepší možnou aproximací daných nepodobností δ_{ij} a tedy aby byla minimální míra vzájemného nesouladu S:

$$S = \sum_{i<j} (\delta_{ij} - d^*_{ij})^2. \tag{3}$$

Na výpočetní algoritmus metrického MDS se podíváme z hlediska klasického modelu W. Torgersona.

2.1 Klasické MDS

Klasický Torgersenův model metrického MDS je založen na předpokladu shody nepodobností δ_{ij} a vzdáleností d_{ij}, které jsou euklidovské, tj.

$$\delta_{ij} = d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}. \tag{4}$$

Proximitní maticí $\mathbf{\Delta}$ je tedy matice vzdáleností $\mathbf{D}^{(n\times n)} = (d_{ij})$, jež je euklidovská.

Matici souřadnic $\mathbf{Y}^{(n\times q)}$ odvodíme z euklidovské matice $\mathbf{D}$ pomocí transformace $\mathbf{D}$ na *matici skalárních součinů* $\mathbf{B}^{(n\times n)}$ a následného spektrálního rozkladu $\mathbf{B}$.

Známe-li matici $\mathbf{X}^{(n\times p)}$, pak je matice $\mathbf{B}$ dána vztahem

$$\mathbf{B} = \mathbf{X}\mathbf{X}^\top \tag{5}$$

a pro její prvky b_{ij} platí

$$b_{ij} = \sum_{k=1}^{p} x_{ik}x_{jk}. \tag{6}$$

Z (4) a (6) plyne, že čtvercové vzdálenosti lze pomocí prvků $\mathbf{B}$ zapsat jako

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}. \tag{7}$$

Za předpokladu, že $\mathbf{X}$ je centrovaná ($\overline{\mathbf{x}} = \mathbf{0}$), je $\sum_{i=1}^{n} x_{ik} = 0$, $\forall k = 1, \ldots, p$ a z (6) vyplývá, že $\sum_{i=1}^{n} b_{ij} = 0$, $\forall j = 1, \ldots, n$.

Sumací (7) přes i, přes j a přes i a j dostaneme

$$\begin{aligned}
\sum_{i=1}^{n} d_{ij}^2 &= \sum_{i=1}^{n} b_{ii} + nb_{jj}, \\
\sum_{j=1}^{n} d_{ij}^2 &= \sum_{i=1}^{n} b_{ii} + nb_{ii}, \\
\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 &= 2n\sum_{i=1}^{n} b_{ii}.
\end{aligned} \tag{8}$$

S využitím (7) a (8) můžeme nyní prvky matice $\mathbf{B}$ vyjádřit pomocí čtvercových vzdáleností jako

$$b_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2), \tag{9}$$

kde

$$\begin{aligned}
d_{i\bullet}^2 &= \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2, \\
d_{\bullet j}^2 &= \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2, \\
d_{\bullet\bullet}^2 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2.
\end{aligned} \tag{10}$$

Označíme-li

$$a_{ij} = -\frac{1}{2}d_{ij}^2, \tag{11}$$

dostaneme

$$b_{ij} = a_{ij} - a_{i\bullet} - a_{\bullet j} + a_{\bullet\bullet}, \tag{12}$$

kde

$$\begin{aligned} a_{i\bullet} &= \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \\ a_{\bullet j} &= \frac{1}{n}\sum_{i=1}^{n} a_{ij}, \\ a_{\bullet\bullet} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}. \end{aligned} \tag{13}$$

Matice $\mathbf{B}$ je pak dvojitě centrovanou verzí matice $\mathbf{A}^{(n\times n)} = (a_{ij})$, tzn.

$$\mathbf{B} = \mathbf{HAH}, \tag{14}$$

kde $\mathbf{H}^{(n\times n)} = \mathbf{I} - \frac{1}{n}\mathbf{J}$, $\mathbf{I}^{(n\times n)}$ je jednotková matice a $\mathbf{J}^{(n\times n)}$ je matice samých jedniček.

Již víme, jak z matice $\mathbf{D}$ získáme matici $\mathbf{B}$ a můžeme přistoupit k problematice spektrálního rozkladu $\mathbf{B}$.

Matice $\mathbf{B}$ je symetrická. Předpokládáme, že matice $\mathbf{X}$ má plnou hodnost, tedy $h(\mathbf{X}) = p$ a v důsledku (5) je i $h(\mathbf{B}) = p$. Proto pro $\mathbf{B}$ platí, že má p nenulových a $n - p$ nulových vlastních čísel. Jelikož je $\mathbf{D}$ euklidovská matice, je $\mathbf{B}$ pozitivně semidefinitní a neobsahuje tak žádná záporná vlastní čísla.

$\mathbf{B}$ tak můžeme rozložit na

$$\mathbf{B} = \mathbf{V_1}\mathbf{\Lambda_1}\mathbf{V_1}^{\top}, \tag{15}$$

kde $\mathbf{\Lambda_1}^{(p\times p)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ je diagonální matice kladných vlastních čísel $\mathbf{B}$, přičemž $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ a $\mathbf{V_1}^{(n\times p)} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$ je matice odpovídajících normovaných vlastních vektorů, a proto $\mathbf{v}_i^{\top}\mathbf{v}_i = 1,\ \forall i = 1, \ldots, n$.

Matici souřadnic, která udává řešení v $p-$rozměrném prostoru vypočítáme podle vztahu

$$\mathbf{X}^{(n\times p)} = \mathbf{V_1}\mathbf{\Lambda_1}^{\frac{1}{2}}. \tag{16}$$

.

Nejlepší $q-$rozměrnou reprezentaci ($q < p$) řešení (16) získáme tak, že v matici $\mathbf{\Lambda_1}$ ponecháme pouze prvních q vlastních čísel a v matici $\mathbf{V_1}$ odpovídající $q-$tici vlastních vektorů, tj.

$$\mathbf{Y}^{(n\times q)} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}, \tag{17}$$

kde $\mathbf{\Lambda}^{(q\times q)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ a $\mathbf{V}^{(n\times q)} = (\mathbf{v}_1, \ldots, \mathbf{v}_q)$.

Pokud proximitní matice obsahuje euklidovské vzdálenosti vypočítané z datové matice $\mathbf{X}$, pak je klasické MDS ekvivalentní metodě hlavních komponent. Proto bývá také nazýváno *analýzou hlavních os*. Souřadnice objektů na q hlavních osách odpovídají komponentním skóre objektů na q hlavních komponentách.

Zároveň jde o projekci objektů na q ortonormálních os při zachování maximální variability dat.

2.2 Méně restriktivní varianta klasického MDS

V praxi máme v proximitní matici zřídkakdy přesné euklidovské vzdálenosti mezi pozorovanými objekty. Např. máme-li dány vzdálenosti mezi městy, jsou to vzdálenosti silniční či letecké, které lze považovat pouze za přibližné vyjádření vzdáleností euklidovských.

Proto existuje další model metrického MDS, jenž na nepodobnosti klade méně omezující podmínky než ten klasický Torgersenův. Předpokládá, že výchozí nepodobnosti δ_{ij} jsou pouhými aproximacemi skutečných euklidovských vzdáleností d_{ij}, tj.

$$\delta_{ij} = d_{ij} + k = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} + k, \tag{18}$$

kde k je nějaká konstanta.

Použijeme stejný výpočetní algoritmus jako v klasickém modelu. Na počátku stojí proximitní matice $\mathbf{\Delta}^{(n\times n)} = (\delta_{ij})$. Transformujeme ji na matici skalárních součinů $\mathbf{B}^{(n\times n)}$ a matici souřadnic $\mathbf{Y}^{(n\times q)}$ získáme opět pomocí spektrálního rozkladu $\mathbf{B}$.

Platí, že minimálně jedno vlastní číslo $\mathbf{B}$ je nulové. Nyní však matice $\mathbf{B}$ nemusí být semidefinitní. Proto může obsahovat i záporná vlastní čísla, jež vedou na imaginární souřadnice.

Tento problém lze vyřešit dvěma způsoby. Buďto můžeme záporná vlastní čísla ignorovat, pokud je prvních q vlastních čísel kladných, nebo lze matici $\mathbf{B}$ změnit na pozitivně semidefinitní přičtením vhodné konstanty c ke všem mimodiagonálním prvkům matice $\mathbf{\Delta}$, tj.

$$\delta_{ij} \leftarrow \delta_{ij} + c \text{ pro } i \neq j. \tag{19}$$

2.3 Volba dimenze metrického řešení

Za jeden z klíčových úkolů MDS považujeme adekvátní zvolení dimenze řešení q. Na jednu stranu se snažíme vzít dostatečně malé q pro jednoduché grafické znázornění. Na druhou stranu ale můžeme výběrem příliš nízkého q získat zkreslené výsledky. Platí, že bez ztráty informace lze dostat řešení v prostoru dimenze $n-1$.

Výběr vhodného q souvisí s velikostí vlastních čísel λ_i matice skalárních součinů. Ta jsou seřazena sestupně. Teoreticky můžeme zvolit takové q, že počínaje λ_{q+1} se vlastní čísla „stabilizují“, tj. hodnota λ_{q+1} je podstatně menší než velikost předešlých vlastních čísel. V praxi obvykle volíme $q = 2$, eventuálně $q = 3$.

Vhodnost volby q pak posuzujeme podle kritéria

$$P_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \lambda_i}, \tag{20}$$

které udává míru vysvětlené variability při použití prvních q rozměrů. Přičemž za dobré řešení je považováno takové, kdy hodnota P_q je minimálně 0,8. (Tuto

hraniční hodnotu samozřejmě nelze brát za jedinou možnou, její volba se může lišit v souvislosti s podstatou řešeného problému.) Je zřejmé, že nejvíc variability skrývá první rozměr řešení, následuje druhý rozměr atd.

Pro matici **B**, jež není pozitivně semidefinitní, upravíme podobu kritéria (20) na

$$P_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} |\lambda_i|} \tag{21}$$

nebo

$$P_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{n} \max(\lambda_i, 0)}. \tag{22}$$

2.4 Ilustrativní příklad k metrickému MDS

Ilustrativní příklad 1 nám bude sloužit k demonstraci výpočetního algoritmu metrického a v další kapitole i nemetrického MDS.

Příklad 1. Vytvořte mapu objektů A, B, C, D a E, máte-li v tabulce 1 k dispozici jejich vzájemné nepodobnosti.

	A	B	C	D	E
A					
B	1,3				
C	1,5	1,8			
D	7,3	6,3	5,1		
E	7,4	6,1	7,1	9,3	

Tabulka 1: Nepodobnosti mezi objekty v příkladu 1

Pro metrický charakter úlohy budeme uvažovat, že nepodobnosti z tabulky 1 představují vzdálenosti (v km) mezi určitými stanovišti A, B, C, D a E. Naším cílem bude vytvořit orientační plán, který by znázorňoval vzájemnou polohu stanovišť.

Při řešení příkladu pomocí metrického MDS se budeme řídit postupem z tabulky 2.

1.	Ze vstupní proximitní matice $\mathbf{\Delta}^{(n\times n)} = (\delta_{ij})$ vytvoříme matici $\mathbf{A}^{(n\times n)} = (a_{ij})$, kde $a_{ij} = -\frac{1}{2}\delta_{ij}^2$.
2.	Dvojitým centrováním $\mathbf{A}$ získáme matici skalárních součinů $\mathbf{B}^{(n\times n)} = \mathbf{HAH}$, kde $\mathbf{H}^{(n\times n)} = \mathbf{I} - \frac{1}{n}\mathbf{J}$, $\mathbf{I}^{(n\times n)}$ je jednotková matice a $\mathbf{J}^{(n\times n)}$ je matice samých jedniček.
3.	Najdeme vlastní čísla $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ a odpovídající normované vlastní vektory $\mathbf{v}_1, \ldots, \mathbf{v}_n$ matice $\mathbf{B}$.
4.	Zvolíme dimenzi řešení q.
5.	Pokud je některé z prvních q vlastních čísel záporné, provedeme v matici $\mathbf{\Delta}$ úpravy $\delta_{ij} \leftarrow \delta_{ij} + c$ pro $i \neq j$ (c je vhodně zvolená konstanta tak, abychom dostali pozitivně semidefinitní matici $\mathbf{B}$) a vrátíme se ke kroku 1. Máme-li q vlastních čísel kladných, spočítáme hodnotu kritéria $P_q = \frac{\sum_{i=1}^{q}\lambda_i}{\sum_{i=1}^{n-1}\lvert\lambda_i\rvert}$. Jestliže je hodnota $P_q < 0{,}8$, provedeme $q \leftarrow q + 1$ a navrátíme se na začátek kroku 5.
6.	Vytvoříme matice $\mathbf{\Lambda}^{(q\times q)} = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ a $\mathbf{V}^{(n\times q)} = (\mathbf{v}_1, \ldots, \mathbf{v}_q)$.
7.	Vypočítáme matici souřadnic pro $q-$rozměrné řešení $\mathbf{Y}^{(n\times q)} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}$.
8.	Vytvoříme mapu objektů podle matice $\mathbf{Y}$.

Tabulka 2: Praktický výpočetní algoritmus metrického MDS

1. Z proximitní matice

$$\mathbf{\Delta} = \begin{pmatrix} 0{,}0 & 1{,}3 & 1{,}5 & 7{,}3 & 7{,}4 \\ 1{,}3 & 0{,}0 & 1{,}8 & 6{,}2 & 6{,}1 \\ 1{,}5 & 1{,}8 & 0{,}0 & 5{,}1 & 7{,}1 \\ 7{,}3 & 6{,}2 & 5{,}1 & 0{,}0 & 9{,}3 \\ 7{,}4 & 6{,}1 & 7{,}1 & 9{,}3 & 0{,}0 \end{pmatrix}$$

získáme matici $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} 0{,}00 & -0{,}85 & -1{,}13 & -26{,}65 & -27{,}38 \\ -0{,}85 & 0{,}00 & -1{,}62 & -19{,}22 & -18{,}61 \\ -1{,}13 & -1{,}62 & 0{,}00 & -13{,}01 & -25{,}21 \\ -26{,}65 & -19{,}22 & -13{,}01 & 0{,}00 & -43{,}25 \\ -27{,}38 & -18{,}61 & -25{,}21 & -43{,}25 & 0{,}00 \end{pmatrix}.$$

2. Vytvoříme matici skalárních součnů $\mathbf{B}$:

$$\mathbf{B} = \begin{pmatrix} 8{,}24 & 4{,}26 & 4{,}11 & -9{,}17 & -7{,}45 \\ 4{,}26 & 1{,}96 & 0{,}48 & -4{,}89 & -1{,}81 \\ 4{,}11 & 0{,}48 & 2{,}23 & 1{,}46 & -8{,}28 \\ -9{,}17 & -4{,}89 & 1{,}46 & 26{,}69 & -14{,}09 \\ -7{,}45 & -1{,}81 & -8{,}28 & -14{,}09 & 31{,}62 \end{pmatrix}.$$

3. Najdeme vlastní čísla $\mathbf{B}$:

   $\lambda_1 = 44{,}79;\ \lambda_2 = 26{,}88;\ \lambda_3 = 0{,}71;\ \lambda_4 = 0{,}00;\ \lambda_5 = -1{,}63$

   a odpovídající vlastní vektory:

   $\mathbf{v}_1^\top = (-0{,}02;\ 0{,}03;\ -0{,}17;\ -0{,}61;\ 0{,}77)$,

   $\mathbf{v}_2^\top = (-0{,}58;\ -0{,}25;\ -0{,}21;\ 0{,}61;\ 0{,}43)$,

   $\mathbf{v}_3^\top = (-0{,}12;\ 0{,}77;\ -0{,}62;\ 0{,}08;\ -0{,}11)$,

   $\mathbf{v}_4^\top = (-0{,}45;\ -0{,}45;\ -0{,}45;\ -0{,}45;\ -0{,}45)$,

   $\mathbf{v}_5^\top = (0{,}67;\ -0{,}38;\ -0{,}59;\ 0{,}22;\ 0{,}08)$.

4. Vzhledem k velikosti vlastních čísel se jako vhodná volba pro dimenzi řešení jeví $q = 2$.

5. První dvě vlastní čísla jsou kladná, a proto nemusíme upravovat matici $\mathbf{\Delta}$.
   Spočítáme hodnotu kritéria P_2:
   $P_2 = 0{,}97 > 0{,}8$, výsledné řešení tedy bude velmi dobré.

6. Vytvoříme matici $\mathbf{\Lambda}$:
   $$\mathbf{\Lambda} = \begin{pmatrix} 44{,}79 & 0{,}00 \\ 0{,}00 & 26{,}88 \end{pmatrix}$$
   a matici $\mathbf{V}$:
   $$\mathbf{V} = \begin{pmatrix} -0{,}02 & -0{,}58 \\ 0{,}03 & -0{,}25 \\ -0{,}17 & -0{,}21 \\ -0{,}61 & 0{,}61 \\ 0{,}77 & 0{,}43 \end{pmatrix}.$$

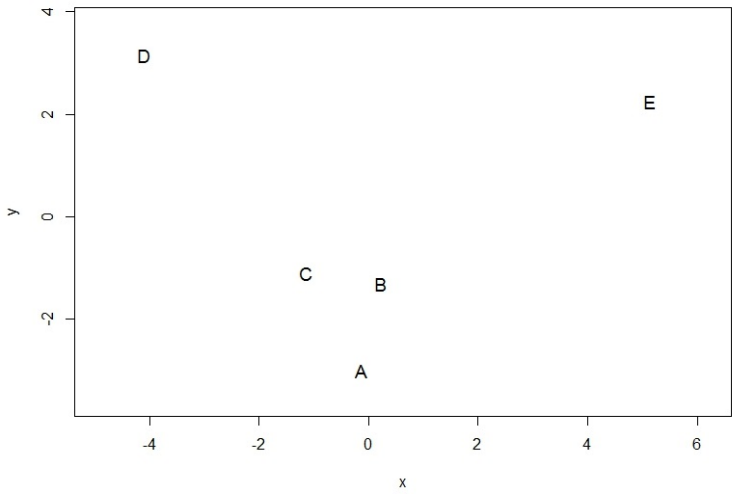
7. Vypočteme matici souřadnic $\mathbf{Y}$:
   $$\mathbf{Y} = \begin{pmatrix} -0{,}12 & -2{,}99 \\ 0{,}23 & -1{,}31 \\ -1{,}15 & -1{,}09 \\ -4{,}10 & 3{,}15 \\ 5{,}15 & 2{,}25 \end{pmatrix}.$$

8. Vytvoříme mapu objektů, jež se nachází na obrázku 1.

Získané řešení je nalezeno tak, že matice euklidovských vzdáleností mezi výslednými body $\mathbf{D}^*$:

$$\mathbf{D}^* = \begin{pmatrix} 0{,}00 & 1{,}71 & 2{,}16 & 7{,}32 & 7{,}44 \\ 1{,}71 & 0{,}00 & 1{,}40 & 6{,}22 & 6{,}08 \\ 2{,}16 & 1{,}40 & 0{,}00 & 5{,}17 & 7{,}14 \\ 7{,}32 & 6{,}22 & 5{,}17 & 0{,}00 & 9{,}30 \\ 7{,}44 & 6{,}08 & 7{,}14 & 9{,}30 & 0{,}00 \end{pmatrix}$$

je nejlepší aproximací vstupní proximitní matice a to z hlediska minimální míry



Obrázek 1: Mapa objektů k metrické verzi příkladu 1

vzájemného nesouladu (3). Neznamená to ale, že by mělo platit: čím menší je původní nepodobnost mezi dvěma objekty, tím blíže u sebe jsou odpovídající body na mapě objektů. Například daná vzdálenost mezi B a C je větší než ta mezi A a B či A a C, avšak u bodů řešení je tomu naopak.

Dosáhnout takového řešení, aby pořadí původních nepodobností pokud možno odpovídalo těm na mapě objektů, je cílem nemetrického MDS, jímž se budeme zabývat v následující kapitole.

3 Nemetrické MDS

Informace použité v této kapitole jsou čerpány z literatury [2], [3], [4] a [5].

Motivací pro vznik nemetrického MDS byly dvě hlavní slabiny metrického MDS: za prvé, definice explicitní funkční závislosti mezi nepodobnostmi a vzdálenostmi a za druhé, omezení na euklidovskou geometrii.

Idea nemetrického MDS spočívá v mírnějším požadavku na vztah mezi nepodobnostmi a vzdálenostmi. Předpokládáme, že skutečné vzdálenosti d_{ij} (tj. vzdálenosti mezi body v relativní $p-$rozměrné konfiguraci) jsou funkčními hodnotami daných nepodobností δ_{ij}:

$$d_{ij} = f(\delta_{ij}), \tag{23}$$

přičemž f je neznámá rostoucí funkce. A tedy je-li $\delta_{ij} < \delta_{kl}$, pak $f(\delta_{ij}) < f(\delta_{kl})$.

Nemetrické MDS se od metrického liší ve způsobu, jakým zachází s nepodobnostmi. Zatímco metrické MDS pracuje s číselnými hodnotami, nemetrické s pořadím hodnot.

Pomocí nemetrického MDS tak řešíme příklady, kdy máme k dispozici nepodobnosti v podobě ordinálních dat. Mezi takové úlohy patří i ty, v nichž je proximitní matice získána dotazováním. A to tak, že jsou respondenti požádáni, aby zhodnotili podobnost (nepodobnost) mezi pozorovanými objekty. V takových případech jsou tázaní schopni posoudit, zda jsou určité objekty podobnější než jiné. Avšak stěží budou moci vyjádřit, o jakou hodnotu se liší.

Máme-li tedy n pozorovaných objektů, pak každý dotazovaný přiřadí jednotlivým dvojicím objektů číslo z určité škály, přičemž nejmenší hodnota škály odpovídá shodným a největší hodnota naprosto odlišným objektům.

Proximitní matici $\mathbf{\Delta}^{(n\times n)} = (\delta_{ij})$ pak dostaneme následovně. Na diagonále budou nuly. Pro každý z $m = n(n-1)/2$ párů objektů spočítáme průměr z přiřazených hodnot. Takto získáme prvky $\delta_{ij},\ i < j$, které se rovnají prvkům δ_{ji}.

Úkolem nemetrického MDS je na základě daných pořadí hodnot nepodobností mezi objekty vytvořit jejich reprezentaci v $q-$rozměrném prostoru ($q < p$). Chceme

najít takovou konfiguraci, v níž pořadí vzdáleností d^*_{ij} mezi body zachovává pořadí původních nepodobností δ_{ij}. Tedy, je-li m nepodobností $\delta_{ij},\ i < j$ seřazeno od nejmenší po největší:

$$\delta_{i_1 j_1} < \delta_{i_2 j_2} < \ldots < \delta_{i_m j_m}, \tag{24}$$

pak pro příslušné vzdálenosti d^*_{ij} má platit:

$$d^*_{i_1 j_1} < d^*_{i_2 j_2} < \ldots < d^*_{i_m j_m}; \tag{25}$$

přitom d^*_{ij} představují Minkowského vzdálenosti, tj.

$$d^*_{ij} = \sqrt[\lambda]{\sum_{k=1}^{p} |y_{ik} - y_{jk}|^\lambda},\ \lambda > 0. \tag{26}$$

Ne vždy bude pro výsledné řešení platit podmínka (25). Ve skutečnosti budeme chtít, aby podmínka monotónnosti (avšak slabé) platila pro tzv. *disparity* $\hat{d}_{ij}$, jež aproximují vzdálenosti d^*_{ij}. Disparity jsou proto určovány tak, aby splňovaly nerovnost:

$$\hat{d}_{i_1 j_1} \leq \hat{d}_{i_2 j_2} \leq \ldots \leq \hat{d}_{i_m j_m}. \tag{27}$$

Matice souřadnic $\mathbf{Y}^{(n \times q)}$ je hledána pomocí iterativního Shepard-Kruskalova algoritmu, který je založen na minimalizaci ztrátové funkce $Stress$, jež je definována:

$$Stress(\mathbf{Y}) = \sum_{i<j} \left(d^*_{ij} - \hat{d}_{ij}\right)^2. \tag{28}$$

Běžně bývá kritérium (28) nahrazeno jeho normovanou verzí (28):

$$Stress(\mathbf{Y}) = \sqrt{\frac{\sum_{i<j} \left(d^*_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} \left(d^*_{ij}\right)^2}}, \tag{29}$$

případně

$$Stress(\mathbf{Y}) = \sqrt{\frac{\sum_{i<j} \left(d^*_{ij} - \hat{d}_{ij}\right)^2}{\sum_{i<j} \left(d^*_{ij} - \bar{d}^*\right)^2}}, \tag{30}$$

kde $\bar{d}^*$ je průměrná vzdálenost d^*_{ij}.

Podle výsledné hodnoty *Stress*u posuzujeme kvalitu celého modelu. Tabulka 3 uvádí Kruskalem doporučené hodnoty.

Stress	Kvalita modelu
$> 0{,}2$	slabá
$0{,}1 - 0{,}2$	uspokojivá
$0{,}05 - 0{,}1$	dobrá
$0{,}025 - 0{,}05$	vynikající
$0 - 0{,}025$	perfektní

Tabulka 3: Hodnocení kvality modelu nemetrického MDS podle *Stress*u

3.1 PAV algoritmus

Disparity $\hat{d}_{ij}$ dostaneme pro vzdálenosti d^*_{ij} provedením monotónní regrese nepodobností δ_{ij} na d^*_{ij} s požadavkem slabé monotónnosti, tj. je-li $\delta_{ij} < \delta_{kl}$, pak $\hat{d}_{ij} \leq \hat{d}_{kl}$. Jednou z nejužívanějších metod k určení disparit je PAV algoritmus (z angl. pool-adjacent violators).

PAV algoritmus je popsán následovně: k m vzestupně seřazeným δ_{ij} přiřadíme odpovídající d^*_{ij}. Postupně procházíme d^*_{ij} a zjišťujeme, zda má jejich uspořádání neklesající charakter. Kdykoli je v nějakém bloku tato podmínka porušena, nahradíme hodnoty dané skupiny prvků jejich průměrem. Přitom jako počátek bloku je brán poslední prvek, který není menší než jeho předchůdce.

Aplikaci PAV algoritmu si ukážeme na následujícím ilustrativním příkladu.

Příklad 2. Pro $n = 5$ objektů je v prvním sloupci tabulky 4 dáno $m = 10$ vzájemných nepodobností δ_{ij}, $i < j$ seřazených od nejmenší po největší. Ve druhém sloupci se nachází jejich odpovídající vzdálenosti d^*_{ij} na mapě objektů. Pomocí PAV algoritmu určete pro vzdálenosti d^*_{ij} disparity $\hat{d}_{ij}$.

δ_{ij}	d^*_{ij}	I	II	III	IV	$\hat{d}_{ij}$
1,1	1,2	1,2	1,2	1,2	1,2	1,2
2,3	3,1	3,1	2,4	2,4	2,4	2,4
2,5	2,8	2,8	2,4	2,4	2,4	2,4
3,4	1,3	1,3	2,4	2,4	2,4	2,4
3,8	2,7	2,7	2,7	2,7	2,7	2,7
4,7	5,2	5,2	5,2	4,4	4,2	4,2
4,9	3,6	3,6	3,6	4,4	4,2	4,2
5,5	3,8	3,8	3,8	3,8	4,2	4,2
6,2	6,5	6,5	6,5	6,5	6,5	5,9
6,8	5,3	5,3	5,3	5,3	5,3	5,9

Tabulka 4: Určování disparit podle PAV algoritmu v příkladu 2

Řešení příkladu demonstruje tabulka 4. V každém ze sloupců I až IV je vyznačena první skupina prvků, které porušují požadavek nesestupného uspořádání. Všem prvkům ze zvýrazněné bloku je v následném sloupci přiřazena jejich průměrná hodnota. Poslední sloupec obsahuje výsledné disparity.

3.2 Shepard-Kruskalův algoritmus

Cílem Shepard-Kruskalova algoritmu je najít takovou konfiguraci bodů $\mathbf{Y}^{(n\times q)}$, pro kterou je hodnota $Stress$u minimální a přitom je splněna podmínka slabé monotónnosti pro disparity (27). Jde o optimalizační iterativní proces, jehož každá iterace se skládá z následujících kroků.

V prvním kroku každé iterace $t = 0, 1, \ldots$ nejprve určíme konfiguraci $\mathbf{Y}_t$ a vypočítáme Minkowského vzdálenosti $d^{*(t)}_{ij}$ mezi body v této konfiguraci. Běžně volíme $\lambda = 2$ a v tom případě $d^{*(t)}_{ij}$ představují vzdálenostií euklidovské.

V dalším kroku nalezneme ke vzdálenostem $d^{*(t)}_{ij}$ odpovídající disparity $\hat{d}^{(t)}_{ij}$. K tomu použijeme PAV algoritmus.

V posledním kroku spočítáme pro danou $\mathbf{Y}_t$ hodnotu funkce $Stress(\mathbf{Y}_t)$. Pokud je tato hodnota minimem funkce, matice $\mathbf{Y} = \mathbf{Y}_t$ představuje výsledné řešení. V opačném případě přejdeme k další iteraci.

Počáteční konfiguraci $\mathbf{Y}_0$ volíme obvykle tak, že na proximitní matici $\mathbf{\Delta}$ aplikujeme metrické MDS nebo využijeme výsledky z předchozí úlohy. Použít lze i generování náhodných čísel.

Každá následná konfigurace pro $t = 1, 2, \ldots$ je získána z předešlé konfigurace užitím metody největšího spádu, jež zaručí, že $Stress$ je minimalizován nejrychlejším možným způsobem. Tato metoda spočívá v tom, že konfigurace je posunuta ve směru určeném parciálními derivacemi funkce $Stress(\mathbf{y})$ podle $\mathbf{y}$, kde

$$\mathbf{y} = (y_{11}, y_{12}, \ldots, y_{1q}, \ldots, y_{nq})^\top. \tag{31}$$

Nové souřadnice získáme výpočtem:

$$\mathbf{y}^t = \mathbf{y}^{t-1} - \alpha_t \mathbf{z}, \tag{32}$$

kde

$$\mathbf{z} = \frac{\partial Stress}{\partial \mathbf{y}} \Bigg/ \left| \frac{\partial Stress}{\partial \mathbf{y}} \right| \tag{33}$$

a α_t je délka kroku v $t-$té iteraci. Pro urychlení algoritmu bývá délka kroku průběžně měněna. Kruskal doporučuje jako startovací hodnotu $\alpha_1 = 0{,}2$.

Celý proces je ukončen tehdy, je-li dosaženo minima funkce $Stress$ (tj. rozdíl v příští hodnotě $Stress$u je zanedbatelně malý). Může ale nastat situace, že konečná hodnota $Stress$u představuje místo globálního minima (pro zvolenou dimenzi q) pouze minimum lokální, neboť vybraná počáteční konfigurace nemusí být optimální. Proto někdy zkoušíme různé $\mathbf{Y}_0$, a dostáváme-li stejné řešení, pak je zřejmě dosaženo globálního minima.

3.3 Volba dimenze nemetrického řešení

Teoreticky vybereme vhodný rozměr řešení tak, že vypočítáme minimální hodnotu $Stress$u pro různá q, kterou označíme S_q, a zobrazíme body (q, S_q). Čím větší

je zvolené q, tím je hodnota S_q menší. Proto, spojíme-li body (q, S_q), dostaneme křivku klesajícho průběhu.

Adekvátní počet rozměrů řešení indikuje tzv. elbow kritérium, což je místo, v němž graf přechází ze strmého na horizontální. A pro vyšší dimenzi se tak hodnota S_q zmenší jen nepatrně.

V praxi chceme, aby byla dimenze řešení co nejmenší. Z toho důvodu, dostaneme-li kvalitní model (dle tabulky 3) pro $q = 2$ nebo $q = 3$, volíme dvou či trojrozměrné řešení.

3.4 Shepardův diagram

Vztah mezi původními nepodobnostmi $\delta_{ij},\ i < j$, vzdálenostmi ve výsledné konfiguraci d^*_{ij} a jím přiřazenými disparitami $\hat{d}_{ij}$ lze graficky znázornit pomocí Shepardova diagramu.

Shepardův diagram má podobu bodového grafu proloženého křivkou monotónní regrese. Stupnice na ose x odpovídá škále nepodobností a stupnice osy y rozpětí vzdáleností a disparit. Každý bod představuje hodnotu d^*_{ij} vzhledem k příslušné δ_{ij}. Schodovitě rostoucí křivka prochází hodnotami disparit.

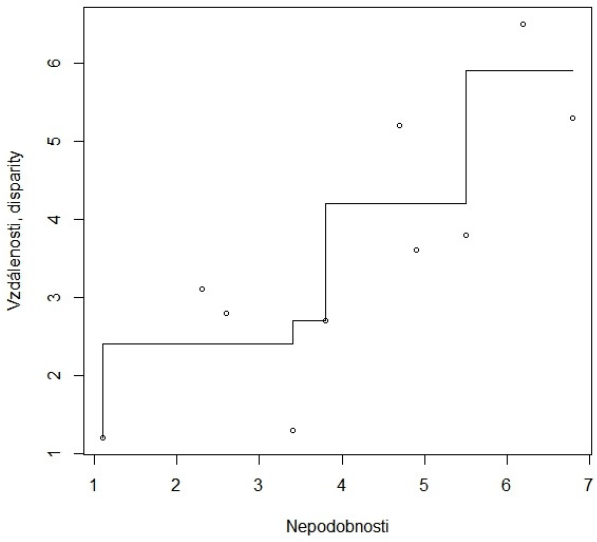
Sečteme-li čtvercové odchylky bodů od křivky, dostaneme hodnotu $Stress$u (28). Proto Shepardův diagramu slouží k posuzování kvality modelu. Čím méně jsou body vzdáleny od křivky, tím je model lepší. V ideálním případě leží všechny body na křivce.

Na obrázku 2 najdeme Shepardův diagram k příkladu 2.

3.5 Ilustrativní příklad k nemetrickému MDS

K demonstraci algoritmu nemetrického MDS nám bude sloužit opět příklad 1.

Pro nemetrický charakter úlohy budeme uvažovat, že nepodobnosti z tabulky 1 představují nepodobnosti mezi výrobky A, B, C, D a E získané dotazováním zákazníků. Každý respondent měl pro $m = 10$ dvojic výrobků zhodnotit jejich



Obrázek 2: Shepardův diagram k příkladu 2

vzájemnou podobnost a přiřadit jim číslo na škále od 0 do 10, kde 0 odpovídá naprosto shodným a 10 zcela odlišným výrobkům. Dané hodnoty pak tvoří průměr hodnot, které zákazníci přiřadili.

Při řešení příkladu pomocí nemetrického MDS se budeme řídit postupem z tabulky 5.

1. Vzestupně uspořádáme nepodobnosti: $\delta_{i,j},\ i < j$:
   $\delta_{12} = 1{,}3 < \delta_{13} = 1{,}5 < \delta_{23} = 1{,}8 < \delta_{34} = 5{,}1 < \delta_{25} = 6{,}2 < \delta_{24} = 6{,}3 < \delta_{35} =$
   $= 7{,}1 < \delta_{14} = 7{,}3 < \delta_{15} = 7{,}4 < \delta_{45} = 9{,}3.$

2. Zvolíme $q = 2$.

3. Položíme $t = 0$.

4. Jako počáteční konfiguraci zvolíme matici souřadnic získanou metrickým MDS:

1.	$m = \frac{1}{2}n(n-1)$ nepodobností $\delta_{i,j},\ i < j$ z proximitní matice $\mathbf{\Delta}^{(n\times n)} = (\delta_{ij})$ seřadíme od nejmenší po největší.
2.	Zvolíme dimenzi řešení q.
3.	Položíme $t = 0$.
4.	Stanovíme konfiguraci bodů $\mathbf{Y}_t^{(n\times q)}$. $\mathbf{Y}_0$ získáme aplikací metrického MDS na matici $\mathbf{\Delta}$. $\mathbf{Y}_t, t = 1, 2, \ldots$ dostaneme z předchozí konfigurace užitím metody nejstrmějšího sestupu.
5.	Vypočítáme (euklidovské) vzdálenosti $d_{ij}^{*(t)},\ i < j$ mezi body v konfiguraci $\mathbf{Y}_t$.
6.	K $d_{ij}^{*(t)}$ určíme disparity $\hat{d}_{ij}^{(t)},\ i < j$ (podle PAV algoritmu).
7.	Spočítáme hodnotu ztrátové funkce $Stress$ pro $\mathbf{Y}_t$: $Stress(\mathbf{Y}_t) = \sqrt{\frac{\sum_{i<j}\left(d_{ij}^{*(t)} - \hat{d}_{ij}^{(t)}\right)}{\sum_{i<j}\left(d_{ij}^{*(t)}\right)^2}}$ Dokud nedostaneme minimum S_q funkce $Stress$, provedeme $t \leftarrow t + 1$ a vrátíme se ke kroku 4.
8.	Pokud $S_q > 0{,}2$, položíme $q \leftarrow q + 1$ a navrátíme se ke kroku 3. V opačném případě představuje matice z poslední iterace $\mathbf{Y}$ výsledné řešení.
9.	Vytvoříme mapu objektů podle matice $\mathbf{Y}$.

Tabulka 5: Praktický výpočetní algoritmus nemetrického MDS

$$\mathbf{Y}_0 = \begin{pmatrix} -0{,}12 & -2{,}99 \\ 0{,}23 & -1{,}31 \\ -1{,}15 & -1{,}09 \\ -4{,}10 & 3{,}15 \\ 5{,}15 & 2{,}25 \end{pmatrix}.$$

5. Vypočítáme euklidovské vzdálenosti $d_{ij}^{*(0)}$ $i < j$:

   $d_{12}^{*(0)} = 1{,}71;\ d_{13}^{*(0)} = 2{,}16;\ d_{23}^{*(0)} = 1{,}40;\ d_{34}^{*(0)} = 5{,}17;\ d_{25}^{*(0)} = 6{,}08;\ d_{24}^{*(0)} = 6{,}22;$
   $d_{35}^{*(0)} = 7{,}14;\ d_{14}^{*(0)} = 7{,}32;\ d_{15}^{*(0)} = 7{,}44;\ d_{45}^{*(0)} = 9{,}30.$

6. K $d_{ij}^{*(0)}$ určíme PAV algoritmem disparity $\hat{d}_{ij}^{(0)}$ $i < j$:

   $\hat{d}_{12}^{(0)} = 1{,}71 \leq \hat{d}_{13}^{(0)} = 1{,}78 \leq \hat{d}_{23}^{(0)} = 1{,}78 \leq \hat{d}_{34}^{(0)} = 5{,}17 \leq \hat{d}_{25}^{(0)} = 6{,}08 \leq \hat{d}_{24}^{(0)} =$
   $= 6{,}22 \leq \hat{d}_{35}^{(0)} = 7{,}14 \leq \hat{d}_{14}^{(0)} = 7{,}32 \leq \hat{d}_{15}^{(0)} = 7{,}44 \leq \hat{d}_{45}^{(0)} = 9{,}30.$

7. Ztrátová funkce $Stress$ má pro $\mathbf{Y}_0$ hodnotu $Stress(\mathbf{Y}_0) = 0{,}029$.

   Změníme t na $t = 1$ a vrátíme se ke kroku 4.

4. Vypočítáme souřadnice pro $\mathbf{Y}_1$:

$$\mathbf{Y}_1 = \begin{pmatrix} -0{,}19 & -2{,}87 \\ 0{,}37 & -1{,}34 \\ -1{,}23 & -1{,}20 \\ -4{,}11 & 3{,}16 \\ 5{,}16 & 2{,}25 \end{pmatrix}.$$

$\vdots$

4. V sedmé iteraci dostaneme:

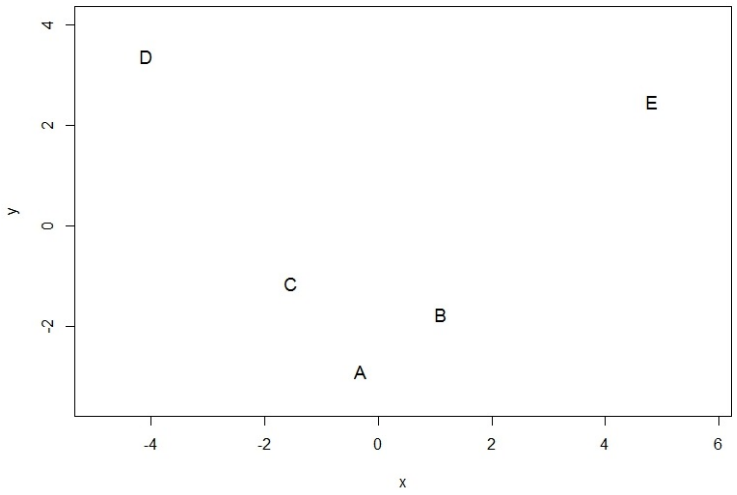
$$\mathbf{Y}_7 = \begin{pmatrix} -0{,}29 & -2{,}91 \\ 1{,}10 & -1{,}78 \\ -1{,}54 & -1{,}16 \\ -4{,}09 & 3{,}38 \\ 4{,}83 & 2{,}46 \end{pmatrix}.$$

$\vdots$

7. Hodnota *Stress*u pro $\mathbf{Y}_7$ je $Stress(\mathbf{Y}_7) = 0{,}000 = S_2$.
   Výsledná matice souřadnic je $\mathbf{Y} = \mathbf{Y}_7$.

8. Podle hodnoty S_2 lze dvourozměrné řešení považovat za perfektní.

9. Vytvoříme mapu objektů, jež se nachází na obrázku 3.



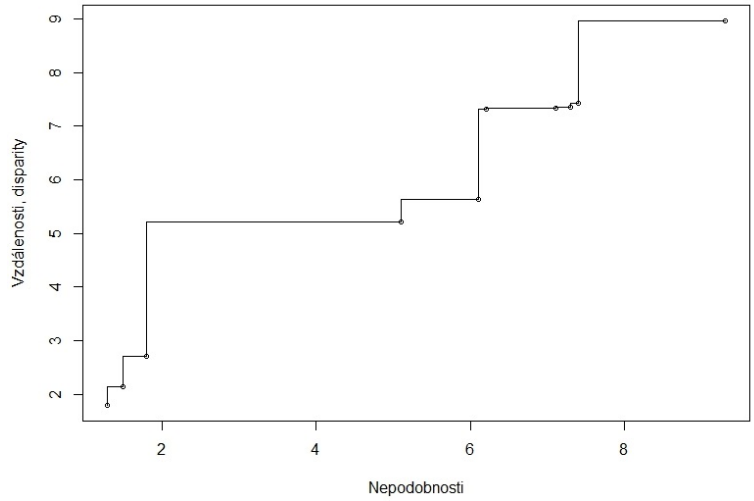
Obrázek 3: Mapa objektů k nemetrické verzi příkladu 1

Matice euklidovských vzdáleností mezi výslednými body $\mathbf{D}^*$:

$$\mathbf{D}^* = \begin{pmatrix} 0{,}00 & 1{,}80 & 2{,}15 & 7{,}35 & 7{,}42 \\ 1{,}80 & 0{,}00 & 2{,}71 & 7{,}32 & 5{,}64 \\ 2{,}15 & 2{,}71 & 0{,}00 & 5{,}21 & 7{,}33 \\ 7{,}35 & 7{,}32 & 5{,}21 & 0{,}00 & 8{,}96 \\ 7{,}42 & 5{,}64 & 7{,}33 & 8{,}96 & 0{,}00 \end{pmatrix}$$

aproximuje vstupní proximitní matici tak, aby hodnota *Stress*u byla minimální. V tomto příkladě odpovídá pořadí nepodobností $\delta_{i,j}$ nejen pořadí disparit $\hat{d}_{ij}$, ale i pořadí vzdáleností d^*_{ij}.

To, že výsledné řešení lze pokládat za perfektní, ilustruje i příslušný Shepardův diagram, jenž nalezneme na obrázku 4.



Obrázek 4: Shepardův diagram k příkladu 1

4 MDS v R

Tato kapitola je věnována tomu, jak lze pro MDS využít statistický software R. Její obsah vychází z nápovědy v R a zdroje [2].

4.1 Metrické úlohy

K řešení metrických úloh MDS klasickým způsobem slouží v R příkaz `cmdscale`. Tato funkce vyžaduje jeden povinný argument, další čtyři jsou nepovinné:

- `d` povinný parametr, ve kterém je zadána vstupní proximitní matice; může jít o maticí vzdáleností vypočítanou funkcí `dist`, máme-li k dispozici datovou matici, nebo o jakoukoli symetrickou matici nepodobností typu $n \times n$;
- `k` v tomto nepovinném argumentu zadáváme zvolenou dimenzi řešení, tedy q v našem značení; musí být z množiny $\{1, 2, \ldots, n-1\}$; defaultní hodnota je nastavena na 2;
- `eig` parametr, jemuž přiřadíme logickou hodnotu `TRUE`, pokud chceme mít na výstupu kromě výsledných souřadnic i vlastní čísla matice skalárních součinů $\mathbf{B}$; výchozí hodnotou je u tohoto, i u zbývajících nepovinných argumentů, `FALSE`;
- `add` u tohoto parametru lze nastavit to, aby byla k mimodiagonálním prvkům proximitní matice přičtena konstanta c tak, abychom u matice $\mathbf{B}$ dostali jen nezáporná vlastní čísla;
- `x.ret` argument, jenž určuje, zda má být na výstupu i dvojitě centrovaná matice čtvercových nepodobností (vynásobením této matice číslem $-\frac{1}{2}$ dostaneme matici $\mathbf{B}$).

Máme-li tedy pod názvem `Delta` uloženou proximitní matici, pak dvojrozměrné metrické řešení dostaneme zadáním příkazu:

```
> cmdscale(Delta)
```

což je ekvivalentní s:

```
> cmdscale(Delta, k = 2, eig = FALSE, add = FALSE,
+          x.ret = FALSE)
```

Výše uvedeným způsobem získáme pouze výsledné souřadnice řešení. Pokud je ale alespoň u jednoho z parametrů logických hodnot zadáno `TRUE`, na výstupu dostaneme seznam obsahující následující položky:

- `$points` tento prvek seznamu obsahuje matici s počtem sloupců rovným `k`, jejíž řádky udávají souřadnice bodů na mapě objektů;
- `$eig` v této položce jsou vypsána vlastní čísla matice $\mathbf{B}$, pokud `eig = TRUE`; jinak je vrácena hodnota `NULL`;
- `$x` při volbě `x.ret = TRUE` se v tomto prvku seznamu skrývá dvojitě centrovaná matice čtvercových nepodobností, v opačném případě hodnota `NULL`;
- `$ac` pokud zadáme `add = TRUE`, tato položka nám ukáže, jaká konstanta byla připočtena k mimodiagonálním prvkům proximitní matice; jestli nevybereme pravdivou logickou hodnotu, je vrácena `0`;
- `$GOF` poslední prvek seznamu představuje dvourozměrný vektor, jehož první složka udává hodnotu kritéria (21) a druhá hodnotu (22) pro q rovno `k`.

Vraťme se nyní k ilustrativnímu příkladu 1 z kapitoly 2 a podívejme se na jeho řešení v R.

Nejprve vytvoříme příslušnou proximitní matici `Delta` a pojmenujeme její řádky a sloupce. Dostaneme:

```
> Delta
    A   B   C   D   E
A 0.0 1.3 1.5 7.3 7.4
B 1.3 0.0 1.8 6.2 6.1
C 1.5 1.8 0.0 5.1 7.1
D 7.3 6.2 5.1 0.0 9.3
E 7.4 6.1 7.1 9.3 0.0
```

Přejdeme k řešení pomocí funkce `cmdscale`. Necháme si zobrazit i vlastní čísla matice $\mathbf{B}$:

```
> reseni = cmdscale(Delta, eig = TRUE)
> reseni
$points
        [,1]       [,2]
A -0.1232445 -2.992941
B  0.2254902 -1.314891
C -1.1523629 -1.094363
D -4.1032885  3.151575
E  5.1534056  2.250619

$eig
[1]  4.478854e+01  2.688197e+01  7.131198e-01 -1.554312e
   -15 -1.625633e+00

$x
NULL

$ac
[1] 0

$GOF
[1] 0.9683992 0.9901481
```

Nakonec vytvoříme mapu objektů, která se nachází na obrázku 1:

```
> Y = reseni$points
> x = Y[,1]
> y = Y[,2]
> plot(x, y, type = "n", xlab = "x", ylab = "y",
+      xlim = range(x)*1.2, ylim = range(y)*1.2)
> text(x, y, label = colnames(Delta))
```

Zajímá-li nás podoba matice vzdáleností mezi výslednými body, provedeme:

```
> D_q = as.matrix(dist(Y))
> D_q
          A         B         C         D         E
A 0.000000 1.713904 2.159556 7.320917 7.438948
B 1.713904 0.000000 1.395389 6.219939 6.082533
C 2.159556 1.395389 0.000000 5.170682 7.138040
```

```
D 7.320917 6.219939 5.170682 0.000000 9.300436
E 7.438948 6.082533 7.138040 9.300436 0.000000
```

4.2 Nemetrické úlohy

Nemetrické úlohy MDS lze v R řešit pomocí funkce `isoMDS`, která je součástí knihovny `MASS`. Má jeden povinný argument a šest nepovinných:

- `d` povinný parametr představující vstupní proximitní matici;
- `y` v tomto prvním nepovinném argumentu zadáme počáteční konfiguraci, pokud nechceme, aby jí bylo klasické řešení; jinak je automaticky zvoleno `y = cmdscale(d, k)`;
- `k` parametr, jímž lze vybrat jinou dimenzi řešení než je defaulitní hodnota `k = 2`;
- `maxit` tímto argumentem s výchozí hodnotou `maxit = 50` můžeme změnit nastavení maximálního počtu iterací;
- `trace` jde o parametr logických hodnot; pokud nezměníme jeho defaultní hodnotu `TRUE`, tak na výstupu dostaneme počáteční hodnotu funkce *Stress* a hodnotu *Stress*u po každé páté iteraci; dále je nám poskytnuta informace o tom, proč byl ukončen výpočet, tj. buď kvůli provedení maximálního počtu iterací, nebo díky dosažení minima *Stress*u (podle pravidel konvergence zabudovaných v programu);
- `tol` v tomto argumentu lze upravit konvergenční toleranci; automaticky nastavená hodnota je `tol = 1e-3`;
- `p` poslední nepovinný argument s výchozí hodnotou `p = 2`, jenž představuje parametr λ ze vzorce (26) pro výpočet Minkowského vzdáleností v konfiguraci.

Standardní dvojrozměrné nemetrické řešení pro vstupní proximitní matici `Delta` pak dostaneme:

```
> isoMDS(Delta)
```

což je ekvivalentní s:

```
> isoMDS(Delta, y = cmdscale(Delta, k), k = 2,
+        maxit = 50, trace = TRUE, tol = 1e-3, p = 2)
```

Na výstupu dostáváme seznam obsahující dvě položky:

- `$points` v tomto prvku se nachází výsledná matice souřadnic;
- `$stress` tato položka obsahuje finální hodnotu funkce *Stress* (v procentech).

Řešení nemetrické verze příkladu 1 v R získáme takto:

```
> library("MASS")
> reseni = isoMDS(Delta)
initial  value 2.854420
iter   5 value 0.416743
final  value 0.000000
converged
> reseni
$points
         [,1]       [,2]
A -0.2925252 -2.907899
B  1.1005277 -1.775666
C -1.5431499 -1.161326
D -4.0908855  3.382435
E  4.8260329  2.462455

$stress
[1] 2.088284e-14
```

Mapu objektů, jež se nachází na obrázku 3, dostaneme stejným zůsobem, jaký je uveden v předchozí podkapitole.

Ke tvorbě příslušného Shepardova diagramu využijeme funkcí `Shepard`, jejímiž vstupními argumenty jsou dolní trojuhelník proximitní matice a výsledná matice souřadnic. Na výstupu dostaneme seznam, který obsahuje po řadě vzestupně

uspořádané nepodobnosti, vzdálenosti ve výsledné konfiguraci a jejich odpovídající disparity:

```
> Shep = Shepard(Delta[lower.tri(Delta)], Y)
> Shep
$x
 [1] 1.3 1.5 1.8 5.1 6.1 6.2 7.1 7.3 7.4 9.3

$y
 [1] 1.795146 2.148158 2.714120 5.209292 5.642788
    7.318249 7.327911 7.348187 7.418918 8.964251

$yf
 [1] 1.795146 2.148158 2.714120 5.209292 5.642788
    7.318249 7.327911 7.348187 7.418918 8.964251
```

Shepardův diagram, jenž se nachází na obrázku 4, pak získáme následovně:

```
> plot(Shep, xlab="Nepodobnosti", ylab="Vzdalenosti,
+      disparity", xlim=range(Shep$x),
+      ylim=range(Shep$y), cex=0.7)
> lines(Shep$x, Shep$yf, type="S")
```

5 Praktické příklady

V této kapitole si ukážeme využití metody MDS na reálných datech. Příklady budeme řešit pomocí statistického softwaru R.

5.1 Města České republiky

V prvním příkladu použijeme metrické MDS k tomu, abychom s využitím informace o vzdálenostech mezi městy vytvořili jejich mapu.

Pomocí zdroje [6] získáme letecké vzdálenosti mezi deseti největšími městy České republiky a vyrobíme příslušnou matici vzdáleností, jejíž řádky a sloupce budou náležitě pojmenovány:

```
> D_mestaCR
                 Pra Brn Ost Plz Lib Olo UnL CBu HKr Par
Praha              0 186 278  84  88 210  69 124 102  97
Brno             186   0 140 242 207  65 246 157 126 112
Ostrava          278 140   0 353 252  79 316 292 180 181
Plzen             84 242 353   0 165 280 112 117 183 175
Liberec           88 207 252 165   0 204  73 204  83  96
Olomouc          210  65  79 280 204   0 259 213 123 117
Usti nad Labem    69 246 316 112  73 259   0 190 137 142
Ceske Budejovice 124 157 292 117 204 213 190   0 169 151
Hradec Kralove   102 126 180 183  83 123 137 169   0  19
Pardubice         97 112 181 175  96 117 142 151  19   0
```

Užitím příkazu `cmdscale` dostaneme souřadnice bodů ve dvourozměrném prostoru, jež reprezentují daná města. Volbou argumentu `eig = TRUE` dostaneme i vlastní čísla a velikost kritéria P_2:

```
> reseni_mestaCR = cmdscale(D_mestaCR, eig = TRUE)
> reseni_mestaCR
$points
                       [,1]        [,2]
Praha              80.12350   -5.113914
Brno              -92.34648   64.673772
Ostrava          -196.72350  -29.320446
Plzen             148.65183   43.820539
```

```
Liberec             48.45610 -87.283943
Olomouc           -129.56868  12.437961
Usti nad Labem     117.63635 -63.591065
Ceske Budejovice    56.33419 116.561715
Hradec Kralove     -16.57082 -35.921623
Pardubice          -15.99249 -16.262996

$eig
 [1]  1.124234e+05  3.394727e+04  1.820858e+02  1.422177
    e+02  2.240113e+01  1.335820e-12 -3.656561e+00
    -9.122059e+01 -1.235020e+02 -2.139892e+02

$x
NULL

$ac
[1] 0

$GOF
[1] 0.9947056 0.9976369
```

Kritérium P_2 je téměř rovno jedné a dvourozměrné řešení je tedy takřka perfektní.

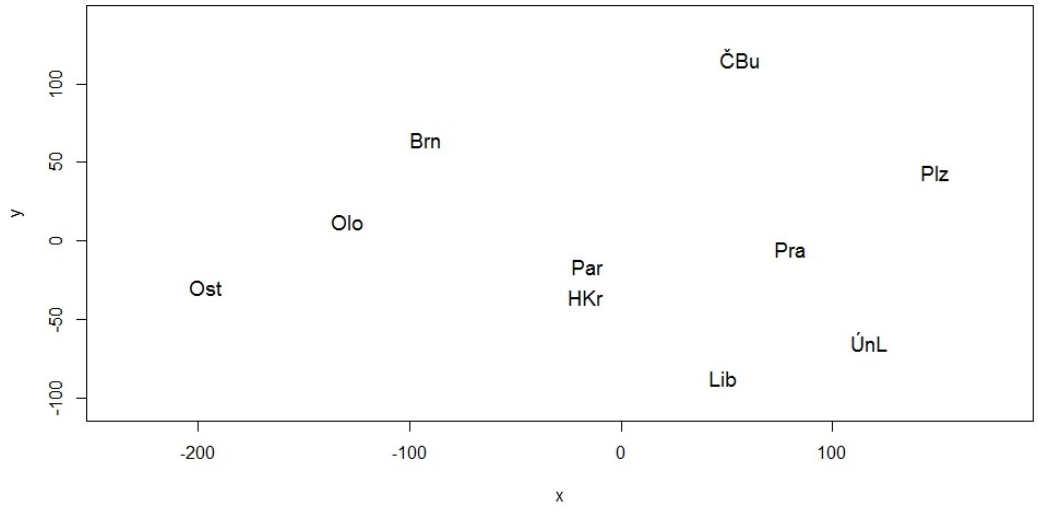
Ze získaných souřadnic vytvoříme mapu objektů, která se nachází na obrázku 5.

Poloha daných měst se více liší ve směru východ-západ než sever-jih. První rozměr řešení tak souvisí se zeměpisnou délkou a druhý se zeměpisnou šířkou měst. Avšak proto, aby se v horní (dolní) části grafu nacházela města, která leží severněji (jižněji), je třeba výslednou konfiguraci rotovat. To provedeme změnou znamének u souřadnic řešení a dostaneme mapu, jež se nachází na obrázku 6.

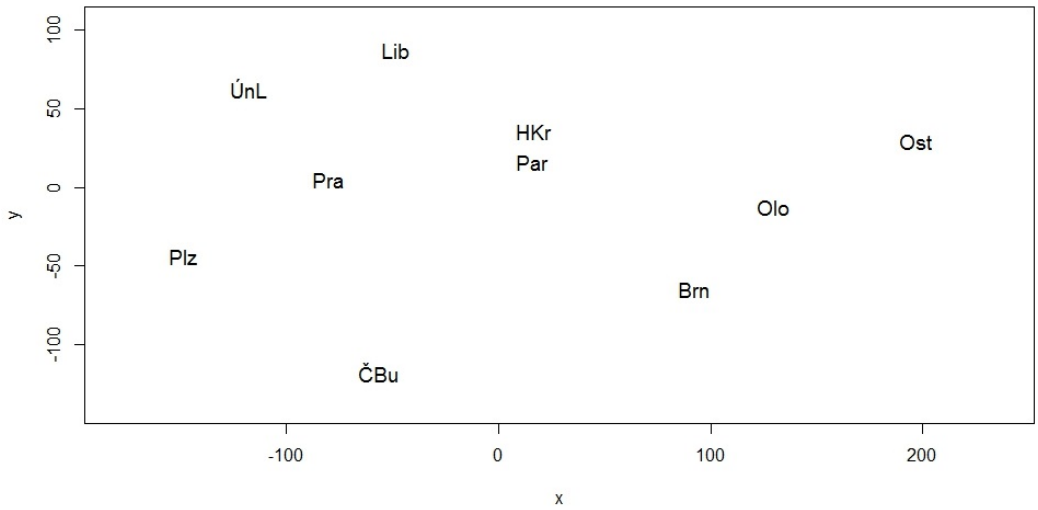
Vidíme, že body na mapě z obrázku 6 přibližně odpovídají pozicím měst na mapě České republiky.

5.2 Politické strany v Poslanecké sněmovně

Aplikaci nemetrického MDS si ukážeme na příkladu, jenž se týká politických stran v Poslanecké sněmovně Parlamentu České republiky. Budeme pracovat s ma-



Obrázek 5: Mapa objektů k příkladu o městech ČR (bez rotace)



Obrázek 6: Mapa objektů k příkladu o městech ČR (po rotaci)

ticí nepodobností mezi těmi stranami, které v parlamentních volbách 2013 dostaly více než 5% hlasů a získaly tak křesla v Poslanecké sněmovně (tj. ČSSD, ANO, KSČM, TOP 09, ODS, ÚSVIT a KDU-ČSL). Naším cílem bude pokusit se klasifikovat nově vzniklá seskupení ANO a ÚSVIT v rámci sněmovních stran.

Ke tvorbě proximitní matice využijeme volební kalkulačku [7]. Její součástí bylo pětatřicet otázek, jejichž seznam lze najít v příloze A.1, a odpovědi, které k daným otázkám poskytly kandidující strany. Sněmovní strany ke zmíněným otázkám vyslovily tato stanoviska:

```
> odpovedi
    CSSD   ANO  KSCM TOP 09   ODS USVIT KDU-CSL
1    Pro   Pro   Pro  Proti Proti   Pro   Proti
2    Pro Proti     -    Pro   Pro   Pro     Pro
3    Pro   Pro   Pro    Pro   Pro   Pro     Pro
4    Pro Proti   Pro    Pro   Pro Proti     Pro
5    Pro     -     -    Pro Proti Proti   Proti
6  Proti     - Proti    Pro Proti Proti   Proti
7    Pro Proti   Pro    Pro Proti Proti     Pro
8      -   Pro   Pro  Proti Proti Proti   Proti
9  Proti Proti     -  Proti Proti Proti   Proti
10     - Proti Proti  Proti Proti   Pro   Proti
11   Pro   Pro   Pro  Proti Proti   Pro     Pro
12   Pro Proti   Pro  Proti Proti     -     Pro
13   Pro   Pro   Pro  Proti Proti     -     Pro
14 Proti     - Proti    Pro   Pro     -   Proti
15 Proti Proti Proti    Pro   Pro   Pro     Pro
16   Pro Proti   Pro  Proti Proti Proti   Proti
17 Proti   Pro Proti    Pro   Pro   Pro   Proti
18   Pro Proti   Pro  Proti Proti Proti     Pro
19     -   Pro     -    Pro Proti   Pro     Pro
20   Pro   Pro   Pro    Pro     -     -     Pro
21     -   Pro   Pro    Pro Proti   Pro     Pro
22   Pro   Pro   Pro    Pro Proti   Pro     Pro
23   Pro   Pro   Pro    Pro     - Proti     Pro
24 Proti Proti Proti    Pro Proti     -   Proti
25   Pro   Pro   Pro    Pro Proti     -     Pro
26 Proti Proti Proti  Proti Proti Proti   Proti
27     -   Pro     -  Proti Proti Proti   Proti
```

```
28   Pro   Pro   Pro    Pro Proti Proti     Pro
29   Pro   Pro     -      - Proti Proti       -
30 Proti Proti     -  Proti Proti Proti   Proti
31   Pro   Pro   Pro  Proti Proti Proti     Pro
32   Pro   Pro   Pro  Proti   Pro Proti   Proti
33 Proti   Pro Proti    Pro     -   Pro       -
34   Pro   Pro Proti    Pro   Pro   Pro     Pro
35   Pro   Pro   Pro    Pro   Pro   Pro     Pro
```

Tyto odpovědi nám poslouží k vyjádření vzájemných nepodobností mezi stranami. Nepodobnosti budeme uvažovat na škále od 0 do 35, kde 0 znamená, že strany na všechny otázky odpověděly úplně stejně, a 35, že ke každému problému zaujímají naprosto odlišný postoj. Pokud strany k některé otázce svůj názor nevyjádřily, budeme předpokládat, že se k danému problému staví napůl pro a napůl proti. Nepodobnost mezi dvěma stranami pak určíme takto. Budeme procházet jejich odpovědi a kdykoli budou u nějaké otázky zcela protichůdné, zvýšíme nepodobnost o jedničku. Bude-li odpověď jedné strany pro nebo proti a u druhé strany nebude uvedena žádná odpověď, nepodobnost zvýšíme o jednu polovinu.

V R budeme postupovat tak, že nejprve matici `odpovedi`, jež se složena z `"Pro"`, `"Proti"` a `"- "`, změníme na číselnou matici `odpovedi2`, kde `1` bude zastupovat odpovědi `"Pro"`, `0` odpovědi `"Proti"` a `0.5` neuvedené odpovědi `"-"`:

```
> odpovedi = as.matrix(odpovedi)
> odpovedi2 = matrix(c(rep(0,nrow(odpovedi)*
+             ncol(odpovedi))),
+             nrow(odpovedi),ncol(odpovedi))
> for(i in 1:nrow(odpovedi2))
+   for(j in 1:ncol(odpovedi2))
+     if (odpovedi[i,j] == "Pro"){
+       odpovedi2[i,j]=1
+     } else if (odpovedi[i,j] == "Proti"){
+       odpovedi2[i,j]=0
+     } else {
+       odpovedi2[i,j]=0.5
+     }
```

Matici nepodobností pak získáme následovně:

```
> Delta_stranyPS = matrix(c(rep(0,ncol(odpovedi2)^2)),
+                  ncol(odpovedi2),ncol(odpovedi2))
> for(i in 1:nrow(Delta_stranyPS))
+   for(j in 1:ncol(Delta_stranyPS))
+     for(k in 1:nrow(odpovedi2))
+     {Delta_stranyPS[i,j] = Delta_stranyPS[i,j] +
+      abs(odpovedi2[k,i] - odpovedi2[k,j])}
> colnames(Delta_stranyPS) = rownames(Delta_stranyPS) =
+                            colnames(odpovedi)
> Delta_stranyPS
        CSSD  ANO KSCM TOP 09  ODS USVIT KDU-CSL
CSSD     0.0 12.0  5.0   17.0 20.0  18.5     8.5
ANO     12.0  0.0 12.0   14.0 18.0  13.5    13.5
KSCM     5.0 12.0  0.0   19.0 22.0  20.5     9.5
TOP 09  17.0 14.0 19.0    0.0 12.0  12.5    10.5
ODS     20.0 18.0 22.0   12.0  0.0  11.5    15.5
USVIT   18.5 13.5 20.5   12.5 11.5   0.0    13.0
KDU-CSL  8.5 13.5  9.5   10.5 15.5  13.0     0.0
```

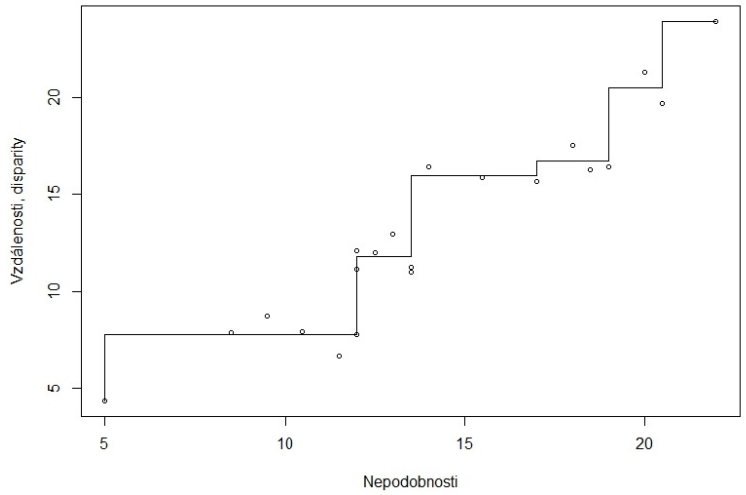
Využitím funkce `isoMDS` získáme dvourozměrné řešení:

```
> library("MASS")
> reseni_stranyPS = isoMDS(Delta_stranyPS)
initial  value 10.649730
iter   5 value 5.281802
iter  10 value 4.435196
iter  15 value 4.322084
iter  20 value 4.226686
final  value 4.225005
converged
> reseni_stranyPS
$points
              [,1]       [,2]
CSSD     -8.396533  1.4010964
ANO      -3.242085  7.1758347
KSCM    -10.992625 -2.0913510
TOP 09    4.605881 -7.2857362
ODS      12.864873  0.2342304
USVIT     7.636253  4.3490228
KDU-CSL  -2.475763 -3.7830971

$stress
```

```
[1] 4.225005
```

Vidíme, že minima ztrátové funkce *Stress* bylo dosaženo mezi dvacátou a pětadvacátou iterací. Hodnota S_2 je rovna 0,042 a dvourozměrné řešení je tak dle tabulky 3 vynikající.

Kvalita modelu je zřejmá i z příslušného Shepardova diagramu, jenž se nachází na obrázku 7.



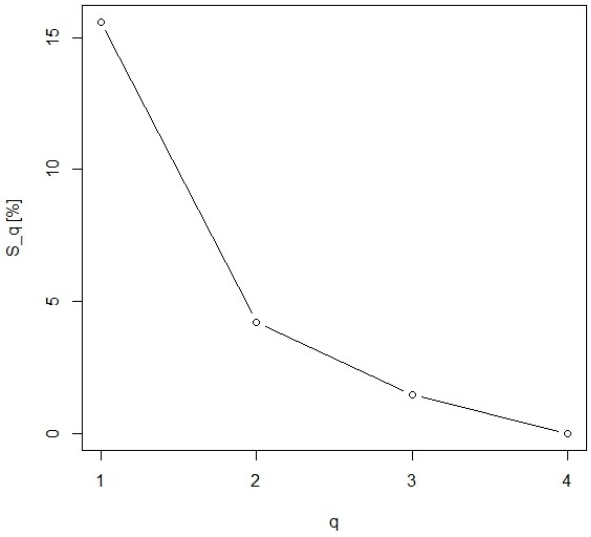
Obrázek 7: Shepardův diagram k příkladu o politických stranách v PS

Dimenze řešení rovna dvěma se jeví jako vhodně zvolená i podle elbow kritéria (viz obrázek 8).

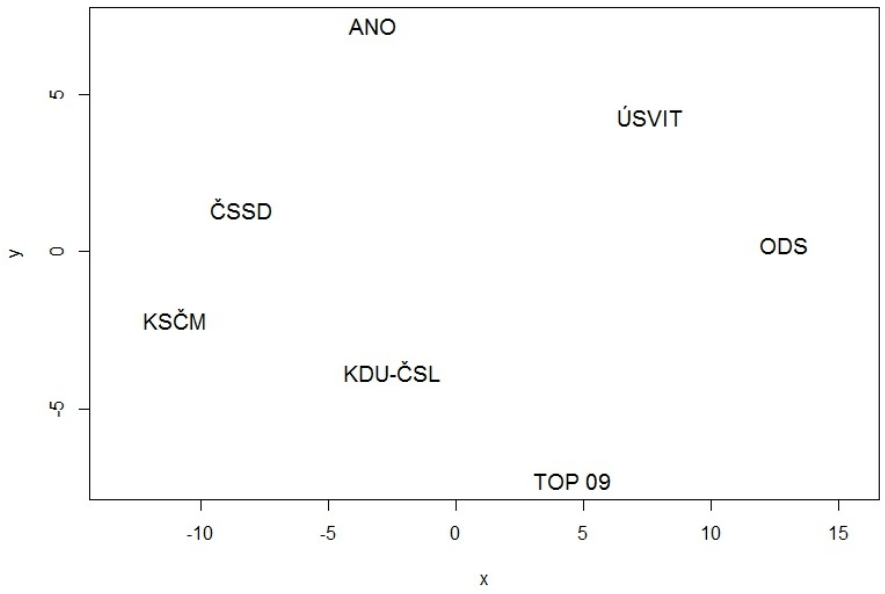
Nyní pomocí získaných souřadnic vytvoříme mapu objektů, která se nalézá na obrázku 9.

Nakonec se zaměříme na interpretaci výsledného řešení. Podíváme-li se na rozmístění stran vzhledem k prvnímu směru, vidíme, že v levé části grafu leží levicová KSČM a ČSSD, v pravé pravicové ODS a TOP 09 a mezi nimi středová KDU-ČSL. Z tohoto pohledu bychom pak ÚSVIT mohli považovat za pravicovou stranu a ANO za středovou, případně středolevicovou.

Co se týče druhého směru řešení, jeho objasnění už tak zjevné nebude. Můžeme si však všimnout, že v horní části grafu leží nově vzniklé strany ANO a ÚSVIT



Obrázek 8: Elbow kritérium k příkladu o politických stranách v PS



Obrázek 9: Mapa objektů k příkladu o politických stranách v PS

a v dolním konzervativní TOP 09 a KDU-ČSL. Projdeme-li si jejich odpovědi a to, u jakých otázek se lišily, pak bychom mohli říct, že poloha stran v tomto směru zhruba souvisí s jejich populistickým charakterem.

6 Korespondenční analýza

Informace obsažené v této kapitole, jež se zabývá využitím klasického MDS v mnohorozměrné statistické metodě s názvem korespondenční analýza, jsou čerpány z literatury [2] a [5].

Korespondenční analýza slouží jako nástroj pro analýzu vztahů mezi kategoriálními proměnnými v *kontingenčních tabulkách*. Úkolem metody je najít souřadnice v nízko-dimenzionálním (dvou či trojrozměrném) prostoru reprezentující řádkové a sloupcové kategorie, které mají pomoci objasnit vzájemnou (ne)závislost mezi proměnnými a vztahy mezi jednotlivými kategoriemi.

Jednou z variant *jednoduché korespondenční analýz*y (tj. pro dvourozměrné kontingenční tabulky) je *duální škálování*, jež spočívá v aplikaci klasického MDS zvlášť na matici sloupcových a zvlášť na matici řádkových vzdáleností.

Obecně lze dvourozměrnou kontingenční tabulku pro řádkovou proměnnou X s r kategoriemi a sloupcovou proměnnou Y s s kategoriemi zapsat tak, jak je to uvedeno v tabulce 6.

X/Y	1	2	...	s	$\sum$
1	n_{11}	n_{12}	...	n_{1s}	n_{1+}
2	n_{21}	n_{22}	...	n_{2s}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	...	n_{rs}	n_{r+}
$\sum$	n_{+1}	n_{+2}	...	n_{+s}	n

Tabulka 6: Kontingenční tabulka

Kontingenční tabulka s r řádky a s sloupci má rs buněk. Všech n pozorovaných objektů je klasifikováno podle řádkové a sloupcové kategorie. Každá $ij-$tá buňka obsahuje číslo n_{ij}, které udává napozorovanou četnost v $i-$té řádkové a $j-$té sloupcové kategorii. Marginální četnosti značíme $n_{i+} = \sum_{j=1}^{s} n_{ij},\ i = 1, \ldots, r$ a $n_{+j} = \sum_{i=1}^{r} n_{ij},\ j = 1, \ldots, s$. Platí $n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i+} = \sum_{j=1}^{s} n_{+j}$.

6.1 Duální škálování

Body zastupující jednotlivé kategorie proměnných dostaneme pomocí duálního škálování následovně. Nejprve z kontingenční tabulky zkonstruujeme *řádkové poměry* :

$$p_{ij}^{(r)} = \frac{n_{ij}}{n_{i+}},\ i = 1, \ldots, r,\ j = 1, \ldots, s \tag{34}$$

a *sloupcové poměry*:

$$p_{ij}^{(s)} = \frac{n_{ij}}{n_{+j}},\ i = 1, \ldots, r,\ j = 1, \ldots, s. \tag{35}$$

Poté s využitím řádkových a sloupcových poměrů získáme *chí-kvadrát vzdálenosti mezi řádky i a j* (řádkové vzdálenosti), jež jsou definovány:

$$d_{ij}^{(r)} = \sum_{k=1}^{s} \frac{1}{p_{+k}} \left(p_{ik}^{(r)} - p_{jk}^{(r)} \right)^2,\ i, j = 1, \ldots, r, \tag{36}$$

kde

$$p_{+k} = \frac{n_{+k}}{n} \tag{37}$$

a *chí-kvadrát vzdálenosti mezi sloupci i a j* (sloupcové vzdálenosti), jež jsou definovány:

$$d_{ij}^{(s)} = \sum_{k=1}^{r} \frac{1}{p_{k+}} \left(p_{ik}^{(s)} - p_{jk}^{(s)} \right)^2,\ i, j = 1, \ldots, s, \tag{38}$$

kde

$$p_{k+} = \frac{n_{k+}}{n}. \tag{39}$$

Řádkové a sloupcové vzdálenosti představují vážené euklidovské vzdálenosti s váhami $1/p_{+k}$, respektive $1/p_{k+}$. Podle tvaru vah vidíme, že kategorie s nižší četností mají na vzdálenosti větší vliv než ty s četností vyšší .

Nyní přejdeme k duálnímu škálování, což znamená, že nejdříve aplikujeme klasické MDS na matici řádkových vzdáleností a pak na matici sloupcových vzdáleností. Souřadnice (ve dvou, případně třídimenzionálním prostoru) odpovídající

řádkovým kategoriím a souřadnice pro sloupcové kategorie zobrazíme do jednoho grafu. Tento graf, jenž obsahuje $r + s$ bodů, je označován jako *korespondenční mapa*. Je žádoucí, aby body reprezentující řádkové kategorie (řádkové body) byly označeny odlišným symbolem či barvou než ty patřící ke kategoriím sloupcovým (sloupcové body). Pokud jsou řádkové nebo sloupcové kategorie určitým způsobem seřazeny, např. podle věku, je vhodné pro lepší vizualizaci body těchto kategorií v příslušném pořadí propojit

Výsledná korespondenční mapa je interpretována podle rozestavění bodů, jež reflektuje vztahy mezi jednotlivými kategoriemi. Řádkové (sloupcové) body, které jsou blízko sebe indikují řádkové (sloupcové) kategorie s podobnými procentními podíly (podmíněnou pravděpodobností). Řádkové body, jež leží blízko sloupcovým bodům, naznačují, že příslušná kombinace se objevuje častěji, než by bylo v případě nezávislosti proměnných X a Y. Obecně bychom však neměli konstatovat, že je-li určitý řádkový bod velmi blízko některému sloupcovému bodu, pak jsou příslušné kategorie spolu spřízněné a naopak.

Platí, že pro kontingenční tabulku s r řádky a s sloupci lze přesné řešení najít v prostoru dimenze rovné $\min\{r-1, s-1\}$. Jsou-li tak r i s větší než 3, pak korespondenční mapa ve dvourozměrném prostoru, jež se standardně užívá, představuje pouhou aproximaci skutečného řešení.

6.2 Barva vlasů a očí

V knize [1], v kapitole o kontingenčních tabulkách, najdeme příklad, ve kterém je využitím chí-kvadrát testu nezávislosti dokázáno, že existuje závislost mezi barvou očí a barvou vlasů. My se na tentýž příklad podíváme pohledem korespondenční analýzy.

K dispozici máme data týkající se $n = 6800$ mužů, u kterých byla zjišťována barva očí a vlasů. Získané údaje jsou uvedeny v tabulce 7.

Příklad budeme řešit v softwaru R. Nejprve pod názvem `tab` vytvoříme tabulku odpovídající dané kontingenční tabulce. Dostaneme:

oči/vlasy	světlé	kaštanové	černé	zrzavé	$\sum$
modré	1768	807	189	47	2811
šedé nebo zelené	946	1387	746	53	3132
hnědé	115	438	288	16	857
$\sum$	2829	2632	1223	116	6800

Tabulka 7: Kontingenční tabulka k příkladu o barvě vlasů a očí

```
> tab
                  svetle kastanove cerne zrzave
modre               1768        807   189      47
sede nebo zelene     946       1387   746      53
hnede                115        438   288      16
```

Pomocí následně definované funkce `f` získáme matice řádkových a sloupcových vzdáleností:

```
> f = function(x){
+     a = t(t(x)/colSums(x))
+     ret = sqrt(colSums((a[, rep(1:ncol(x), ncol(x))]-
+         a[, rep(1:ncol(x), rep(ncol(x), ncol(x)))])^2*
+         sum(x)/rowSums(x)))
+     matrix(ret, ncol = ncol(x))}
> Dr = f(t(tab))
> Dr
          [,1]      [,2]     [,3]
[1,] 0.0000000 0.6943064 1.058330
[2,] 0.6943064 0.0000000 0.364952
[3,] 1.0583299 0.3649520 0.000000
>
> Ds = f(tab)
> Ds
          [,1]      [,2]      [,3]      [,4]
[1,] 0.0000000 0.6716967 1.0007090 0.4738400
[2,] 0.6716967 0.0000000 0.3297843 0.2014926
[3,] 1.0007090 0.3297843 0.0000000 0.5275901
[4,] 0.4738400 0.2014926 0.5275901 0.0000000
```

Užitím příkazu `cmdscale` dostaneme souřadnice řádkových a sloupcových bodů:

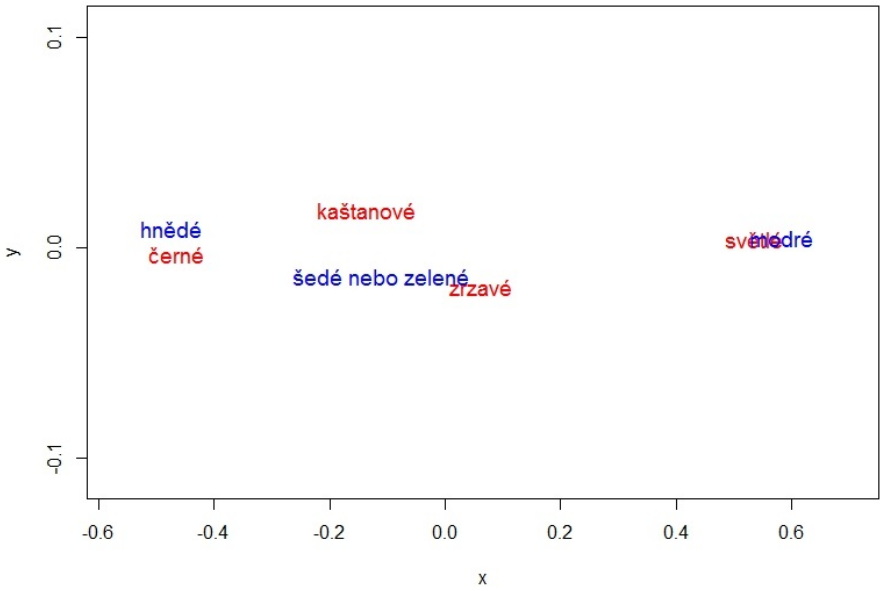
```
> r = cmdscale(Dr, eig = TRUE)
```

```
> r
$points
            [,1]          [,2]
[1,]  0.5841290  0.004682886
[2,] -0.1099365 -0.013605814
[3,] -0.4741925  0.008922928

$eig
[1]  5.781512e-01  2.866662e-04 -4.079141e-17

$x
NULL

$ac
[1] 0

$GOF
[1] 1 1

> s = cmdscale(Ds, eig = TRUE)
> s
$points
             [,1]          [,2]
[1,]  0.53638878  0.003894160
[2,] -0.13516706  0.017649465
[3,] -0.46429547 -0.003139508
[4,]  0.06307375 -0.018404117

$eig
[1]  5.255316e-01  6.752361e-04  3.469447e-17 -4.258374e
   -18

$x
NULL

$ac
[1] 0

$GOF
[1] 1 1
```

Prvky vektoru `$GOF` odpovídají tomu, že souřadnice ve dvourozměrném prostoru poskytují přesné řešení, neboť $\min\{r-1, s-1\} = \min\{3-1, 4-1\} = 2$.

Nakonec vytvoříme korespondenční mapu, kterou najdeme na obrázku 10:

```
> plot(s$points, xlim = range(s$points[,1],r$points[,1])
+      *1.2, ylim = range(s$points[,2],r$points[,2])*5
+      *1.2, type = "n", xlab = "x", ylab = "y",
+      yaxp = c(-0.1,0.1,2))
> text(s$points, labels = colnames(tab), col = "red",
+      cex = 1.2)
> text(r$points, labels = rownames(tab), col = "blue",
+      cex = 1.2)
```



Obrázek 10: Korespondenční mapa k příkladu o barvě vlasů a očí

Z obrázku 10 je patrné, že existuje závislost mezi barvou očí a barvou vlasů, neboť přinejmenším kombinace černé vlasy a hnědé oči, světlé vlasy a modré oči i zrzavé vlasy a šedé nebo zelené oči se objevují častěji, než by tomu bylo v případě vzájemné nezávislosti.

Závěr

Záměrem této diplomové práce bylo objasnit mnohorozměrnou statistickou metodu s názvem vícerozměrné škálování (MDS). Jde o metodu, která nám na základě znalosti o vzájemných nepodobnostech mezi objekty umožní získat jejich grafické zobrazení v nízkodimenzionálním prostoru. Matice daných nepodobností bývá označována jako proximitní matice a grafu s výslednými body říkáme mapa objektů.

Základním dělením MDS je dělení na metrické a nemetrické metody. V metrických úlohách pracujeme se vzdálenostmi mezi objekty a při hledání řešení využíváme teorii spektrálního rozkladu matice. Snažíme se získat takovou mapu objektů, ve které by vzdálenosti mezi body co nejlépe aproximovaly vzdálenosti původní. Nemetrické MDS používáme tehdy, máme-li k dispozici informaci o pořadí hodnot nepodobností mezi objekty, a řešení získáme pomocí optimalizačního iterativního algoritmu. Chceme docílit toho, aby vzdálenosti mezi body ve výsledné konfiguraci zachovávaly pořadí příslušných nepodobností.

MDS může sloužit jako nástroj pro jinou statistickou metodu, a to korespondenční analýzu. Jde o metodu, která pracuje s kontingenčními tabulkami. Duální škálování lze použít k vizualizaci kategorií proměnných a ke zjištění, zda jsou proměnné vzájemně závislé či nezávislé.

Literatura

[1] Anděl, J.: *Základy matematické statistiky.* MATFYZPRESS, Praha, 2002.

[2] Everitt, B., Hothorn, T.: *An Introduction to Applied Multivariate Analysis with R.* Springer, Heidelberg, 2011.

[3] Härdle, W. K., Simar, L.: *Applied Multivariate Statistical Analysis.* Springer, Berlín, 2003.

[4] Hebák, P., Hustopecký J., Pecáková I., Průša M., Řezanková H., Svobodová A., Vlach P.: *Vícerozměrné statistické metody 3.* INFORMATORIUM, Praha, 2007.

[5] Izenman, A. J.: *Modern Multivariate Statistical Techniques.* Springer, New York, 2008.

[6] *Kalkulátor leteckých vzdáleností mezi dvěma městy* [online]. [cit. 2015-11-02]. Dostupné z: `http://cs.thetimenow.com/distance-calculator.php`.

[7] *Volební kalkulačka k volbám do Poslanecké sněmovny 2013* [online]. [cit. 2015-11-02]. Dostupné z: `http://volebnikalkulacka.cz/poslanecka-snemovna-2013/`.

A Příloha

A.1 Otázky z volební kalkulačky

Seznam otázek ve volební kalkulačce [7].

1 Zavedení celostátních referend

2 Zachování Senátu

3 Přímá volba prezidenta

4 Prezidentské milosti

5 Instituce po celé ČR

6 Výpověď bez udání důvodů

7 Přijetí eura

8 Stop církevním restitucím

9 Daň na škodlivé potraviny

10 Zachování poplatku za prázdná média

11 Vlastnická struktura na internetu

12 Progresivní daň

13 DPH na potraviny

14 Privatizace Budvaru

15 Zdravotnické poplatky

16 Zvýšení minimální mzdy

17 Zachování 2. pilíře

18 Zavedení registračních pokladen

19 Otevřená výběrová řízení

20 Soudci doživotně

21 Smlouvy povinně na internetu

22 Možnost zabavit neprokázaný majetek

23 Zachování biosložek v palivech

24 Školné na VŠ

25 Místo ve školce

26 Očkování dobrovolné

27 Adopce homosexuálními páry

28 Romští asistenti

29 Zákaz kouření v restauracích

30 Legalizace marihuany

31 České potraviny povinně

32 Garantované připojení k internetu

33 Zachování limitů těžby uhlí

34 Členství v NATO

35 Členství v EU