# A többdimenziós skálázáson alapuló statisztikai módszerek

Diplomamunka

Írta: Lőrincz Géza

Alkalmazott matematikus szak

Témavezető:

Pröhle Tamás, egyetemi tanársegéd
Valószínűségelméleti és Statisztika Tanszék
Eötvös Loránd Tudományegyetem, Természettudományi Kar

Eötvös Loránd Tudományegyetem
Természettudományi Kar

2011. január 6.

# Tartalomjegyzék

**1. Bevezetés** **3**

**2. A többdimenziós skálázás matematikai modellje** **4**
2.1. A `stress` függvény . . . . . 8
2.2. Prokrusztész analízis . . . . . 12

**3. A módszerhez alkalmazott főbb `R` függvények** **17**

**4. Biplot** **19**
4.1. A matematikai modell . . . . . 19
4.2. Dimenzió csökkentés a biplotok alkalmazásához . . . . . 23
4.3. Biplot alkalmazása a lineáris regresszión keresztül . . . . . 27
4.4. Az MDS és a biplot kapcsolata . . . . . 29

**5. A módszer gyakorlati alkalmazása** **30**
5.1. Klaszterezés . . . . . 32
5.2. Biplotok alkalmazása . . . . . 42

**6. Összegzés** **47**

**7. Mellékletek** **48**

# 1. Bevezetés

Gyakran a megfigyeléseink olyan diszkrét valószínűségi változókra vonatkoznak, amelyek struktúráját valamilyen hasonlósági mérőszámmal jellemezhetjük. Ilyenek lehetnek például bizonyos szociológiai megfigyelések, amelyekben tulajdonságokat kell egy adott skálán pontozni vagy banki csődelőrejelzési vizsgálatok esetén az ügyfelek paramétereire vonatkozó adatok. A kategóriákat a számunkra leginkább értelmezhető környezetben azaz kettő vagy három dimenziós térben szeretnénk elhelyezni, majd a kapott ábra alapján esetleg olyan rejtett kapcsolatokra rálátni, amely az értékelés és a kategóriák egyes tulajdonságai között fennállhatnak.

Tehát adott $n$ objektumunk melyek $p$ darab tulajdonságára egy-egy megfigyelést végeztünk, ebből elkészíthetünk egy $n \times p$-es adatmátrixot, melyben a mátrix $n$ sora reprezentálja az objektumokat (pontokat), $p$ oszlopa pedig a pontokra adott megfigyeléseket. Célunk, hogy az objektumainkat egy euklideszi térbe ágyazzuk bele. Ha kettő elemünk bizonyos értelemben távol volt egymástól az eredeti $p$-dimenziós terünkben, akkor azt szeretnénk, hogy ebben a $k$-dimenziós euklideszi térben is távol essenek egymástól, azaz az elemeink a redukált térben is - bizonyos hiba határon belül - ugyanolyan távol legyenek egymástól. A fenti adatredukciós eljárás egyik matematikai eszköze a többdimenziós skálázás (*Multidimensional Scaling - MDS*), amely Torgerson-Gower skálázás néven is ismert. Legelső alkalmazásai, melyek többek között a módszer fejlődését is indukálták a magatartástudományokban, pszihometrikában jelentek meg, majd ezt követően vált népszerűvé és kezdett széles körben elterjedni.

A szakirodalomban leggyakrabban metrikus, nem metrikus, súlyozott és általánosított MDS-ről olvahsatunk. A szakdolgozatban a továbbikaban csak a metrikus és nem-metrikus esettel fogok foglakozni. A matematikai modell leírásával kezdem, majd megvizsgálom, hogy milyen értékekkel lehet a hiba nagyságát becsülni és hogyan lehet több megoldást összehasonlítani (Prokrusztész analízis). Mindezek után bevezetem az úgynevezett biplot ábrákat, amelyek szintén egy adatelemzési technika elemei, amellyel bizonyos objektumok és az azokhoz köthető változók kapcsolatát jól szemléltető ábrákat kaphatunk. A lényege, hogy egy kettő vagy három dimenziós térben ábrázolt pontok mellé a tér főtengelyein kívül elhelyezzünk további ún. biplot tengelyeket, amelyekre a pontok vetítésével kiolvashatjuk meghatározott változók értékeit. Majd megvizsgálom annak egy alkalmazását a lineáris regresszióban, továbbá rámutatok az MDS és a biplot közös felhasználhatóságára is.

Mindezek után egy valós gazdasági feladat vizsgálatára vállalkoztam, aminek a segítségével jól tudom szemléltetni az elméleti eredményeket a gyakorlatban is, és megmutathatom a segítségével annak gyakorlatiasságát. A számítások elvégzéséhez az `R` statisztikai programcsomagot használtam.

# 2. A többdimenziós skálázás matematikai modellje

Első lépésben a mintavételi eljárás után, elkészítjük az $X$ adatmátrixot és abból az ún. $\Delta = (\delta_{ij})$ különbözőségi (vagy éppen hasonlósági) mátrixot, ahol $\delta_{ij}$ jelöli az $i$. és a $j$. elem különbözőségét vagy másnéven különbözőségi koefficiensét. A mintavételezésünk eredménye egy $X$-el jelölt $n \times p$-es adatmátrix, mely sorai reprezentálják az objektumokat oszlopai pedig a megfigyelések számát, hogy ebből hogyan is készítjük el a különbözőségi mátrixot az már korántsem olyan egyértelmű. A megfigyelési változóink lehetnek folytonos (kvantitatív) és diszkrét vagy másnéven kategorikus változók, a diszkrét változókon belül megkülönböztetünk csoportosító és minősítő (kvalitatív) változókat, e kettő között a különbség annyi, hogy előbbinél számít a sorrend utóbbinál pedig nem. Kategorikus változók esetén azt mondjuk, hogy az $i$. és $j$. elem különbözősége 0, ha azonos kategóriába tartoznak és 1 különben. Bináris változókról akkor beszélünk, ha az csak kettő értéket vehet fel. Bináris különbözőség-mátrixból könnyen kapunk a triviális $s_{ij} = 1 - \delta_{ij}$ transzformációval hasonlósági mátrixot. A következő 1. táblázat mutat néhány példát arra, hogy miképpen lehet ezeket a különbözőséget kvantitatív illetve hasonlóságot a bináris változót tartalmazó adatmátrixokból nyerni.

| Különbözőségi mérték (Kvantitatív változókra) | $\delta_{ij}$ | Hasonlósági mérték (Bináris változókra) | $s_{ij}$ |
|---|---|---|---|
| Euklideszi távolság | $\left\{\sum_{t=1}^{p}(x_{it}-x_{jt})^2\right\}^{1/2}$ | Braun, Blanque | $\frac{a}{max\{(a+b),(a+c)\}}$ |
| Súlyozott Euklideszi távolság | $\left\{\sum_{t=1}^{p} w_t(x_{it}-x_{jt})^2\right\}^{1/2}$, $w_i$ súlyokkal | Czekanowski, Sorensen, Dice | $\frac{2a}{2a+b+c}$ |
| "City Block" metrikus | $\sum_{t=1}^{p} \mid x_{it}-x_{jt} \mid$ | Hamman | $\frac{a-(b+c)+d}{a+b+c+d}$ |
| Minkowski metrikus | $\left\{\sum_{t=1}^{p} w_t(x_{it}-x_{jt})^{\lambda}\right\}^{1/\lambda}$, $\lambda \geq 0$ | Jaccard koefficiens | $\frac{a}{a+b+c}$ |
| Bhattacharyya metrikus | $\sqrt{\sum_{t=1}^{p}\left(\sqrt{x_{it}}-\sqrt{x_{jt}}\right)^2}$ | Kulczynski | $\frac{a}{b+c}$ |
| Wave Hedges | $\sum_{t=1}^{p}\left(1-\frac{min(x_{it},x_{jt})}{max(x_{it},x_{jt})}\right)$ | Kulczynski | $\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$ |
| Soergel | $\frac{1}{p}\frac{\sum_{t=1}^{p}\lvert x_{it}-x_{jt}\rvert}{\sum_{t=1}^{p} max(x_{it},x_{jt})}$ | Rogers, Tanimoto | $\frac{a+d}{a+2b+2c+d}$ |

, ahol $x_{it}$ az $X$ adatmátrix i. sorának t. eleme, $a = \mid \{t : x_{it} = x_{jt} = 1, t = 1 \ldots p\} \mid$; $b = \mid \{t : x_{it} = 1, x_{jt} = 0, t = 1 \ldots p\} \mid$ ; $c = \mid \{t : x_{it} = 0, x_{jt} = 1, t = 1 \ldots p\} \mid$ és $d = \mid \{t : x_{it} = x_{jt} = 0, t = 1 \ldots p\} \mid$

1. Táblázat

Ha egy $W$ adatmátrixunk többféle típusú megfigyelést tartalmaz akkor vegyes adatmátrixról beszélünk. Gower (1971) definiál egy általános hasonlósági koefficienst

$$s_{ij} = \frac{\sum_{t=1}^{p} w_{ijt} s_{ijt}}{\sum_{t=1}^{p} w_{ijt}},$$

ahol $s_{ijt}$ jelöle az i. és j. objektum hasonlóságát a $t$. változóban, továbbá $w_{ijt} = 1$, ha az $i$. és a $j$. elem összehasonlítható a $t$. változó szempontjából , különben 0.

Gower a kvantitatív változókra a következőt javasolja:

$$s_{ijt} = 1 - \frac{\mid x_{it} - x_{jt} \mid}{R_t},$$

ahol $R_t$ a $t$. változóhoz tartozó megfigyelések terjedelme, azaz $\max_i x_{it} - \min_i x_{it}$.

Bináris változók esetén

$$s_{ijt} = \begin{cases} 1 & , \textit{ha az i. és j. elem a t. változóban } 1 \\ 0 & , \textit{különben} \end{cases}$$

$$w_{ijt} = \begin{cases} 0 & , \textit{ha az i. és j. elem a t. változóban } 0 \\ 1 & , \textit{különben} \end{cases}$$

Kategorikus változó esetén

$$s_{ijt} = \begin{cases} 1 & , \textit{ha az i. és j. elem egyazon kategóriában sorolható} \\ 0 & , \textit{különben} \end{cases}$$

A megfigyeléseink természetesen lehetnek más-más mértékegységűek, ezért az összehasonlításuk csak akkor lehetséges ha azonos dimenziójúak, ezért ha különböző mértékegységűek az adataink, akkor szükséges a dimenzió mentesítés azaz, hogy azonos nagyságrendűre hozzuk azokat, amelynek legegyszerűbb módja a százalék számítás. Vagy történhet úgy is, hogy az adatokat egy közös dimenzióban fejezzük ki, amennyiben lehetséges, ez utóbbi lehetőség eléggé nehézkes is lehet és erősen feladat függő.

Adott n objektumunk és a rajtuk értelmezett $\delta_{ij}$ különbözőségek egy $p$-dimenziós térben. Az MDS megpróbálja megtalálni egy alacsonyabb dimenziós térnek azon $n$ pontját, amelyek az objektumokat reprezentálják és a térbeli távolságok monoton függvényei legyenek az adatok közti különbözőségeknek. Vagyis ha $\delta_{ij} < \delta_{kl}$ akkor az $i$. és $j$. illetve a $k$. és $l$. elem közötti $d_{ij}$, $d_{kl}$ távolságokra fennáll a $d_{ij} \leq d_{kl}$. A gyenge monotonitási kritériumot teljesítő MDS módszereket nem-metrikus módszereknek nevezzük. Ha a pontok közötti $d_{ij}$ távolságok a következőképpen van definiálva:

**Definíció.** $f$ legyen egy folytonos paraméterű *monoton* függvény, ekkor ha a

$$d_{ij} = f(\delta_{ij}).$$

függvénnyel definiált ($f$ lehet maga az identitás is.) Akkor metrikus többdimenziós skálázásról beszélünk.

Célunk olyan $\mathbb{R}^n$-beli $k$-dimenziós altér meghatározása, amelyre ortogonális vetítéssel a pontjaink távolság mátrixa a lehető legkisebb mértékben tér el az eredeti távolságmátrixtól. Ebben az altérben a szokásos módon értelmezzük az euklideszi távolságot, az $X_i$ és a $X_j$ elem távolsága

$$\hat{d}_{ij} = \|X_i - X_j\| = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2} \quad (i, j = 1 \ldots n)$$

A klasszikus MDS *(Classical Multidimensional Scaling)* matematikai modell tárgyalásánál szükségünk lesz a következő definíciókra.

**Definíció.** a $D \in \mathbb{R}^{nxn}$ mátrixot *távolság-mátrixnak* nevezzük, ha eleget tesz a következő tuljadonságoknak:

$(i)$ $d_{ii} = 0$, $i \in \{1 \ldots n\}$
$(ii)$ $d_{ij} = d_{ji} \geq 0$, $1 \leq i < j \leq n$
$(iii)$ $d_{ik} \leq d_{ij} + d_{jk}$ $i, j, k \in \{1 \ldots n\}$, azaz teljesül a háromszög-egyenlőtlenség.

**Definíció.** a fenti D távolság-mátrixot *euklideszi*nek nevezzük, ha rögzített $p$ egész mellett, léteznek olyan $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$vektorok, melyekre fennáll a következő összefüggés

$$d_{ij} = \| x_i - x_j \| \quad (i, j = 1, \ldots, n)$$

**Definíció.** $H_n$ az ún. *centráló mátrix.*

$$H_n = I_n - \frac{1}{n} 1_n 1_n^T$$

Ha n rögzített akkor $H := H_n$. A centráló mátrix segítségével belátható az álábbi tétel:

**1.1 Tétel.** a $D \in \mathbb{R}^{nxn}$ távolság-mátrix akkor és csak akkor euklideszi, ha a $B = HAH$ mátrix pozitív szemidefinit, ahol az $A$ mátrix elemire: $a_{ij} = \frac{1}{2} d_{ij}^2$.
**Bizonyítás.** [4]□

A gyakorlati problémák, feladatok kezelésében a mintavételi eljárását követően sok esetben eleve úgy készítjük el a távolság-mátrixot, hogy az euklideszi legyen, de többször a $D$ távolság-mátrixunk nem is euklideszi. Az utóbbi esetre konstruálhatunk egy olyan $\widetilde{D}$- al jelölt mátrixot, mely már teljesíti *(i)-(iii)* és $D$-nek egyik legjobb közelítésének tekinthető. Tehát a továbbiakban tegyük fel, hogy $D$ euklideszi, szeretnénk az objektumainkat egy alacsonyabb (például a szemléletes 1-, 2-, 3-dimenziós térbe) beágyazni, ehhez keressük azt a $k$-dimenziós alteret, melybe az objektumainkat ortogonálisan

vetítve az ottani távolság-mátrix a lehető legkisebb mértékben tér el az eredeti $D$ mátrixtól. Jelöljük X-el az adatmátrixunkat. $X \in \mathbb{R}^{n\times p}$.

**Állítás.** Ha $\widehat{D}$-al jelöljük az alacsonyabb $k$-dimenzióban vett távolságokat, akkor fennáll

$$\hat{d}_{ij}^2 \leq d_{ij}^2.$$

Az ortogonális vetítéshez tekintsük a $V = (V_1, V_2) \in \mathbb{R}^{p\times p}$ ortogonális mátrixot , ahol $V_1 = (v_1, \ldots, v_k) \in \mathbb{R}^{p\times k}$ illetve $V_2 = (v_{k+1}, \ldots, v_p) \in \mathbb{R}^{p\times(p-k)}$. $V_1$ az a mátrix, mely oszlopai által kifeszített altérre ($k$-dimenziós) szeretnénk vetíteni. A vetülethez tartozó távolság-mátrix elemeire fennáll

$$\hat{d}_{ij}^2 = \sum_{s=1}^{k} \left(x_i^T v_s - x_j^T v_s\right)^2$$

V ortogonalitása miatt az eredeti $X$ és a $XV$ mátrixok objektumokat reprezentáló sorai között a távolságok ugyanazok.

$$d_{ij}^2 = \sum_{s=1}^{p} (x_{is} - x_{js})^2 = \sum_{s=1}^{p} \left(x_i^T v_s - x_j^T v_s\right)^2.$$

Így $\hat{d}_{ij}^2 \leq d_{ij}^2$.

Definiáljuk az eredeti és a $k$-dimenziós térben mért távolságok eltérését a következő módon:

$$\psi(D, \widehat{D}) := \sum_{i=1}^{n} \sum_{j=1}^{n} \left(d_{ij}^2 - \hat{d}_{ij}^2\right) \qquad (\geq 0)$$

Tekintsük az $X^T X$ spektrálfelbontását: $\widehat{V}\Lambda\widehat{V}^T$, ahol $\Lambda$ diagonális és a főátlóbeli elemek a $B$ mátrix $\lambda_1 \geq \ldots \geq \lambda_p$ pozitív sajátértékei, ugyanis $B$ pozitív szemidefinit, ekkor fennáll a következő tétel

**1.2 Tétel**. A $\psi(D, \widehat{D})$ eltérést az $X^T X$ mátrix $k$ legnagyobb sajátértékeihez tartozó $\hat{v_1}, \ldots, \hat{v_k}$ sajátvektorokból, mint oszlopvektorokból álló $V_1 = \widehat{V_1}$ mátrix minimalizálja.
**Bizonyítás.**[ 4] □

A következő algoritmus előállítja a konfiguráció pontjainak vektorait. $f = id$ (metrikus esetre algorimtus).

1. Elkészítjük a $A := \left[-\frac{1}{2}\delta_{ij}^2\right]$ mátrixot, $i = 1 \ldots n$, $j = 1 \ldots n$.
2. $B := HAH$, ahol $H = I - \frac{1}{n}1_n 1_n^T$ és $1_n = (1, \ldots, 1)$ $n$-dimenziós vektor.
3. Állítsuk elő a spektrál felbontását $B$-nek: $B = V\Lambda V^T$, ahol $\Lambda$ diagoinlási mátrix a diagonálisában pedig a $B$ mátrix sajátértékei szerepelnek $V$ pedig a megfelelő sajátvektorok.

4. Ha a pontjainkat a $k$-dimenziós térben szeretnénk ábrázolni, akkor az első $k$ sajátértéke $B$-nek nem nulla a rákövetkező $n-k$ nulla, hagyjuk el ezeket $\Lambda$-ból és a megfelelő sajávektorokat $V$-ből, jelöljük az előbbit $\Lambda_1$ az utóbbit $V_1$-el.

5. $Y := V_1\Lambda_1^{1/2}$, ekkor az $Y$ sorai adják az objketumokat reprezentált pontok koordinátáit.

## 2.1. A `stress` függvény

Nyilvánvaló, hogy $m$-dimenziós pontokat $m$-dimenzióban tudunk a lehető legtökéletesebben ábrázolni, mivel a célunk éppen a dimenzió csökkentésre irányul, ezért egy alacsonyabb térbe való beágyazásakor mindeképpen kell hibával számolnunk, ebben a szakaszban erre térek ki részletesebben. A `stress` függvényre úgy érdemes gondolni mint függvények egy halmazára, amelyek különféle megközelítés szerint vizsgálják, hogy mekkora az eltérés az eredeti és az alacsonyabb dimenziós térben az elemek közötti távolságokra nézve. Tehát az illeszkedés pontosságának a mérésére több út is kínálkozik. Egyrészt van lehetőség, hogy a megkapott eredmény „jó illeszkedését" a $B$ sajátértékeinek a segítségével fejezzük ki.

$$\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{n-1}\mid\lambda_i\mid} \qquad vagy \qquad \frac{\sum_{i=1}^{p}\lambda_i}{\sum_{\lambda_i>0}\lambda_i}.$$

Másfelől a közelítés hatékonyságának egy másik szempotból való megközelítése, hogy az un. veszteségfüggvényeket mennyire tudjuk minimalizálni, a fent már definiált veszteségfüggvény mellett egy további legkézenfekvőbb függvény a következő lehet

$$\sigma_{direkt}(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\delta_{ij} - d_{ij}\left(Y\right)\right)^2 \qquad (1)$$

A gyakorlatban persze sokszor előfordul, hogy az adataink hiányosak, ekkor természetesen $\delta_{ij}$ is hiányos, pontosabban az adott $\delta_{ij}$-t nem tudjuk értelmezni. Ezen probléma kiküszöbölésére vezessük be a következő bináris változót

$$\omega_{ij} = \begin{cases} 1 & , ha\, létezik\, \delta_{ij} \\ 0 & , ha\, hiányzik\, \delta_{ij} \end{cases}$$

ekkor (1)-et a következőképpen írhatjuk át

$$\begin{aligned} \sigma^{\omega}_{direkt}(Y) &= \sum_{i<j}\omega_{ij}\left(\delta_{ij} - d_{ij}(Y)\right)^2 \\ &= \sum_{i<j}\omega_{ij}\delta_{ij}^2 + \sum_{i<j}\omega_{ij}d_{ij}^2(Y) - 2\sum_{i<j}\omega_{ij}\delta_{ij}d_{ij}(Y) = \\ &= \eta_\delta^2 + \eta^2(Y) - 2\rho(Y), \end{aligned}$$

ahol $\eta_\delta^2$ csak $\omega_{ij}$ és $\delta_{ij}$ től függ, amik fix értékűek így $\eta_\delta^2$ konstans, $\eta^2(Y)$ a $d_{ij}(Y)$ távolságok súlyozott négyzetösszege, végül $-2\rho(Y)$ a $d_{ij}(Y)$ távolságok súlyozott összege.

Önmagában $\sigma_{direkt}$ nem túl informatív, ugyanis például ha vesszük az objektumok közötti távolságokat valamilyen mértékegység szerint -például kilométerben- majd utána egy másik mértékenységben viszgáljunk, $\sigma_{direkt}$ megváltozik - a példánknál megmaradva ha méterben akkor a $\sigma_{direkt}$ 1000-szeresére változik- ami persze nem azt jelenti, hogy az egyik rosszabb konfiguráció volna mint a másik, éppen ezért normálva $d_{ij}$-vel skála-invariánsá tesszük

$$\sigma_1^2 = \frac{\sigma_{direkt}}{\sum_{ij} d_{ij}^2(Y)} = \frac{\sum_i \sum_j \left(\delta_{ij} - d_{ij}\left(Y\right)\right)^2}{\sum_{ij} d_{ij}^2(Y)} \qquad (2)$$

Mivel a gyakorlatban $\sigma_1$ elég kicsi, ezért a négyzetgyökét szokás venni, ezt nevezzük *Stress-1-nek* (Kruskal).

$$\sigma_1 = \sqrt{\frac{\sum_i \sum_j \left(\delta_{ij} - d_{ij}\left(Y\right)\right)^2}{\sum_{ij} d_{ij}^2(Y)}}$$

Ugyanígy létezik $\sigma_1^2$-nek egy $\left(\sigma_1^w\right)^2$ jelölt , súlyozott alakja is

$$\left(\sigma_1^w\right)^2 = \frac{\sum_{i<j} \omega_{ij} \left(\delta_{ij} - d_{ij}(Y)\right)^2}{\sum_{i<j} \omega_{ij} d_{ij}^2(Y)} = \frac{\sigma_{direkt}^{\omega}\left(Y\right)}{\eta^2\left(Y\right)}$$

Hasonlóan, ha a fenti $\sigma_{direkt}^{\omega}(Y)$-t normalizáljuk de most $\delta_{ij}$-vel, akkor *Normalizált Stress* értékről beszélünk

$$\sigma_n(Y) = \frac{\sum_{i<j} \omega_{ij} \left(\delta_{ij} - d_{ij}(Y)\right)^2}{\sum_{i<j} \omega_{ij} \delta_{ij}^2} = \frac{\sigma_{direkt}^{\omega}}{\eta_\delta^2}$$

Világos, hogyha $\sum_{i<j} \omega_{ij} \delta_{ij}^2 = 1$ akkor $\sigma_n(Y) = \sigma_{direkt}^{\omega}(Y)$.

Mivel a súlyozott esetnek speciális esete a súlyozatlan - speciálisan ha valamennyi $\omega_{ij} = 1$ - ezért a jelölések egyszerüsítése érdekében a további vizsgálatokkor a felső $\omega$ indexet, amely a súlyozottságot jelöli elhagyom, de mindvégig az általánosabb súlyozott esetet értem alatta. Vizsgáljuk meg $\sigma_1$ és $\sigma_n$ kapcsolatát. Mindenekelőtt ehhez szükségunk lesz egy megfigyelésre és egy állításra. Megfigyelésünk a következő, tegyük fel, hogy $Y^*$ lokális minimuma $\sigma_{direkt}(Y)$-nak, ekkor $X^* = \frac{1}{b} Y^*$ is lokális minumum, hiszen a koordináta mátrixot egy konstansal, az un. skála faktorral szoroztuk.

**Állítás.**

$$\sigma_n\left(bX^*\right) = 1 - \left(\frac{\rho\left(X^*\right)}{\eta_\delta \eta\left(X^*\right)}\right)^2$$

**Bizonyítás.** Felhasználva, hogy az euklideszi távolság homogén függvény abban a térben amelyre redukálunk, azaz $d_{ij}(bX^*) = bd_{ij}(X^*)$, ha $b \geq 0$, ekkor

$$\begin{aligned}
\sigma_{direkt}(bX^*) &= \sum_{i<j} \omega_{ij} (\delta_{ij} - d_{ij}(bX^*))^2 = \\
&= \sum_{i<j} \omega_{ij}\delta_{ij}^2 + b^2 \sum_{i<j} \omega_{ij} d_{ij}^2(X^*) - 2b \sum_{i<j} \omega_{ij}\delta_{ij} d_{ij}(X^*) = \\
&= \eta_\delta^2 + b^2\eta^2(X^*) - 2b\rho(X^*)
\end{aligned}$$

Deriváljuk $b$ szerint majd tegyük egyenlővé 0-val

$$\frac{\partial \sigma_{direkt}(bX^*)}{\partial b} = 2b\eta^2(X^*) - 2\rho(X^*) = 0$$

Így az optimális b$^* = \frac{\rho(X^*)}{\eta^2(X^*)}$ , helyettesítsük vissza $\sigma_{direkt}(bX^*)$-ba

$$\sigma_{direkt}(b^*X^*) = \eta_\delta^2 - \left(\frac{\rho(X^*)}{\eta(X^*)}\right)^2$$

Mindkét oldalt osztva $\eta_\delta^2$-vel kapjuk, hogy

$$\sigma_n(bX^*) = \frac{\sigma_{direkt}(b^*X^*)}{\eta_\delta^2} = 1 - \left(\frac{\rho(X^*)}{\eta_\delta \eta(X^*)}\right)^2$$

Nézzük meg milyen kapcsolat van a normalizált és az stress-1 értéke között. De Leeuw és Heiser bizonyította, hogy ha $Y^*$ lokális minimum hely, akkor $\eta^2(Y^*) = \rho(Y^*)$ így a fenti állításból következik, hogy

$$\sigma_n(Y^*) = 1 - \frac{\eta^2(Y^*)}{\eta_\delta^2} \Leftrightarrow \frac{\eta_\delta^2}{\eta^2(Y^*)} = \frac{1}{1 - \sigma_n(Y^*)} \qquad (*)$$

Vegyük $\sigma_1^2$-nek egy hasonló kifejezését mint $\sigma_{direkt}$-nek és használjuk fel újra De Leeuw és Heiser eredményét miszerint $\eta^2(Y^*) = \rho(Y^*)$

$$\begin{aligned}
\sigma_1^2(Y^*) &= \frac{\sum_{i<j} \omega_{ij} (\delta_{ij} - d_{ij}(Y^*))^2}{\sum_{i<j} \omega_{ij} d_{ij}^2(Y^*)} = \\
&= \frac{\sum_{i<j} \omega_{ij}\delta_{ij}^2 + \sum_{i<j} \omega_{ij} d_{ij}^2(Y^*) - 2\sum_{i<j} \omega_{ij}\delta_{ij} d_{ij}(Y^*)}{\sum_{i<j} \omega_{ij} d_{ij}^2(Y^*)} = \\
&= \frac{\eta_\delta^2 + \eta^2(Y^*) - 2\rho(Y^*)}{\eta^2(Y^*)} = \\
&= \frac{\eta_\delta^2 - \eta^2(Y^*)}{\eta^2(Y^*)} = \frac{\eta_\delta^2}{\eta^2(Y^*)} - 1
\end{aligned}$$

Felhasználva (*)-ot kapjuk , hogy

$$\sigma_1^2\left(Y^*\right)=\frac{1}{1-\sigma_n\left(Y^*\right)}-1=\frac{\sigma_n\left(Y^*\right)}{1-\sigma_n\left(Y^*\right)}$$

Azonban az $Y^*$ skálázása nem optimuma $\sigma_1$-nek, ezért vegyük figyelembe a $b$ skála faktort is. Ennek megfelelően felhasználva ismét, hogy $\eta^2\left(Y^*\right)=\rho\left(Y^*\right)$

$$\sigma_1^2\left(bY^*\right)=\frac{\eta_\delta^2+b^2\eta^2\left(Y^*\right)-2b\rho\left(Y^*\right)}{b^2\eta^2\left(Y^*\right)}=\frac{\eta_\delta^2+\left(b^2-2b\right)\eta^2\left(Y^*\right)}{b^2\eta^2\left(Y^*\right)} \qquad (**)$$

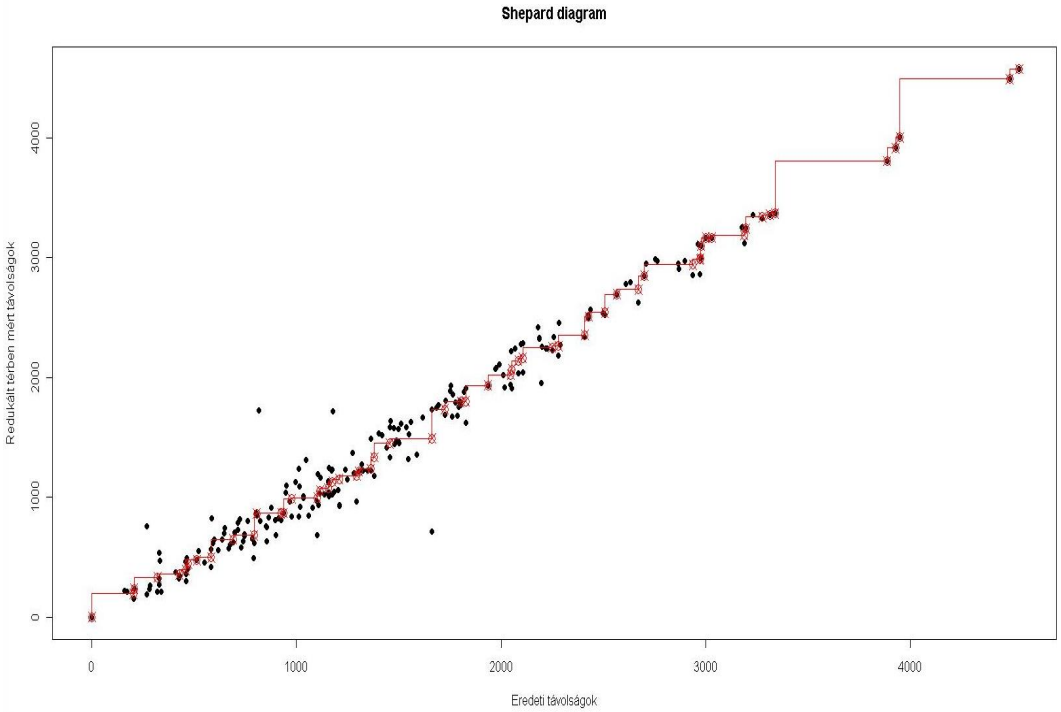
Deriválva $b$ szerint, majd egyenlővé téve 0-val, az optimális $b^*$ érték

$$\begin{aligned}\frac{\partial\sigma_1^2\left(bY^*\right)}{\partial b} &= \frac{b^2\left(2b-2\right)\eta^4\left(Y^*\right)-2b\eta^2\left(Y^*\right)\left[\eta_\delta^2+\left(b^2-2b\right)\eta^2\left(Y^*\right)\right]}{b^4\eta^4\left(Y^*\right)}= \\ &= \frac{2b\eta^2\left(Y^*\right)-2\eta_\delta^2}{b^3\eta^2\left(Y^*\right)}=0\end{aligned}$$

Így $b^*=\eta_\delta^2/\eta^2(Y^*)$ vissza helyettesítve (**)-ba kapjuk, hogy

$$\begin{aligned}\sigma_1^2\left(b^*Y^*\right) &= \frac{\eta_\delta^2+\frac{\eta_\delta^4}{\eta^4(Y^*)}\eta^2(Y^*)-2\frac{\eta_\delta^2}{\eta^2(Y^*)}\eta^2(Y^*)}{\frac{\eta_\delta^4}{\eta^4(Y^*)}\eta^2(Y^*)}= \\ &= \frac{\frac{\eta_\delta^2}{\eta^2(Y^*)}-1}{\frac{\eta_\delta^2}{\eta^2(Y^*)}}=1-\frac{\eta^2(Y^*)}{\eta_\delta^2}=\sigma_n\left(Y^*\right)\end{aligned}$$

Tehát az stress-1 és a normalizált stress érték megegyezik a lokális minimumban, feltéve ha a $b$ skála faktor a megfelelően van megválasztva. Az úgynevezett Shepard diagram elősegíti a kiértékelését a vizsgálatainknak. Ebben egy izotónikus regressziót (monoton regresszió) hajtunk végre az eredeti távolság-mátrix elemei és a redukált térben a nekik megfelelő távolságok között. Az alábbi ábrán erre láthatunk egy példát, ahol az európai nagyvárosok közti távolságok, mint távolság-mátrixra elvégzett többdimenziós skálázás outputjára végeztem el az izotonikus regressziót.




1. Ábra. *Az izotonikus regresszió az európai nagyvárosok többdimenziós skálázására*

Összességében a `stress`-érték több faktortól függ nézzünk ezek közül néhányat, hogy miként hatnak rá

1. Magasabb $n$ pontszám általában magsabb `stress` értéket ad.
2. Ha a redukált tér dimenziója $m$, akkor magass $m$ alacsonyabb `stress`-t generál.
3. Több hiba az adatokban nagyobb `stress`-t indukál.
4. A hiányzó adatok magas száma alacsony `stress` értéket ad.

## 2.2. Prokrusztész analízis

Az általánosított Prokrusztész analízis (Generalised Procrustes Analysis) egy módszer az adathalmazoknak egymáshoz képesti összeillés vizsgálatához különféle transzformációkkal, elsősorban rotációval. A módszernek alapvetően a lineáris és a nem-lineáris változata ismert, az optimális skálázás esetén a nem lineáris változatát használják. Az alkalmazásokból látható, hogy a különböző MDS módszerek hasonló de végtére is különböző konfigurációkat (megoldásokat) adnak. Ezen konfigurációk tetszőleges transzlációjával, rotációjával és reflexiójával azonos értékű megoldásokhoz jutunk. Így szükségünk van valamilyen értelemben vett összehasonlításukra a különböző megoldásoknak, szükséges van egyfajta összevetésre a két konfiguráció között. A Prokrusztész analízis egy izotróp dilatációt (nyújtás), transzlációt (eltolás), rotációt (forgatás) és reflexiót (tükrözés) talál a két konfiguráció közötti legjobb összeilléshez. Az elemzés részletes leírását és hasonló módszereket Gower és Dijksterhuis (2004) dolgozott ki.

Tegyük fel, hogy az $n$ pontú konfigurációnk pontonkénti koordinátái az $n \times q$ méretű $X$ mátrixal adott a $q$-dimenziós térben, egy másik $n$ pontú konfigárció koordinátaái a

$p$-dimenziós térben egy $n \times p$ méretű $Y$ mátrixal meghatározott, ahol $p \geq q$. Egészítsük ki $X$-et $\underline{0}$ oszlopvektorokkal, azzal a céllal , hogy $X$ és $Y$ dimenziója ugyanaz legyen. A diszkrepancia (eltérés) nagysága a következőképpen definiált

$$R^2 = \sum_{r=1}^{n} (y_r - x_r)^T (y_r - x_r),$$

ahol $X = [x_1, \ldots, x_n]^T$, $Y = [y_1, \ldots, y_n]^T$ és $x_r$, $y_r$ az $r$. pont koordináta vektorai a két térben.

Legyen $x_r^{'}$ az $r$. pont koordináta vektora egy tetszőleges nyújtása, forgatása, eltolása és tükrözése után $x_r$-nek, ekkor $x_r^{'}$ a következőképpen írható

$$x_r^{'} = \rho \phi^T x_r + b,$$

ahol $\rho$ a dilatáció mértéke, $\phi$ a forgatás és esetleg a reflexió mátrixa, $b$ pedig az eltolás vektor.

A fentiek optimális értékei, amelyeket elvégzeve $X$-en, az $R^2$-et minimalizálja $X$ és $Y$ között:

(1) Optiális transzláció : Mind a kettő konfiguráció középpontja a az origóban van.

(2) Optimális rotáció : határozzuk meg az $\phi = \left(X^T Y Y^T X\right)^{\frac{1}{2}} \left(Y^T X\right)^{-1}$ és forgassuk $X-$et $X\phi-$ba.

(3)Optimális dilatáció: Az $X$ konfiguráció minden egyes pontját szorozzuk meg a $\rho = \frac{tr\left(X^T Y Y^T X\right)}{tr(X^T X)}$

Majd az $R^2$ értékét a következőképpen számoljuk

$$0 \leq R^2 = 1 - \frac{\left(tr\left(X^T Y Y^T X\right)^{\frac{1}{2}}\right)^2}{tr\left(X^T X\right) tr\left(Y^T Y\right)} \leq 1,$$

ahol a triviális módon a 0 érték jelöli a legjobb összeillést, és minél nagyobb $R^2$ annál kevésbé jó a hasonlóság. Tehát ha adott $X$, $Y$ konfiguráció akkor az $Y$-t $R^2$ szerint a legjobban $X$-nek a fenti műveletek elvégzése utáni alakja közelíti.

Vizsgáljuk meg a kérdéskört általánosabban. Legyen $A$ a cél konfiguráció és $B$ a vizsgált konfiguráció. Jelöljük most a forgatás és tükrözés mátrixát $T$-vel a dilatációt $s$-el. Feltehető, hogy $TT^{'} = T^{'}T = I$. A koordináták közti különbséget $A - sBT$ mátrix adja. Tekintsük a következőt

$$P = (A - sBT)^{'} (A - sBT)$$

A fenti $P$ mátrix főátlóbeli elemei az $(A - sBT)$ $i$. oszlopának és a $(A - sBT)^{'}$ $i$. sorának a skalárszorzata, azaz a $P$ főátlóbeli elemek az $(A - sBT)$ $i$. oszlop elemeinek a négyzetösszege. A két konfiguráció közti különbséget a $P$ főátlóbeli elemek négyzetösszegével tökéletesen tudjuk modellezni. A főátlón kívüli elemek most nem érdekesek, éppen ezért hasznos nekünk a nyom függvény.

Legyen

$$L(A,B,s,T) := trP = tr\left[(A-sBT)^{'}(A-sBT)\right] \qquad (1)$$

$trP$ a mátrix nyomát, azaz a főátlóbeli elemek összegét jelöli. Célunk, hogy találjunk olyan $s$ és $T$ értékeket, amelyek minimalizálják $L(A,B,s,T)$. Felhasználva a mátrix nyomára vonatkozó $trABC = trCAB = trBCA$ tulajdonságot, továbbá a $T^{'}T = TT^{'} = I$ összefüggést.

$$trT^{'}B^{'}BT = trB^{'}BTT^{'} = trB^{'}B$$

így rögzített $A$, $B$ és $s = 1$ mellett (1) a következőképpen írható

$$\begin{aligned} L(T) &= tr(A-BT)^{'}(A-BT) = trA^{'}A + trT^{'}B^{'}BT - 2trA^{'}BT = \\ &= trA^{'}A + trB^{'}B - 2trA^{'}BT \end{aligned}$$

Mivel sem $trA^{'}A$ sem $trB^{'}B$ nem függ $T$-től ezért a fenti $L(T)$ minimalizálása ekvivalens a következő minimalizálásával

$$L(T) = c - 2trA^{'}BT,$$

ahol $T^{'}T = I$ és a $c$ konstans nem függ $T$-től.
Tegyük fel, hogy már létezik egy $L(T) \geq h$ alsó korlátunk, ha minimalizálni szeretnénk $L(T)$-t $T$-ben, és ha találunk egy olyan $T$-t amelyre $L(T) = h$ akkor az bizonyosan minimalizálni fogja.

**Állítás.(Kristof)** Ha $Y$ diagonális mátrix nem-negatív elemekkel és $T$ ortonormált mátrix, akkor fennáll a következő egyenlőtlenség

$$-trTY \geq -trY$$

egyenlőség akkor és csak akkor áll fenn ha $T = I$.
**Bizonyítás.** Mivel $Y$ diagonális ezért az egyenlőtlenség átfogalmazható

$$-trTY = -\sum_{i=1}^{n} t_{ii}y_{ii} \geq -\sum_{i=1}^{n} y_{ii} = -trY$$

$TT^{'} = T^{'}T = I$, így $t_i^{'}t_i = \sum_i t_{ii}^2 = 1$ ennélfogva $-1 \leq t_{ii} \leq 1$, ezért $-t_{ii}y_{ii} \geq -y_{ii}$ és ebből már következik az egyenlőtlenség. Nyilvánvaló, hogy egyenlőség csak akkor állhat ha $t_{ii} = 1$.

**Definíció.** Legyen $S$ egy tetszőleges $n \times m$ méretű mátrix, ekkor az

$$S = P\Phi Q^{'},$$

felbontást az $A$ mátrix szinguláris érték felbontásának (*Singular Value Decomposition*) nevezzük, ahol $P$ egy $n \times m$ mátrixa az ún. baloldali szinguláris vektoroknak,

$\Phi$ egy $m \times m$ méretű diagonális mátrix, melyben a főátló beli elemeket szinguláris értékeknek nevezzük, $Q$ egy $m \times m$ méretű, amely oszlopai az ún. jobb oldali szinguláris vektorok.

Vegyük az $A'B$ szinguláris felbontását: $A'B = P\Phi Q'$, ahol $PP' = I$, $Q'Q = I$, $A$ a cél konfiguráció és $B$ a vizsgált konfiguráció, ismét felhasználva a nyom függvény invariáns tulajdonságát

$$\begin{aligned} L(T) &= c - 2trA'BT = \\ &= c - 2trP\Phi Q'T = c - 2trQ'TP\Phi \geq \\ &\geq c - 2tr\Phi \end{aligned}$$

Utóbbi egyenlőtlenség éppen a Kristof egyenlőtlenségből következik, hiszen $(Q'TP)' Q'TP = P'T'QQ'TP = I$, azaz ortonormált és $\Phi$ diagonális. Felhasználva, hogy a fenti egyenlőtlenség egyenlőséggel teljesül - azaz az alsó korlát mint minimum felvétetik- ha $Q'TP = I$ teljesül, azaz

$$T = QP'$$

ugyanis $Q'TP = Q'QP'P = I$, így

$$L(T) = c - 2tr\Phi.$$

Ezeddig meghatároztuk az optimális forgatás, tükrözés mátrixát, de a nyújtás ($s$) és az eltolás ($t$) mértékét fixnek, nevezetesen egynek és nullának vettük. Legyen újra $A$ és $B$ a két konfiguráció, $B$ transzfomációja $sBT + 1t'$, ahol $s$ a nyújtás mértéke, $T$ a forgatás-tükrözés, $t$ pedig az eltolás nagysága. A mi olvasatunkban tehát, rögzített $A$ és $B$ mellett minimalizálni kell

$$L(s,t,T) = tr\left[A - (sBT + 1t')\right]'\left[A - (sBT + 1t')\right], \qquad (2)$$

ahol $T'T = I$.

Az optimális $t$ eltolás vektort és az $s$ dilatációs skalárt az $L(s,t,T)$ parciális deriváltjaiból kapjuk, azaz

$$\begin{aligned} \frac{\partial L(s,t,T)}{\partial t} &= 2nt - 2A'1 + 2sT'B'1 = 0 \\ t &= \frac{(A - sBT)'1}{n} \end{aligned}$$

Helyettesítsük be (2)-be ezt az optimális $t$ értéket

$$\begin{aligned} L(s,T) &= tr\left[(A - sBT) - \frac{11'}{n}(A - sBT)\right]'\left[(A - sBT) - \frac{11'}{n}(A - sBT)\right] = \\ &= tr\left[(A - sBT)\left(I - \frac{11'}{n}\right)\right]'\left[(A - sBT)\left(I - \frac{11'}{n}\right)\right] = \\ &= tr\left[JA - sJBT\right]'\left[JA - sJBT\right], \end{aligned}$$

ahol $J$ a már korábban definiált $H$ centráló mátrix.
Így

$$\begin{aligned}\frac{\partial L(s,T)}{\partial s} &= 2s\left(trB^{'}JB\right) - 2trA^{'}JBT = 0\\ s &= \frac{trA^{'}JBT}{trB^{'}JB}.\end{aligned}$$

Behelyettesítve az optimális $t$ és $s$ értékeket (2)-be kapjuk, hogy

$$\begin{aligned}L(T) &= tr\left[JA - \frac{trA^{'}JBT}{trB^{'}JB}JBT\right]^{'}\left[JA - \frac{trA^{'}JBT}{trB^{'}BY}JBT\right] =\\ &= trA^{'}JA + \frac{\left(trA^{'}JBT\right)^2}{trB^{'}JB} - 2\frac{\left(trA^{'}JBT\right)^2}{trB^{'}JB} =\\ &= trA^{'}JA - \frac{\left(trA^{'}JBT\right)^2}{trB^{'}JB}\end{aligned}$$

Tehát a fenti $L(T)$ minimalizálása ekvivalens $-trA^{'}JBT$ minimalizálásával $T$-ben. Felhasználva az eddigi eredményeket már tudjuk, hogy ez akkor minimális ha $T$-t az $A^{'}JB$ szinguláris érték felbontásában szereplő mátrixok segítségével írjuk fel, azaz a Prokrusztész hasonlósági transzformáció a következő képpen állítható elő:

1. Számítsuk ki a $C = A^{'}JB$ mátrixot.
2. Állítsuk elő $C$-nek a szinguláris érték felbontását, azaz $C = P\Phi Q^{'}$.
3. Az optimális forgatás mátrix: $T = QP^{'}$.
4. Az optimális nagyítás mértéke: $s = \left(trA^{'}JBT\right)/\left(trB^{'}JB\right)$.
5. Az optimális eltolás vektora: $t = \frac{1}{n}\left(A - sBT\right)^{'}1$.

A sokdimenziós skálázás módszerének outputja egy olyan térbeli ábra, mely szemléletesen vizualizálja a pontok geometriai alakzatát. Ez a megjelenés az adataink belső struktúráját jeleníti meg és ezáltal segíti az adatok kapcsolódási rendszerére való alapos rálátást. Mint láttuk az input nem egyértelmű és így kettő tetszőleges konfiguráció összehasonlítása sem nyilvánvaló, de könnyen meghatározható az a transzformáció, amelyet az egyik koordináta mátrixra elvégezve a leginkább megközelíthető a másik, abban az értelemben, hogy az $L(A, B, s, t, T)$ különbség függvényt minimalizáljuk.

# 3. A módszerhez alkalmazott főbb R függvények

A szakdolgozatomban tárgyalt modellek mindegyikének a számításai a nagyméretű adatállomány miatt csak számítógépes programokkal lehetséges, valamennyi számításhoz az R programot használtam, a program vitathatatlan nagy előnye, hogy bárki számára hozzáférhető és letölhető, és az alkalmazáshoz szükséges függvényeket megtaláljuk benne, ezért esett a választásom erre a statisztikai software-re.

Az R szoftverben találhatunk egy `cmdscale` nevű beépített függvényt

```
cmdscale(d, k = 2, eig = FALSE, add = FALSE, x.ret = FALSE),
```

melyben az argumentumai közül `d` jelöli a távolságmátrixot, `k` azt, hogy hány dimenziós térbe vetítünk, `eig` jelöli a sajátértékeket, mely esetében megadhatjuk, hogy akarjuk-e szerepeltetni (`TRUE`) vagy sem (`FALSE`) az outputban, `add=TRUE` argumentummal megadhatjuk, hogy az eredeti $d$ mátrixban adott távolságok helyett egy $d + c11'$ mátrixal számoljon, az `x.ret=TRUE` input paraméterrel egy kitüntetett szerepű szimmetrikus mátrixot kaphatunk vissza. Alapértelmezett esetben `k=2, eig=FALSE, add=FALSE, x.ret=FALSE`.

A $d$ távolság mátrixot a `dist` fügvénnyel készítjük el

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2),
```

amelyben az `x` az adatmátrix, mely sorai jelentik az elemeket oszlopai pedig a változóknak felelnek meg, a `method` input paraméterrel adhatjuk meg, hogy melyik távolság formulát választjuk meg a távolság-mátrix kiszámításában, az R-ben lehetőségünk van `euclidean`, `maximum` ($\max_{i=1}^{p}(x_{li}, x_{ki})$) , `manhattan`( abszolút távolság kettő elem között), `canberra` ($\sum_{i=1}^{p} \frac{|x_{ki}-x_{li}|}{|x_{ki}+x_{li}|}$) , `binary` és `minkowski` távolság formulák közül választhatunk, `diag` logikai érték mellyel a távolság-mátrix főátlóbeli elemeit kérhetjük vissza, `upper` szintén egy logikai érték, amivel a távolság-mátrixot felső háromszögmátrixként kaphatjuk vissza, `p` jelöli a Minkowski-féle együtthatót. Az alapértelmezett értékek `method = "euclidean", diag = FALSE, upper = FALSE, p = 2`. A megfelelő távolság formula kiválasztása feladat függő. Elkészítettem az R-ben az úgynvezett s-stress *(square test)* nevű függvényt:

$$\Phi_1 = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(d_{ij} - \hat{d_{ij}}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2}}.$$

Melynek a parancsa az `illeszkedes(x,y)`, ahol `x`, `y` a bemenő paraméterek, `x` jelenti az eredeti $n$-dimenziós térben vett pontok közötti euklideszi távolságokból nyert távolságmátrix, `y` pedig a `cmdscale` parancs outputjaként megjelenő koordináta mátrix. Evvel hasonlítom össze a kapott MDS ábráknak a pontosságát, hogy azok mennyire használhatóak, szükséges-e a dimenziószám növelése.

A fenti két definícióból következik, hogy ha tökéletes a konfiguráció az eredeti adatokhoz képest akkor a számláló értéke 0 és így $\Phi_1 = 0$. Tehát a közelítésünk annál jobb

minnél kisebbek $\Phi_1$ érétke. Az álábbi táblázatban foglaltuk össze, mekkora értékkel milyen a konfigurációink pontossága:

| s-stress értéke: x | Pontosság mértéke |
|---|---|
| $0 \leq x \leq 0.05$ | Szignifikáns eredményt kaptunk |
| $0.05 < x \leq 0.1$ | További értelmezésekre alkalmas |
| $0.1 < x \leq 0.2$ | Elfogadható, de fenntartással kezelendő |
| $0.2 < x$ | Szükséges a dimenzió szám növelés |

2. Táblázat

A többdimenziós skálázás ábrái mellett elvégzünk egy klaszteranalízist is a `hclust` függvénnyel

```
hclust(d, method = "complete", members=NULL)
```

A d argumentum egy távolság-mátrix a módszer (method) alapértelmezett értéken `"complete"`, de választhatunk a következők közül: `"ward"`, `"single"`, `"average"`, `"mcquitty"`, `"median"`, `"centroid"`. A `members` értéke alapételmezetten `NULL`, különben pedig vektor értékű, amelyben a megfigyelések száma szerepel klaszterenként. Outputja egy olyan $h_0$, $h_1$,..., $h_l$ magassággal jellemzett fastruktúrája `d`-nek (dendrogramm), hogy minden $h_i$ $(i = 1, \ldots l)$ magassághoz tartozik egy neki megfelelő partíciója `d`-nek, továbbá minden objektumpárhoz azt a szintet rendeljük hozzá, amelyen legelőször egyesülnek, azaz $h_0 \longmapsto \{x_1\}, \{x_2\}, \ldots, \{x_n\}$ és $h_l \longmapsto \{x_1, x_2, \ldots x_n\}$.
A biplot ábrák elkészítéséhez saját készítésű programokat használtam.

# 4. Biplot

A biplotok az úgynevezett szórás ábrák általánosításának tekinthetők. Ha adottak bizonyos elemekhez változók, akkor abból két változót kiválasztva az elemek ábrázolhatók egyszerű módon egy grafikonon, az x tengelynek választva az egyik, az y tengelynek választva a másik változót, ezt hívjuk szórás ábrának. Ezzel egy szemléletes képet kapunk a két változó összehasonlításának a vizsgálatához. Természetesen megtehetjük mindezt három dimenzióban is, három változót választva magunknak. Mindemellett a mi célunk, hogy valamilyen módon egy ábrában helyezzük el az elemeket és a változókat jól reprezentáló vektoraikat, oly módon hogy az elemek és a változók közötti kapcsolatot a lehető legjobban szemléltesse az adott ábra. Ezt a fajta ábrázolási módot nevezzük biplotnak, amely így egy hasznos vizuális segédeszköznek bizonyulhat az adatelemzésben.

## 4.1. A matematikai modell

A modell felépítéséhez egészen le egyszerüsítjük a valós életbeli problémánkat, hogy tisztán a matematikájára koncentrálhassunk. Tegyük fel, hogy adott a következő tetszőleges $n \times m$ méretű, $r$ rangú $S$ mátrixunk

$$S = XY^T,$$

ahol $X$ $n \times r$, $Y$ $r \times m$ méretű és mindkettő rangja $r$. Ez a faktorizálás nem egyértelmű. Egy lehetséges út a felbontáshoz például, hogy $X$ $r$ oszlopát válasszuk meg, mint az $S$ oszlopai által kifeszített tér egy ortonormált bázisát, majd $Y = S^T X$, így egy megfelelő felbontáshoz jutunk, ugyanis

$$X\left(S^T X\right)^T = XX^T S = S$$

Az $S$ mátrix $i.$ sorának $j.$ eleme $s_{ij}$ az $X$ $i.$ sorának és az $Y$ $j.$ oszlopának a skalárszorzata. Így egy bármilyen mátrix valamennyi sora és oszlopa repzentálható vektorokkal. Nevezetesen az $X$-nek az $n$ darab sorvektora, és $Y$-nak az $m$ darab oszlopvektora lesz az. Az $x_1$,$x_2$,...,$x_n$ sorvektorokat hívjuk ún. sor hatásoknak, míg az $y_1^T$,$y_2^T$,...,$y_m^T$ oszlopvektorokat oszlop hatásoknak. Tegyük fel, hogy $S$ rangja kettő, ábrázolhatjuk a síkban ezt az $n + m$ darab vektort, amely informatív képet ad az $S$ $nm$ darab eleméről - amelyek a megfelelő sor és oszlopvektorok skalárszorzatai - olyan módon, hogy az $x_1$,$x_2$,...,$x_n$ pontok az $y_1^T$,$y_2^T$,...,$y_m^T$ irányú egyenesekre vett merőleges vetületei lesznek az $S$ megfelelő elemei. Ezeket az ábrákat nevezzük biplotoknak, ezáltal a biplot az adatok egy gyors vizuális kiértékelését teszi lehetővé számunkra. Az $S$ mátrix nem egyértelmű faktorizációja miatt természetesen a biplot sem egyértelmű, viszont a forgatás, tükrözés és nyújtás erejéig már igen. Mivel a gyakorlatban általában két dimenziós biplotokat jelenítünk meg, ezért a fenti egyértelműséget is csak két dimenzióban mutatom meg.

$$S = \left(XR^T\right)\left(YR^{-1}\right)^T,$$

ahol $R$ tetszőleges nem-szinguláris mátrix és alkalmas megteremteni a kapcsolatot tetszőleges két különböző faktorizáció megfelelő elemei között.
Határozzuk meg $R^T$ szinguláris érték felbontását

$$R^T = V^T \theta W,$$

ahol $V$ és $W$ $r \times r$-es ortonormált mátrixok, $\theta$ diagonális mátrix, főátlóbeli elemei $\theta_1$ és $\theta_2$
$R$ inverze

$$R^{-1} = V^T \theta^{-1} W,$$

ugyanis $RR^{-1} = \left(R^T\right)^T R^{-1} = \left(W^T \theta V\right)\left(V^T \theta^{-1} W\right) = W^T \theta I \theta^{-1} W = W^T W = I$. Mivel $V$ és $W$ ortonormáltak.

Vegyük az $X \longmapsto XR^T$és $Y \longmapsto YR^{-1}$ transzformációkat, mind a kettő transzformáció tartalmaz egyrészt egy forgatást, amelyet $V^T$ határoz meg, majd egy nyújtást és egy esetleges tükrözést, melyet $X$ transzformációja esetén $\theta_1$, $\theta_2$ illetve $Y$ esetében $1/\theta_1$, $1/\theta_2$ határoz meg, végezetül egy újabb forgatás, amely $W$-el meghatározott. Mindenerre nézzünk meg egy példát. Legyen

$$S = \begin{pmatrix} -1 & 3 & 2 \\ -2 & 1 & -1 \end{pmatrix},$$

a cél mátrixunk, melynek a rangja kettő.
Vegyük ennek kettő különböző faktorizációját

$$\begin{aligned} \begin{pmatrix} -1 & 3 & 2 \\ -2 & 1 & -1 \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} -1 & 3 & 2 \\ -1 & -2 & -3 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} -1.5 & 2 & 0.5 \\ 2 & -1 & 1 \end{pmatrix}, \end{aligned}$$

$$X_1 = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, X_2 = \begin{pmatrix} 2 & 1 \\ 0 & -1 \end{pmatrix}$$

$$Y_1^T = \begin{pmatrix} -1 & 3 & 2 \\ -1 & -2 & -3 \end{pmatrix}, Y_2^T = \begin{pmatrix} -1.5 & 2 & 0.5 \\ 2 & -1 & 1 \end{pmatrix}$$

szereposztással.
Keressük azt az $R^T$ mátrixot, amely megteremti a kapcsolatot $X_1$ és $X_2$, illetve $Y_1$ és $Y_2$ között.Vagyis $X_2 = X_1 R^T$ és $Y_2 = Y_1 R^{-1}$, a fenti példánál

$$R^T = \begin{pmatrix} 2 & 1 \\ -2 & -2 \end{pmatrix}.$$

Vegyük $R^T$ szinguláris érték felbontását

$$\begin{aligned} R^T &= \begin{pmatrix} 2 & 1 \\ -2 & -2 \end{pmatrix} \\ &= \begin{pmatrix} -0.62 & -0.79 \\ 0.79 & -0.62 \end{pmatrix} \begin{pmatrix} 3.56 & 0 \\ 0 & 0.56 \end{pmatrix} \begin{pmatrix} -0.79 & -0.62 \\ -0.62 & 0.79 \end{pmatrix}, \end{aligned}$$

ahol a jelöléseket megtartva a fentiekkel

$$V^T = \begin{pmatrix} -0.62 & -0.79 \\ 0.79 & -0.62 \end{pmatrix}$$

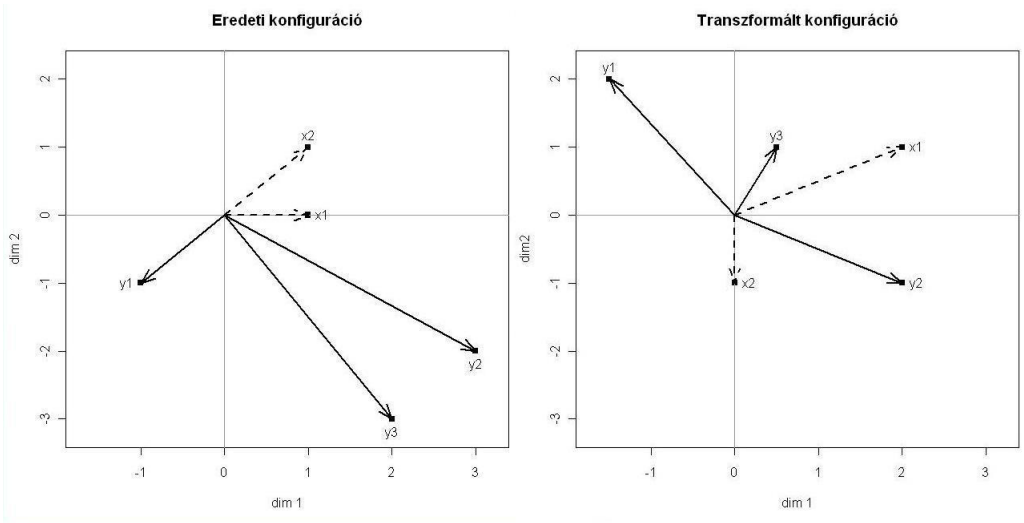
$$\theta = \begin{pmatrix} 3.56 & 0 \\ 0 & 0.56 \end{pmatrix}$$

$$W = \begin{pmatrix} -0.79 & -0.62 \\ -0.62 & 0.79 \end{pmatrix}.$$

Így

$$R^{-1} = V^T \theta^{-1} W = \begin{pmatrix} 1 & 0.5 \\ -1 & -1 \end{pmatrix}$$

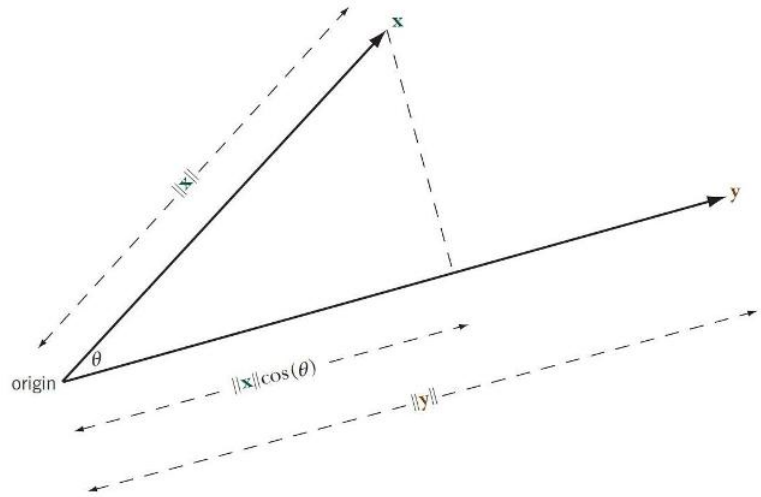
Az alábbi ábra mutatja az eredeti $X_1$ és $Y_1$ konfigurációt és annak a transzformáltját $X_2$ és $Y_2$-t.




2. Ábra: *Az $S$ célmátrix kettő különböző felbontásához tartozó biplotok*

Az $S$ mátrixot mint fentebb már említettük cél mátrixnak, $X$-et bal oldali, $Y$-t jobb oldali mátrixnak nevezzük. A bal oldali mátrix minden $x_1$,$x_2$,..., $x_n$ sorát, az $Y$ minden

$y_1$,$y_2$,...,$y_m$ oszlopát, mint pontot ábrázoljuk egy euklideszi térben, ahol a pontok koordinátái az $X$ sor illtve $Y$ oszlop vektorai. Az origóból az $y_i$ pontokba húzott vektorok irányával megegyező tengelyeket nevezzük biplot tengelyeknek, $y_i$-t pedig biplot vektornak, $x_i$-t biplot pontoknak. Az $X$ $i$. sorának és Y $j$. oszlopának skalárszorzata nem más mint az $x_i$ vetülete az $y_j$ vektorhoz tartozó biplot tengelyre szorozva az $\| y_j \|$-val.




3. Ábra: *Az $x$ biplot pont vetülete az $y$-hoz tartozó biplot tengelyre*

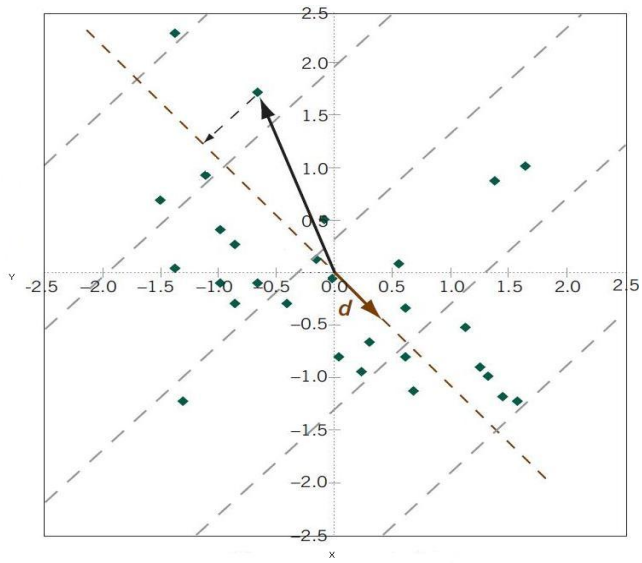
Így ha már meghatároztuk a biplotot akkor ahhoz, hogy közvetlenül leolvashassuk a cél mátrixunk elemeit a tengelyekről, természetsen szükségünk van arra, hogy mekkora egy egység hossza a biplot tengelynek. Mivel

$$\begin{array}{rcl} \begin{array}{c} cél\,mátrix\,értéke \\ az\,i.\,sorban \end{array} & = & \begin{pmatrix} x_i\,vetületeének\,a\,hossza \\ a\,biplot\,tengelyre \end{pmatrix} \begin{pmatrix} a\,meg f elelő\,biplot\,vekor \\ hossza \end{pmatrix} \\ & = & \| x \| \, cos\,(\Theta) \, \| y \| \end{array}$$

ezért a cél mátrix értékét 1-nek választva kapjuk, hogy

$$\begin{array}{c} x_i\,vetületeének\,a\,hossza \\ a\,biplot\,tengelyre \end{array} = 1/\,(a\,meg f elelő\,biplot\,vekor\,hossza)$$

és ezáltal elérjük, hogy cél mátrix megfelelő értékét olvashassuk le az $x$ pont $y$-hoz tartozó biplot tengelyre vett vetítésekor.




4. Ábra: *A biplot skálázása*

Tehát ha a cél mátrix oszlopai változók és a sorok az egyes objektumok, akkor a cél mátrix biplot reprezentációja azt fogja jelenteni, hogy a változókat mint tengelyeket fogjuk ábrázolni, melyek olyan irányba mutatnak, hogy az objektumok értéke az adott változón, megkapható legyen az egyes objektumok adott változóra való vetítésével. Az értékei az adott változónak nem olyan fontosak, mint inkább azt látni, hogy hogyan sorakoznak az egyes objektumok a változók mentén. És ha két biplot tengely ugyanabban az irányban fekszik, következtetni tudunk arra, hogy az objektumoknak ugyanaz a relatív poziciója a két változó mentén. Ami magas, változók közti korrelációt jelent. Így a biplotnál a „bi" az nem a két dimenzióra utal, annak ellenére, hogy legtöbbször két dimenziós biplotokat ábárzolunk, hanem arra, hogy két halmaz pontjait jelenítjük meg.

## 4.2. Dimenzió csökkentés a biplotok alkalmazásához

Mint láttuk fentebb, hogy a biplothoz mint elemzési eszközhöz nyúlhassunk, szükséges, hogy az ábrázolni kívánt mátrix rangja kettő vagy esetleg három legyen, így az ennél magasabb dimenziószámú mátrixok esetén szükségünk van dimenzió csökkentésre. Az egyik lehetséges eljárás az, hogy az $S$ $r$ rangú mátrixunkat a szinguláris érték felbontás segítségével közelítjük, egy olyan $S_{(s)}$ jelölt mátrixal, mely rangja $s$. Minimalizálni akarjuk a

$$\| S - S_{(s)} \|^2 := tr\left(\left(S - S_{(s)}\right)\left(S - S_{(s)}\right)^T\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(s_{ij} - s_{ij}^{(s)}\right)^2 \qquad (1)$$

Ehhez tekintsük $S$ szinguláris érték felbontását

$$S = \sum_{i=1}^{r} \lambda_i p_i q_i^T = (p_1, p_2, \ldots, p_r) \begin{pmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_r \end{pmatrix} \begin{pmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_r^T \end{pmatrix} = PD_\lambda Q^T,$$

ahol $\lambda_i$ a szinguláris értékek, $p_i$ a szinguláris oszlopok, $q_i^T$ a szinguláris sorok, ekkor belátható, hogy fenállnak a következő összefüggések

$$p_i^T S = \lambda_i q_i^T$$

$$S q_i = \lambda_i p_i$$

$$S S^T p_i = \lambda_i^2 p_i$$

$$S^T S q_i = \lambda_i^2 q_i$$

$$\lambda_1 \geq \ldots \geq \lambda_r > 0$$

$$p_i^T p_j = q_i^T q_j = \delta_{ij},$$

ahol $\delta_{ij}$ az un. Kronecker-féle delta.
A legkisebb négyzetek módszere adja $S_{(s)}$, amely a fenti $S$-t a legjobban közelíti, abban az értelemben, hogy (1) minimális.

$$S_{(s)} = \sum_{i=1}^{s} \lambda_i p_i q_i^T = P_{[s]} D_{\lambda[s]} Q_{[s]}^T,$$

ahol $S_{(s)}$ $s$ rangú mátrix, $[s]$ jelöli, hogy az első $s$ komponensét vesszük.
A közelítés hibáját $\epsilon_{(s)}$-el jelölve kapjuk, hogy

$$\begin{aligned} \epsilon_{(s)} &= 1 - \frac{\| S - S_{(s)} \|^2}{\| S \|^2} = 1 - \frac{\lambda_{s+1}^2 + \ldots + \lambda_r^2}{\lambda_1^2 + \ldots + \lambda_r^2} = \\ &= \frac{\sum_{i=1}^{s} \lambda_i}{\sum_{i=1}^{r} \lambda_i}. \end{aligned}$$

Alkalmazva mindezt $s = 2$-re

$$S_{(2)} = \begin{pmatrix} p_{11} & p_{21} \\ \vdots & \vdots \\ p_{1n} & p_{2n} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} q_{11} & \ldots & q_{1m} \\ q_{21} & \ldots & q_{2m} \end{pmatrix}$$

$$\epsilon_{(2)} = \left(\lambda_1^2 + \lambda_2^2\right) / \sum_{i=1}^{r} \lambda_i$$

Ez alapján az eredeti $S = XY^T$ faktorizációját az $S_{(s)} = \widehat{X}\widehat{Y}^T$ faktorizációjával közelítjük, amely így

$$\widehat{X} = \begin{pmatrix} \sqrt{\lambda_1}p_{11} & \sqrt{\lambda_2}p_{21} \\ \vdots & \vdots \\ \sqrt{\lambda_1}p_{1n} & \sqrt{\lambda_2}p_{2n} \end{pmatrix}; \quad \widehat{Y}^T = \begin{pmatrix} \sqrt{\lambda_1}q_{11} & \dots & \sqrt{\lambda_1}q_{1m} \\ \sqrt{\lambda_2}q_{21} & \dots & \sqrt{\lambda_2}q_{2m} \end{pmatrix}$$

Nézzünk egy példát, tekintsük az

$$S = \begin{pmatrix} 5 & 4 & 8 \\ 7 & 12 & 30 \\ 2 & 7 & 28 \end{pmatrix} = \begin{pmatrix} 3 & 3 & -1 \\ 3 & -1 & 5 \\ 1 & -4 & 5 \end{pmatrix} \begin{pmatrix} 1 & 2 & 7 \\ 1 & 1 & -4 \\ 1 & 2 & 1 \end{pmatrix}$$

faktorizációt, ekkor az $S$ mátrix - melyet ábrázolni kívánunk egy biplotban - három rangú, de mi ezt a faktorizációt az $S_{(2)} = \widehat{X}\widehat{Y}^T$-al közelítjük.

$S$ szinguláris érték felbontása

$$S = \begin{pmatrix} -0.21 & 0.64 & 0.74 \\ -0.74 & 0.39 & -0.55 \\ -0.64 & -0.66 & 0.39 \end{pmatrix} \begin{pmatrix} 44.75 & 0 & 0 \\ 0 & 5.56 & 0 \\ 0 & 0 & 1.15 \end{pmatrix} \begin{pmatrix} -0.17 & -0.32 & -0.93 \\ 0.83 & 0.47 & -0.31 \\ 0.54 & -0.82 & 0.18 \end{pmatrix}$$

ebből $s = 2$-re kapjuk, hogy

$$\begin{aligned} S_{(2)} &= \widehat{X}\widehat{Y}^T \\ &= \begin{pmatrix} \sqrt{\lambda_1}p_{11} & \sqrt{\lambda_2}p_{21} \\ \vdots & \vdots \\ \sqrt{\lambda_1}p_{1n} & \sqrt{\lambda_2}p_{2n} \end{pmatrix} \begin{pmatrix} \sqrt{\lambda_1}q_{11} & \dots & \sqrt{\lambda_2}q_{1m} \\ \sqrt{\lambda_2}q_{21} & \dots & \sqrt{\lambda_2}q_{2m} \end{pmatrix} \\ &= \begin{pmatrix} -1.43 & 1.5 \\ -4.93 & 0.92 \\ -4.29 & -1.56 \end{pmatrix} \begin{pmatrix} -1.12 & -2.12 & -6.24 \\ 1.95 & 1.1 & -0.73 \end{pmatrix} \\ &= \begin{pmatrix} 4.54 & 4.7 & 7.84 \\ 7.34 & 11.48 & 30.12 \\ 1.76 & 7.37 & 27.92 \end{pmatrix} \end{aligned}$$

a hiba nagysága

$$\begin{aligned} \epsilon_{(2)} &= \frac{44.75 + 5.56}{44.75 + 5.56 + 1.15} \\ &= 0.98 \end{aligned}$$

azaz elég nagy pontossággal tudjuk az eredeti $S$-t közelíteni.

Gyakran előfordul, hogy érdemes és/vagy kell súlyozni az objektumokat és a változókat. Például ha egy felmérésben túl sok nőt kérdeztünk meg, akkor nem lesz a mintánk reprezentatív a populációra nézve, ezért a nők által adott értékekeket egy alacsonyabb súllyal súlyozzuk, így ellensúlyozva a megfelelő reprezentativitás hiányát, vagy ha egy vizsgálatban bizonyos változók hangsúlyosabbak, akkor azokat megfelelő súlyozással hangsúlyosabbá tesszük. Lássuk most $S$ közelítését egy $S_{(s)}$ két rangú mátrixal szintén, csak éppen a súlyozott esetre fókuszálva, amely tulajdonképpen a fentiek általánosításának tekinthető.

Tegyük fel, hogy adottak a sorokhoz (objektumokhoz) tartozó $w_1$,$w_2$,...,$w_n$, illetve az oszlopkhoz (változókhoz) tartozó $q_1$,$q_2$,...,$q_m$ súlyok. Feltehető, hogy $\sum_{i=1}^{n} w_i = 1$ és $\sum_{i=1}^{n} q_i = 1$. Most a minimalizálandó függvényünk

$$\| S - S_{(s)} \|^2 := tr\left(D_w \left(S - S_{(s)}\right) D_q \left(S - S_{(s)}\right)^T\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_i q_j \left(s_{ij} - s_{ij}^{(s)}\right)^2, \quad (2)$$

ahol $D_w$ =diag$(w_1, w_2, \ldots, w_n)$ és $D_q$ =diag$(q_1, q_2, \ldots, q_m)$.
Ehhez először elkészítjük azt a $T$ mátrixot $S$ -ből, amely már a megfelelő módon figyelembe veszi a súlyokat

$$T = D_w^{1/2} S D_q^{1/2},$$

ahol $D_w^{1/2}$ =diag$\left(\sqrt{w_1}, \sqrt{w_2}, \ldots, \sqrt{w_n}\right)$ illetve $D_q^{1/2}$ =diag$\left(\sqrt{q_1}, \sqrt{q_2}, \ldots, \sqrt{q_m}\right)$.
Meghatározzuk a $T$ szinguláris érték felbontását

$$T = \sum_{i=1}^{r} \sigma_i p_i q_i^T = (p_1, p_2, \ldots, p_r) \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \ldots & 0 & \sigma_r \end{pmatrix} \begin{pmatrix} q_1^T \\ q_2^T \\ \vdots \\ q_r^T \end{pmatrix} = P D_\sigma Q^T$$

Definiáljuk $\widetilde{P}$ és $\widetilde{Q}$ a következőképpen

$$\widetilde{P} = D_w^{-1/2} P \quad \widetilde{Q} = D_q^{-1/2} Q$$

így fennáll, hogy $\widetilde{P}^T D_w \widetilde{P} = I$ és $\widetilde{Q}^T D_q \widetilde{Q} = I$.
Ekkor $S_{(s)}$ , amely minimalizálja (2)-t

$$S_{(s)} = \widetilde{P}_{[s]} D_{\sigma[s]} \widetilde{Q}_{[s]}$$

Valamennyi módszer egy általános keretbe foglalható, melyek speciális eseteiként adódnak belőle, ez a témakör az un. főkomponens analízis (*Principal Components Analysis*).

## 4.3. Biplot alkalmazása a lineáris regresszión keresztül

A regresszió analízis feladata, hogy két vagy több megfigyelési változó között egy lineáris képletet adjon adott mérési vagy megfigyelési adatokból. Azaz kiválasztunk egy változót és a képlet segítségével meghatározzuk, hogy az mekkora mértékben függ más változóktól. Az előbbit függő változóknak hívjuk, utóbbit pedig független változóknak vagy szokás még magyarázó változóknak is nevezni. Jelöljük $\widehat{Y}$-al, a kiválasztott változók regreszió analízis által becsült értékét, $X$-el a magyarázó változók mátrixát, $B$-vel a lineáris regresszió által adott ún. regressziós koefficienseket, amelyek a legkisebb négyzetes módszer által meghatározottak, azaz

$$\widehat{Y} = XB$$

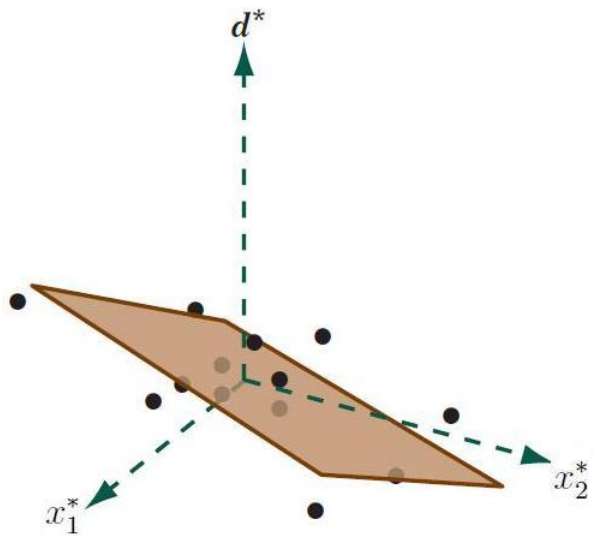
Tegyük fel, hogy adott egy $n$ elemű független minta és mindegyik mintán $p$ számú megfigyelést végeztünk. Például rögzítsünk egy $d$ függő változót, illetve két (esetleg három) alkalmas $x_1$,$x_2$ magyarázó változót, amely vélhetően jól karakterizálja majd $d$-t. Itt jegyezném meg azt, hogy nem véletlen az, hogy két vagy három magyarázó változót választunk, hiszen a biplotokat is a legjobb értelmezhetőség kedvéért kettő vagy három dimenzióban ábrázoltuk, és ezen a ponton előre megemlítem, hogy a magyarázó változókhoz tartozó regressziós koefficiensek jelentik majd számunkra a biplot vektorokat. Ekkor

$$\widehat{d} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

ahol $\beta_0$,$\beta_1$,$\beta_2$ jelölik a regressziós koefficiensek.
A statisztikai szoftverek minden esetben szolgáltatnak egy olyan értékkel, ami a modell magyarázó erejét mutatja, leggyakrabban $R^2$-el jelölt , 0 és 1 közé eső értékről van szó. A $\beta_1$, $\beta_2$ regressziós koefficiensek az $x_1$, $x_2$ által meghatározott sík meredekségét adja meg. Sajnos azonban e kettő együttható meghatározása függ a $d$ és a magyarázó változók skálázásától, így azok objektív összehasonlítása nem lehetséges, éppen ezért standardizálást hajtunk végre mind a $d$ és az $x_1$, $x_2$ változókon, mindegyik változóból kivonjuk az átlagát és elosztjuk a szórásnégyzetükkel, ezáltal kiküszöböljük a skálázás okozta problémát. Jelöljük a három standardizált változót $d^*$,$x_1^*$,$x_2^*$. Így erre is alkalmazva a lineáris regressziót kapjuk, hogy

$$\hat{d}^* = \beta_1^* x_1^* + \beta_2^* x_2^*$$

Most $\beta_1^*$, $\beta_2^*$ már összhasonlítható és eldönthetjük, hogy melyik magyarázó változónak erősebb a magyarázó ereje $d^*$-ra vonatkozóan, azaz ha $\beta_1^* > \beta_2^*$ akkor $x_1^*$ magyarázó változó hangsúlyosabban szerepel $d^*$ számításában, mint $x_2^*$. A $\beta_1^*$, $\beta_2^*$ együtthatók, ugyanúgy a standardizált változók által definiált sík merdekségét adja meg.




$\frac{\partial \hat{d}^*}{x_1^*} = \beta_1^* \; \frac{\partial \hat{d}^*}{x_2^*} = \beta_2^*$

5. Ábra: *A regressziós keofficiensek szemléltetése*

Készítsük el $x_1^*$, $x_2^*$ szórás ábráját és rajzoljuk bele a $(\beta_1^*, \beta_2^*)$ gradiens vektort, az így kapott ábra nem más mint a fenti 5.ábra vetülete az $x_1^*$, $x_2^*$ síkba. Ebben húzzuk be a gradiens vektornak megfelelő $d$-hez tartozó biplot tengelyt. Azt szeretnénk elérni, hogy a biplot tengelyt megfelelően skálázva a minta elemeit - melyeket az $x_1^*$, $x_2^*$változói szerint ábrázoltunk egy két dimenziós szórás ábrában - mint pontokat erre a tengelyre vetítve leolvashassuk az eredeti nem standardizált $d$ változó értékeit. Természtesen a regresszió magyarázó ereje nagy valószínűséggel nem lesz tökéletes ezért ennek a hibának a következtében az esetek többségében csak közelítést tudunk adni. Tehát az egyik kulcskérdés, hogy hogyan skálázzuk megfelelően ezt a biplot tengelyt. Legyen $\sigma$ a $d$ szórása, ekkor egy egység hossza a biplot tengelyen

$$1/\sqrt{(\sigma\beta_1^*)^2 + (\sigma\beta_2^*)^2} \quad = \quad 1/\sigma\sqrt{\beta_1^{*2} + \beta_2^{*2}}$$

azaz a stanadardizált változókhoz tartozó biplot vektor hosszának és a szórásnak a szorzatának vegyük a reciprokát. Miután meghatároztuk a biplot tengely egy egységének a hosszát, el kell döntenünk hogy ezen tengelyen hol van a 0 érték, ugyanis ennek hiányában nem tudjuk még megfelelően kiolvasni az egyes mintákhoz tartozó $d$ értékeket. A biplot ábra középpontjában a biplot tengely vegye fel a $d$ értékeinek az átlagát és a biplot vektor iránya szabja meg, hogy a tengely milyen irányban növekszik. Nézzük meg, hogy a fenti esetben mi felel meg a cél-, jobb-, illetve bal oldali mátrixnak $n$ elemű minta esetén, egy függő változóval

$$\begin{pmatrix} \hat{d}_1^* \\ \hat{d}_2^* \\ \vdots \\ \hat{d}_n^* \end{pmatrix} = \begin{pmatrix} x_{11}^* & x_{12}^* \\ x_{21}^* & x_{22}^* \\ \vdots & \vdots \\ x_{n1}^* & x_{n2}^* \end{pmatrix} \begin{pmatrix} \beta_1^* \\ \beta_2^* \end{pmatrix},$$

ahol a baloldalon szerepel a cél mátrix, a jobb oldalon a két magyarázó változó értékeiből álló mátrix a bal oldali, míg a koefficiensekből álló mátrix pedig a jobb oldali mátrixot jelenti.
A biplot ábrában ha több lineáris regressziót hajtunk végre akkor a fenti cél mátrix és jobb oldali mátrix a további megfelelő függő változókhoz tartozó oszlopokkal bővülnek, azaz ha $m$ darab függő változót vizsgálunk egyszerre akkor

$$\begin{pmatrix} \hat{d^*_{11}} & \hat{d^*_{12}} & \dots & \hat{d^*_{1m}} \\ \hat{d^*_{21}} & \hat{d^*_{22}} & \dots & \hat{d^*_{2m}} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{d_{n1}} & \hat{d_{n2}} & \dots & \hat{d_{nm}} \end{pmatrix} = \begin{pmatrix} x^*_{11} & x^*_{12} \\ x^*_{21} & x^*_{22} \\ \vdots & \vdots \\ x^*_{n1} & x^*_{n2} \end{pmatrix} \begin{pmatrix} \beta^*_{11} & \beta^*_{12} & \dots & \beta^*_{1m} \\ \beta^*_{21} & \beta^*_{22} & \dots & \beta^*_{2m} \end{pmatrix}$$

Az $m$ darab biplot vektor iránya informál bennünket arról, hogy a magyarázó változók hasonló viszonyban vannak-e a vizsgált függő változókhoz, azaz ha két vektornak közel azonos az iránya akkor azon biplot vektorokhoz tartozó függő változók hasonló kapcsolatot mutanak a magyarázó változókkal.

## 4.4. Az MDS és a biplot kapcsolata

Mint láttuk az MDS ábrából leolvasható távolságok és az eredeti távolságok között hibával kell számolnunk, ennek a hibának a mérésére a legegyszerűbb út a fentieken kívül a regresszió analízis. Minnél magasabb dimenzióra redukálunk annál jobb lesz a „minősége" az MDS ábrának és a dimenziószám növelésével mindig egyre kissebb mértékben ugyan, de nőni fog a variancia értéke. Mint láttuk az MDS ún. irány tengelyeinek (*principal axes*) az értelmezése elég szubjektív és korántsem egyszerű feladat, azok ugyanis nem kettő vagy három megfigyelt változókhoz, hanem az ún. látens változókhoz tartozó tengelyek lesznek. Tegyük fel, hogy az $n$ ponthoz adott $p$ számú megfigyelést követően már elkészítettük az MDS ábrát, és az $n$ ponthoz további $t$ darab változó értékeit rendeltük hozzá. Szeretnénk megvizsgálni, hogy ezek az extra változók, milyen viszonyban vannak az objektumokkal, és ezeket a változókat mint biplot tengelyeket szeretnénk a kapott MDS ábrába bele illeszteni. Lineáris regressziót készítünk az MDS pontjainak a koordinátáin az új változókra, ez összesen $t$ darab lineáris regressziót jelent, és mint előbb minden egyes regresszióra a regressziós koefficiensek alkotják majd a biplot vektorokat. Nem csak újabb megfigyeléseket mint biplotvektorokat lehet bele venni az ábrába, hanem azokat a változókat is amelyekből a többdimenziós skálázást származtatjuk. Az ilyen biplotoknak az összhibája kettő hibából tevődik össze, egyrészt a dimenziós csökkentésből származó hibából ( s-tress), másfelől amely a lineáris regresszióból adódik ($R^2$).

# 5. A módszer gyakorlati alkalmazása

Mint láttuk a fenti elméleti módszerek eredményeinek értelmezhetősége elég nehéz is lehet, sőt olyan is előfordulhat, hogy az adataink semmitmondóak ezen módszerek alkalmazásakor. A modellek alkalmasak arra, hogy az adatainkat összességében is vizsgáljuk, ne csak bizonyos részeit. Igyekeztem a szakdolgozatomban ezeknek a módszereknek olyan alkalmazását választani, amellyel mindez megvalósítható és valós életbeli problémát elemez. Callcenter üzemeltetésével foglalkozó vállalatok/szervezetek megbízást adnak ezen rendszerek fejlesztésével és kiépítésével foglalkozó cégnek olyan callcenter kiépítésére, amely képes az ügyfélhez beérkező hívásokat a nála foglalkoztatott ügyintézőkhöz a kívánt stratégia alapján kiosztani, illetve egy adatbázis alapján, automatikusan, kimenő hívásokat kezdeményezni az ügyfelek irányába. Mind a kimenő mind a bejövő hívásokról minden elérhető információ rögzítésre kerül egy adatbázisban. Ha a megbízó vállalatnak szüksége van az adatokra vagy az igénye az, hogy készüljön azokból bármilyen statisztika akkor a kiépítő céget megbízza ennek megvalósításával. A konkrét példámban egy főleg kimenő hívásokkal foglalkozó cég (bér callcenter), kérte fel a kivitelező céget, hogy saját megoldásával váltsa le a megrendelő meglévő rendszerét. Az új callcenter megoldás, a hívásokról minden adatot egy SQL adatbázisban tárol, amelyből az ügyintézőkhöz tartozó legkülönbözőbb adatok hívhatók le. A vállalat kérése a kivitelezőtől, hogy készítsen egy statisztikát az ügyintézőkről.

Jó néhány cégnél az ügyfelekkel foglalkozó ügyintéző egyik szerepe az elégedettség megőrzése, ügyfelek megtartása, a "gondoskodás" az ügyfelekről. A másik szerep, pedig, hogy nagyfokú akarattal keresse a lehetőséget újabb és újabb termék értékesítésére. Kérdés az, hogy tudnak majd sikeresen új szolgáltatást/ terméket (továbbiakban termék) értékesíteni meglevő ügyfeleiknek. Bizonyos cégek, úgy gondolják, az ügyfelekkel foglalkozó kollegák csak maradjanak meg saját klasszikus munkakörüknél, gondoskodjanak az ügyfelekről, oldják meg annak problémáit, hiszen ehhez értenek és egy különálló profi értékesítési csapat majd meg fogja oldani az termék értékesítését. Azonban ha jobban belegondolunk, nagyobb potenciál az új termék eladásra, mint az ügyfél által kezdeményezett személyes találkozó kihasználása sosem lesz, így az ügyintézés, ügyfélszolgálat művelése mellett, új termék értékesítése is elvárás lehet. mindezek arra motiválják a vállalatokat, hogy folyamatos monitoringja legyen az ő munkájuknak, ezzel ellenőrizve a dolgozók hatékonyságát.

A mi konkrét példánkban az operátorok közül csak kevesen és azok is csak kevésszer foglalkoznak ügyintézéssel, mint inkább értékesítéssel. A megbízó vállalat célja, hogy alapos rálátása legyen az értékesítők munkájára ahhoz, hogy az egész értékesítési osztály hatékonyságát növelni tudja, továbbá hogyan tudja költségeit minimalizálni az értékesítők munkáján keresztül. Az értékesítők minden hívás előtt átnézik az ügyintézői felület által megjelenített adatokat a hívandó ügyfélről majd a hívás gomb megnyomásával elindul a hívási folyamat, amely mindaddig tart, amíg az ügyfél vagy az ügyintéző nem bontja a vonalat és egy úgynevezett terminációs kóddal le nem zárja a hívást. A felhívandó ügyfél adatainak megjelenése a képernyőn és a hívás gomb megnyomása között eltelt időt felkészülési időnek hívjuk. A terminációs kódok kampányonként változnak. Kampánynak nevezzük egy adatbázist a hívandó ügyfelekkel, melyet a bércallcenter az őt megbízóktól kap. Megkülönböztetünk sikertelen, sikeres

és eladással zárult hívásokat. Sikertelen hívásnak nevezzük azokat a hívásokat, amelyhez tartozó felhívott ügyfél nem esik bele a célközönségbe, vagy nem sikerült elérni. Sikeres hívások azok a hívások, amelyhez tartozó ügyfél megfelelt a kritériumoknak, de nem történt értékesítés, eladással zárult hívások értelemszerűen azok a hívások amelyek sikeresek és történt értékesítés. Lehetőség van arra, hogy az ügyfél visszahívást kérjen az értékesítőtől, ha érdeklődik a termék/szolgáltatás iránt de pillanatnyilag nem ér rá, ekkor visszahívásról beszélünk. Miután az operátor megszakította a hívást, onnantól a terminációs kód beállításáig eltelt idő hívjuk utómunka időnek, amely így eladással zárult hívás esetén általában nagyobb mint sikeres vagy sikertelen híváskor. A rendelkezésemre álló adattáblákból a következő adatok hívhatók le

1. *Eladásokkal zárult* egész másodpercben mért *híváshossz.*
2. *Sikeresen zárult* egész másodpercben mért *híváshossz.*
3. *Utómunka idő*, amely a hívás megszakítása és a terminációs kóddal való munka folyamat lezárás között eltelt (terminációs kódok kampányonként változhatnak) , szintén egész másodpercben mért híváshossz.
4. Az ügyintéző számára megjelent ügyfél adatok megjelenése és a hívásgomb elindítása között eltelet egész másodpercben mért *felkészülési idő.*
5. *Szüneten tölött idő*, , amely a szünet gomb megnyomásától a szünet vége gomb megnyomásáig tart.
6. *Tetszőleges idő intervallumon belül a sikertelen hívások darabszáma.*
7. *Tetszőleges idő intervallumon belül a sikeres hívások darabszáma.*
8. *Tetszőleges idő intervallumon belül a eladással zárult hívások darabszáma.*
9. *Visszahívások darabszáma.*
10. *Csengetési idő.*
11. *Tetszőleges időintervallumon belül befejezett sikeres hívások száma.*
12. *Tetszőleges időintervallumon belül befejezett eladással zárult hívások száma.*
13. *Foglalt hívások darabszáma.*
14. *Azon hívások darabszáma, melyeket nem vettek fel.*
15. *Azon hívások száma, ahol nem létezett a telefon száma.*
16. *Azon hívások száma, melyeknél a telefonszámon fax jelentkezett.*

A 13, 14, 15, 16 paraméterek a vizsgálatok szempontjából érdektelenek, hiszen azok nem az ügyintéző munkáját minősítik, míg a többi közvetve vagy közvetlenül azt. Valamennyi változó kvantitatív változó.

Az elemzés során megpróbálom az ügyintézőket klaszterezni, ehhez több út is kínálkozik aszerint, hogy melyik változó mentén tesszük ezt meg, több MDS ábrát készítve kiválasztom melyik az, amelyik legjobban karakterizálhatja az ügyintézőket, és azok mekkora mértékben különböznek egymástól, hisz láttuk, hogy az MDS outputjai nem egyértelműek. Az `R` program lehetőséget biztosít egy, az MDS-től eltérő, másfajta modellen alapuló klaszterezési eljárásra is, amely által kapott eredményeket összevetjük a többdimenziós skálázás eredményeivel, hogy azok egybe esnek-e vagy ha nem, akkor annak mi lehet az oka. Ezek után elkészítek egy biplot ábrát, amelyben a biplot tengelyek lesznek az MDS konfigurációk elkészítésében szereplő változók, megkísérlek olyan

változókat keresni, amely releváns a vállalat működése szempontjából és mint biplot tengelyek beleilleszthetők. Azok az ügyintézők, akik leginkább kitűnnek valamilyen szempontból, azokat külön-külön is megvizsgálom.

## 5.1. Klaszterezés

A megbízó cég érdeke a profit maximalizálása, költségeinek a minimalizálása és az ügyintézők hatékonyságának növelése, ezért olyan változók jelentékenyek a számunkra, melyek e kettő céllal erősen összefüggnek, ezért úgy kell megválasztanunk azokat, hogy a többdimenziós skálázást azon dimenziókra elvégezve a lehető legtöbb információt nyerhessük az adatainkból. A profit maximalizáláshoz legszorosabban köthető változó az eladásszám, a költség minimalizáláshoz a híváshossz, a hatékonysághoz pedig a szüneten töltött idő, nyilvánvaló, hogy a vállalat leghatékonyabb működéséhez szükséges a minnél nagyobb eladásszám, a lehetséges legrövidebb hívások, és a lehető legjobb kihasználása a munkaidőnek. Az ügyintézők munkaideje sajnos nem homogén, vannak részmunkaidős dolgozók is, és egy adott kampányban is más-más intenzitással vesznek részt az operátorok, így a legkézenfekvőbb megválasztása a változóknak, hogy a napi nyolc órás munkaidejüket nyolc egyforma részre osztva és ezen idő intervallumokhoz rendelve a különböző megfigyelések átlagát készítsük el a többdimenziós skálázást, nem kivitelezhető. Három különböző adatmátrixból készítettem el az MDS ábrákat, összesen 19 ügyintézőt vizsgáltam, akiket egy hozzárendelt azonosítóval (user_id) láttunk el. Valamennyi mátrixnak 19 sora és 11 oszlopa van és adatait egy teljes kampányból vettem. Az első mátrix $i$. sorának $j$. eleme

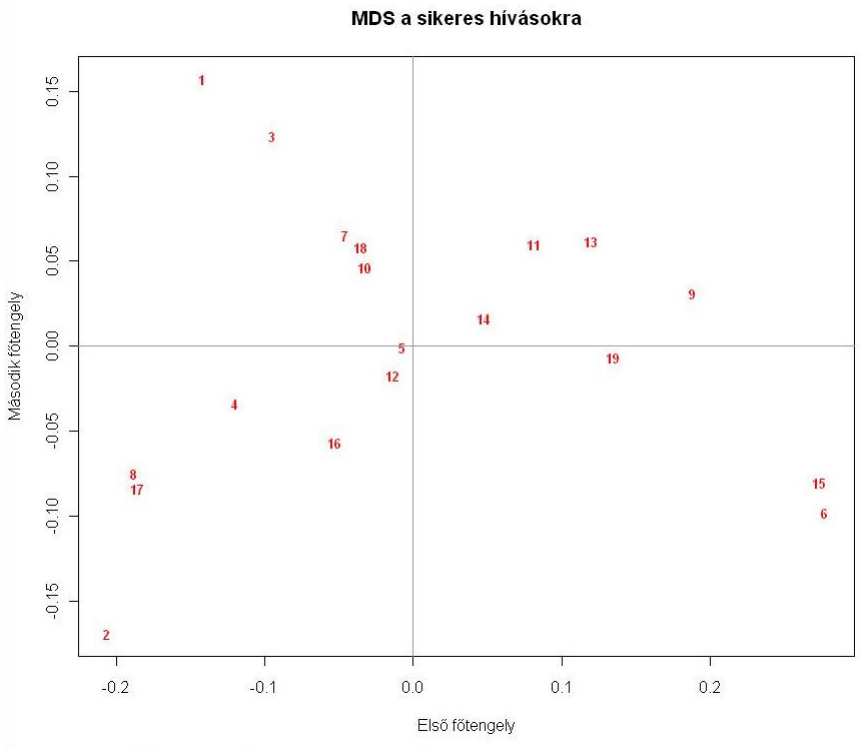
$$\begin{aligned} x_{ij} &= i.\,\text{ügyintéző}\,(j-1)*40.\,\text{és}\,j*40.\,\text{másodperc között befejezett} \\ &\quad \text{sikeres hívások száma} \end{aligned}$$

Azaz az első oszlopban a 0-40 másodperc között befejezett sikeres hívások relatív darabszáma, második oszlopban a 40-80 másodperc között befejezetteké és így tovább a 11. oszlopban a 400 másodpercen felüliek. A 40 másodpercenkénti idősávokat úgy választottam meg, hogy minden egyes oszlopban megfelelő mennyiségű adat kerüljön, majd elkészítettem az MDS ábráját és megvizsgáltam kik azok az ügyintézők akik leginkább érdekesek lehetnek a callcentert üzemeltető cég számára. A felhasznált programok a mellékletben találhatók. Az adatmátrixunkat az alábbi táblázat tartalmazza.

| | | x másodpercen belül befejezett sikeres hívások relatív darabszáma | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **sorszám** | **user_id** | *0-40* | *40-80* | *80-120* | *120-160* | *160-200* | *200-240* | *240-280* | *280-320* | *320-360* | *360-400* | *440+* |
| 1 | *14* | 0,0256 | 0,077 | 0,2051 | 0,28205 | 0,1795 | 0,07692 | 0,02564 | 0,0769 | 0,02564 | 0 | 0,0256 |
| 2 | *16* | 0,1429 | 0,071 | 0,0238 | 0,14286 | 0,0714 | 0,11905 | 0,09524 | 0 | 0,02381 | 0,1667 | 0,1667 |
| 3 | *18* | 0,1429 | 0,143 | 0,1429 | 0,28571 | 0,1429 | 0 | 0,28571 | 0 | 0 | 0 | 0 |
| 4 | *19* | 0,1014 | 0,135 | 0,1284 | 0,12162 | 0,1892 | 0,09459 | 0,05405 | 0,0541 | 0,06081 | 0,0338 | 0,0676 |
| 5 | *26* | 0,1642 | 0,224 | 0,1493 | 0,14925 | 0,194 | 0,08955 | 0,0597 | 0,0299 | 0,04478 | 0 | 0 |
| 6 | *29* | 0,3546 | 0,396 | 0,2133 | 0,12742 | 0,0332 | 0,0277 | 0,01385 | 0,0111 | 0 | 0 | 0,0055 |
| 7 | *40* | 0,0854 | 0,191 | 0,1809 | 0,21608 | 0,1357 | 0,07035 | 0,06533 | 0,0402 | 0,01508 | 0,0201 | 0,0251 |
| 8 | *44* | 0,0533 | 0,08 | 0,1467 | 0,06667 | 0,1467 | 0,2 | 0,08 | 0,04 | 0,08 | 0,0267 | 0,08 |
| 9 | *46* | 0,1967 | 0,295 | 0,3513 | 0,13349 | 0,0843 | 0,02342 | 0,01171 | 0,0094 | 0,00468 | 0,007 | 0,0023 |
| 10 | *47* | 0,1104 | 0,185 | 0,1881 | 0,2 | 0,1313 | 0,07463 | 0,06269 | 0,0269 | 0,01791 | 0,0149 | 0,0269 |
| 11 | *51* | 0,1357 | 0,211 | 0,325 | 0,15714 | 0,0964 | 0,05 | 0,02857 | 0,0071 | 0,01429 | 0,0071 | 0,0071 |
| 12 | *52* | 0,17 | 0,17 | 0,18 | 0,13 | 0,12 | 0,07 | 0,07 | 0,07 | 0,02 | 0,01 | 0,02 |
| 13 | *56* | 0,1452 | 0,315 | 0,2339 | 0,20968 | 0,1129 | 0,02419 | 0,00806 | 0,0081 | 0 | 0 | 0,0081 |
| 14 | *57* | 0,1099 | 0,275 | 0,2198 | 0,13187 | 0,1209 | 0,05495 | 0,07692 | 0,022 | 0,01099 | 0,022 | 0,011 |
| 15 | *58* | 0,3119 | 0,367 | 0,2936 | 0,08257 | 0,0459 | 0,02752 | 0,00917 | 0 | 0,00917 | 0 | 0,0092 |
| 16 | *59* | 0,1429 | 0,143 | 0,1786 | 0,07143 | 0,1429 | 0 | 0,10714 | 0,0357 | 0 | 0,0714 | 0,0714 |
| 17 | *61* | 0,08 | 0,093 | 0,1067 | 0,09333 | 0,1333 | 0,13333 | 0,12 | 0,0267 | 0,10667 | 0,0267 | 0,1067 |
| 18 | *62* | 0,1453 | 0,134 | 0,2137 | 0,21652 | 0,1368 | 0,10541 | 0,03989 | 0,0228 | 0 | 0,0142 | 0,0085 |
| 19 | *63* | 0,2419 | 0,258 | 0,2742 | 0,16129 | 0,0968 | 0,08065 | 0,01613 | 0 | 0,01613 | 0 | 0,0161 |

1. Táblázat: *Sikeres hívásokhoz tartozó adatmátrix*

Ebből elkészíttettem a távolság-mátrixot, amelyben az euklideszi távolságok szerepelnek. Ezek a változók kvantitatívak és azonos dimenziójúak így dimenzió mentesítésre szerencsére nincsen szükség, ennek alapján készítettem el az MDS ábrát, ahol a két dimenziós térbe való vetítést állítottam be, így kaptam a következőt.



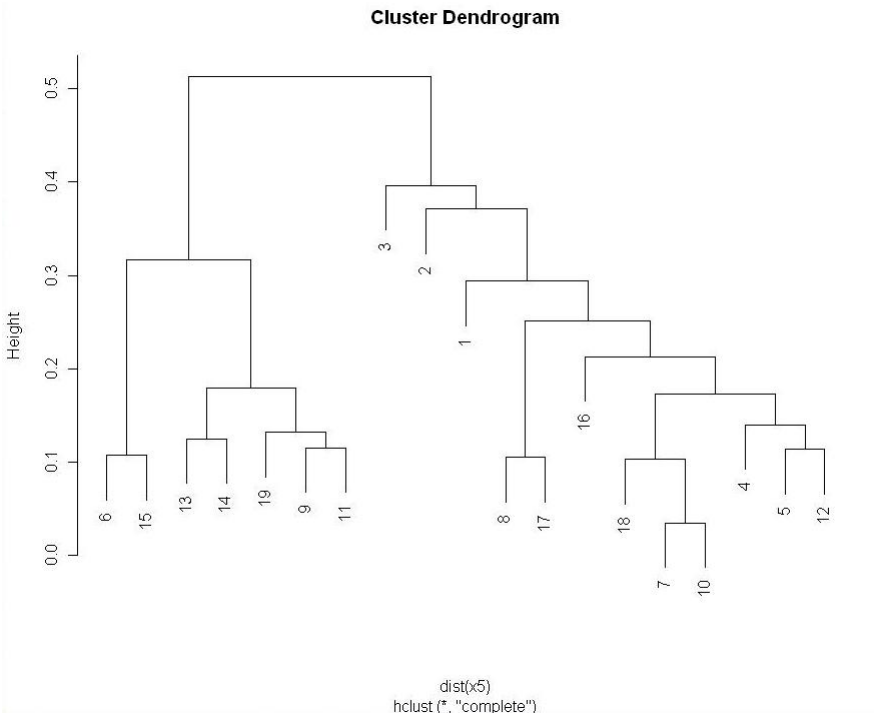

6. Ábra: *Sikeres hívások többdimenziós skálázása*

Ahol 1-nek az 1.tábla első sorához tartozó ügyintéző felel meg, 2 a másodiknak és így tovább. Láthatjuk, hogy a sikeres hívások tekintetében az 1, 3, 2, 8, 17, 15 és 6-os

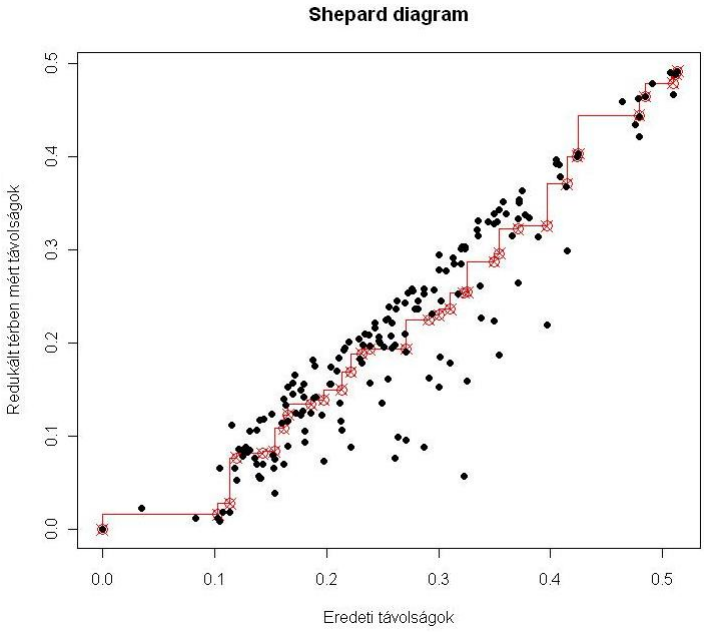
sorszámú operátorok térnek el az átlagostól. Nézzük meg milyen klasztereket tudunk kiolvasni az adatokból és ábrázoljuk őket egy dendrogramban mekkora hibával hoztuk létre a dimenzió csökkentést, és szemléltessük azt a Shepard diagrammal.

```
> illeszkedes(d5,cmds5)
[1] 0.1343946
```

Azaz az eredményt elemezhetjük, de azt némiképp fenntartásokkal kell kezelnünk és a kapott konklúziónkat alaposabban kell átgondolnunk. Elkészítve a fent említett dendrogramot `method = "complete"` beállítást választva, láthatjuk, hogy a fastruktúrája 5 szintű a legközelebbi két operátor e szerinti elemzés szerint a 7. és 10. ( 40 és 47 user_id-ú) van a legközelebb egymáshoz, és kettő markánsan elkülönülő klasztert fedezhetünk fel benne, amelyek az MDS ábránkon a második főtengelytől jobbra illetve balra helyezkednek el. A két osztályt $A$ és $B$-vel jelölve user_id-ként kaptuk, hogy $A = \{29, 46, 51, 56, 57, 58, 63\}$ és $B = \{14, 16, 18, 19, 26, 40, 44, 47, 52, 59, 61, 62\}$.




7. Ábra: *Sikeres hívások klaszter dendrogramja*




8. Ábra: *Sikeres hívások Shepard diagramja*

Vizsgáljuk meg, hogy a számunkra érdekesebb eladással zárult sikeres hívásokra, hogyan klaszterezhetők az ügyintézők. Az adatmátrixunk nagyon hasonló mint az előbb azzal a különbséggel, hogy most az idő intervallumok 60 másodpercesek, az adatmátrixunk $i$. sorának $j$. eleme

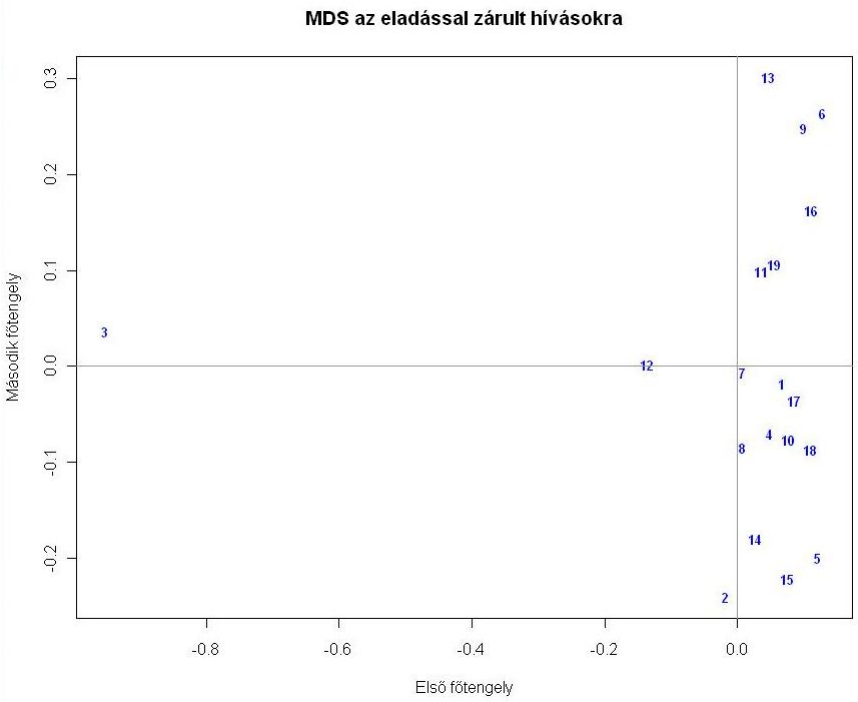
$$\begin{aligned} x_{ij} &= i.\,\text{ügyintéző}\,(j-1)*60.\,\text{és}\,j*60.\,\text{másodperc között befejezett} \\ &\quad \text{eladással zárult hívások relatív darabszáma} \end{aligned}$$

Az adatmátrixunk a következő.

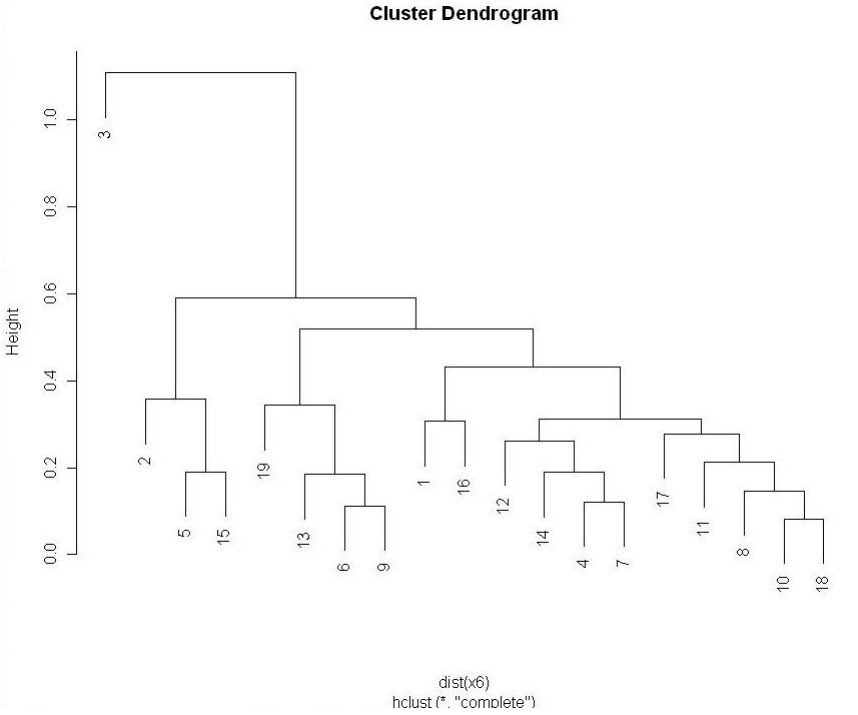
| | | x másodpercen belül befejezett eladással zárult hívások relatív darabszáma | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **sorszám** | **user_id** | *0-60* | *60-120* | *120-180* | *180-240* | *240-300* | *300-360* | *360-420* | *420-480* | *480-540* | *540-600* | *600+* |
| 1 | *14* | 0,1538 | 0,0769 | 0,07692 | 0,07692 | 0,15385 | 0 | 0 | 0,23077 | 0,07692 | 0,07692 | 0 |
| 2 | *16* | 0 | 0 | 0 | 0,09091 | 0,18182 | 0,27273 | 0,09091 | 0,18182 | 0,18182 | 0 | 0 |
| 3 | *18* | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | *19* | 0,0179 | 0,0893 | 0,07143 | 0,19643 | 0,125 | 0,10714 | 0,03571 | 0,08929 | 0,05357 | 0,05357 | 0,0179 |
| 5 | *26* | 0 | 0 | 0,04348 | 0,34783 | 0,17391 | 0,21739 | 0 | 0,13043 | 0,04348 | 0 | 0 |
| 6 | *29* | 0,0484 | 0,129 | 0,41935 | 0,24194 | 0,08065 | 0,04839 | 0,01613 | 0 | 0 | 0 | 0 |
| 7 | *40* | 0,0547 | 0,0938 | 0,14844 | 0,16406 | 0,14844 | 0,13281 | 0,09375 | 0,07031 | 0,02344 | 0,03125 | 0,0156 |
| 8 | *44* | 0,075 | 0,025 | 0,1 | 0,25 | 0,15 | 0,175 | 0,1 | 0,025 | 0,025 | 0,025 | 0 |
| 9 | *46* | 0,023 | 0,0805 | 0,4023 | 0,1954 | 0,12644 | 0,01149 | 0,03448 | 0,04598 | 0,03448 | 0 | 0 |
| 10 | *47* | 0,0339 | 0,0678 | 0,10169 | 0,27966 | 0,22034 | 0,13559 | 0,04237 | 0,01695 | 0,01695 | 0,04237 | 0 |
| 11 | *51* | 0,0465 | 0,093 | 0,25581 | 0,23256 | 0,13953 | 0,09302 | 0,09302 | 0,02326 | 0 | 0 | 0 |
| 12 | *52* | 0,0385 | 0,0769 | 0,11538 | 0,15385 | 0,15385 | 0,07692 | 0,23077 | 0,07692 | 0 | 0 | 0,0385 |
| 13 | *56* | 0 | 0,08 | 0,44 | 0,12 | 0,16 | 0 | 0,08 | 0 | 0 | 0,04 | 0 |
| 14 | *57* | 0,1 | 0,0667 | 0 | 0,2 | 0,16667 | 0,2 | 0,06667 | 0,13333 | 0,03333 | 0,03333 | 0 |
| 15 | *58* | 0 | 0 | 0,04545 | 0,31818 | 0,18182 | 0,31818 | 0,04545 | 0 | 0 | 0,04545 | 0,0455 |
| 16 | *59* | 0,15 | 0,1 | 0,3 | 0,15 | 0 | 0,05 | 0 | 0,25 | 0 | 0 | 0 |
| 17 | *61* | 0,0455 | 0,1364 | 0,04545 | 0,27273 | 0,09091 | 0 | 0 | 0,09091 | 0,04545 | 0 | 0,0455 |
| 18 | *62* | 0,0185 | 0,0556 | 0,12963 | 0,27778 | 0,24074 | 0,18519 | 0,01852 | 0,03704 | 0,01852 | 0 | 0 |
| 19 | *63* | 0,1176 | 0,2353 | 0,17647 | 0,17647 | 0,23529 | 0 | 0,05882 | 0 | 0 | 0 | 0 |

2. Táblázat: *Az eladással zárult hívásokhoz tartozó adatmátrix*

Az R-ben használt függvényeknek ugyanolyan beállításokat választottam mint előbb, elkészítve a fentiekhez hasonlóan az elemzéseket a következő eredményeket kaptuk.



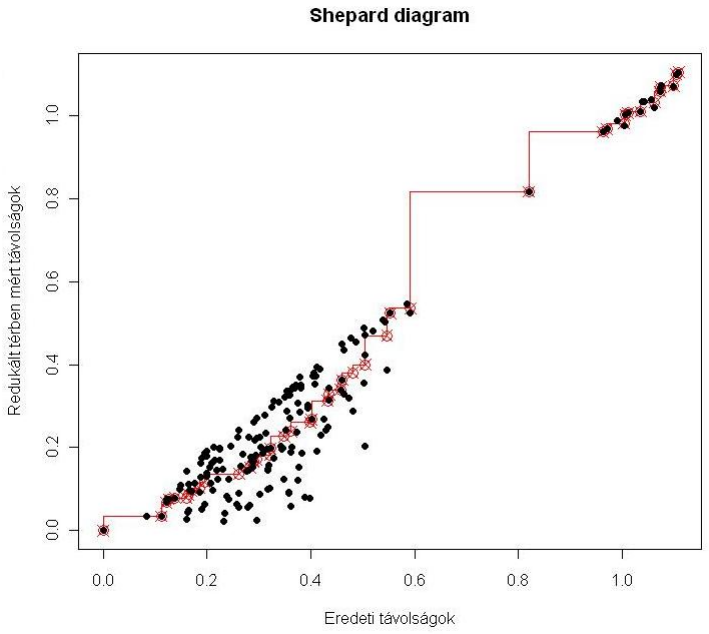

9. Ábra: *Az eladással zárult hívások többdimenziós skálázása*




10. Ábra: *Az eladással zárult hívásokhoz tartozó klaszter dendrogram*

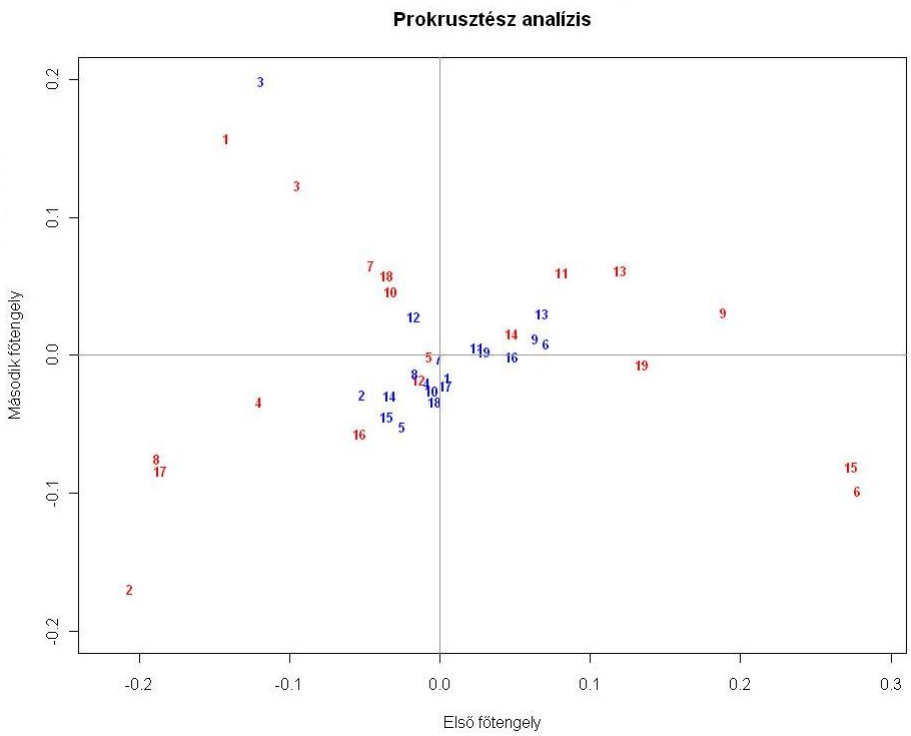
A dimnezió csökkentés hibája

```
> illeszkedes(d6,cmds6)
[1] 0.1892521
```




11. Ábra: *Shepard diagram az eladással zárult hívásokra*

A kapott eredmények alapján szükséges, hogy kellő fenntartással vizsgáljuk tovább az 9. ábrát. Jól látható a 3-as ügyintéző szegregált helyzete, ennek oka jól látható visszanézve az adatmátrixra, hogy mindösszesen csak egyetlen eladással rendelkező hívással bír. Három jól kivehető csoportot olvashatunk ki, amit nagyrészben a dendrogramban is észrevehetünk. Az a hipotézisünk lehet, hogy az eladásokhoz tartozó 9. ábra kissebb eltérést mutat az ügyintézők között, mintha csak a sikeres hívásokat vizsgálnánk, ugyanakkor a megbízó cégnek az eladással zárult hívásokhoz köthető változók alapján elkészített dimenzió csökkentés fontosabb lehet az átlagos nagyobb híváshossz miatt. Éppen ezért mivel az MDS outputja a forgatás, tükrözés, eltolás és skála invariáns, ezért elvégeztem a matematikai modellben vázolt prokrusztész analízist, hogy összehasonlíthassam valóban helytálló-e az az elképzelésem, miszerint az eladással zárult hívások alapján egy homogénebb csoportot alkotnak az értékesítők.



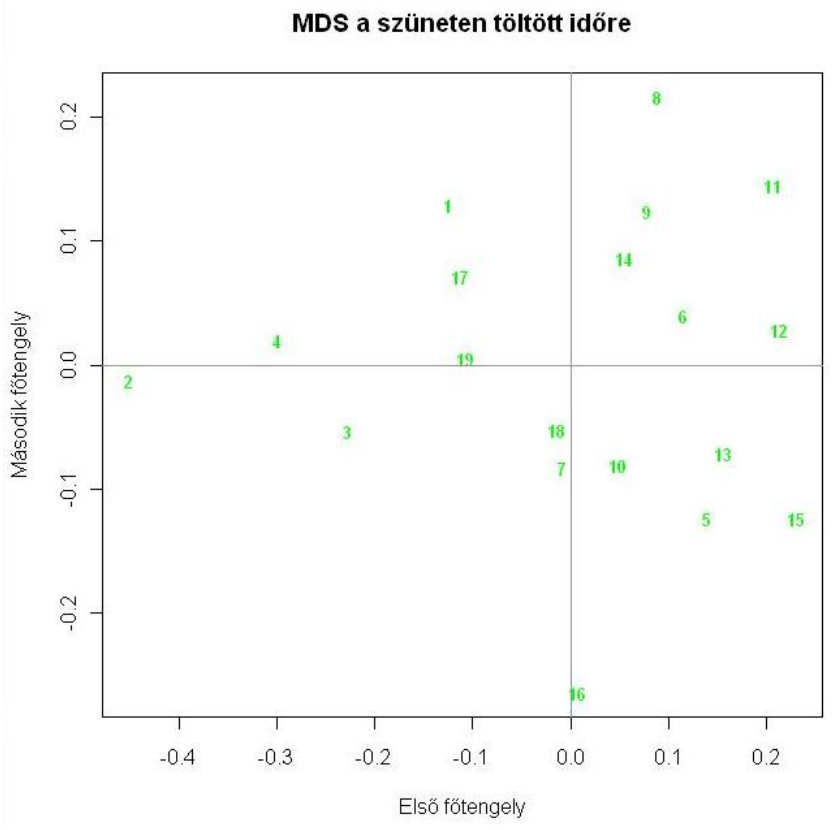

12.Ábra: *Az eladással zárult és a sikeres hívások MDS outputjainak összehasonlítása*

Azt láthatjuk, hogy az eladással zárult hívásokra készített dimenzió csökkentés az ügyintézők között egy homogénebb képet ad mint a sikeres hívásokéra, ugyanakkor észrevehetünk több olyan operátort, akik mind a kettő ábrában ugyanúgy periférián helyezkednek el, ilyenek a 2, 3, 13, 6, és a 9-es sorszámúak. Nézzük ezek után, hogy a hatékonyság szempontjából lényeges változó, a szüneten töltött idő alapján, hogyan néznek ki az ügyintézők a kettő dimenziós térben, vajon ott is hasonlóan, az átlagostól jobban eltérnek-e ezek az ügyintézők.

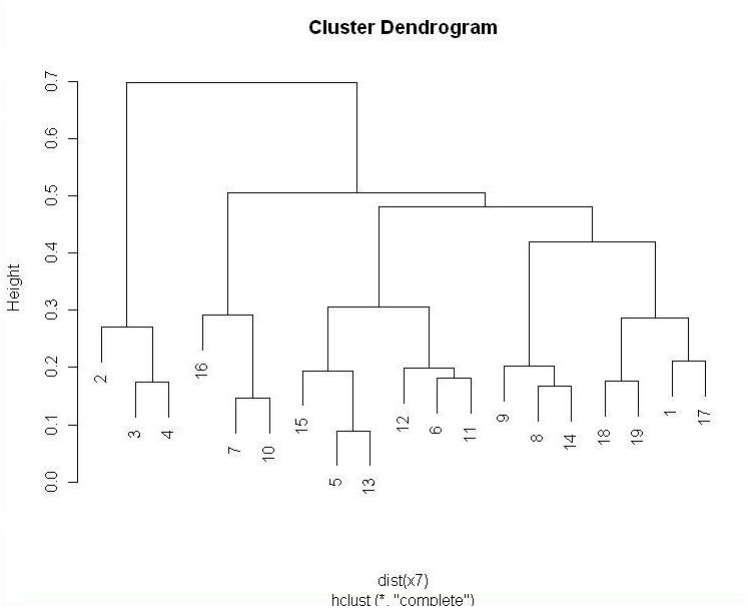
$$x_{ij} = i.\,\ddot{u}gyint\acute{e}z\H{o}\,(j-1)*90.\,\acute{e}s\,j*90.\,m\acute{a}sodperc\,k\ddot{o}z\ddot{o}tt\,befejezett\,sz\ddot{u}neten\,t\ddot{o}lt\ddot{o}tt\,id\H{o}\,relat\acute{i}v\,darabsz\acute{a}ma$$

| | | x másodpercen belül befejezett szünetek relatív darabszáma | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *sorszám* | **user_id** | *0-90* | *90-180* | *180-270* | *270-360* | *360-450* | *450-540* | *540-630* | *630-720* | *720-810* | *810-900* | *900+* |
| 1 | *14* | 0 | 0 | 0,05882 | 0 | 0 | 0 | 0,23529 | 0,11765 | 0,05882 | 0,17647 | 0,3529 |
| 2 | *16* | 0,148 | 0,0494 | 0,01235 | 0,01235 | 0,0206 | 0,01235 | 0,04938 | 0,03292 | 0,02058 | 0,01646 | 0,6255 |
| 3 | *18* | 0,15 | 0,0531 | 0,04134 | 0,03937 | 0,0315 | 0,05906 | 0,11417 | 0,05906 | 0,04528 | 0,03346 | 0,374 |
| 4 | *19* | 0,077 | 0,0361 | 0,02062 | 0,03093 | 0,0412 | 0,05155 | 0,15979 | 0,04639 | 0,00515 | 0,01546 | 0,5155 |
| 5 | *26* | 0,192 | 0,1154 | 0,07692 | 0 | 0 | 0 | 0,42308 | 0,03846 | 0 | 0 | 0,1538 |
| 6 | *29* | 0,067 | 0,0167 | 0,05 | 0,03333 | 0 | 0 | 0,41667 | 0,11667 | 0,08333 | 0,01667 | 0,2 |
| 7 | *40* | 0,134 | 0,0488 | 0,07317 | 0 | 0,0366 | 0,14634 | 0,21951 | 0,08537 | 0,07317 | 0,0122 | 0,1707 |
| 8 | *44* | 0,038 | 0 | 0,03846 | 0,03846 | 0 | 0 | 0,26923 | 0,38462 | 0,03846 | 0,03846 | 0,1538 |
| 9 | *46* | 0,017 | 0 | 0,01724 | 0,03448 | 0 | 0,12069 | 0,32759 | 0,24138 | 0,03448 | 0,01724 | 0,1897 |
| 10 | *47* | 0,081 | 0,0233 | 0,04651 | 0,04651 | 0,0349 | 0,23256 | 0,2907 | 0,05814 | 0,03488 | 0 | 0,1512 |
| 11 | *51* | 0,02 | 0 | 0 | 0,02041 | 0 | 0,02041 | 0,5102 | 0,22449 | 0 | 0,02041 | 0,1837 |
| 12 | *52* | 0,074 | 0 | 0 | 0 | 0 | 0 | 0,59259 | 0,07407 | 0,03704 | 0 | 0,2222 |
| 13 | *56* | 0,143 | 0,1071 | 0,03571 | 0 | 0 | 0,03571 | 0,42857 | 0,07143 | 0,03571 | 0 | 0,1429 |
| 14 | *57* | 0,136 | 0,0455 | 0 | 0 | 0 | 0 | 0,27273 | 0,27273 | 0,04545 | 0,04545 | 0,1818 |
| 15 | *58* | 0,265 | 0 | 0 | 0,02941 | 0 | 0 | 0,47059 | 0,08824 | 0,05882 | 0 | 0,0882 |
| 16 | *59* | 0,333 | 0 | 0,08333 | 0,04167 | 0 | 0,125 | 0,25 | 0 | 0 | 0 | 0,1667 |
| 17 | *61* | 0,023 | 0 | 0 | 0 | 0,0455 | 0,04545 | 0,31818 | 0,06818 | 0,06818 | 0,02273 | 0,4091 |
| 18 | *62* | 0,177 | 0,0692 | 0,04615 | 0,01538 | 0,0462 | 0,00769 | 0,25385 | 0,11538 | 0,03077 | 0,01538 | 0,2231 |
| 19 | *63* | 0,188 | 0 | 0 | 0 | 0 | 0,0625 | 0,1875 | 0,1875 | 0,0625 | 0 | 0,3125 |

3. Táblázat: *Szüneten töltött időhöz tartozó adatmátrix*

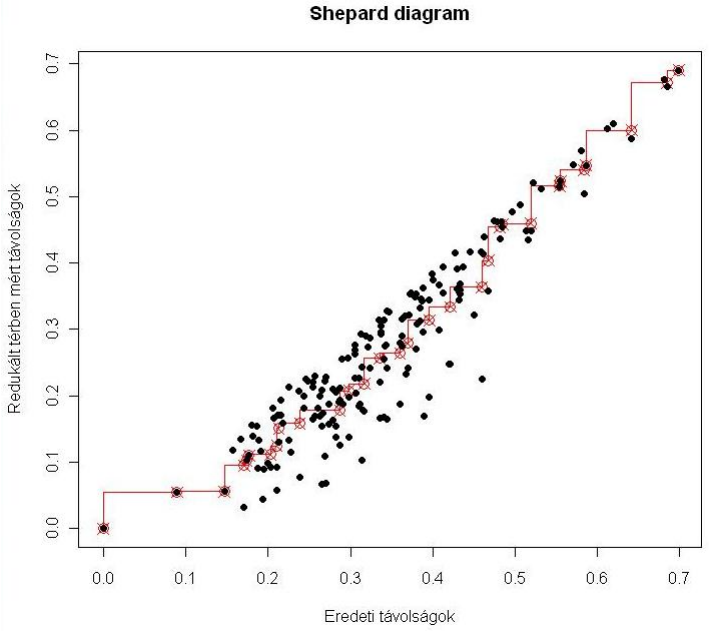



13. Ábra: *A szüneten töltött idő alapján a többdimenziós skálázás*




14.Ábra: *A szüneten töltött időhöz tartozó klaszter dendrogram*

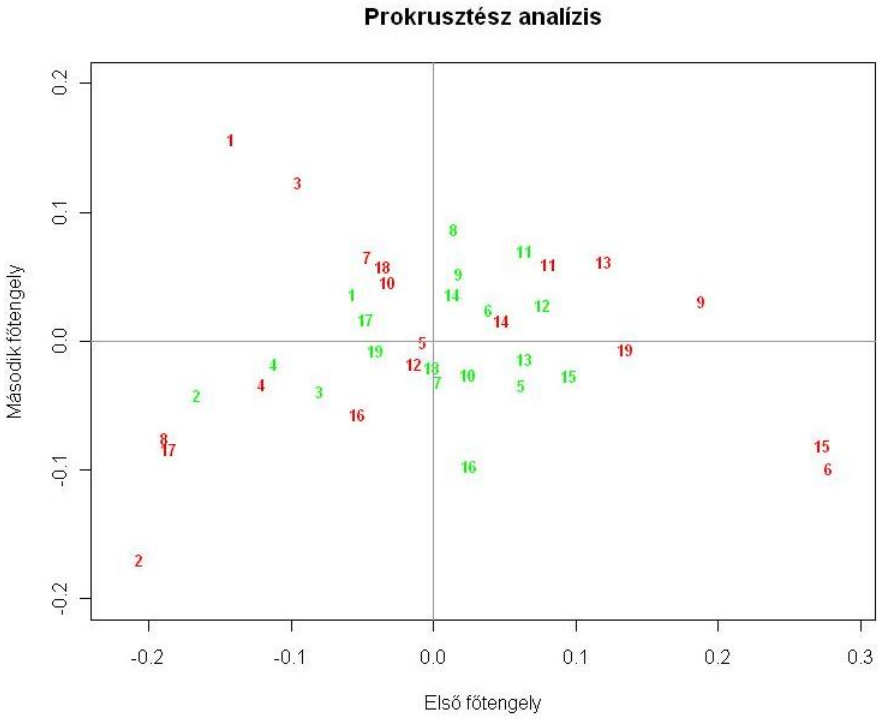
```
> illeszkedes(d7,cmds7)
[1] 0.1504381
```



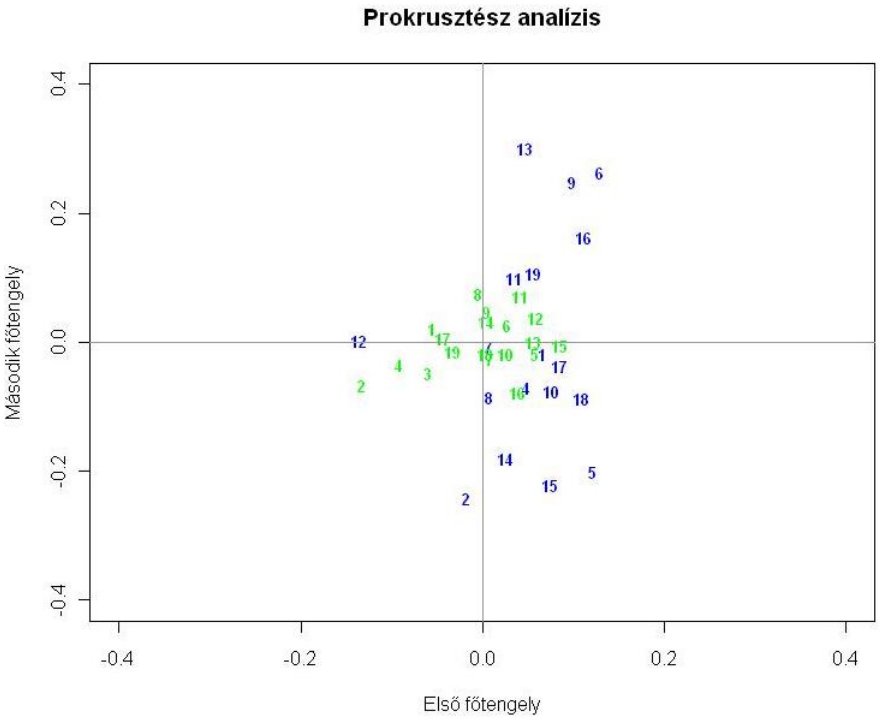

15. Ábra: *Shepard diagram a szüneten töltött időre*

Az 0.15 érték mutatja, hogy elégséges a többdimenziós skálázás outputja, viszont az 13. és 14. ábrából láthatjuk, hogy nincsenek olyan egyértelmű klaszterek a dolgozók között mint az előzőekben, érdekes lehet azonban megvizsgálni azt, hogy az eladással zárult és a sikeres hívásokra készített 6. és 9. ábra szélső pontjai ugyanúgy szélső

pontok-e ebben az esetben is. Végezzük el a prokrusztész analízisét 13.ábrának az eladással zárult és a sikeres hívások alapján elkészített MDS ábrákra.




16.Ábra: *Sikeres hívások és a szüneten tölött idő MDS outputjainak összehasonlítása*

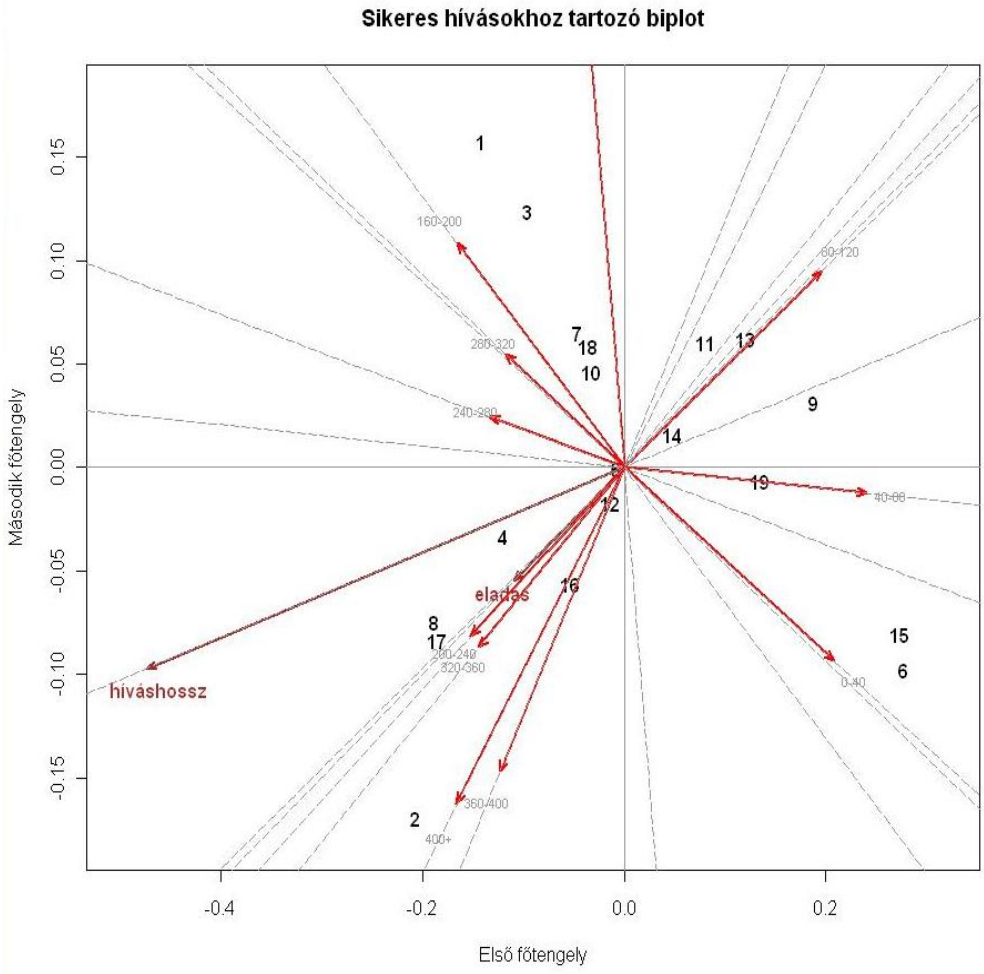



17.Ábra: *Eladással zárult hívások és a szüneten tölött idő MDS outputjainak összehasonlítása*

Nagyobb hasonlóságot az 6. ábra és az 13. ábra között látunk, nevezetesen a 2, 1, 15 és némiképp a 16-os sorszámú ügyintéző többeikhez képest elfoglalt pozíciója hasonló, sajnos az eladással zárult hívások és a szüneten töltött idő alapján elkészített MDS képe között a 2-es sorszámot viselő ügyintézőtől eltekintve hasonlóságot nem fedezhetünk fel.

## 5.2. Biplotok alkalmazása

Felhasználva a dimenzió csökkentés eredményeit, megvizsgáltam, hogy a változók amik mentén elkészítettük az adattábláink kétdimenziós térbe való vetítését, azok hogyan illeszthetők bele az ábrákban mint biplot tengelyek, továbbá az eladásszámra és az átlagos híváshosszra vonatkozó biplot tengelyekre is kitérek, evvel elősegíthetjük a látens változók megkeresését, egyfajta tágabb értelmezést adhatunk az MDS ábráknak. Mind a három esetre (sikeresre, eladásra, szünetre) az adatgyűjtéshez ugyanazt a kampányt használtuk fel, ehhez a kampányhoz tartozóan, lekértem az egyes ügyintézőkhöz tartozó relatív eladásszámot és átlagos híváshosszt. A modellben említett standardizálást valamennyi változóra megtettem és az `R`-ben megírt programot futtattam rájuk. Elsőként nézzük a sikeres hívásokra vonatkozó eredményünket. Az ehhez tartozó MDS ábrába 13 biplot tengelyt (eladás, híváshossz, MDS-t létrehozó 11 változó) helyezünk el, a biplot vektorokat felcímkéztem, hogy melyik változónak melyik tengely felel meg. Majd az eladással zárult hívásokra és a szüneten töltött idő alapján elkészített MDS ábrákba is analóg módon ugyanezt a 13 biplot tengelyt illesztjük bele. Némelyik biplotvektor a standardizálás ellenére nem fért bele az ábrába a jobb láthatóság érdekében




18. Ábra: *Sikeres hívásokhoz tartozó MDS és az őt létrehozó változók az eladásszám és az átlagos híváshossz biplottengelyek kapcsolata*

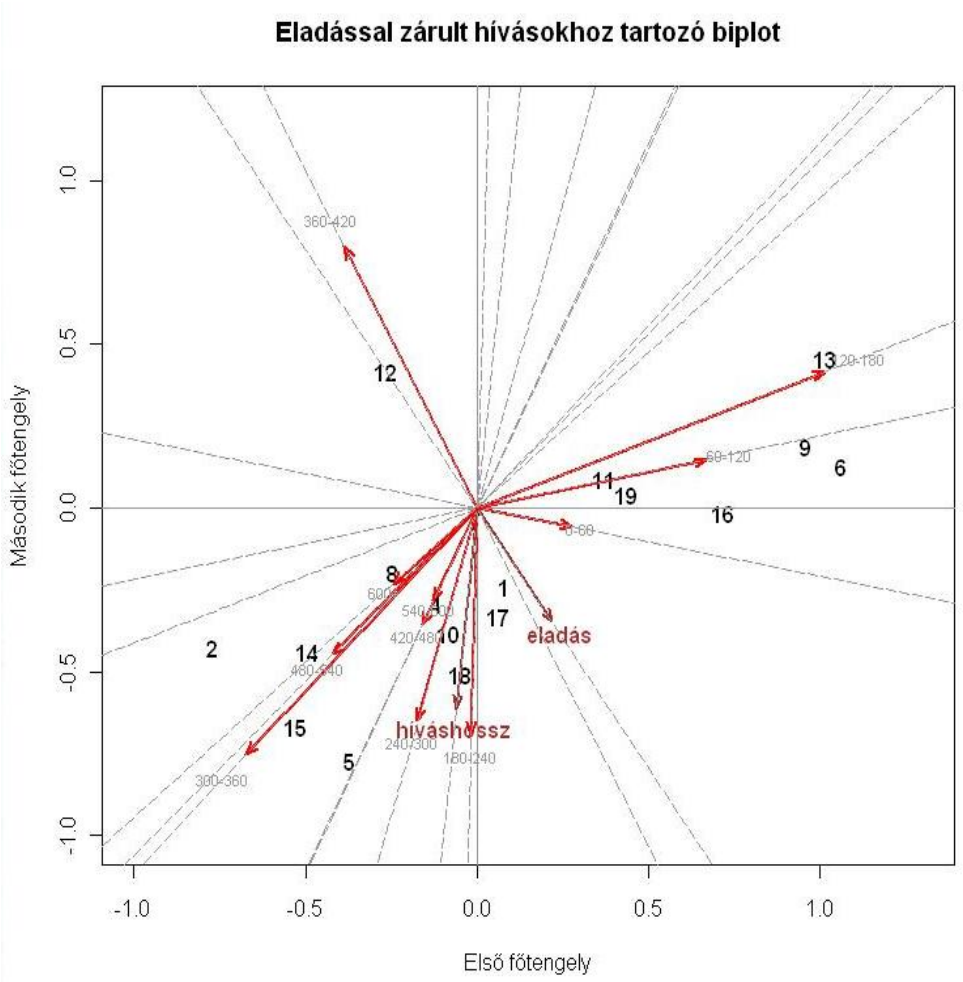
Nézzük meg mekkora a modellünk magyarázó ereje biplot tengelyenként, mindemellett ne feledkezzünk a dimenzió csökkentéséből származó hibáról sem, amikor az ábra kiértékelésébe kezdünk. A megfelelő $R^2$ értékeket az alábbi táblázatban foglaltuk össze

| Változó | Eladás | Híváshossz |
|---|---|---|
| $R^2$ | 0.01512 | 0.2343 |

| Változó | 0-40 | 40-80 | 80-120 | 120-160 | 160-200 | 200-240 | 240-280 | 280-320 | 320-360 | 360-400 | 400+ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R^2$ | 0.832 | 0.930 | 0.758 | 0.687 | 0.628 | 0.476 | 0.293 | 0.268 | 0.454 | 0.587 | 0.865 |

4.Táblázat

Láthatjuk, hogy az eladáshoz tartozó biplot tengely nem karakterizálja az ügyintézőket ebben a két dimenziós térben és a híváshosszhoz tartozó tengely sem túl jól írja le a kapcsolatot a hiváshossz és az ügyintézők MDS-beli pozíciója között. Ellenben az adatmátrixunk oszlopai jól jellemzik a többdimenziós skálázás eredményét, felfigyelhetünk arra a tényre, hogy a 0-40, 40-80, 80-120 időintervallumokhoz tartozó biplot tengelyekeknek a legerősebb a magyarázó erejük továbbá az ezekhez tartozó értékek a második főtengelytől jobbra eső irányba növekvő, ezért a sikeres hívásokat a leggyorsabban a 6, 9, 11, 13, 14, 15, 19-es sorszámú ügyintézők zárják le, tehát az MDS ábrához az első főtengely egyik lehetséges értelmezése az lehet, hogy mennyire hatékonyak az ügyintézők a híváshossz lerövidítésében a sikeres hívásokra vonatkozóan. Nézzük meg ugyanezeket az eredményeket az eladással zárult hívásokra, ebben az esetben a prokrusztész analízishez felhasznált transzformált ábrájába pórbáljuk meg beleilleszteni ugyanezeket a biplot tengelyeket.



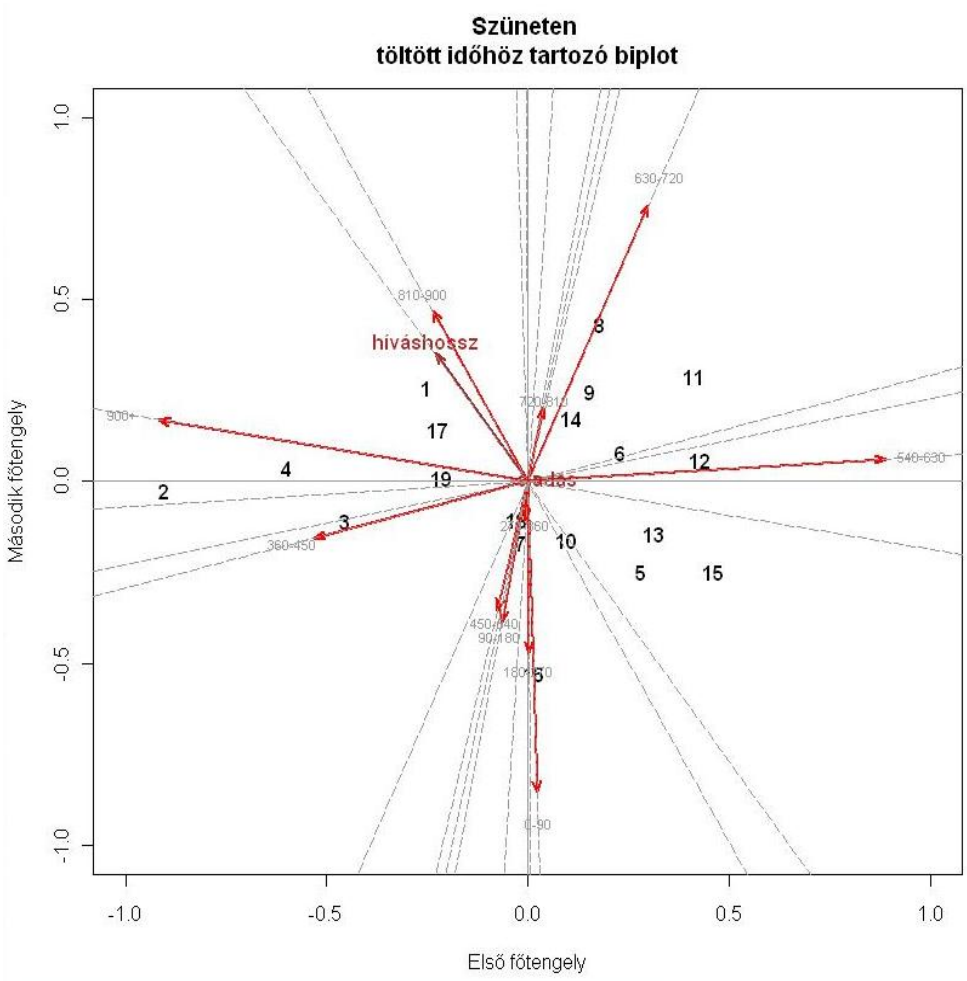

19. Ábra: *Eladással zárult hívásokhoz tartozó MDS és az őt létrehozó változók az eladásszám és az átlagos híváshossz biplottengelyek kapcsolata*

| Változó | Eladás | Híváshossz |
|---|---|---|
| $R^2$ | 0.220 | 0.35 |

| Változó | 0-60 | 60-120 | 120-180 | 180-240 | 240-300 | 300-360 | 360-420 | 420-480 | 480-540 | 540-600 | 600+ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R^2$ | 0.089 | 0.402 | 0.915 | 0.469 | 0.377 | 0.679 | 0.994 | 0.113 | 0.250 | 0.071 | 0.075 |

5. Táblázat

Azt vehetjük észre, hogy az eladáshoz és a hiváshosszhoz tartozó biplot tengelyek magyarázó ereje nőt ugyan, de nem kellő mértékben, hogy az releváns legyen a számunkra és a további változóknak pedig romlott az elfogadhatósága, azonban megfigyelhető, hogy az ábrának hasonló pozíciójában találjuk a 0-60, 60-120 és 120-180 változókra vonatkozóan a 6, 9, 11, 13, 19-es sorszámú ügyintézők ugyanúgy a rövid idősávokhoz tartozó biplot tengelyen veszik fel a legnagyobb értékeket. Így az előzőekhez hasonlóan a második főtengelynek is ugyanazt az értélemezést tudjuk adni. Nézzük most meg, hogy a szüneten tölött idő alapján elkészített dimenzió csökkentés esetén, hogyan alakul a modell használhatósága, tudunk-e további összefüggéseket kiolvasni belőle.




20. Ábra: *Szüneten töltött időhöz tartozó MDS és az őt létrehozó változók az eladásszám és az átlagos híváshossz biplottengelyek kapcsolata*

Nézzük meg, hogy ehhez a paraméterhez tartozó biplotot (20. Ábra) mekkora magyarázó erővel jellemezhetünk, sajnos $R^2$-re vonatkozóan olyan kicsi értékeket kaptunk, hogy abból számunkra lényeges új információt nem tudunk kiolvasni.

| Változó | Eladás | Híváshossz |
|---|---|---|
| $R^2$ | 0.001 | 0.177 |

| Változó | 0-690 | 90-180 | 180-270 | 270-360 | 360-450 | 450-540 | 540-630 | 630-720 | 720-810 | 810-900 | 900+ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $R^2$ | 0.733 | 0.153 | 0.224 | 0.011 | 0.307 | 0.131 | 0.796 | 0.668 | 0.043 | 0.277 | 0.870 |

6. Táblázat

# 6. Összegzés

A munkáltató szempontjából leginkább jelentős három paramétert választottam ki az ügyintézőkre nézve, ezekre vonatkozó kvantitatív változókat képeztem az idő tengely diszkretizálása alapján, amelyekből meghatározott távolság-mátrixok képezték a többdimenziós skálázás inputját. Mind a három esetben a kettő dimenziós térbe vetítettem a 11 dimenziós adatmátrixunkat, amelyekből külön-külön is leolvashatók kik azok az ügyintézők akik eltérnek az átlagostól, illetve milyen csoportokat fedezhetünk fel a választott paraméterek mentén. A prokrusztész analízissel megvizsgáltam, hogy melyik paraméter alapján klaszterezhetők jobban az ügyintézők és megfigyelhetővé vált, hogy a három MDS ábra valójában mennyire is tér el egymástól. Ahhoz, hogy megkönnyítsük annak a leolvasásását, hogy az ábra mely pozícióiban helyezkednek el a zömében rövid híváshosszokkal illetve szüneten töltött időkkel rendelkezők, elkészítettem azt a biplotot, ahol a meglévő MDS ábrába illesztettem bele a 11 változónknak megfelelő tengelyt. Majd az alapján megállapíthattuk, hogy a sikeres hívásokra elkészített dimenzió csökkentés esetén az első főtengely egyik lehetséges értelmezése lehet az, hogy mennyire hatékony az ügyintéző abban, hogy a híváshosszokat minimalizálja a sikeres hívásokra vonatkozóan. Ezt az eredményt elfogadhatjuk, a modell magyarázó ereje kellő mértékű volt. Ebből a szempontból a leginkább hatékony operátorok user_id-ként az $A = \{29, 46, 51, 56, 57, 58, 63\}$ klaszter, az eladással zárult hívásokra ugyanez az értelmezés jóval gyengébb magyarázó erővel ugyan, de kijött, sajnos azonban a szüneten töltött időre hasonló következtetéseket már nem tudtam levonni ebből az adatsorból az igen alacsony $R^2$ értékek miatt. Megkíséreltem az eladásszámot és az átlagos híváshosszt is biplot tengelyenként az ábrákba illeszteni, de szintén az alacsony $R^2$ miatt az eredményünk nem elfogadható.

Összegzésképpen elmondható, hogy az adatsorunkra a sikeres és eladással zárult hívásokra elkészített dimenzió redukció volt a leghasznosabb az adathalmazunk áttekintése szempontjából és a biplotok felhasználásával részleges értelmezését is tudtuk adni a látens változóknak. A módszereink hasznossága akkor mutatkozik meg igazán, amikor lényegesen nagyobb adatsorokra futtatjuk ezeket a modelleket, a gyakorlatban akár 200 fős ügyintéző csoportal is találkozhatunk, így az elsőre áttekinthetetlen adathalmazból egy közérthetőbb, vizuális elemzéssel szolgálhatunk a menedzsment számára.

# 7. Mellékletek

## Klaszterezés

A sikeres hívásokhoz tartozó adatokból az alábbiak szerint készítjük el az MDS ábrát és a hozzá tartozó klaszter analízist. Az eladással zárult hívások és a szüneten töltött idő alapján elkészített többdimenziós skálázás (9. Ábra, 13. Ábra) és klaszter analízise (10. Ábra, 14. Ábra) teljesen analóg módon történik.

```
x5=read.csv("c:/5.csv",sep=";",dec=",",header=F)
```
# x5: az 1.Táblában szereplő adatok beolvasása egy 5.csv nevű excel fájlból
```
d5=as.matrix(dist(x5))
```
# d5: az 1. Táblához tartozó távolság-mátrix
```
cmds5=cmdscale(d5)
```
# cmds5: többdimenziós skálázás outputja az objektumok kétdimenziós koordináta mátrixa
```
x_sikeres=cmds5[,1]
```
# x_sikeres: output x tengelyéhez tartozó koordináta vektora
```
y_sikeres=cmds5[,2]
```
# y_sikeres: output y tengelyéhez tartozó koordináta vektora

Az MDS grafikai megjelenése (6. Ábra)
```
plot(x_sikeres,y_sikeres,t="n",xlab="Első főtengely",
 ylab="Második főtengely",main="MDS a sikeres hívásokra")
text(x_sikeres,y_sikeres,cex=0.8,col="red",font=2)
abline(h=0,v=0,col="gray60")
```

Dendrogram elkészítése (7.Ábra)
```
plot(hclust(dist(x5),"complete"))
```

## Shepard diagram

A Shepard diagramok elkészítéséhez egy `shepard` nevű függvényt írtam, amely bemenő paramétere a megfelelő távolság-mátrix.

```
shepard<-function(tavolsag_matrix)
  {d=NULL
  dk=NULL
  d=as.matrix(tavolsag_matrix)
```
# d: eredeti térben mért távolság-mátrix
```
  dk=as.matrix(dist(as.matrix(cmdscale(tavolsag_matrix))))
```
# dk: redukált térben mért távolság-mátrix

```
  plot(isoreg(d,dk),main="Shepard diagram",xlab="Eredeti távolságok",
   ylab="Redukált térben mért távolságok")
   # Izotónikus regresszió a d és dk szórás ábrájára és annak kirajzolása
  points(d,dk,pch=16)}
   # d és dk koordinátájú pontok beillesztése

shepard(eurodist)
   # 1. Ábra
shepard(d5)
   # 8. Ábra, ahol d5 a sikeres hívásokhoz tartozó távolság-mátrix
```

Az eladással zárult hívásokhoz (11. Ábra) és a szüneten töltött időhöz (15. Ábra) tartozó Shepard diagram teljesen analóg módon történik.

## A dimenziócsökkentés hibája

A dimenzió csökkentéssel kapott grafikonok elemezhetőségének mérése az `illeszkedés` fügvénnyel történik, amelynek kettő inputja közül az első a távolság-mátrix, második az MDS ábra koordináta mátrixa.

```
illeszkedes<- function(d,mdskoord)
  {f=NULL
  dkalap=as.matrix(dist(mdskoord))
   # dkalap: a redukált térben mért távolság-mátrix
  for(i in 1:ncol(d))
   # for ciklus egytől az eredeti távolság-mátrix sor  számáig  ( ncol(d)=nrow(d))
   {
     for (j in 1:ncol(d))
   # for ciklus egytől az eredeti távolság-mátrix oszlop számáig
      {f=rbind(f,(d[i,j]-dkalap[i,j])^2)}}
   # f elemei a távolság-mátrixok megfelelő eleminek a különbségének a négyzete
  s=sqrt(sum(f)/sum(d))
```

$$\# \; s= \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(d_{ij}-\hat{d_{ij}}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}^2}}$$

```
   print(s)
   # s kiírása
  }

illeszkedes(d5,cmds5)
   # a sikeres hívások MDS grafikonjának elemezhetőségének mérőszáma
```

Az eladáshoz és a szüneten töltött időre vonatkozó mérőszámok teljesen analóg módon határozandók meg.

## Prokrusztész analízis

A következőkben a sikeres hívások és az eladással zárult hívások dimenzió csökkentés utáni grafikonjaira végezzük el a prokusztész analízist, az utóbbit transzformálva, meghatározva az optimális forgatás-tükrözés, eltolás és nyújtás nagyságát (12. Ábra). A szüneten töltött időhöz tartozó kettő prokrusztész elemzés (16. ábra, 17. Ábra) teljesen analóg módon történik. A jelöléseket megtartva az $A$ konfiguráció jelenti a sikeres hívások MDS outputjának koordináta mátrixát, míg $B$ az eladással zárult hívásokét.

```
I=diag(c(1),nrow=19,ncol=19)
   # I: a megfelelő méretű egység mátrix
J=I-(1/19)*matrix(c(1),nrow=19,ncol=19)
   # J: a centráló mátrix
C=t(cmds5)%*%J%*%cmds6
d=diag(c(svd(C)$d),nrow=2,ncol=2)
   # a C szinguláris érték felbontásában szereplő diagonális mátrix
u=as.matrix(svd(C)$u)
   # a C szinguláris érték felbontásában szereplő bal oldali szinguláris érték mátrix
v=as.matrix(svd(C)$v)
   # a C szinguláris érték felbontásában szereplő jobb oldali szinguláris érték mátrix
T=v%*%t(u)
   # az optimális forgatás mátrix
s=sum(diag(t(cmds5)%*%J%*%cmds6%*%T))/sum(diag(t(cmds6)%*%J%*%cmds6))
   # az optimális nyújtás nagysága
t=(1/19)*t((cmds5-s*cmds6%*%T))%*%matrix(c(1),nrow=19,ncol=1)
   # az optimális eltolás mértéke
cmds6_tr=s*cmds6%*%T+(matrix(c(1),nrow=19,ncol=1)%*%t(t))
   # A B konfiguráció transzformációja
x_eladas_tr=cmds6_tr[,1]
   # a transzformáció outputjának x koordinátai
y_eladas_tr=cmds6_tr[,2]
   # a transzformáció outputjának y koordinátai
```

Ábrázuljuk $A$-t és $B$ transzformáltját egy ábrában (12. Ábra)

```
plot(x_eladas_tr,y_eladas_tr,t="n",xlab="Első főtengely",
ylab="Második főtengely",main="MDS az eladással zárult hívásokra")
text(x_eladas_tr,y_eladas_tr,cex=0.8,col="blue",font=2)
plot(x_sikeres,y_sikeres,xlim=c(-0.22,0.29),ylim=c(-0.2,0.2),t="n",
xlab="Első főtengely",ylab="Második főtengely",
main="Prokrusztész analízis")
text(x_sikeres,y_sikeres,cex=0.8,font=2,col="red")
text(x_eladas_tr,y_eladas_tr,cex=0.8,font=2,col="blue")
abline(h=0,v=0,col="gray60")
```

## Biplot

Elkészítjük a sikeres hívásokhoz tartozó biplot ábrát, amelyhez a dimenzió csökkentés grafikonját (6. Ábra) használjuk fel és abba illesztjuk bele az ügyintézőkhöz tartozó relatív eladásszámot és átlagos híváshosszt, továbbá az MDS-t létrehozó változókat, így összesen 13 biplot tengelyt illesztünk az ábrába és rendre kiszámítjuk a hozzájuk tartozó $R^2$ értékeket.

```
x6a<-read.csv("c:/6a.csv",header=T,sep=";",dec=",")
```
# beolvassuk a relatív eladásszámokat tartalmazó 6a.csv nevű excel fájlt
```
x6b<-read.csv("c:/6b.csv",header=T,sep=";",dec=",")
```
# beolvassuk az átlagos híváshosszt tartalmazó 6b.csv nevű excel fájlt
```
hivasszam=x6a[,2]
```
# `hivasszam`: kimentjük azon hívások számát amelyek eladással zárultak
```
hivashossz=x6b[,2]
```
# `hivashossz`: kimentjük a teljes kampányra vonatkozó átlagos híváshosszt

Standardizáljuk a változóinkat, hogy az objektív összehasonlításuk lehetségessé váljon.

```
x_sikeres=(x_sikeres-mean(x_sikeres))/sd(x_sikeres)
y_sikeres=(y_sikeres-mean(y_sikeres))/sd(y_sikeres)
hivasszam=(hivasszam-mean(hivasszam))/sd(hivasszam)
hivashossz=(hivashossz-mean(hivashossz))/sd(hivashossz)
```

Az eladásszámra és az átlagos híváshosszra egy lineáris regressziót készítünk az MDS outputjának koordinátáit választva magyarázó változóknak.

```
summary(lm(hivasszam~x_sikeres+y_sikeres))
```
# leolvashatjuk a magyarázó erejét a modellnek az eladásszámra
```
eh=lm(hivasszam~x_sikeres+y_sikeres)$coefficients
```
# `eh`: kimentjük a regressziós koefficienseket ebbe a vektorba
```
summary(lm(hivashossz~x_sikeres+y_sikeres))
```
# leolvashatjuk a magyarázó erejét a modellnek a híváshosszra
```
eh2=lm(hivashossz~x_sikeres+y_sikeres)$coefficients
```
# `eh`: kimentjük a regressziós koefficienseket ebbe a vektorba

18. ábra elkészítése első lépésében létrehozzuk a biplotot a sikeres hívásokhoz tartozó MDS ábra és az eladásszám illetve átlagos híváshosszra mint biplot tengelyre vonatkozóan.

```
plot(cmds5[,1],cmds5[,2],t="n",ylim=c(-0.18,0.18),xlim=c(-0.5,0.32),
 xlab="Első főtengely",ylab="Másodikfőtengely"
  ,main="Sikeres hívásokhoz tartozó biplot")
text(cmds5[,1],cmds5[,2],font=2)
```
# redukált térben az objektumok beillesztése a koordinátáik alapján

```
arrows(0,0,eh[2],eh[3],length=0.1,angle=20,lwd=2.5,col="brown")
   # biplot irány az eladásra
text(1.1*eh[2],1.1*eh[3],labels="eladás",col="brown",font=2)
   #  biplot címke az eladásra
abline(a=0, b=(eh[3])/(eh[2]), col = "gray60",lwd=1,lty=5)
   #  biplot tengely az eladásra
arrows(0,0,eh2[2],eh2[3],length=0.1,angle=20,lwd=2.5,col="brown")
   # biplot irány a híváshosszra
text(1.1*eh2[2]+0.06,1.1*eh2[3],labels="híváshossz",col="brown",font=2)
   # biplot címke a híváshosszra
abline(a=0, b=(eh2[3])/(eh2[2]), col = "gray60",lwd=1,lty=5)
   # biplot tengely a híváshosszra
abline(h=0,v=0,col = "gray60",lwd=1.5)
   # x, y koordináta tengelyek  kirajzolása
```

Standardizáljuk az 1.táblához tartozó 11 oszlop adatait.

```
for(i in 1:11)
{ x5[,i]=(x5[,i]-mean(x5[,i]))/sd(x5[,i])}
```

Beillesztjük a már meglévő ábránkba az MDS-t létrehozó változókat mint biplot tengelyeket.

```
l=c("0-40","40-80","80-120","120-160","160-200","200-240","240-280",
 "280-320","320-360","360-400","400+")
   # a 11 tengely címkéi
coefs=NULL
r=NULL
for(i in 1:11)
{ coefs=rbind(coefs,lm(x5[,i]~x_sikeres+y_sikeres)$coefficients)
   # rendre kimentjük a coefs vektorba a változókhoz tartozó koefficienseket
r=rbind(r,summary(lm(x5[,i]~x_sikeres+y_sikeres))$r.squared)
   # R^2 értékek kimentése
arrows(0,0,0.25*coefs[i,2],0.25*coefs[i,3],length=0.1,angle=20,lwd=2.5,
 col="red")
   # biplot irányok
text(1.1*0.25*coefs[i,2],1.1*0.25*coefs[i,3],labels=l[i],col="gray60",
 font=1,cex=0.7)
   # biplot címkék
abline(a=0, b=coefs[i,3]/coefs[i,2], col = "gray60",lwd=1,lty=5)}
   # biplot tengelyek
```

Az eladással zárult hívásokhoz (19. Ábra) és a szüneten töltött idő alapján elkészített MDS-hez tartozó (20. Ábra) biplotok teljesen analóg módon készítendők.

# Hivatkozások


[1] Ingwer Borg, Patrick Groenen: *Modern Multidimensional Scaling Theory and applications*, Springer, 1997

[2] Michael Greenacre: *Biplots in Practice*, Fundación BBVA, 2010

[3] Barna Ildikó, Székelyi Mária: *Túlélőkészlet az SPSS-hez*, Typotex, 2003

[4] Bolla Marianna, Krámli András: *Statisztikai következtetések elmélete*, Typotex, 2005

[5] Michael A.A.Cox, Trevor F. Cox: *Handbook of Data Visualization*, 315-347, Springer, 2008

[6] K. Gabriel: *Biplot*, Biometrika 58: 453-467, 1971

[7] Timm, Neil H.: *Applied Multivariate Analysis*, Springer, 2002

[8] Michael J. Crawley: *The R Book*, J Wiley, 2007

[9] Füstös László, Meszéna György, Simonné Mosolygó Nóra: *A sokváltozós adatelemzés statisztikai módszerei*, Akadémiai Kiadó, 1986

# Köszönetnyilvánítás

Köszönettel tartozom elsősorban konzulensemnek és tanáromnak Pröhle Tamásnak, hogy rendelkezésemre állt amikor szükségem volt rá és hasznos tanácsaival segítette a szakdolgozatom irányait, illetve felhívta a figyelmem a dolgozatban szereplő pontatlanságokra. Továbbá Willinger Krisztiánnak híradástechnikai rendszermérnöknek, aki az adatbányászatban segítségemre volt és az adatokat a rendelkezésemre bocsátotta.