Analiza korespodencije

Jerebić, Vladimir

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Vladimir Jerebić

ANALIZA KORESPONDENCIJE

Diplomski rad

Voditelj rada:
doc. dr. sc. Lavoslav Čaklović

Zagreb, srpanj, 2015.godina

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

Prvenstveno se zahvaljujem svom mentoru doc.dr.sc Lavoslavu Čakloviću na velikom trudu i pomoći oko izrade ovog diplomskog rada. Hvala mojim roditeljima što su me potakli na studij matematike i omogućili mi ugodno studiranje, hvala baki Ani na njezinoj srdačnoj potpori tokom cijelog studija i hvala Karli koja mi je bila podrška tijekom pisanja ovog rada.

Sadržaj

Uvod

Vizualizacija podataka je integralni dio statističkog analitičkog procesa, štoviše, ne postoji niti jedan statistički alat korisniji od dobro napravljenog grafa. Grafički prikaz ne samo da prikazuje informacije sadržane u podacima nego služi kao podloga za dobivanje novih informacije do kojih bi teško došli osnovnom parametarskom multivarijatnom analizom. Dijagram razmještaja (eng. scatterplot) je jedan od grafičkih alata pogodnih za vizualizaciju podataka. To je dvodimenzionalni (trodim.) prikaz skupa podataka. Prvenstveno nam pruža informacije o međusobnim odnosima pojedinih varijabli, a u slučaju kada se uzorak sastoji od dvije ili tri varijable moguće je precizno prikazati podatke na dvodimenzionalnom odnosno trodimenzionalnom dijagramu razmještaja. Činjenica što su ljudi ograničeni s tri dimenzije predstavlja problem i postavlja se pitanje kako vizualizirati višedimenzionalne podatke. Odgovor na to pitanje dan je redukcijom dimenzionalnosti podataka. Osnovna ideja je prikazati podatke u nižedimenzionalnom potprostoru u kojem je gubitak informacija nastao smanjenjem dimenzije najmanji a upravo je to osnova biplota. Biplot možemo opisati kao grafički alat koji nam omogućava analizu dvosmjerne interakcije u tablici s n objekata i p varijabli kako bi se strukturne povezanosti između pojedinih objekata i varijabli lakše uočile i procijenile. Pri tome naglasimo kako prefiks "bi" u nazivu biplota nije aluzija na ideju dvodimenzionalnosti biplota, već na mogućnost istovremenog prikaza redaka i stupaca tablice podataka kao dva odvojena skupa podataka.

Analiza korespondencije (eng. Correspondence analysis) ili skraćeno CA je metoda vizualizacije podataka koja je primjenjiva na kontigencijske tablice. Iako se teorijska pozadina metode razvila prije pedesetak godina, pravi poticaj modernoj primjeni analizi korespodencije daje francuski lingvist Jean-Paul Benzécri šezdesetih godina prošlog stoljeća. Metoda se proširila van Francuske i ušla u različite znanstvene grane zahvaljujući knjizi *Theory and Application of Correspondence Analysis* Michaela Greenacrea iz 1984. godine jer su pomoću te knjige znanstvenici upoznali metodu koja im omogućava grafički prikaz kompleksnih tablica s nenumeričkim podacima. U današnje vrijeme analiza korespondencije se najviše koristi u sociologiji, psihologiji, ekologiji, marketingu i biomedicinskom istraživanju.

Cilj ovog diplomskog rada je izložiti osnove teorije analize korespondencije, odnosno njezine primjene u praksi. U Poglavlju 1 ovog rada dajemo teoreme koji će biti potrebni

u radu kako bi mogli detaljnije upoznati biplotove, njihovu konstrukciju te predstaviti različite vrste biplota. U Poglavlju 2 predstavljamo analizu korespondencije, definiramo osnovne pojmove i objašnjavamo osnovni algoritam analize korespodencije. Kasnije ćemo za kategorijske podatke poopćiti ovu metodu u višestruku analizu korespodencije (eng. Multiple Correspondence analysis) odnosno skraćeno MCA koja je vrlo popularna u analizi anketa. Također ćemo se upoznati s kanonskom analizom korespodencije (eng. Canonical Correspondence analysis) odnosno skraćeno CCA čija je prednost vizualiziranje podataka u restringiranom prostoru kovarijabli. Za kraj ćemo pokazati kako se podaci koji nisu u kontigencijskoj tablici mogu svesti na oblike pogodne za analizu korespondencije. U poglavlju 3 sam naveo osnovne kodove programskog jezika R koje sam koristio pri obradi podataka za ovaj diplomski.

Poglavlje 1

Biplot

Uvod

U ovom radu će se koristi brojni rezultati linearne algebre pri čemu su oni bitniji objašnjeni u ovom poglavlju. Također, kasnije ćemo u ovom poglavlju objasniti što je to biplot i demonstrirati njegovu konstrukciju, prikazati različite vrste biplota na primjerima te ukratko objasniti metodu analize glavne komponente.

1.1 Potrebni rezultati linearne algebre

Pretpostavimo da je $\boldsymbol{X}$ matrica dimenzija $n \times p$, pri čemu bez smanjenja općenitosti možemo pretpostaviti da vrijedi $n \geq p$. Dekompozicija singularnih vrijednosti nam omogućava sljedeći rastav matrice $\boldsymbol{X}$:

$$\boldsymbol{X} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^T,$$

gdje su $\widetilde{\boldsymbol{U}}_{n\times n}$ i $\widetilde{\boldsymbol{V}}_{p\times p}$ ortogonalne matrice i $\widetilde{\boldsymbol{\Sigma}}$ je $n \times p$ matrica sa singularnim vrijednostima σ_j za $j = 1, \ldots, s$ koje su sortirane padajuće na glavnoj dijagonali. Pretpostavimo da je rang matrice $\mathbf{X}$ je jednak s i $s \leq p$, stoga:

$$\widetilde{\boldsymbol{\Sigma}} = \begin{matrix} \\ s \\ n-s \end{matrix} \begin{matrix} \begin{matrix} s & p-s \end{matrix} \\ \begin{pmatrix} \boldsymbol{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \end{matrix}, \tag{1.1.1}$$

tako da je $\boldsymbol{\Sigma}$ dijagonalna $s \times s$ matrica sa singularnim vrijednostima na dijagonali. $\widetilde{\boldsymbol{\Sigma}}$ je dijagonalna blok-matrica što vidimo iz (1.1.1). Matrica $\widetilde{\boldsymbol{U}}$ se sastoji od n ortogonalnih svojstvenih vektora $\boldsymbol{X}\boldsymbol{X}^T$ koje još nazivamo lijevim singularnim vektorima. Stupce matrice $\widetilde{\boldsymbol{V}}$ čine p ortogonalnih svojstvenih vektora matrice $\boldsymbol{X}^T\boldsymbol{X}$ i njih nazivamo desnim

singularnim vektorima. Napomenimo još kako su obje matrice ortonormalne. Pokaže se kako matrice $\boldsymbol{U}_{n\times s}$ odnosno $\boldsymbol{V}_{p\times s}$ sastavljenje od prvih s stupaca matrice $\widetilde{\boldsymbol{U}}$ odnosno $\widetilde{\boldsymbol{V}}$ omogućuju SVD matrice $\boldsymbol{X}$:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{1.1.2}$$

Primjetimo kako su matrice $\boldsymbol{U}$ i $\boldsymbol{V}$ ortonormalne i kako dekompoziciju matrice $\boldsymbol{X}$ možemo zapisati ovako:

$$x_{ij} = \sum_{t=1}^{s} \sigma_t u_{it} v_{jt}, \tag{1.1.3}$$

što implicira da je $3 \times s$ elemenata dovoljno za prikaz jednog elemenata matrice $\boldsymbol{X}$. Ovaj zapis će nam koristiti kasnije kada budemo konstruirali biplote više-dimenzionalnih matrica.

Teorem 1.1.1. *(Eckart-Young) Neka je $n \times p$ matrica $\boldsymbol{X}$ ranga s. $\boldsymbol{X}$ se može aproksimirati $n \times p$ matricom $\hat{\boldsymbol{X}}_{[r]}$ ranga r tako da vrijedi $r \leq s$. Aproksimacija se temelji na minimizaciji sljedeće Frobeniusove norme:*

$$\left\|\mathbf{X} - \hat{\boldsymbol{X}}_{[r]}\right\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \hat{x}_{ij})^2} = \left(tr\left(\boldsymbol{X} - \hat{\boldsymbol{X}}_{[r]}\right)^T\left(\boldsymbol{X} - \hat{\boldsymbol{X}}_{[r]}\right)\right)^{1/2}$$

uz uvjet da je rang $(\hat{\boldsymbol{X}}_{[r]}) = r$. SVD matrice $\boldsymbol{X}$ nam nudi rješenje:

$$\hat{\boldsymbol{X}}_{[r]} = \boldsymbol{U}\boldsymbol{\Sigma}_{[r]}\boldsymbol{V}^T,$$

gdje je $\boldsymbol{\Sigma}_{[r]}$ matrica koja na mjestu $s-r$ najmanjih singularnih vrijednosti matrice $\boldsymbol{\Sigma}$ ima 0*.*

Dokaz. Trebamo minimizirati $\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}_{[r]}\right\|_F$ uz uvjet da je rang $(\hat{\boldsymbol{X}}_{[r]}) = r$. Neka je singularna dekompozicija matrice $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ matrice $\boldsymbol{X}$ takva da je $\boldsymbol{\Sigma} = \boldsymbol{U}^T\boldsymbol{X}\boldsymbol{V}$. Frobeniusova norma je unitarno invarijatna, odnosno za svaku matricu $\boldsymbol{A}$ vrijedi $\|\boldsymbol{A}\|_F = \|\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}\|_F$ gdje su $\boldsymbol{U}$ i $\boldsymbol{V}$ unitarne matrice. Prisjetimo se, za realnu matricu $\boldsymbol{U}$ kažemo da je unitarna ako vrijedi $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{I}$. Zbog svojstva unitarne invarijatnosti Frobeniusove norme slijedi da je minimizacija $\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}_{[r]}\right\|_F$ ekvivalenta minimizaciji $\left\|\boldsymbol{\Sigma} - \boldsymbol{U}^T\hat{\boldsymbol{X}}_{[r]}\boldsymbol{V}\right\|_F$.

Kako je matrica $\boldsymbol{\Sigma}$ $s \times s$ dijagonalna matrica tada je matrica $\boldsymbol{U}^T\hat{\boldsymbol{X}}_{[r]}\boldsymbol{V}$ također dijagonalna. Definirajmo dijagonalnu matricu $\boldsymbol{S} = diag(s_i)$ za $i = 1, \ldots, s$ takvu da je $\boldsymbol{U}^T\hat{\boldsymbol{X}}_{[r]}\boldsymbol{V} = \boldsymbol{S}$ i $\hat{\boldsymbol{X}}_{[r]} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$. Problem se svodi na traženje sljedećeg minimuma :

$$\min_{s_i} \|\boldsymbol{\Sigma} - \boldsymbol{S}\|_F = \min_{s_i}\left(\sum_{i=1}^{s}(\sigma_i - s_i)^2\right)^{1/2}$$

S obzirom na ograničenje ranga, minimum gornjeg izraza se postiže kada je $\sigma_i = s_i$ za $i = 1, \ldots, r$ i odgovarajući singularni vektori su jednaki onima iz dekompozicije matrice $\boldsymbol{X}$. □

Korolar 1.1.2. *Nek je $c \in \Re^p$ i $\boldsymbol{X}$ je $n \times p$ matrica. Tada je izraz*

$$\left\|\boldsymbol{X} - \mathbf{1}\boldsymbol{c}^T\right\|^2 = \left\|\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\boldsymbol{X}\right\|^2 + n\left\|\frac{1}{n}\mathbf{1}\boldsymbol{X}^T - \boldsymbol{c}^T\right\|^2$$

minimalan za $\boldsymbol{c}^T = \frac{1}{n}\mathbf{1}^T\boldsymbol{X}$. Odnosno, suma kvadrata centroida je minimalna.

Dokaz. Prvo promotrimo lijevu stranu jednadžbe

$$\begin{aligned}
\left\|\boldsymbol{X} - \mathbf{l}\boldsymbol{c}^T\right\|^2 &= tr\left[(\boldsymbol{X} - \mathbf{l}\boldsymbol{c}^T)(x - \mathbf{l}\boldsymbol{c}^T)^T\right] \\
&= tr\left[\boldsymbol{X}\boldsymbol{X}^T - \boldsymbol{X}\boldsymbol{c}\mathbf{l}^T - \mathbf{l}\boldsymbol{c}^T\boldsymbol{X}^T + \mathbf{l}\boldsymbol{c}^T\boldsymbol{c}\mathbf{l}^T\right] \\
&= tr\left[\boldsymbol{X}\boldsymbol{X}^T - \mathbf{l}^T\boldsymbol{X}\boldsymbol{c} - \boldsymbol{c}^T\boldsymbol{X}^T\mathbf{l} + \mathbf{l}^T\mathbf{l}\boldsymbol{c}^T\boldsymbol{c}\right] \\
&= tr\left[\boldsymbol{X}\boldsymbol{X}^T - \mathbf{l}^T\boldsymbol{X}\boldsymbol{c} - \boldsymbol{c}^T\boldsymbol{X}^T\mathbf{l} + n\boldsymbol{c}^T\boldsymbol{c}\right]
\end{aligned}$$

Sada promotrimo desnu stranu

$$\begin{aligned}
&\left\|\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\boldsymbol{X}\right\|^2 + n\left\|\frac{1}{n}\mathbf{1}\boldsymbol{X}^T - \boldsymbol{c}^T\right\|^2 \\
&= tr\left[(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\boldsymbol{X})(\boldsymbol{X} - \frac{1}{n}\mathbf{1}\mathbf{1}^T\boldsymbol{X})^T + n(\frac{1}{n}\mathbf{1}\boldsymbol{X}^T - \boldsymbol{c}^T)(\frac{1}{n}\mathbf{1}\boldsymbol{X}^T - \boldsymbol{c}^T)^T\right] \\
&= tr\left[\boldsymbol{X}\boldsymbol{X}^T - \frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T\mathbf{1}\mathbf{1}^T - \frac{1}{n}\mathbf{1}\mathbf{1}^T\boldsymbol{X}\boldsymbol{X}^T + \frac{1}{n^2}\mathbf{1}\mathbf{1}^T\boldsymbol{X}\boldsymbol{X}^T\mathbf{1}\mathbf{1}^T\right. \\
&\quad \left. + \frac{1}{n}\mathbf{1}^T\boldsymbol{X}\boldsymbol{X}^T\mathbf{1} - \boldsymbol{c}^T\boldsymbol{X}^T\mathbf{1} - \mathbf{1}^T\boldsymbol{X}\boldsymbol{c} + n\boldsymbol{c}^T\boldsymbol{c}\right] \\
&= tr\left[\boldsymbol{X}\boldsymbol{X}^T - \frac{2}{n}\mathbf{1}^T\boldsymbol{X}\boldsymbol{X}^T\mathbf{1} + \frac{2}{n}\mathbf{1}^T\boldsymbol{X}\boldsymbol{X}^T\mathbf{1} - c^T\boldsymbol{X}^T\mathbf{1} - \mathbf{1}^T\boldsymbol{X}\boldsymbol{c} + n\boldsymbol{c}^T\boldsymbol{c}\right] \\
&= tr\left[\boldsymbol{X}\boldsymbol{X}^T - \mathbf{l}^T\boldsymbol{X}\boldsymbol{c} - \boldsymbol{c}^T\boldsymbol{X}^T\mathbf{l} + n\boldsymbol{c}^T\boldsymbol{c}\right].
\end{aligned}$$

Iz ovog vidimo da su lijeva i desna strana jednake i dokazujemo jednakost. Također, iz desne strane jednadžbe se vidi da za $\boldsymbol{c}^T = \frac{1}{n}\mathbf{l}^T\boldsymbol{X}$ vrijedi $n\left\|\frac{1}{n}\mathbf{l}\boldsymbol{X}^T - \boldsymbol{c}^T\right\|^2 = 0$. Takav $\boldsymbol{c}$ ujedno minimizira desnu stranu jednadžbe te implicira da se minimum postiže u $\boldsymbol{c}^T = \frac{1}{n}\mathbf{l}^T\boldsymbol{X}$ što nazivamo centroidom matrice $\boldsymbol{X}$. □

Teorem 1.1.3. *(Faktorizacija matrice) Svaka $n \times p$ matrica $\boldsymbol{X}$ ranga q se može rastaviti*

$$\boldsymbol{X} = \boldsymbol{G}\boldsymbol{H}^T$$

na $n \times q$ matricu $\boldsymbol{G}$ i $p \times q$ matricu $\boldsymbol{H}$, koje su nužno ranga q. Ova faktorizacija nije jedinstvena.

Dokaz. Neka je $n \times p$ matrica $\boldsymbol{X}$ ranga q i bez smanjenja općenitosti možemo pretpostaviti da je $p \leq n$. Neka je SVD matrice X dan s $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$. Slično kao u dokazu Eckhart-Young teorema, neka je reducirani SVD matrice $\boldsymbol{X}$ dan s $\boldsymbol{X} = \boldsymbol{U}_q\boldsymbol{D}_q\boldsymbol{V}_q^T$. Rastavimo li matricu $\boldsymbol{D}_q$ na $\boldsymbol{D}_q^\alpha$ i $\boldsymbol{D}_q^{1-\alpha}$, gdje je $0 \leq \alpha \leq 1$, matrica $\boldsymbol{X}$ se može prikazati kao produkt dviju matrica ranga q :

$$\begin{aligned}
\boldsymbol{X} &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \\
&= \boldsymbol{U}_q\boldsymbol{D}_q\boldsymbol{V}_q^T \\
&= \boldsymbol{U}_q\boldsymbol{D}_q^\alpha\boldsymbol{D}_q^{1-\alpha}\boldsymbol{V}_q^T \\
\longrightarrow \boldsymbol{X} &= \boldsymbol{G}\boldsymbol{H} \text{ gdje je } \boldsymbol{G} = \boldsymbol{U}_q\boldsymbol{D}_q^\alpha \text{ and } \boldsymbol{H} = \boldsymbol{D}_q^{1-\alpha}\boldsymbol{V}_q^T.
\end{aligned}$$

S obzirom da su stupci matrice $\boldsymbol{G} = \boldsymbol{U}_q\boldsymbol{D}_q^\alpha$ zapravo reskalirani stupci matrice $\boldsymbol{U}_q$, za koju znamo da je ortogonalna, rang matrice $\boldsymbol{G} = \boldsymbol{U}_q\boldsymbol{D}_q^\alpha$ je jednak q. Analogno, s obzirom da su retci matrice $\boldsymbol{H} = \boldsymbol{D}_q^{1-\alpha}\boldsymbol{V}_q^T$ reskalirani retci matrice $\boldsymbol{V}_q^T$, za koje znamo da su ortogonalni, rang matrice $\boldsymbol{H} = \boldsymbol{D}_q^{1-\alpha}\boldsymbol{V}_q^T$ je jednak q. □

Ovaj korolar pokazuje da se svaka matrica ranga q može prikazati kao skalarni produkt dviju matrica ranga q. Geometrijski, ovo znači da se svaka $n \times p$ matrica ranga q može savršeno prikazati s $n + p$ vektora u q-dimenzionalnom prostoru, odnosno svaki se element matrice može dobiti iz q-dimenzionalnog prikaza tih $n + p$ vektora.

1.2 Konstrukcija biplota

Najprije demonstrirajmo konstrukciju jednostavnog biplota. U sljedećim poglavljima ćemo prikazati korištenje prethodno navedenih alata iz linearne algebre. Za početak definirajmo skalarni produkt:

Definicija 1.2.1. *Za svaka dva vektora $\boldsymbol{x}^{\mathrm{T}} = [x_1, x_2, \ldots, x_m] \in \mathbb{R}^{m\times 1}$ i $\boldsymbol{y}^{\mathrm{T}} = [y_1, y_2, \ldots, y_m] \in \mathbb{R}^{m\times 1}$, definiramo skalarni produkt $\boldsymbol{x}$ i $\boldsymbol{y}$ kao:*

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = x_1y_1 + x_2y_2 + \cdots + x_my_m \tag{1.2.1}$$

Kako bi pokazali biplot, trebamo početnu matricu $\boldsymbol{X}$ izraziti pomoću skalarnog produkta dviju matrica te prikazati njihove točke na zajedničkom scatterplotu.

$$\boldsymbol{X} = \boldsymbol{G}\boldsymbol{H}^{\mathrm{T}} = \begin{pmatrix} \boldsymbol{g}_1^{\mathrm{T}} \\ \boldsymbol{g}_2^{\mathrm{T}} \\ \boldsymbol{g}_3^{\mathrm{T}} \\ \boldsymbol{g}_4^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \boldsymbol{h}_1 & \boldsymbol{h}_2 & \boldsymbol{h}_3 & \boldsymbol{h}_4 \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{h}_1 & \boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{h}_2 & \boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{h}_3 & \boldsymbol{g}_1^{\mathrm{T}}\boldsymbol{h}_4 \\ \boldsymbol{g}_2^{\mathrm{T}}\boldsymbol{h}_1 & \boldsymbol{g}_2^{\mathrm{T}}\boldsymbol{h}_2 & \boldsymbol{g}_2^{\mathrm{T}}\boldsymbol{h}_3 & \boldsymbol{g}_2^{\mathrm{T}}\boldsymbol{h}_4 \\ \boldsymbol{g}_3^{\mathrm{T}}\boldsymbol{h}_1 & \boldsymbol{g}_3^{\mathrm{T}}\boldsymbol{h}_2 & \boldsymbol{g}_3^{\mathrm{T}}\boldsymbol{h}_3 & \boldsymbol{g}_3^{\mathrm{T}}\boldsymbol{h}_4 \\ \boldsymbol{g}_4^{\mathrm{T}}\boldsymbol{h}_1 & \boldsymbol{g}_4^{\mathrm{T}}\boldsymbol{h}_2 & \boldsymbol{g}_4^{\mathrm{T}}\boldsymbol{h}_3 & \boldsymbol{g}_4^{\mathrm{T}}\boldsymbol{h}_4 \end{pmatrix} \tag{1.2.2}$$

Tu smo matricu $\boldsymbol{X}$ prikazali kao skalarni produkt matrica $\boldsymbol{G}$ i $\boldsymbol{H}$. Na primjer, prvi element matrice $\boldsymbol{X}$ je jednak skalarnom produktu točaka prvih redaka matrica $\boldsymbol{G}$ i $\boldsymbol{H}$.

Koordinate točaka smo dobili rastavom početne matrice na dvije matrice. Jednu matricu, odnosno skup točaka, izaberemo za osi biplota i taj skup nazivamo biplot vektori dok drugu matricu, odnosno drugi skup točaka, nazivamo biplot točke. Matricu X nazivamo matricom cilja (eng. Target matrix), matricu G lijevom matricom i matricu H desnom. Konačno, možemo iz skalarnog produkta biplot točke i biplot vektora rekonstruirati aproksimaciju podataka iz početne matrice. Geometrijski, biplot točke se projiciraju na biplot vektore i njihove projekcije se pomnože s dužinom biplot vektora i taj umnožak je jednak odgovarajućoj aproksimativnoj vrijednosti početne matrice.

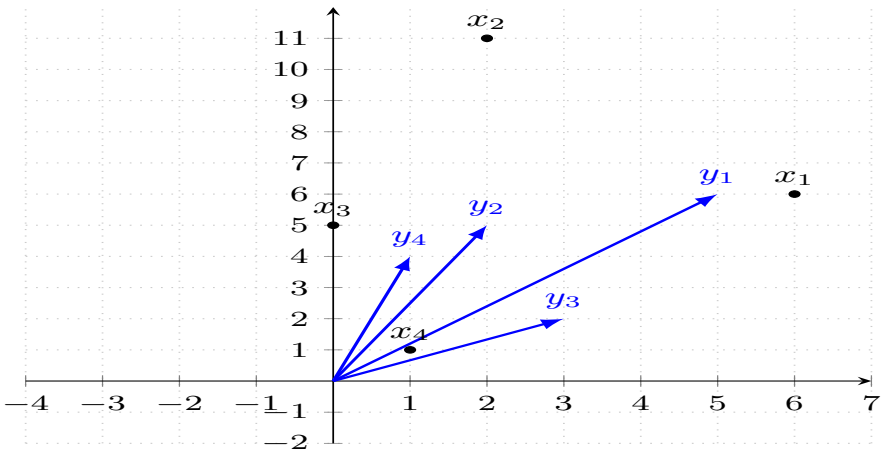
Primjer 1.2.1. Kako bi prikazali biplot na jednostavnom primjeru, neka je matrica $X \in \mathbb{R}^{4\times4}$ te matrice $L \in \mathbb{R}^{4\times2}$ i $R \in \mathbb{R}^{2\times4}$.

$$\boldsymbol{X} = \begin{pmatrix} 66 & 76 & 30 & 11 \\ 42 & 59 & 25 & 7 \\ 18 & 6 & 0 & 3 \\ 30 & 46 & 20 & 5 \end{pmatrix} \quad \boldsymbol{L} = \begin{pmatrix} 5 & 6 \\ 2 & 5 \\ 3 & 0 \\ 1 & 4 \end{pmatrix} \quad \boldsymbol{R} = \begin{pmatrix} 6 & 2 & 0 & 1 \\ 6 & 11 & 5 & 1 \end{pmatrix}$$

Sve matrice su ranga 2 i zadovoljavaju sljedeću jednakost:

$$\begin{pmatrix} 66 & 76 & 30 & 11 \\ 42 & 59 & 25 & 7 \\ 18 & 6 & 0 & 3 \\ 30 & 46 & 20 & 5 \end{pmatrix} = \begin{pmatrix} 5 & 6 \\ 2 & 5 \\ 3 & 0 \\ 1 & 4 \end{pmatrix} \times \begin{pmatrix} 6 & 2 & 0 & 1 \\ 6 & 11 & 5 & 1 \end{pmatrix}.$$

Sukladno s biplot oznakama, matrica X je matrica cilja, matrica L lijeva a matrica R desna matrica. Nacrtamo biplot točke i vektore i dobijemo sljedeći dijagram razmještaja:



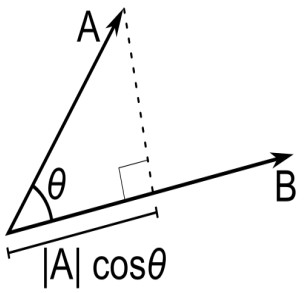
Slika 1.1: Biplot

Biplot se u svojoj osnovi temelji na skalarnom produktu, stoga nam razumijevanje geometrijske interpretacije skalarnog produkta omogućava novu percepciju biplota. Odnosno, geometrijski možemo definirati ovako:

Definicija 1.2.2. *Skalarni produkt 2 vektora je jednak duljini projekcije prvog vektora na drugi vektor pomnoženoj s duljinom drugog vektora, odnosno:*

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = \|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|\cos(\theta) \tag{1.2.3}$$

gdje je θ kut između vektora $\boldsymbol{x}$ i $\boldsymbol{y}$.



Slika 1.2: Skalarni produkt

Sada na primjer 1.2.1 pokažimo geometrijsku interpretaciju skalarnog produkta. Na primjer, skalarni produkt vektora $\boldsymbol{x}_2^T$ i $\boldsymbol{y}_1$ je $2\times5+6\times6=42$, što je jednako elementu $x_{2,1}$ matrice $\boldsymbol{X}$. Kut θ između vektora $\boldsymbol{x}_2$ i $\boldsymbol{y}_1$ iznosi 23.2° pa je $\cos(\theta)=0.9191$ a duljine vektora $\boldsymbol{x}_2$ i $\boldsymbol{y}_1$ su $\|\boldsymbol{x}_2\|=5.3852$ odnosno $\|\boldsymbol{y}_1\|=8.4853$. Prema definiciji skalarnog produkta, umnožak duljina vektora te kosinusa kuta između njih je jednak njihovom skalarnom produktu:

$$\|\boldsymbol{x}_2\|\cdot\|\boldsymbol{y}_1\|\cos(\theta)=41,9984\approx42=\boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{y}_1.$$

Zbog činjenice da je rang matrice X jednak 2 možemo savršeno rekonstruirati matricu X iz biplota. Time smo ujedno geometrijski demonstrirali činjenicu da rang predstavlja dimenzionalnost prikaza.

1.3 Regresijski biplot

Vidjeli smo kako se biplot temelji na dekompoziciji početne matrice na umnožak dviju matrica. Situacija u kojoj ćemo se prirodno susresti s takvom dekompozicijom matrice je u regresijskoj analizi. Osnovna regresijska jednadžba je $\hat{\boldsymbol{Y}}=\boldsymbol{XB}$ gdje je $\boldsymbol{X}$ matrica eksplanatornih varijabli, $\boldsymbol{B}$ matrica procijenjenih regresijskih koeficijenata a $\hat{\boldsymbol{Y}}$ matrica procijenjene varijable odaziva. Dolazimo do ideje regresijskog biplota u kojem bi $\hat{\boldsymbol{Y}}$ predstavljala matricu cilja dok bi $\boldsymbol{X}$ i $\boldsymbol{B}$ predstavljale lijevu odnosno desnu matricu. Regresijski biplot će nam poslužiti kao uvod u aproksimaciju višedimenzionalnih podataka.

Kada radimo linearnu regresiju koju želimo prikazati biplotom, korisno je standardizirati podatke. Standardizacija podataka je postupak kojim od svake varijable oduzmemo njezino očekivanje i tu razliku podijelimo s standardnom devijacijom varijable. Tako se rješavamo konstantog člana u regresiji, jer je tada očekivanje svih varijabli 0. Također, eksplanatorne varijable su usporedive jer nestaje utjecaj reda veličine.

Nakon standardiziranja imamo regresijsku jednadžbu $\hat{\boldsymbol{Y}}=\boldsymbol{XB}$ gdje je $\boldsymbol{X}$ matrica standardiziranih eksplanatornih varijabli punog ranga po stupcima a $\boldsymbol{B}$ matrica standardiziranih regresijskih koeficijenata. Matrica cilja $\hat{\boldsymbol{Y}}$ je predikcija opažanih vrijednosti $\boldsymbol{Y}$. Matrica regresijskih koeficijenata $\boldsymbol{B}^{\mathrm{T}}$ je izračunata sljedećom formulom:

$$\boldsymbol{B}^{\mathrm{T}}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}. \tag{1.3.1}$$

Cijeli postupak regresijskog biplota se može tada zapisati kao:

$$\boldsymbol{Y}\approx\hat{\boldsymbol{Y}}=\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}. \tag{1.3.2}$$

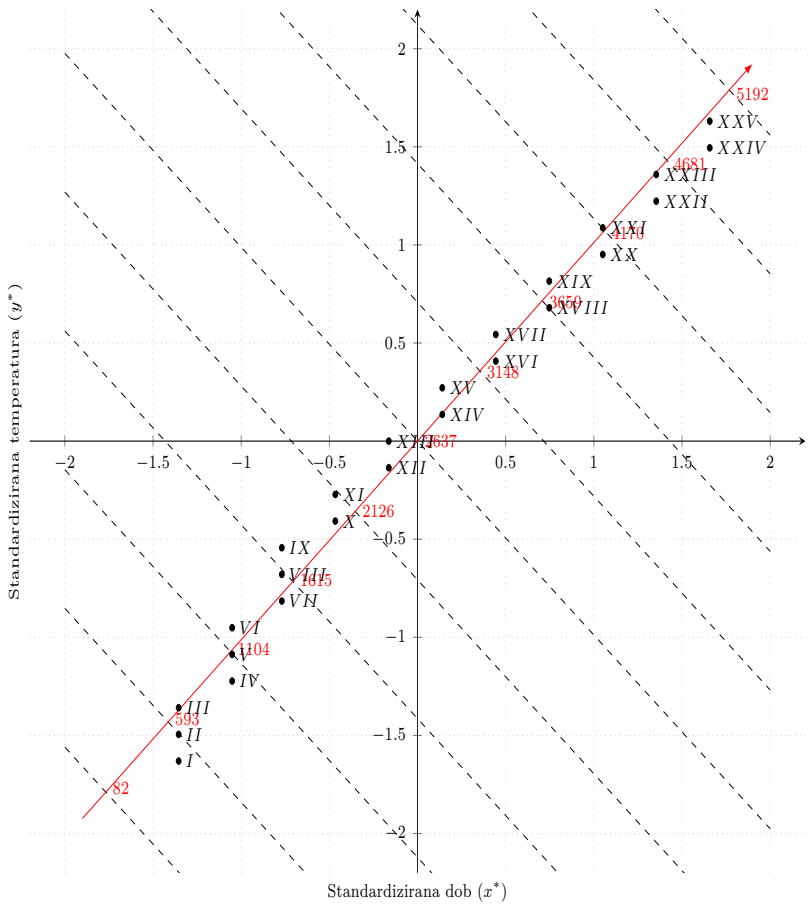
Matrica $\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ se naziva *projekcijskom matricom*. $\boldsymbol{Y}$ je jednak idućem zbroju:

$$\begin{aligned}\boldsymbol{Y} &= \hat{\boldsymbol{Y}} + \left(\boldsymbol{Y} - \hat{\boldsymbol{Y}}\right) \\ \boldsymbol{Y} &= \left(\boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right) + \left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\right)\boldsymbol{Y} \end{aligned} \tag{1.3.3}$$

Prvi pribrojnik u jednakosti je projekcija $\boldsymbol{Y}$ na prostor eksplanatornih varijabli a drugi pribrojnik je projekcija $\boldsymbol{Y}$ na ortogonalni komplement prostoru eksplanatornih varijabli. $\hat{\boldsymbol{Y}}$ možemo prikazati biplotom u skladu s jednadžbom 1.3.2, pri čemu je matrica eksplanatornih varijabli $\boldsymbol{X}$ lijeva matrica a matrica regresijskih koeficijenata $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ desna matrica.

Primjer 1.3.1. U ovom primjeru ćemo prikazati konstrukciju regresijskog biplota. Podaci se sastoje od 25 observacija, pri čemu varijabla *a* predstavlja duljinu ribe u milimetrima, varijabla *x* starost ribe u danima te varijabla *y* temperaturu vode u stupnjevima Celzijevim u spremniku u kojem je živjela promatrana riba. Pretpostavimo kako duljina pojedine ribe ovisi o temperaturi vode te dobi ribe i napravimo linearnu regresiju gdje su eksplanatorne varijable temperatura vode i dob ribe a duljina ribe varijabla odaziva. Na idućoj stranici prikazujemo početne podatke, standardizirane podatke i regresijski biplot .

a	x	y
550	14	15
670	14	16
810	14	17
940	28	18
1,090	28	19
1,240	28	20
1,385	41	21
1,530	41	22
1,675	41	23
1,890	55	24
2,105	55	25
2,320	69	26
2,535	69	27
2,750	83	28
2,965	83	29
3,180	97	30
3,395	97	31
3,610	111	32
3,825	111	33
4,040	125	34
4,255	125	35
4,470	139	36
4,685	139	37
4,900	153	38
5,115	153	39



x^*	y^*
−1.354653	−1.630479
−1.354653	−1.494606
−1.354653	−1.358732
−1.051308	−1.222859
−1.051308	−1.086986
−1.051308	−0.951113
−0.769630	−0.815240
−0.769630	−0.679366
−0.769630	−0.543493
−0.466285	−0.407620
−0.466285	−0.271747
−0.162940	−0.135873
−0.162940	0.000000
0.140406	0.135873
0.140406	0.271747
0.443751	0.407620
0.443751	0.543493
0.747096	0.679366
0.747096	0.815240
1.050441	0.951113
1.050441	1.086986
1.353787	1.222859
1.353787	1.358732
1.657132	1.494606
1.657132	1.630479

Regresijska veza između varijabli je :

$$\left[\hat{\boldsymbol{a}}^*\right] = \begin{bmatrix}\hat{\boldsymbol{x}}^* & \hat{\boldsymbol{y}}^*\end{bmatrix}\begin{bmatrix}0.497\\0.503\end{bmatrix}. \tag{1.3.4}$$

Kako bi znali isčitati procijenjene vrijednosti iz grafa, trebamo preračunati udaljenosti. Što se tiče vrijednosti u ishodištu regresijskog biplota, ona je jednaka očekivanju originalne varijable odaziva što iznosi približno 2637 u našem slučaju. Općenito, udaljenost na biplotu se preračunava:

$$\text{Jedinica varijable odaziva } = 1/\left(\begin{array}{c}\text{standardna devijacija}\\ \text{varijable odaziva}\end{array} \times \begin{array}{c}\text{duljina biplot}\\ \text{vektora}\end{array}\right) \tag{1.3.5}$$

Odnosno, u našem slučaju:

$$\begin{aligned}\text{Jedinica varijable odaziva } &= 1/\left(\begin{array}{c}\text{standardna devijacija}\\ \text{varijable odaziva}\end{array} \times \begin{array}{c}\text{duljina biplot}\\ \text{vektora}\end{array}\right)\\ &= 1/\left(1446.368 \times \sqrt{(0.497)^2 + (0,503^2}\right)\\ &= 1/\left(1446.368 \times \sqrt{0.247 + 0.253}\right)\\ &= 1/(1446.368 \times 0.707568421)\\ &= 0.001\end{aligned}$$

Dakle, 1 milimetar udaljenosti na biplotu predstavlja otprilike 1 milimetar duljine ribe. Sada, ako se prisjetimo geometrijske interpretacije skalarnog produkta iz prethodnog poglavlja te formule 1.3.2, zaključujemo kako je projekcija eksplanatorne vrijednosti na pravac određen koeficijentima regresije jednaka procijenjenoj vrijednosti varijable odaziva. Kako bi olakšali tu vizualnu inspekciju biplota, na biplotu nacrtamo okomite isprekidane linije koje nam služe kao oznake nivoa procijenjenih vrijednosti.

1.4 Biplot reduciranog ranga

U stvarnom životu matrice s podacima su često višedimenzionalne i ne mogu se savršeno prikazati dvodimenzionalnim ili trodimenzionalnim biplotom. Ideja biplota je pronaći točke retka x_i i točke stupca y_j takve da njihov skalarni produkt aproksimira odgovarajuće elemente početne matrice. Na temelju rezultata iz poglavlja 1.2 možemo konstruirati biplot za višedimenzionalne matrice. Prepostavimo da su podaci koje promatramo zapisani u matrici $X \in \mathbb{R}^{n\times p}$ koja je ranga s i podatke želimo prikazati u dvodimenzionalnom prostoru uz minimalni gubitak informacija. To ćemo postići tako što ćemo aproksimirati matricu X s matricom ranga 2 čiju egzistenciju jamči Eckart-Youngov teorem. Konačno, možemo iskoristiti teorem 1.2.3 nad matricom $\hat{X}_{[2]}$ i dobiti matrice $\boldsymbol{G}$ i $\boldsymbol{H}$ koje su ranga 2 i za koje vrijedi $\hat{\boldsymbol{X}}_{[2]} = \boldsymbol{G}\boldsymbol{H}^T$. Pokažimo na sljedećem primjeru biplot reduciranog ranga.

Primjer 1.4.1. U ovom primjeru koristimo podatke iz istraživanja mišljenja studentske populacije na fakultetima iz SAD-a koju je sproveo "PEW Research Centre" 2011.godine. Ispitana su studentska mišljena o različitim izjavama te su zabilježeni odgovori na ljestvici 1-5, pri čemu 1 označava izrazito slaganje s izjavom a 5 izrazito neslaganje. Slijedi tablica s kraticama tih izjava:

Tablica 1.1: Tablica s kraticama izjava

St	Posjedovanje stambene nekretnine mi je bitno.
Ug	Ugodna mirovina mi je bitna.
Os	Ostavština za djecu mi je bitna.
F	Fakultet mi je bitan za životni uspjeh.
Dr	Društvene vještine su bitne za životni uspjeh.
Ra	Radna etika je bitna za životni uspjeh.
Vj	Vještine naučene na poslu su bitne za životni uspjeh.
Pr	Fakultet me je pripremio za karijeru.
Ob	Fakultet me oblikovao kao osobu.
Zn	Fakultet mi je proširio znanje i pomogao u intelektualnom razvoju.
Zad	Zadovoljan sam izborom karijere.

U idućoj tablici možemo pogledati rezultate ankete. Vizualnom inspekcijom tablice je teško izvesti zaključke.

Tablica 1.2: Rezultati ankete

	I	II	III	IV	V
St	35	42	12	7	3
Ug	34	45	16	4	1
Os	20	30	28	16	5
F	42	35	18	4	1
Dr	57	36	6	1	0
Ra	61	35	3	0	1
Vj	43	46	9	1	0
Pr	55	32	9	4	1
Ob	67	24	5	3	1
Zn	72	23	3	1	1
Zad	46	26	18	9	1

Aproksimirajmo ovu tablicu s matricom ranga 2 i prikažimo rezultate ankete na biplotu. Prvo zapišimo ovu tablicu matrično.

$$\boldsymbol{X} = \begin{bmatrix} 35 & 42 & 12 & 7 & 3 \\ 34 & 45 & 16 & 4 & 1 \\ 20 & 30 & 28 & 16 & 5 \\ 42 & 35 & 18 & 4 & 1 \\ 57 & 36 & 6 & 1 & 0 \\ 61 & 35 & 3 & 0 & 1 \\ 43 & 46 & 9 & 1 & 0 \\ 55 & 32 & 9 & 4 & 1 \\ 67 & 24 & 5 & 3 & 1 \\ 72 & 23 & 3 & 1 & 1 \\ 46 & 26 & 18 & 9 & 1 \end{bmatrix} \tag{1.4.1}$$

Sada iskoristimo prethodno dokazane rezultate linearne algebre iz poglavlja 1.2 i napravimo singularnu dekompoziciju matrice $\boldsymbol{X} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}'$.

$$\widetilde{\boldsymbol{\Sigma}} = \begin{bmatrix} 202.599702 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 49.480497 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 0.000000 & 23.741840 & 0.000000 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 5.677762 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 0.000000 & 1.769003 \\ 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \\ 0.000000 & 0.000000 & 0.000000 & 0.000000 & 0.000000 \end{bmatrix}$$

$$\widetilde{U} = \begin{bmatrix} -0.266999 & 0.305240 & -0.211691 & -0.596466 & 0.049662 & -0.439369 & -0.243732 & -0.184125 & -0.194181 & -0.241011 & 0.224374 \\ -0.273259 & 0.378719 & -0.252050 & 0.252859 & 0.040325 & 0.049208 & -0.427354 & 0.102011 & 0.430566 & 0.484207 & 0.193664 \\ -0.190483 & 0.531836 & 0.565741 & -0.282824 & 0.241623 & 0.236809 & 0.229497 & 0.008309 & 0.049459 & 0.150373 & -0.297587 \\ -0.280254 & 0.187638 & 0.103505 & 0.642836 & 0.414089 & -0.240705 & -0.016057 & -0.048051 & -0.155004 & -0.453593 & -0.023463 \\ -0.332431 & -0.104613 & -0.183001 & 0.045038 & -0.242932 & 0.128265 & -0.320181 & -0.290908 & -0.222988 & 0.050292 & -0.726749 \\ -0.343155 & -0.195610 & -0.222298 & -0.147193 & 0.306407 & 0.735300 & -0.012330 & -0.019080 & -0.090614 & -0.220021 & 0.284976 \\ -0.305395 & 0.204256 & -0.453896 & 0.089533 & -0.249476 & -0.061590 & 0.758744 & -0.055363 & 0.032662 & 0.070033 & -0.014632 \\ -0.317117 & -0.085809 & 0.039901 & -0.078242 & -0.182291 & -0.081793 & -0.059372 & 0.894506 & -0.132182 & -0.106170 & -0.109680 \\ -0.340324 & -0.363129 & 0.200066 & -0.130308 & -0.051204 & -0.162248 & 0.048306 & -0.130135 & 0.749929 & -0.279632 & -0.080672 \\ -0.355510 & -0.460563 & 0.171360 & -0.021308 & 0.325556 & -0.286396 & 0.124104 & -0.086816 & -0.267744 & 0.576936 & 0.127225 \\ -0.273701 & 0.055428 & 0.448298 & 0.178315 & -0.639592 & 0.120696 & -0.073587 & -0.201605 & -0.199701 & -0.030559 & 0.422762 \end{bmatrix}$$

$$\widetilde{\boldsymbol{V}} = \begin{bmatrix} -0.817668 & -0.539499 & 0.199353 & 0.023978 & 0.006626 \\ -0.546243 & 0.617525 & -0.556427 & -0.097809 & -0.033243 \\ -0.169179 & 0.518248 & 0.641217 & 0.522139 & 0.137874 \\ -0.063558 & 0.234775 & 0.476155 & -0.722727 & -0.437932 \\ -0.019431 & 0.062481 & 0.112983 & -0.441470 & 0.887726 \end{bmatrix}$$

Nadalje, izvršimo redukciju prethodnih matrica na matrice ranga 2 tako što ćemo izdvojiti prva dva stupca iz svake matrice, poštujući pritom dimenzije matrica tako da vrijedi $\boldsymbol{X}_{[2]} = \boldsymbol{U}_{[2]}\boldsymbol{D}_{[2]}\boldsymbol{V}_{[2]}'$.

$$\boldsymbol{U}_{[2]} = \begin{bmatrix} -0.266999 & 0.305240 \\ -0.273259 & 0.378719 \\ -0.190483 & 0.531836 \\ -0.280254 & 0.187638 \\ -0.332431 & -0.104613 \\ -0.343155 & -0.195610 \\ -0.305395 & 0.204256 \\ -0.317117 & -0.085809 \\ -0.340324 & -0.363129 \\ -0.355510 & -0.460563 \\ -0.273701 & 0.055428 \end{bmatrix} \tag{1.4.2}$$

$$\boldsymbol{V}_{[2]} = \begin{bmatrix} -0.8176676 & -0.5394990 \\ -0.5462426 & 0.6175246 \\ -0.1691791 & 0.5182479 \\ -0.0635581 & 0.2347752 \\ -0.0194311 & 0.0624807 \end{bmatrix} \tag{1.4.3}$$

$$\boldsymbol{\Sigma}_{[2]} = \begin{bmatrix} 202.599702 & 0.000000 \\ 0.000000 & 49.480497 \end{bmatrix} \tag{1.4.4}$$

Konačno, dobijamo aproksimiranu matricu $\boldsymbol{X}_{[2]}$:

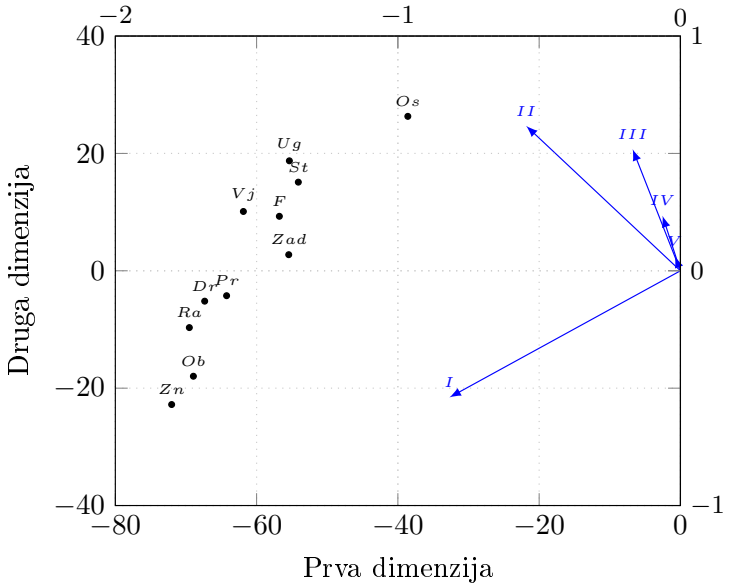
$$\boldsymbol{X}_{[2]} = \begin{bmatrix} 36.082556 & 38.875119 & 16.978871 & 6.984011 & 1.994774 \\ 35.158056 & 41.813068 & 19.077665 & 7.918208 & 2.246583 \\ 17.358016 & 37.330917 & 20.166872 & 8.631044 & 2.394089 \\ 41.417739 & 36.748707 & 14.417537 & 5.788548 & 1.683383 \\ 57.862861 & 33.593174 & 8.711680 & 3.065398 & 0.985272 \\ 62.068586 & 31.999588 & 6.745825 & 2.146405 & 0.746169 \\ 45.139031 & 40.038808 & 15.705392 & 6.305329 & 1.833732 \\ 54.823938 & 32.472943 & 8.668978 & 3.086645 & 0.983120 \\ 66.071424 & 26.567621 & 2.353051 & 0.163908 & 0.217124 \\ 71.188038 & 25.271075 & 0.375043 & -0.772415 & -0.024318 \\ 43.861421 & 31.983700 & 10.802618 & 4.168299 & 1.248846 \end{bmatrix} \tag{1.4.5}$$

Vrijedi $\boldsymbol{G} = \boldsymbol{U}_{[2]}\boldsymbol{\Sigma}_{[2]}$ i $\boldsymbol{H} = \boldsymbol{V}_{[2]}^T$ odnosno $\boldsymbol{X}_{[2]} = \boldsymbol{G}\boldsymbol{H}$:

$$\boldsymbol{G} = \begin{bmatrix} -54.093903 & 15.103410 \\ -55.362143 & 18.739188 \\ -38.591729 & 26.315506 \\ -56.779416 & 9.284446 \\ -67.350418 & -5.176297 \\ -69.523181 & -9.678856 \\ -61.873034 & 10.106680 \\ -64.247755 & -4.245847 \\ -68.949551 & -17.967800 \\ -72.026172 & -22.788866 \\ -55.451694 & 2.742604 \end{bmatrix} \tag{1.4.6}$$

$$\boldsymbol{H} = \begin{bmatrix} -0.817668 & -0.546243 & -0.169179 & -0.063558 & -0.019431 \\ -0.539499 & 0.617525 & 0.518248 & 0.234775 & 0.062481 \end{bmatrix} \tag{1.4.7}$$

Pokazali smo primjer aproksimacije matrice $\boldsymbol{X}$ matricom $\boldsymbol{X}_{[2]}$ ranga 2 te smo matricu $\boldsymbol{X}_{[2]}$ prikazali kao produkt dviju matrica ranga 2. Sada možemo konstruirati biplot vektor gdje su stupci matrice $\boldsymbol{G}$ točke a stupci matrice $\boldsymbol{H}$ vektori:



Slika 1.3: Biplot za primjer 1.4.1

Generalizirana analiza glavne komponente

Idući pojam s kojim ćemo se susresti je analiza glavne komponente (eng. Principal Component Analysis) odnosno skraćeno PCA. PCA je multivarijatna metoda redukcije ranga i vjerojatno najpoznatija metoda takve vrste. Za početak objasnimo generaliziranu PCA. Postupak opisan u prethodnom poglavlju može se generalizirati uvođenjem vektora težina. Svi biploti bi se mogli svesti na jedan generalizirani postupak, ovisno o definiciji matrice $\boldsymbol{Y}$ koju aproksimiramo i težinama $\boldsymbol{w}$ i $\boldsymbol{q}$ koje pridružujemo retcima odnosno stupcima. Ovako poopćena metoda generalizira skoro sve tehnike koje ćemo koristi u nastavku ovog rada i nazivamo je generalizirana analiza glavne komponente. Recimo zbog jednostavnosti da se $\boldsymbol{X} \in \mathbb{R}^{n\times m}$ sastoji od n točaka iz m-dimenzionalnog prostora odnosno :

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \boldsymbol{x}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{pmatrix} \tag{1.4.8}$$

Neka je $\boldsymbol{w} \in \mathbb{R}^{n\times 1}$ vektor težina čije su težine pozitivne i u sumi iznose 1. Udaljenosti u m-dimenzionalnom prostoru su definirane težinskom metrikom gdje su pojedine dimenzije pomnožene s težinskim vektorom $\boldsymbol{q} \in \mathbb{R}^{n\times 1}$, na primjer kvadratima udaljenosti između i-tog i i'-tog retka x_i i $x_{i'}$ odnosno: $(x_i - x_{i'})^T \boldsymbol{D}_{[q]}(x_i \text{ - } x_{i'})$. Cilj metode je pronaći aproksimaciju redaka matrice $\boldsymbol{X}$ koji su najbliži originalnim retcima u smislu težinskih udaljenosti

najmanjih kvadrata. Prisjetimo se kako korolar 1.1.2 implicira da će najbolja ravnina, u smislu minimalnog gubitka informacija, prolaziti kroz centroid podataka. To nam sugerira da najprije obavimo postupak *centriranja* matrice, odnosno $\boldsymbol{X}$ ćemo zamijeniti s

$$\begin{gathered}\boldsymbol{Y} = \boldsymbol{X} - \mathbf{1}\boldsymbol{w}^T\boldsymbol{X} = \left(\boldsymbol{I} - \mathbf{1}\boldsymbol{w}^T\right)\boldsymbol{X} \\ \text{gdje je } \boldsymbol{Y} \text{ centroid zbog } \mathbf{1}\boldsymbol{w}^T\boldsymbol{Y} = 0\,.\end{gathered} \tag{1.4.9}$$

Zbog toga što je $\boldsymbol{Y}$ centroid matrica $(\boldsymbol{I} - \mathbf{1}\boldsymbol{w}^T)$ se naziva *centrirajuća matrica*. Kako bi pronašli aproksimacijsku matricu $\hat{\boldsymbol{Y}}$ ranga $p \leq m$, odnosno retke koji su najbliže retcima $\boldsymbol{Y}$ u smislu težinske sume kvadrata udaljenosti, trebamo riješiti problem:

$$\underset{\hat{\boldsymbol{Y}} \text{ ranga } p}{\text{Minimiziraj za}} \sum_{i=1}^{n} w_i(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^{\mathrm{T}}\boldsymbol{D}_q(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i) \tag{1.4.10}$$

što je jednako idućem problemu :

$$\underset{\hat{\boldsymbol{Y}} \text{ ranga } p}{\text{Minimiziraj za}} \operatorname{tr}\left[\boldsymbol{D}_w(\boldsymbol{Y} - \hat{\boldsymbol{Y}})\boldsymbol{D}_q(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T\right] \tag{1.4.11}$$

Dakle, rješenje se dobije pomoću generaliziranog SVD-a, danog formulom :

$$\begin{aligned}&\boldsymbol{S} = \boldsymbol{D}_w^{1/2}\boldsymbol{Y}\boldsymbol{D}_q^{1/2} \\ &\tilde{\boldsymbol{U}} = \boldsymbol{D}_w^{1/2}\boldsymbol{U} \text{ i } \tilde{\boldsymbol{V}} = \boldsymbol{D}_q^{1/2}\boldsymbol{V}.\end{aligned} \tag{1.4.12}$$

Matrica ranga p koja minimizira izraz (1.4.10) se dobije iz dekompozicije početne matrice preko matrica $\tilde{\boldsymbol{U}}$ i $\tilde{\boldsymbol{V}}$:

$$\hat{\boldsymbol{Y}} = \tilde{\boldsymbol{U}}_{[p]}\boldsymbol{D}_{\beta[p]}\tilde{\boldsymbol{V}}_{[p]}^T \tag{1.4.13}$$

Koordinate redaka u niže-dimenzionalnom prostoru definirane s $\boldsymbol{F} = \tilde{\boldsymbol{U}}_{[p]}\boldsymbol{D}_{\beta[p]}$ se nazivaju *glavne koordinate* (eng. *principal coordinates*) i čine lijevu matricu biplota $\boldsymbol{G}$. Desnu matricu biplota koja predstavlja stupce čini matrica $\boldsymbol{H} = \tilde{\boldsymbol{V}}_{[p]}^T$ i nju nazivamo *standardnim koordinatama*.

1.5 Analiza glavne komponente

U prethodnom poglavlju smo definirali generalizanu formu PCA gdje smo retcima i stupcima dodijelili određene težine. Sada ćemo na primjeru (1.4.1) demonstrirati jednu varijantu analize glavne komponente koja se svodi na *uprosječivanje*, odnosno vektori težine $\boldsymbol{w} = (1/n)\mathbf{1}$ i $\boldsymbol{q} = (1/m)\mathbf{1}$ dodjeljuju svim stupcima i retcima matrice jednake težine.

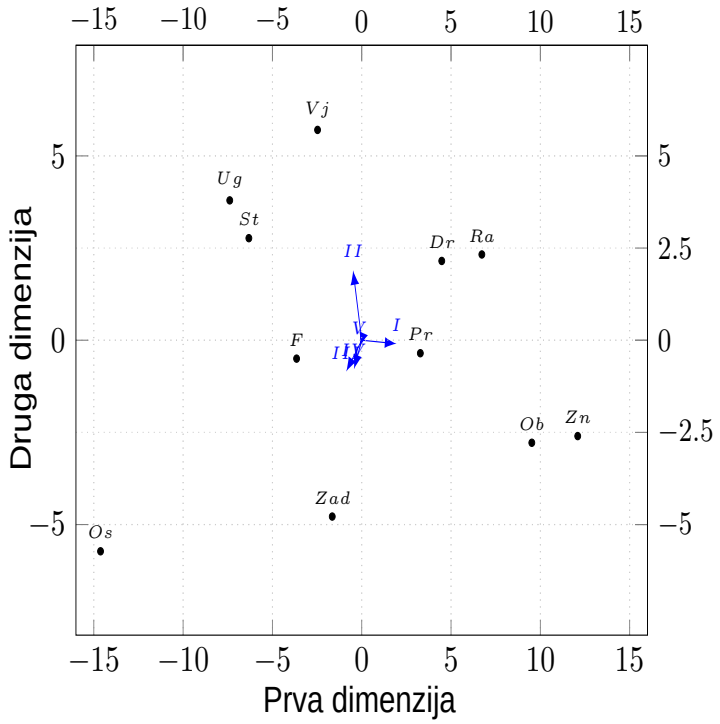
Računski postupak je sljedeći:

$$\text{Centriranje:} \quad \boldsymbol{Y} = [\boldsymbol{I} - (1/n)\mathbf{1}\mathbf{1}^T]\boldsymbol{X} \tag{1.5.1}$$

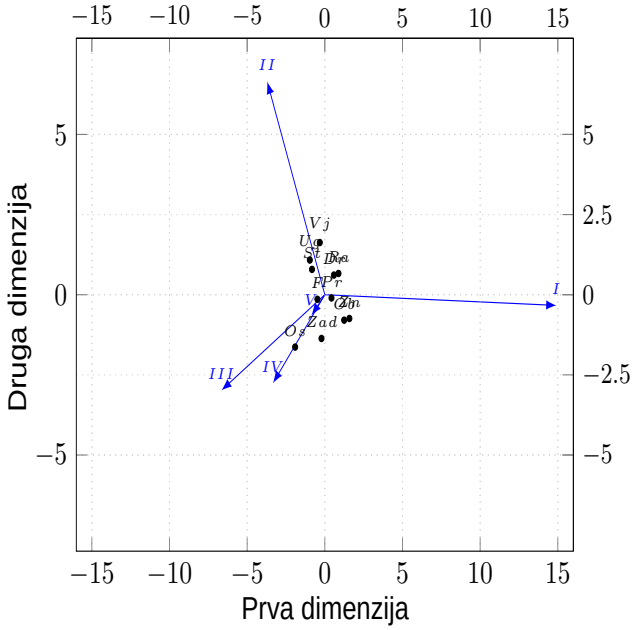
SVD dekompozicija s težinama: $\boldsymbol{S} = (1/n)^{\frac{1}{2}}\boldsymbol{Y}(1/m)^{\frac{1}{2}} = (1/nm)^{\frac{1}{2}}\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}_\beta\boldsymbol{V}^T$ (1.5.2)

Izračun koordinata: $\boldsymbol{G} = \boldsymbol{I}^{\frac{1}{2}}\boldsymbol{U}\boldsymbol{D}_\beta$ i $\boldsymbol{H} = \boldsymbol{I}^{\frac{1}{2}}\boldsymbol{V}^T$ (1.5.3)

Postoje 2 varijante biplota, ovisno o tome množimo li lijevu $\boldsymbol{U}$ ili desnu $\boldsymbol{V}$ matricu s matricom $\boldsymbol{D}_\beta$ singularnih vrijednosti. U prvom slučaju, kada lijeve singularne vektore množimo s odgovarajućim singularnim vrijednostima dobijamo matricu *glavnih koordinata* dok desni singularni vektori čine matricu *standardnih koordinata*. Analogno, kada desne singularne vektore pomnožimo s odgovarajućim singularnim vrijednostima dobijamo matricu *glavnih koordinata* dok lijevi singularni vektori čine matricu *standardnih koordinata*. Iduća dva grafa predstavljaju te dvije verzije PCA biplota.



Slika 1.4: PCA biplot - prva varijanta



Slika 1.5: PCA biplot - druga varijanta

Vidimo da prvi slučaj biplota, u kojem su retci *glavne koordinate*, omogućava bolju preglednost redaka, dok drugi slučaj kada su stupci *glavne koordinate* omogućava bolju preglednost pojedinih stupaca. Primjetimo, PCA biplot nam nudi više informacija nego osnovni biplot iz poglavlja 1.2 PCA biplot jer nam omogućuje grupiranje pitanja iz ankete na temelju njihovih odgovora u određene skupine. Naprimjer, u prvoj varijanti vidimo kako PCA grupira pitanja o važnosti obitelji i znanja dok je pitanje ostavštine djeci najudaljenije od svih te se stav ispitanika o tom pitanju najviše razlikuje. U drugoj varijanti biplota iz grafa možemo zaključiti na pitanja u anketi "Izrazito se slažem" i "Slažem se" prilično razlikuju, dok se ostali odgovori ne razlikuju značajno. Kako kvalitetnije vizualizirati podatke te iz njihovog grafičkog prikaza izvući što više informacija ćemo saznati u idućem poglavlju koje nas uvodi u analizu korespodencije.

Poglavlje 2

Analiza korespondencije

Uvod

Analiza korespondencije (eng. Correspondence analysis) odnosno skraćeno CA je metoda vizualizacije podataka koja je primjenjiva na kontigencijske tablice. U ovom poglavlju ćemo definirati osnovne pojmove analize korespodencije i demonstrirati metodu na par primjera. Za kategorijske podatke ćemo generalizirati ovu metodu u višestruku analizu korespodencije (eng. Multiple Correspondence analysis) ili MCA koja je popularna metoda za analizu anketa. Također ćemo se upoznati s kanonskom analizom korespodencije (eng. Canonical Correspondence analysis) ili CCA čija je prednost vizualiziranje podataka u restringiranom prostoru kovarijabli. Za kraj ćemo pokazati kako se podaci koji nisu u kontigencijskoj tablici mogu svesti na oblike pogodne za analizu korespondencije.

2.1 Osnovna analize korespodencije

Od više načina na koje je moguće definirati analizu korespodencije, mi ćemo se zadržati na sadašnjoj metodologiji te predstaviti analizu korespodencije kao generaliziranu PCA. Definirajmo sada osnovne pojmove i formule te ih prikažimo na primjeru.

Označimo s $\boldsymbol{N}$ matricu dimenzija $n \times m$ s nenegativnim elementima koja predstavlja našu kontigencijsku tablicu s podacima.

Definicija 2.1.1. *Za neku matricu* $\boldsymbol{N}$ *definiramo grand total* $n = \sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij} = \mathbf{1}^T\boldsymbol{N}\mathbf{1}$ *i matricu korespodencije* $\boldsymbol{P}$ *:*

$$\boldsymbol{P} = \frac{1}{n}\boldsymbol{N}$$

Dodatno, definiramo vektore retčanih i stupčanih masa:

$$r_i = \sum_{j=1}^{m} p_{ij} \; i \; c_j = \sum_{i=1}^{n} p_{ij}$$
$$\boldsymbol{r} = \boldsymbol{P}\mathbf{1} \; i \; \boldsymbol{c} = \boldsymbol{P}^T\mathbf{1}$$

I odgovarajuće dijagonalne matrice $\boldsymbol{D}_r = diag(\boldsymbol{r})$ *i* $\boldsymbol{D}_c = diag(\boldsymbol{c})$

Tablica 2.1: Rezultati ankete o pušenju

Zaposlenici	Pušačke navike					*Mase*
	Nikakve	*Lagane*	*Srednje*	*Teške*	*Zbroj*	*redaka*
Stariji menadžment	4	2	3	2	*11*	*0.057*
SM	(0.364)	(0.182)	(0.273)	(0.182)		
Mlađi menadžment	4	3	7	4	*18*	*0.093*
JM	(0.222)	(0.167)	(0.389)	(0.222)		
Stariji zaposlenici	25	10	12	4	*51*	*0.279*
SE	(0.490)	(0.196)	(0.235)	(0.078)		
Mlađi zaposlenici	18	24	33	13	*88*	*0.456*
JE	(0.205)	(0.273)	(0.375)	(0.148)		
Tajnice	10	6	7	2	*25*	*0.130*
SC	(0.400)	(0.240)	(0.280)	(0.080)		
Zbroj	61	45	62	25	*193*	
Prosječni profil retka	(0.316)	(0.233)	(0.321)	(0.130)		

Demonstrirajmo sada te pojmove na gornjem primjeru. Ovaj izmišljeni primjer se smatra osnovnim testnim primjerom analize korespondencije te se nalazi unutar svih važnijih komercijalnih statističkih paketa. Služi kao odličan uvod u dvodimenzionalnu analizu korespodencije te se često citira u znanstvenim radovima. Odnosi se na rezultate ankete 193 zaposlenika jedne kompanije, koja je sprovedena svrhom određivanje kompanijinog stava prema pušenju. Zaposlenici su podijeljeni u 5 kateogorija ovisno o njihov poziciji na poslu i njihove pušačke navike su kategorizirane u 4 grupe. Zbog toga što se radi o matrici dimenzija 5×4 može se savršeno prikazati u trodimenzionalnom prostoru.

Dakle, grand total n = 193 što je ujedno ukupan broj anketiranih zaposlenika. Zapišimo sada matricu korespodencije $\boldsymbol{P}$.

$$\boldsymbol{P} = \begin{bmatrix} 0.021 & 0.010 & 0.016 & 0.010 \\ 0.021 & 0.016 & 0.036 & 0.021 \\ 0.130 & 0.052 & 0.062 & 0.021 \\ 0.093 & 0.124 & 0.171 & 0.067 \\ 0.052 & 0.031 & 0.036 & 0.010 \end{bmatrix}$$

Prisjetimo li se generalizirane PCA, vidimo kako se profili redaka i stupaca prirodno nameću kao vektori težina u analizi korespodenciji i zbog svoje važnosti imaju i posebno ime *mase*. Primjetimo kako je profil retka jednak masi stupca i obratno. Dakle, naši vektori masa su :

$$r = \begin{bmatrix} 0.057 \\ 0.093 \\ 0.279 \\ 0.456 \\ 0.130 \end{bmatrix} \quad i \quad c = \begin{bmatrix} 0.316 \\ 0.233 \\ 0.321 \\ 0.130 \end{bmatrix}$$

Udaljenosti između profila računamo χ^2-metrikom koja je temelj analize korespondencije, no najprije se prisjetimo χ^2 -statistike.

$$\chi^2 = \sum \frac{(f_o - f_t)^2}{f_t}$$

pri čemu f_o označava opažene frekvencije, a f_t očekivane (teoretske) frekvencije. Sada napravimo jednu modifikaciju, podijelimo χ^2-statisku s *grand total-om*. Izraz χ^2/n u analizi korespodencije nazivamo *ukupna incercija* ili samo *inercija*. Inercija je mjera varijance u našoj kontigencijskoj tablici. Svaki redak i stupac imaju svoj doprinos ukupnoj inerciji koji nazivano *inercija retka* odnosno *inercija stupca*. Metoda analize korespodencije bi se mogla shvatiti, u smislu inercije, kao metoda čiji je cilj maksimizirati udio inercije koji ćemo poslije prikazati biplotom. Preciznije, tražimo prvu dimenziju koja će "objasniti" najveći dio inercije zatim drugu dimenziju koja će "objasniti" najveći dio preostale inercije i tako dalje. Pod pojmom "objasniti" mislimo na to koliki će udio ukupne inercije iznositi inercija svih točaka projekciranih u taj prostor.

U sljedećoj tablici ćemo prikazati rastav tablice 2.1 na opažene i očekivane relativne frekvencije te kako ih njih izračunati χ^2-statistiku. Matricu očekivanih relativnih frekvencija $\boldsymbol{O}$ dobijamo množenjem vektora masa $\boldsymbol{r}$ i $\boldsymbol{c}$:

$$\boldsymbol{O} = \boldsymbol{r}\boldsymbol{c}^T = \begin{bmatrix} 0.018 & 0.013 & 0.018 & 0.007 \\ 0.029 & 0.022 & 0.030 & 0.012 \\ 0.084 & 0.062 & 0.085 & 0.034 \\ 0.144 & 0.106 & 0.146 & 0.059 \\ 0.041 & 0.030 & 0.042 & 0.017 \end{bmatrix}$$

Matrično, formulu za izračun inercije možemo zapisati ovako:

$$\text{Inercija} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$$

Sljedeću tablicu nazivamo matricom standardiziranih reziduala $\boldsymbol{S}$ i njezini elementi predstavljaju doprinos ukupnoj inerciji odgovarajućih elemenata početne matrice korespondencije $\boldsymbol{P}$. U našem slučaju, ukupna inercija iznosi 0.0852.

Tablica 2.2: Rastav inercije po elementima

	Inercija				
	Nikakve	*Lagane*	*Srednje*	*Teške*	Ukupno
SM	0.0004	0.0006	0.0004	0.0012	0.0027
JM	0.0026	0.0018	0.0013	0.0062	0.0119
SE	0.0254	0.0016	0.0061	0.0053	0.0383
JE	0.0179	0.0031	0.0041	0.0012	0.0263
SC	0.0029	0.0000	0.0007	0.0025	0.0061
Ukupno	0.0492	0.0071	0.0126	0.0163	0.0852

Kada smo definirali sve potrebne pojmove i matrice možemo objasniti algoritam za računanje analize korespodencije:

1. korak – Izračunajmo matricu standardiziranih reziduala $\boldsymbol{S}$:

$$\boldsymbol{S} = \boldsymbol{D}_r^{-\frac{1}{2}}\left(\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}^{\mathrm{T}}\right)\boldsymbol{D}_c^{-\frac{1}{2}} \tag{2.1.1}$$

2. korak – Napravimo singularnu dekompoziciju matrice $\boldsymbol{S}$:

$$\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}_\alpha\boldsymbol{V}^{\mathrm{T}} \text{ gdje je } \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I} \tag{2.1.2}$$

gdje je $\boldsymbol{D}_\alpha$ dijagonalna matrica pozitivnih singularnih vrijednosti u padajućem poretku: $\alpha_1 \geq \alpha_2 \geq \cdots$

3. korak – Standardne koordinate redaka $\boldsymbol{\Phi}$:

$$\boldsymbol{\Phi} = \boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U} \tag{2.1.3}$$

4. korak – Standardne koordinate stupaca $\boldsymbol{\Gamma}$:

$$\boldsymbol{\Gamma} = \boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{V} \tag{2.1.4}$$

5. korak – Glavne koordinate redaka $\boldsymbol{F}$:

$$\boldsymbol{F} = \boldsymbol{D}_r^{-\frac{1}{2}}\boldsymbol{U}\boldsymbol{D}_\alpha = \boldsymbol{\Phi}\boldsymbol{D}_\alpha \tag{2.1.5}$$

6.korak – Glavne koordinate stupaca $\boldsymbol{G}$ *:*

$$\boldsymbol{G} = \boldsymbol{D}_c^{-\frac{1}{2}}\boldsymbol{V}\boldsymbol{D}_\alpha = \boldsymbol{\Gamma}\boldsymbol{D}_\alpha \tag{2.1.6}$$

Sada se vratimo na pojam ukupne inercije matrice. Neka je rang matrice S jednak k. Vrijedi :

$$\text{Inercija } = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{p_{ij}-r_ic_j}{\sqrt{r_ic_j}} = tr\left(\boldsymbol{S}\boldsymbol{S}^T\right) = tr\left(\boldsymbol{U}\boldsymbol{D}_\alpha\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{D}_\alpha\boldsymbol{U}^{\mathrm{T}}\right)$$

Zbog toga što je $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}$ i $\boldsymbol{D}_\alpha$ dijagonalna matrica slijedi

$$\text{Inercija } = tr\left(\boldsymbol{D}_\alpha\boldsymbol{D}_\alpha\right) = \sum_{i=1}^{k}\alpha_i^2.$$

Kvadrati signularnih vrijednosti dobivenih dekompozicijom matrice $\boldsymbol{S}\boldsymbol{S}^T$ su jednaki svojstvenim vrijednostima te matrice pa slijedi

$$\text{Inercija} = \sum_{i=1}^{k}\alpha_i^2 = \sum_{i=1}^{k}\lambda_i.$$

Dakle, ukupna inercija je jednaka sumi svojstvenih vrijednosti matrice $\boldsymbol{S}\boldsymbol{S}^T$, zbog čega se još u literaturi pojedine inercije nazivaju svojstvenim vrijednostima. Prisjetimo se konstrukcije regresijskog biplota, tamo uzimamo prvih r vektora iz matrica $\tilde{\boldsymbol{U}}$,$\tilde{\boldsymbol{\Sigma}}$ i $\tilde{\boldsymbol{V}}$ što znači da uzimamo i prvih r svojstvenih vrijednosti iz matrice $\tilde{\boldsymbol{\Sigma}}$. Zbog te činjenice i povezanosti svojstvenih vrijednosti s inercijom, zaključujemo da će "objašnjena" inercija u biplotu analize korespondencije biti jednaka kvadratu prvih r singularnih vrijednosti dobivenih dekompozicijom matrice $\boldsymbol{S}$.

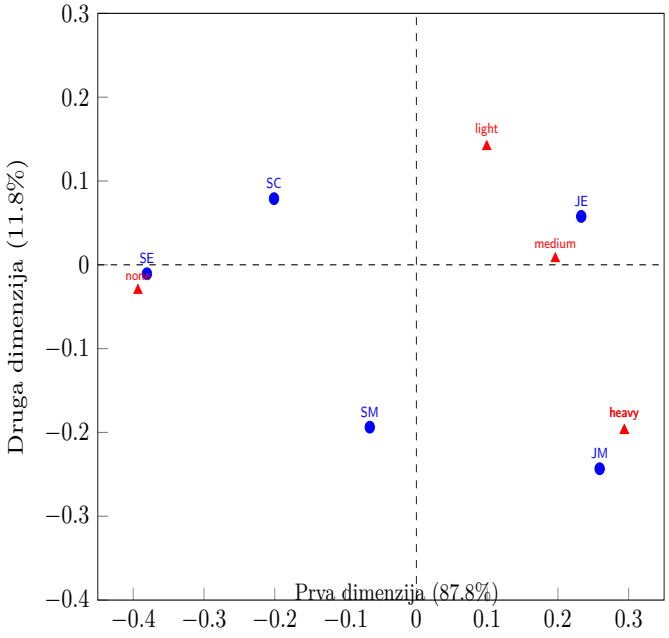
U primjeni se koriste 3 kombinacije koordinata za prikaz biplotova koji se u analizi korespondencije još nazivaju i *mape*. Simetrična mapa je najpopularniji prikaz rezultata Analize korespodencije iz koje dobijamo informacije o međusobnom odnosu pojedinih redaka i stupaca, pri čemu su oni jednako skalirani te je prikaz više estetičan nego u ostala dva asimetrična tipa prikaza. U asimetričnom prikazu su retci ili stupci skalirani drugačije, odnosno retci su u glavnim koordinatama a stupci u strandardnim ili obratno. Ovakve mape su korisne kad su varijable redaka odnosno stupaca od posebnog značaja pa ih želimo preglednije vizualizirati.

Geometrijski, ovisno o pojedinom prikazu, udaljenosti između redaka predstavljaju aproksimativnu χ^2 udaljenost između pojedinih redaka odnosno stupaca ili oboje u slučaju

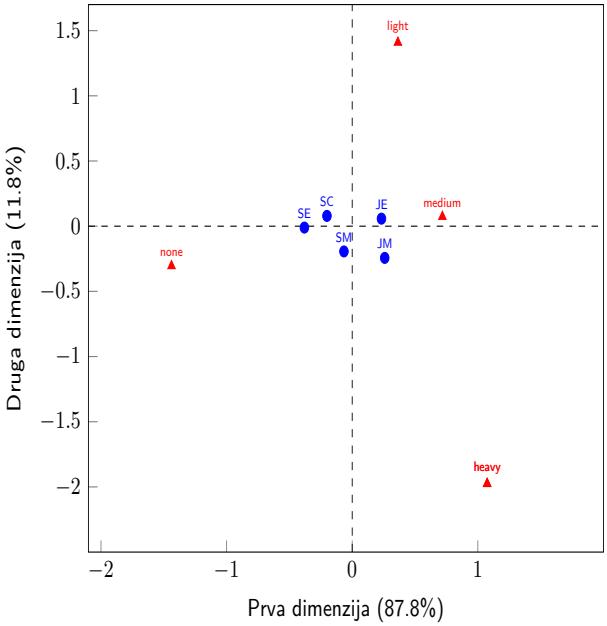
simetričkog grafa. χ^2-udaljenost između redaka računamo :

$$\chi^2\text{-udaljenost između redaka i i k} = \sum_{j=1}^{m} \left(\frac{p_{ij}}{r_i} - \frac{p_{kj}}{r_k} \right)^2 / c_j$$

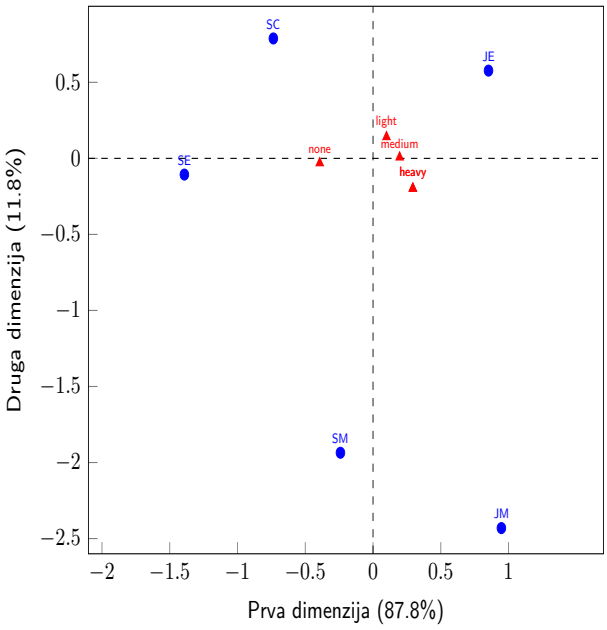
χ^2-udaljenost stupaca se računa analogno. Sada pokažimo CA biplote. Prvo promatramo graf simetričnih koordinata, zatim promatramo prikaz u kojem su retci glavne koordinate, na engleskom se naziva *row-principal map* i na kraju prikaz u kojem su stupci glavne koordinate, koji se naziva *column-principal map*.



Slika 2.1: CA - Simetrična mapa



Slika 2.2: CA - Retci su glavne koordinate

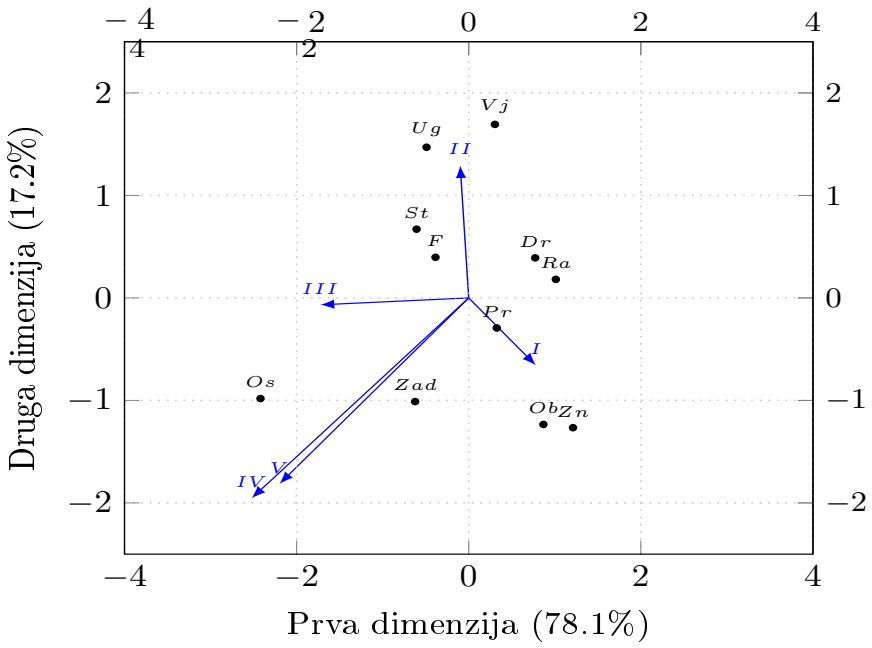


Slika 2.3: CA - Stupci su glavne koordinate

Najprije primjetimo kako su točke s glavnim koordinatama unutar konveksne ljuske određene točkama sa standardnim koordinatama. Nadalje, vidimo kako su mlađi zaposlenici te mlađi menađeri na desnoj strani te stariji zaposlenici na lijevoj strani najudaljeniji te zaključujemo kako je najveća udaljenost u pušačkim navikama između tih skupina. Stariji menađeri leže između mlađih menađera i starijih zaposlenika, dok su tajnice blizu starijih zaposlenika. Kako bi objasnili sličnosti i razlike između pozicija ovih grupa zaposlenika trebamo proučiti njihove pozicije u odnosu na stupce koji predstavljaju biplot vektore. Kako se sve tri pušačke kategorije nalaze na desno a nepušačka kategorija na lijevo, zaključujemo da smjer lijevo-desno predstavlja razliku između pušača i nepušača. Dakle, zaključujemo kako se grupe mlađih zaposlenika i mlađih menađera razlikuju od starijih zaposlenika jer imaju relativno veći broj pušača a stariji zaposlenici relativno manji broj pušača. Prisjetimo se kako u se u središtu biplota nalazi prosječni profil, tako bi udaljenost pojedinih grupa zaposlenika od ishodišta mogli smatrati odstupanjem od prosječnog profila zaposlenika u određenom smjeru, pri čemu smo maloprije interpretirali smjer lijevo-desno.

Sada kada smo objasnili osnovnu analizu korespondencije te crtanje njezinih rezultata možemo iskoristiti to znanje na primjeru 1.3.1. i demonstrirati prednost analize korespodencije nad PCA u analiziranju anketa. Postupak konstrukcije biplota je jedak kao u prethodnom primjeru stoga ćemo ga preskočiti i odmah nacrtati simetričnu mapu.

Primjer 2.1.1.



Slika 2.4: Simetrični CA biplot

Prvo što primjećumo je visok udio inercije objašnjen s ovim biplotom. Naime, prve dvije dimenzije objašnjavaju 95.3% ukupne inercije što je dobro jer će se zaključci izvedeni iz grafa temeljiti na skoro savršenim informacijama dobivenim iz ankete. Možemo primjetiti kako se na desnoj strani nalaze izjave s kojima se studenti relativno više slažu, dok su na lijevoj strani izjave s kojima se studenti relativno manje slažu. Također, možemo primjetiti grupiranje pojedinih izjava, kao što su "Društvene vještine su bitne za životni uspjeh" i "Radna etika je bitna za životni uspjeh" ili "Fakultet me je oblikovao kao osobu" i "Fakultet mi je proširio znanje i pomogao u intelektualnom razvoju". Zanimljiva informacija je kako se studenti uglavnom ne slažu s izjavom "Ostavština za djecu mi je bitna" koja je odvojena od svih. Nadalje, ako promatramo stupce koje predstavljaju biplot vektori, vidimo kako se pojedine izjave nalaze u smjeru određenih odgovora. Na primjer, izjave poput "Fakultet me je oblikovao kao osobu" se nalaze u smjeru odgovora "Izrazito se slažem" dok se izjava "Zadovoljan sam izborom karijere" nalazi u smjeru odgovora "Uglavnom se ne slažem" i "Izrazito se ne slažem". Usporedimo li sliku 2.4 sa slikama 1.3 i 1.4 na str.17 odnosno str.19 vidimo kako nam analiza korespodencije nudi više informacija o međusobnim vezama objekata i varijabli koje promatramo. Pregled podataka je bolji te je manje informacija izgubljeno aproksimacijom podataka iz nižedimenzionalnog prostora.

Suplementarne točke

Često je slučaj kako postoje dodatni stupci i retci podataka koji nisu od primarnog interesa za istraživanje ali su korisni za interpretiranje nekih otkrića dobivenih iz početnih podataka. Bilo koji dodatni redak ili stupac se može staviti na mapu, dokle god je taj novi stupac ili redak usporediv s početnim podacima koji određuju mapu. Takvi dodatni podaci se nazivaju *suplementarne točke*. Oni ne utječu na analizu korespodencije već se naknadno projiciraju na postojeću mapu dobivenu analizom korespodencije. Kako bi sumplementarne točke smjestili na mapu koristimo tranzicijske jednadžbe. Na primjer, ako imamo suplementarni redak $\boldsymbol{h}$ dimenzija $1 \times m$ tada podijelimo $\boldsymbol{h}$ s njegovim totalom $\mathbf{1}^T\boldsymbol{h}$ kako bi dobili profil retka $\tilde{\boldsymbol{h}} = (1/\mathbf{1}^T\boldsymbol{h})\boldsymbol{h}$. Konačno, koordinate suplementarnog retka dobijamo skalarnim produtkom njegovog profila te standardnih koordinata stupca. Navodimo idući primjer kako bi bolje pokazali interpretativnost biplota analize korespondencije a i korisnost suplementarnih točaka.

Primjer 2.1.2. Sljedeća tablica je rezultat ispitivanja nad uzorkom od 100 domaćinstava. Ispitanici su trebali odgovoriti povezuju li pojedinu izjavu s bilo kojom od 8 vrsta doručka pri čemu su dopušteni višestruki odgovori. Izjave i doručci se nalaze u sljedećoj tablici:

Tablica 2.3: Kratice izjava i doručaka

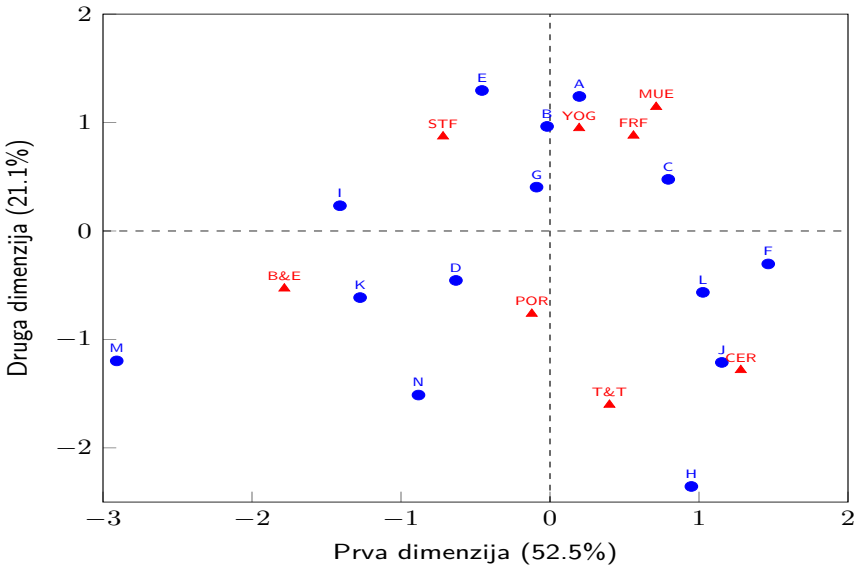
Kratica	Izjava	Doručak	Kratica
A	Zdravo	Žitarice	CER
B	Hranjivo	Muesli	MUE
C	Zimsko jelo	Kaša	POR
D	Ljetno jelo	Slanina i jaja	B&E
E	Skupo	Tost i čaj	T&T
F	Brzo i jednostavno	Svježe voće	FRF
G	Ukusno	Kuhano voće	STF
H	Ekonomično	Jogurt	YOG
I	Poslastica		
J	Tijekom radnog tjedna		
K	Za vikend		
L	Bezukusno		
M	Predugo se sprema		
N	Obiteljski favorit		

U tablici 2.4 na idućoj stranici vidimo rezultate ispitivanja.

Tablica 2.4: Rezultati ispitivanja doručaka

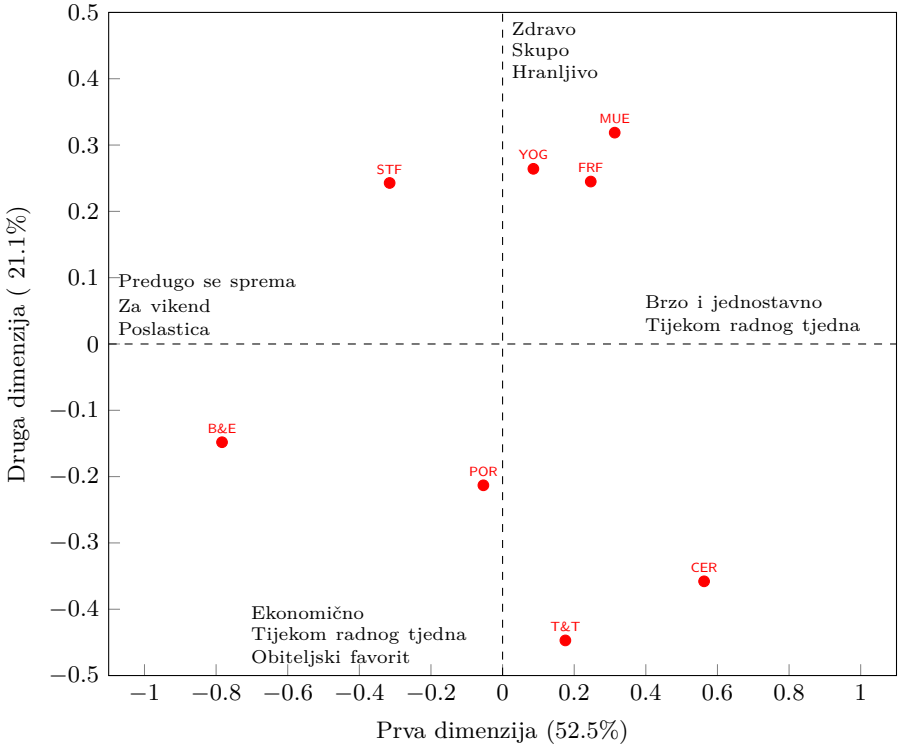
	CER	MUE	POR	B&E	T&T	FRF	STF	YOG
A	14	38	25	18	8	31	28	34
B	14	28	25	25	7	32	26	31
C	42	22	11	13	7	37	16	35
D	10	10	32	26	6	11	19	8
E	6	33	5	27	3	9	18	10
F	54	33	8	2	15	26	8	20
G	24	21	16	34	11	33	26	26
H	24	3	20	3	16	7	3	7
I	5	3	3	31	4	4	16	17
J	47	24	15	9	13	11	6	10
K	12	5	8	56	16	10	23	18
L	8	6	2	2	0	0	2	1
M	0	0	9	35	1	0	10	0
N	14	4	10	31	5	7	2	5

Sada analizom korespondencije prikažimo odnose među ovim varijablama na sljedećem biplotu. Za glavne koordinate izaberemo doručke a za standardne koordinate izjave o doručcima.



Slika 2.5: CA - Simetrična mapa

Ukupni udio inercije objašnjen s dvijema dimenzijama, od ukupno u 7 u kojima se nalaze rezultati anketa, iznosi 73.6%. Pogledajmo koordinate izjava na 1. osi i izdvojimo ekstreme. Primjetimo da su najviše lijevo izjave ”Predugo se sprema”, ”Za vikend”,”Poslastica” a najviše desno izjave ”Brzo i jednostavno” i ”Tijekom radnog tjedna”. Prema rasporedu izjava po prvoj osi možemo zaključiti kako ta os predstavlja pristupačnost doručka. Tako je doručak koji se nalazi na desnoj strani biplota svakodnevan i jednostavan za pripremu, dok su oni na lijevoj strani zahtjevniji i treba im više vremena za pripremu. Sada promotrimo izjave na drugoj osi. Najniže se nalaze izjave ”Ekonomično”, ”Tijekom radnog tjedna” i ”Obiteljski favorit” dok se iznad nalaze izjave ”Zdravo”, ”Skupo” i ”Hranjivo”. Malo je teže pronaći jedinstvenu karakteristiku s kojom bi opisali drugu os ali možemo reći da bi jedna osobina bila cijena. Tako će doručci koji se nalaze ispod prve osi biti jeftiniji od onih koji se nalaze iznad. Sada prikažimo grafički doručke te novodobivene zaključke.



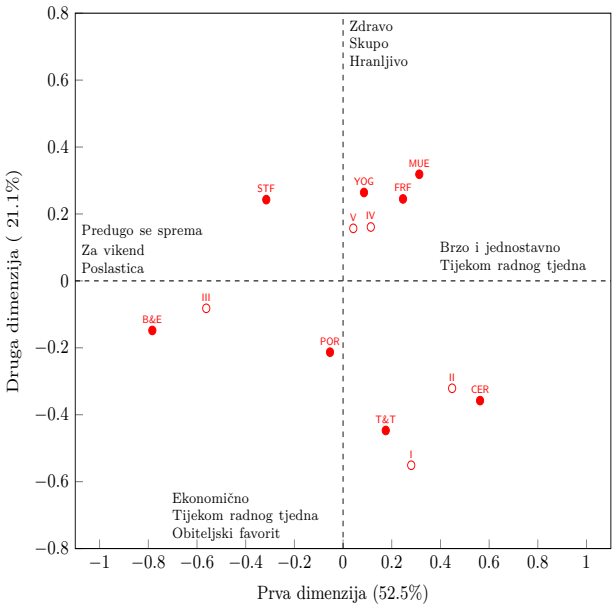
Slika 2.6: Prikaz doručka i dobivenih zaključaka na CA biplotu

Pretpostavimo da smo prikupili podatke o frekvenciji konzumiranja pojedine vrste doručka, pri čemu I označava dnevnu konzumaciju, II više puta tjedno, III više puta mjesečno, IV svakih par mjeseci te V nikad.

Tablica 2.5: Rezultati ispitivanja konzumacije pojedine vrste doručka

	CER	MUE	POR	B&E	T&T	FRF	STF	YOG
I	24	3	4	8	18	2	9	11
II	58	15	8	13	16	10	10	29
III	6	10	12	46	8	14	15	8
IV	2	4	28	9	4	47	4	2
V	10	68	48	24	54	27	62	50

Sada projiciramo suplementarne točke na prethodnu sliku. Vidimo da su najčešće konzmirani doručci žitarice te tost i čaj. Općenitije, možemo zaključiti kako se doručci koji se nalaze niže na grafu češće konzumiraju. Ovime smo pokazali korisnost suplementarnih točaka u dodatnoj interpretaciji prethodno dobivenih rezultata analizom korespodencije.



Slika 2.7: Prikaz stupaca sa suplementarnim točkama

2.2 Višestruka analiza korespondencije

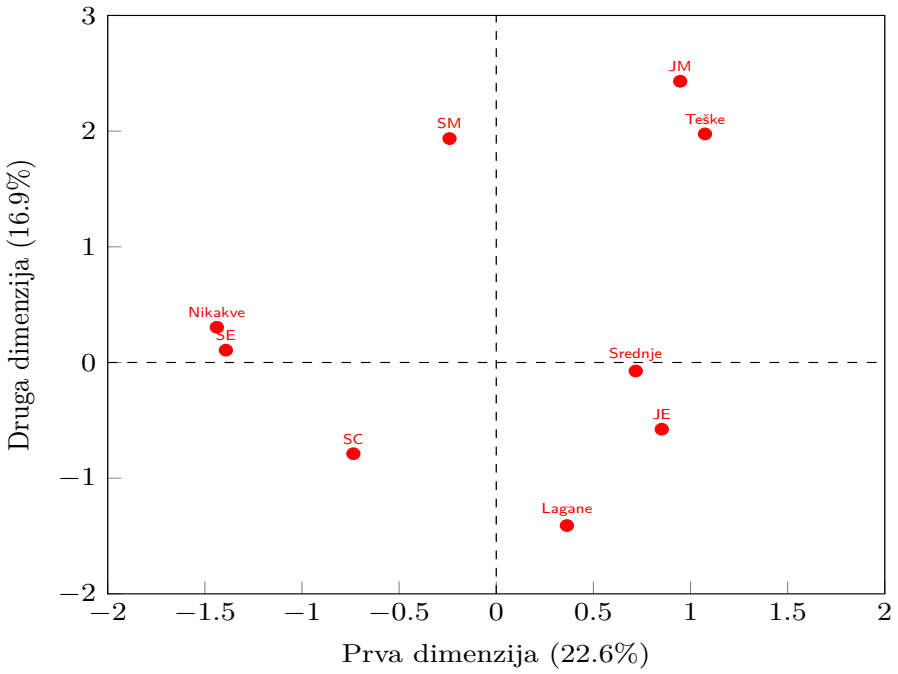
Dosad smo pokazali odnos između dvije kategorijske varijable ili dva skupa kategorijskih varijabli. Sada ćemo promatrati povezanost unutar jednog skupa varijabli, gdje nas zanima na koji način i koliko snažno su te varijable povezane. Postoje dva klasična načina pristupanja ovom problemu kojeg nazivamo višestruka analiza korespondencije (eng. Multiple correspondence analysis) odnosno skraćeno MCA. Jedan način je da zamislimo MCA kao analizu čitavog skupa podataka unutar kojeg imamo *dummy* varijable. Takav skup podataka nazivamo indikatorska matrica. Drugi način je analiza međusobnog odnosa (eng. cross-tabulations) između svih varijabli, što prikazujemo *Burtovom* matricom. Burtova matrica je kontigencijska tablica svih mogućih kombinacija varijabli. Ona je jednaka skalarnom produktu indikatorske matrice same sa sobom. Označimo li s $\boldsymbol{Z}$ indikatorsku matricu te s $\boldsymbol{B}$ Burtovu matricu tada vrijedi relacija $\boldsymbol{B} = \boldsymbol{Z}^T\boldsymbol{Z}$. Iz ove relacije vidimo kako analiza korespondencije daje jednake standardarne koordinate za obje tablice. Razlikuju se jedino glavne koordinate jer Burtova matrica ima kvadratnu inerciju indikatorske matrice. Zbog različitih inercija su glavne koordinate dobivene analizom korespodencije indikatorske matrice i Burtove matrice drukčije skalirane.

Ova metoda se najčešće koristi u vizualizaciji anketa, gdje ispitanici odgovaraju na niz pitanja odgovorima na različitim skalama. Rezultati ankete se spremaju u matricu gdje osim rezultata također i zapisujemo demografske podatke ispitanika tako da se svaka observacija zajedno s odgovarajućim indikatorima zapisuje u svoj redak. Dakle, ova metoda se svodi na osnovnu analizu korespondencije indikatorske matrice ali u primjeni se zbog preglednosti češće koristi Burtova matrica. Zasad demonstrirajmo MCA na indikatorskoj matrici. Zbog jednostavnosti pokazat ćemo indikatorsku matricu za primjer 2.1 za pušače:

Tablica 2.6: Indikatorska matrica za primjer 2.1.

	Grupe zaposlenika					Pušačke navike			
Observacija	SM	JM	SE	ME	SC	Nikakve	Lagane	Srednje	Teške
1	1	0	0	0	0	1	0	0	0
2	1	0	0	0	0	1	0	0	0
3	1	0	0	0	0	1	0	0	0
4	1	0	0	0	0	1	0	0	0
5	1	0	0	0	0	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
191	0	0	0	0	1	0	0	1	0
192	0	0	0	0	1	0	0	0	1
193	0	0	0	0	1	0	0	0	1

Svaki od ova 193 retka predstavlja anketne odgovore jednog zaposlenika poduzeća gdje jedinica u tablici označava pripadnost pojedinoj skupini odnosno nula označava nepripadnost skupini. Na primjer, prva observacija predstavlja jednog starijeg menađera koji je nepušač. Ovakav oblik pogodan je za istraživanje rezultata ankete jer u slučaju dodatnih pitanja u anketi, na primjer pitanje o spolu, samo dodajemo jedan stupac koji bi označavao pripadnost muškom odnosno ženskom spolu. Pokažimo rezultat analize korespondencije indikatorske matrice za ovaj primjer. Primjetimo kako nismo prikazali retke na biplotu jer ih ima 193.



Slika 2.8: MCA biplot indikatorske matrice

Primjetimo kako ovaj biplot ima značajno slabije objašnjenu varijancu, stoga je preporučljivo u slučaju dviju varijabli koristiti osnovnu analizu korespondencije zbog manjeg gubitka informacija.

Kad imamo više varijabli u primjeni koristit ćemo iduću tablicu čija analiza korespodencije daje isti rezultat kao i analiza korespondencije indikatorske matrice. Odgovarajuća Burtova tablica za prethodnu indikatorsku matricu je sljedeća :

Tablica 2.7: Burtova tablica za primjer 2.1

	Zaposlenici					Pušačke navike			
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)
(1) Stariji menađeri	11	0	0	0	0	4	2	3	2
(2) Mlađi menađeri	0	18	0	0	0	4	3	7	4
(3) Stariji zaposlenici	0	0	51	0	0	25	10	12	4
(4) Mlađi zaposlenici	0	0	0	88	0	18	24	33	13
(5) Tajnice	0	0	0	0	25	10	6	7	2
(1) Nikakve	4	4	25	18	10	61	0	0	0
(2) Lagane	2	3	10	24	6	0	45	0	0
(3) Srednje	3	7	12	33	7	0	0	62	0
(4) Teške	2	4	4	13	2	0	0	0	25

Struktura Burtove matrice se odmah isčitava iz tablice. U slučaju dviju kategorijskih varijabli kao u gornjem slučaju, Burtova matrica se sastoji od 4 blok matrice. Prva se odnosi na međusobni odnos prve varijable same sa sobom, druga za prvu varijablu s drugom, treća za drugu varijablu s prvom te četvrta na međusobni odnos druge varijable same sa sobom. Primjetimo da je tablica simetrična i suma dijagonalnih elemenata u prvoj i četvrtoj blok matrici je jednaka ukupnom broju ispitanika. Kao što smo prethodno napisali, rezultat analize korespondencije Burtove matrice je jednak rezultatu analize korespondencije indikatorske matrice stoga ne provodimo analizu.

2.3 Kanonska analiza korespondencije

U poglavlju 1.3 smo se upoznali s regresijskim biplotom i postupkom regresije eksplanatornih varijabli na niždimenzionalni prostor. U ovom poglavlju nas zanima suprotna ideja, točnije kako umjesto regresije varijabli na prostor napraviti regresiju dimenzija na eksplanatorne varijable. Naime, osnovna analiza korespondencije optimizira glavne osi kako bi objasnila najveći mogući dio inercije dok kanonska analiza korespondencije optimizira glavne osi u ograničenom potprostoru. U restringiranom prostoru osnovni CA algoritam nalazi najbolje glavne osi. Ograničeni potprostor je određen dodatnim suplementarnim varijablama koje se nalaze u matrici $\boldsymbol{X}$. Matrica standardiziranih reziduala $\boldsymbol{S}$ se projicira u potprostor te se nad tom projekcijom $\boldsymbol{S}^*$ napravi osnovna analiza korespondencije.

Algoritam kanonske analize korespodencije:

1.korak – Izračunajmo matricu standardiziranih reziduala $\boldsymbol{S}$ kao i u osnovnom algoritmu:

$$\boldsymbol{S} = \boldsymbol{D}_r^{-\frac{1}{2}} \left(\boldsymbol{P} - \boldsymbol{r}\boldsymbol{c}^{\mathrm{T}} \right) \boldsymbol{D}_c^{-\frac{1}{2}} \tag{2.3.1}$$

2.korak – Izračunajmo projekcijsku matricu ranga m, koja nam služi kao projektor u restringirani potprostor:

$$\boldsymbol{Q} = \boldsymbol{D}_r^{\frac{1}{2}} \boldsymbol{X} \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{D}_r \boldsymbol{X} \right)^{-1} \boldsymbol{D}_r^{\frac{1}{2}} \tag{2.3.2}$$

3.korak – Projiciramo matricu standardiziranih reziduala kako bi dobili ograničenu matricu:

$$\boldsymbol{S}^* = \boldsymbol{Q}\boldsymbol{S} \tag{2.3.3}$$

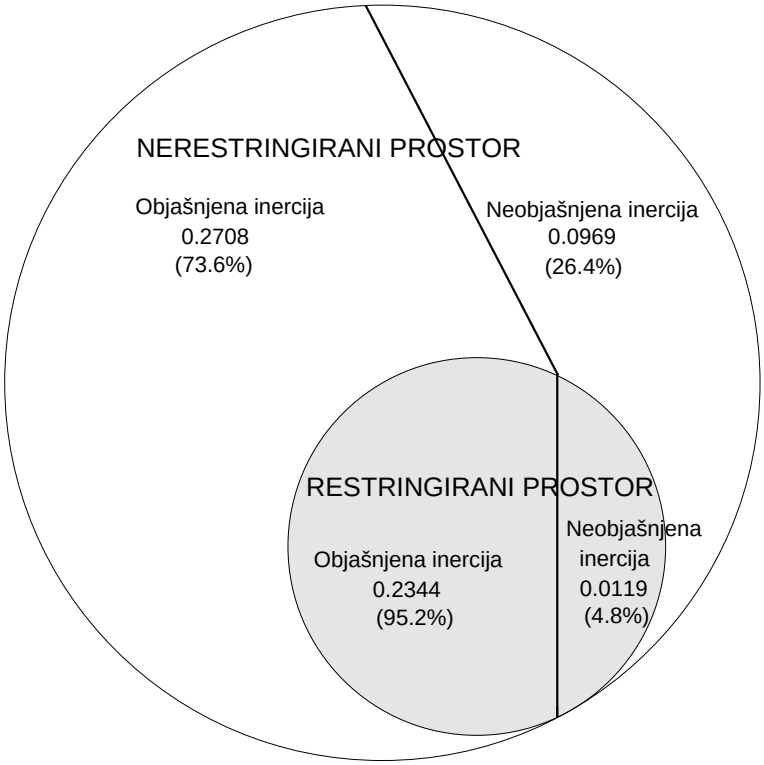
4.korak – Primjenimo osnovnu analizu korespondencije (str 23.-24.) na matricu $\boldsymbol{S}^*$:

5.korak – Glavne inercije λ_k^* u ograničenom potprostoru:

$$\lambda_k^* = \alpha_k^2, \quad k = 1, 2, \ldots, m \tag{2.3.4}$$

Primjer 2.3.1. Prisjetimo se primjera 2.1.2 i na njemu prikažimo kanonsku analizu korespodencije. Zanima nas prikaz rezultata ankete u potprostoru frekvencija konzumiranja pojedine vrste doručka i to u skupu sljedećih vrijednosti: I što označava dnevnu konzumaciju, III više puta mjesečno te V nikad. U primjeru 2.1.2 smo frekvencije konzumacije obradili kao suplementarne točke i pritom objasnili 73.6% inercije. Kanonskom analizom ćemo postići veći udio objašnjenje inercije unutar restringiranog potprostora. Za ovaj primjer je ukupna inercija jednaka 0.3678 i dijeli se na 0.2464 za restringirani potprostor te 0.1213 za nerestringirani prostor. Nadalje, analizom korespodencije na restringiranom

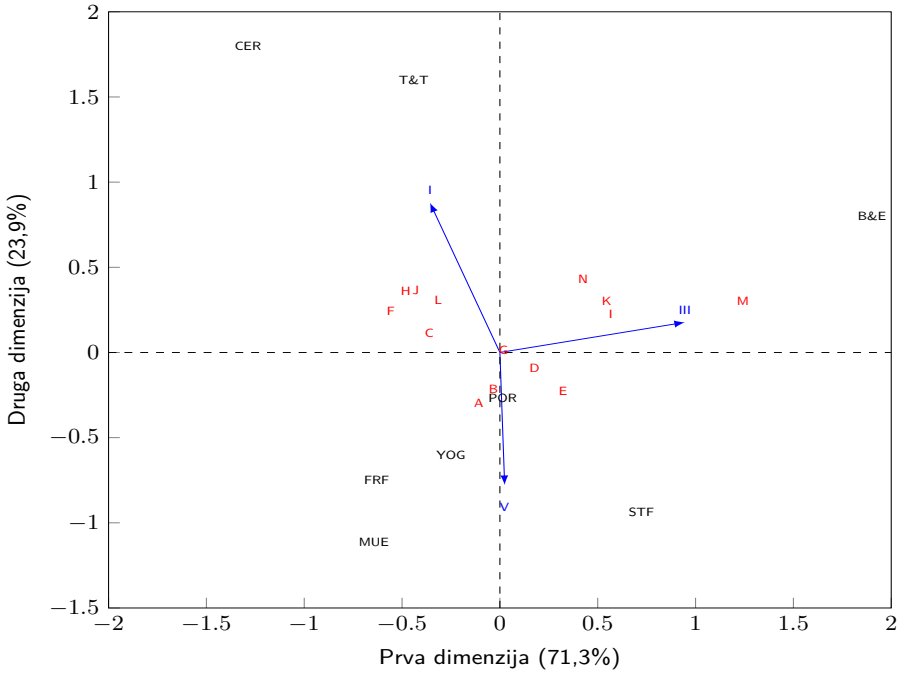
prostoru dobijamo glavne inercije 0.17568, 0.05879 i 0.01192. Izabiremo prve dvije te objašnjavamo 71.3% inercije s glavnom osi odnosno 29.9% preostale inercije drugom osi što znači da smo ukupno objasnili 95.2% inercije što je poboljšanje u odnosu na 73.6% kad smo iste varijable promatrali kao suplementarne točke.



Slika 2.9: Rastav inercije na potprostore

Gornji shematski dijagram prikazuje dekompoziciju inercije za naš primjer između restringiranog i nerestringiranog prostora, pri čemu su linijom odijeljeni dijelovi inercije koji su objašnjeni u dvodimenzionalnom prikazu odnosno dijelovi inercije koji nisu objašnjeni. Vidimo da CCA objašnjava manje ukupne inercije nego CA zato što traži rješenje u restringiranom potprostoru ali pritom objašnjava veći udio inercije u potprostoru koji je od posebnog interesa.

Sada slijedi triplot rezultata kanonske analize korespodencije. Naziva se triplot jer prikazuje dva osnovna skupa podataka kao biplot te dodatni treći skup podataka sačinjen od eksplanatornih varijabli. U ovom triplota možemo jasno vidjeti kako su ”Žitarice” i ”Tost i čaj” česta vrsta doručka, ”Slanina i jaja” se jedu ponekad a ostale vrste doručka rijetko.



Slika 2.10: CCA triplot

2.4 Priprema podataka

U prethodnim poglavljima smo objasnili kako obraditi podatke iz jedne ili više frekvencijskih tablica. Sada ćemo pokazati kako se ostale vrste podataka mogu rekodirati ili transformirati na takav način da možemo primjeniti analizu korespondencije kako bi ih vizualizirali. Tipovi podataka koje ćemo ovdje obraditi su ljestvice ocjena, usporedbe u parovima i rangiranje neprekidnih podataka.

Krenimo prvo s ljestvicom ocjena:

☐	☐	☐	☐	☐
Izrazito se slažem	Slažem se	Niti se slažem niti se ne slažem	Ne slažem se	Izrazito se ne slažem

Metoda za rekodiranje ljestvice ocjena je *dupliranje*. Ideja dupliranja je prebrojiti koliko ima boljih odnosno lošijih ocjena na skali. To ćemo napraviti tako što ćemo redefinirati svaku ocjenu na ljestvici u par ocjena na komplementarnim ljestvicama, pri čemu ćemo jednu ocjenu označiti s ”-” a drugu s ”+”. Na primjer, podatke koji se nalaze na ljestvici 1 - 5 ćemo prebaciti na novu ljestvicu 0 - 4 tako što ćemo svaku ocjenu smanjiti za 1. Nove ocjene na ljestvici 0-4 zovemo ”negativne” ocjene. Komplementarnu ljestvicu

s "pozitivnim" ocjenama ćemo dobiti tako što ćemo od maksimuma, u našem slučaju 5, oduzeti ocjenu na početnoj ljestvici. Pokažimo na sljedećem primjeru jedne ankete kako bi izgledao navedeni postupak dupliranja ocjena:

Primjer 2.4.1. Pretpostavimo da je ovo tablica odgovora na 4 pitanja petorice ispitanika s mogućim odgovorima na ljestvici 1 - 5. Na tablici se vide njihovi odgovori i redefinirane vrijednosti:

Tablica 2.8: Primjer redefiniranja odgovora u anketi

Pitanja				*Pitanje A*		*Pitanje B*		*Pitanje C*		*Pitanje D*	
A	*B*	*C*	*D*	*A*−	*A*+	*B*−	*B*+	*C*−	*C*+	*D*−	*D*+
2	3	4	3	1	3	2	2	3	1	2	2
3	4	2	3	2	2	3	1	1	3	2	2
2	3	2	4	1	3	2	2	1	3	3	1
2	2	2	2	1	3	1	3	1	3	1	3
3	3	3	3	2	2	2	2	2	2	2	2

Sada ćemo pokazati kako redefinirati *usporedbe u parovima*:

Usporedbe u parovima su slobodnija forma preferencijskog rangiranja. Na primjer, svaki od 15 mogućih parova od 6 proizvoda A,B,C,D,E i F je predstavljen ispitaniku koji odabire onaj proizvod koji više preferira u tom paru. Duplirani podaci za svakog ispitanika se formiraju ovako:

$A+$: broj puta koliko je proizvod A preferiran u odnosu na ostale proizvode

$A-$: broj puta koliko su drugi proizvodi preferirani u odnosu na $A(= 5 - A+)$

Za ostale proizvode analogno dobijemo duplirane podatke.

Za kraj, pokazat ćemo primjer s neprekidnim podacima koji objedinjuje gornje dva. Naime, kod neprekidnih podataka prvo trebamo rangirati podatke tako da onome s najvećom (najboljom) vrijednosti dodijelimo prvo mjesto a onom s najmanjom (najlošijom) posljednje mjesto. Ako imamo objekte koji dijele mjesto, recimo treće mjesto tada im dodijelimo vrijednosti 3.5 kao prosjek trećeg i četvrtog mjesta. Na idućem primjeru ćemo pokazati makroekonomske varijable 10 zemalja Europske unije i obraditi ih osnovnom analizom korespondencije.

Tablica 2.9: Makroekonomski pokazatelji za 10 država članica EU za 2014.godinu

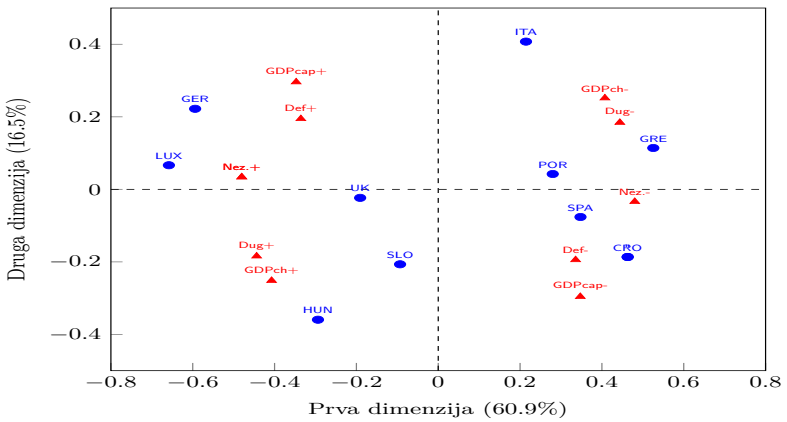
	*O*riginalni podaci					*R*angirani podaci				
Države	*BDP (Δ%)*	*BDP/stan*	*Dug*	*Def*	*Nez*	*BDP (Δ%)*	*BDP/stan*	*Dug*	*Def*	*Nez*
Ujedinjeno Kraljestvo	2.6	39 511	89.4	−5.7	5.4	3.5	3	6	8.5	2
Španjolska	1.4	33 711	97.7	-5.8	22.7	6	5	7	10	9
Slovenija	2.6	29 658	80.9	−4.9	9.3	3.5	6	4	7	5
Portugal	0.9	26 975	130.2	−4.5	13	7	7	8	6	7
Luksemburg	2.8	92 049	23.6	0.6	5.7	2	1	1	2	3
Italija	−0.4	35 486	132.1	−3.0	12.4	9.5	4	9	4	6
Mađarska	3.6	24 942	76.9	−2.6	7.3	1	9	3	3	4
Grčka	0.8	25 859	177.1	−3.5	25.4	8	8	10	5	10
Njemačka	1.6	45 888	74.7	0.7	4.7	5	2	2	1	1
Hrvatska	-0.4	20 889	85	−5.7	17.5	9.5	10	5	8.5	8

BDP (Δ%) = godišnja stopa rasta bruto domaćeg proizvoda u US$, *BDP/stan* = bruto domaći proizvod po glavi stanovnika u US$, *Dug* = omjer javnog duga i BDP-a u postotcima, *Def* = omjer državnog deficita i GDP-a u postotcima, *Nez*= stopa nezaposlenosti u postotcima. Svi brojevi su na godišnjoj razini, važeći za 2014.godinu.

Tablica 2.10: Rekodirana tablica 2.9

	Duplirani rangovi									
Države	BDPch-	BDPch+	BDPcap-	BDPcap+	Dug+	Dug-	Def-	Def+	Nez.-	Nez.+
UK	2.5	6.5	2	7	5	4	7.5	1.5	1	8
SPA	5	4	4	5	6	3	9	0	8	1
SLO	2.5	6.5	5	4	3	6	6	3	4	5
POR	6	3	6	3	7	2	5	4	6	3
LUX	1	8	0	9	0	9	1	8	2	7
ITA	8.5	0.5	3	6	8	1	3	6	5	4
HUN	0	9	8	1	2	7	2	7	3	6
GRE	7	2	7	2	9	0	4	5	9	0
GER	4	5	1	8	1	8	0	9	0	9
CRO	8.5	0.5	9	0	4	5	7.5	1.5	7	2

Iznad vidimo tablicu dupliranih rangova a ispod biplot koji je rezultat analize korespondencije tablice 2.10. Primjetimo kako je ukupna inercija objašnjena biplotom je 77,4%. Iz biplota možemo zaključiti kako se makroekonomske varijable formiraju u 2 skupine, u prvoj su BDP po glavi stanovnika, omjer deficita i BDP-a te nezaposlenost a u drugoj stopa rasta BDP-a te omjer javnog duga i BDP-a. Što se tiče zemalja, primjećujemo kako su se formiraju 4 skupine zemalja. U detaljnije analize ekonomskih situacija pojedinih zemljama nećemo ulaziti jer nam je ovaj primjer služio samo kao demostracija rekodiranja podataka. Napomenimo zbog toga što smo analizirali rangove a ne originalnih vrijednosti ova metoda je otporna na outliere.



Slika 2.11: Biplot makroekonomskog primjera

Poglavlje 3

Dodatak

Navodimo kod u R-u korišten u ovom diplomskom radu.

Za poglavlje 1.3 i izračun regresijskog biplota smo koristili sljedeći kod:

```
Y=(Y-mean(Y))/sd(Y)
X[,1]=(X[,1]-mean(X[,1]))/sd(X[,1])
X[,2]=(X[,2]-mean(X[,2]))/sd(X[,2])
B=t((solve(t(X)%*%X))%*%t(X)%*%Y)
Ykapa=X%*%t(B)
plot(X,xlim=c(-2,2),ylim=c(-2,2),xlab="Standardizirana dob (x*)", ylab="
    Standardizirana temperatura (y*)")
arrows(0,0,B[1,1],B[1,2],col='red')
grid(NULL,NULL)
```

U poglavlju 1.5 kod izračuna PCA biplota smo koristili sljedeći kod te naredbu **SVD** za singularnu dekompoziciju:

```
P<-N/sum(N)
rm<-apply(P,1,sum)
cm<-apply(P,2,sum)
Dr<-diag(rm)
Dc<-diag(cm)
Z<- diag(sqrt(1/rm))%*%(as.matrix(P)-rm%*%t(cm))%*%diag(sqrt(1/cm))
SvdZ<-svd(Z)
row<-diag(1/sqrt(rm))%*%SvdZ$u
col<-diag(1/sqrt(cm))%*%SvdZ$v%*%diag((SvdZ$d))
plot(row[,1:2],xlim=c(-3,3),ylim=c(-3,3),xlab="Prva dimenzija",ylab="Druga
    dimenzija",main="PCA biplot - prva varijanta")
points(col[,1:2],col="red")
text(row[,1:2],rownames(N), cex=0.6, pos=4, col="blue")
text(col[,1:2],colnames(N), cex=0.6, pos=4, col="blue")
row<-diag(1/sqrt(rm))%*%SvdZ$u%*%diag((SvdZ$d))
col<-diag(1/sqrt(cm))%*%SvdZ$v
plot(row[,1:2],xlim=c(-3,3),ylim=c(-3,3),xlab="Prva dimenzija",ylab="Druga
    dimenzija",main="PCA biplot - druga varijanta")
points(col[,1:2],col="red")
text(row[,1:2],rownames(N), cex=0.6, pos=4, col="blue")
```

```
text(col[,1:2],colnames(N), cex=0.6, pos=4, col="blue")
```

Za analizu korespodencije postoji posebni R paket naredbi **ca** koji sadrži gotove naredbe za metodu analize korespodencije. Na primjer, prilikom poziva naredbe **ca** za primjer 2.1.1 dobijamo sljedeći prikaz:

```
> ca(N)

 Principal inertias (eigenvalues):
           1       2        3        4
Value      0.13214 0.029186 0.006687 0.001188
Percentage 78.1%   17.25%   3.95%    0.7%


 Rows:
              St         Ug         Os          F         Dr         Ra
Mass    0.090164  0.091075  0.090164  0.091075  0.091075  0.091075
ChiDist 0.300649  0.311253  0.897614  0.215147  0.290462  0.377784
Inertia 0.00815   0.008823  0.072646  0.004216  0.007684  0.012998
Dim. 1  -0.606029 -0.490242 -2.419809 -0.386156 0.772258  1.012323
Dim. 2  0.671016  1.470295  -0.981833 0.396729  0.391154  0.181221

                Vj         Pr         Ob         Zn        Zad
Mass       0.090164  0.091985  0.091075  0.091075  0.091075
ChiDist    0.311275  0.130406  0.379857  0.492477  0.317019
Inertia    0.008736  0.001564  0.013141  0.022089  0.009153
Dim. 1     0.304804  0.325568  0.867852  1.211872  -0.622908
Dim. 2     1.693266  -0.293244 -1.234394 -1.266053 -1.011401


 Columns:
                1          2          3          4          5
Mass      0.484517  0.340619  0.115665  0.045537  0.013661
ChiDist   0.305570  0.226748  0.646228  0.985754  1.008502
Inertia   0.045241  0.017513  0.048303  0.044249  0.013894
Dim. 1    0.781004 -0.098903 -1.726718 -2.524064 -2.200548
Dim. 2   -0.658524  1.293807 -0.067193 -1.955590 -1.815734
```

Prvo vidimo silazno sortiran glavne inercije kojih ima ukupno 4 zato što je i tablica 1.1 ranga 4. Odavde možemo isčitati doprinos svake dimenzije ukupnoj inerciji. U nastavku koda vidimo sve podatke koji su nam potrebni prilikom analize korespodencije za svaku varijablu odnosno objekt. Na primjer, izjava o stambenom pitanju ima masu jednaku 0.090164, χ^2-udaljenost 0.300649, inerciju 0.00815 te koordinate u dvodimenzionalnom prikazu (-0.606029, 0.671016). Gornji kod je bio izračun simetričnog biplota, međutim kada bi htjeli napraviti asimetričnu mapu trebali bi dodati u funkciju **ca** argument map="rowprincipal" odnosno map="colprincipal". Za suplementarne retke odnosno stupce dodajemo argument suprow=s odnosno supcol=s pri čemu je s vektor indeksa suplementarnih redaka odnosno suplementarnih stupaca tablice.

Za kanonsku analizu korespodencije smo koristili R paket **vegan** i naredbu **cca**. Ova je funkcija za ulazne parametre ima početnu tablicu te suplementarne stupce. Slijedi poziv funkcije za 2.3.1. Primjetimo kako smo transponirali početnu tablicu i suplementarne retke kako bi prilagodili ulazne parametre funkciji **cca**.

```
> cca(t(N),t(X[c(1,3,5),]))
Call: cca(X = t(N), Y = t(X[c(1, 3, 5), ]))

              Inertia Proportion Rank
Total          0.3678     1.0000
Constrained    0.2464     0.6699    3
Unconstrained  0.1214     0.3301    4
Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:
   CCA1    CCA2    CCA3
0.17568 0.05879 0.01192

Eigenvalues for unconstrained axes:
    CA1     CA2     CA3     CA4
0.05966 0.04390 0.01501 0.00284
```

Nakon poziva funkcije vidimo udio inercije u restringiranom i nerestringiranom potprostoru. Za crtanje biplota analize korespodencije ili kanonske analize korespodencije koristimo sljedeće naredbe.

```
plot(ca(N),map = "rowprincipal",arrows=c(FALSE,TRUE))
plot(cca(t(N),t(X[c(1,3,5),])))
```

Za kraj navodimo kod korišten za dupliranje rangova.

```
EUd<-cbind(EU-1,10-EU)
colnames(EUd)<-c(paste(colnames(EU),'-',sep=''),paste(colnames(EU),'+',sep=''))
rownames(EUd)<-c("UK","SPA","SLO","POR","LUX","ITA","HUN","GRE","GER","CRO")
```

Bibliografija

[1] M. Greenacre, *Correspondence analysis in practice* , CRC press, Boca Raton, 2007.

[2] M. Greenacre, *Biplots in practice* , Fundación BBVA, Bilbao, 2010.

[3] M. Greenacre, *Theory and applications of correspondence analysis* , Academic Press, London, 1984.

[4] O. Nenadić i M. Greenacre, *Correspondence Analysis in R, with two- and three-dimensional graphics: The ca package*, Journal of Statistical Software, 20 (3), 1984, str. 1-13.

[5] Helmut Spaeth, *Mathematical Algorithms for Linear Regression*, Academic Press, London, 1991, str. 305, dostupno na `http://people.sc.fsu.edu/~jburkardt/datasets/regression/x06.txt` (svibanj 2015.).

[6] Pew Research Center, *Is College Worth it?*, Washington, 2011. dostupno na `http://www.pewsocialtrends.org/files/2011/05/higher-ed-report.pdf` (travanj 2015.).

[7] Mike Bendixen, *A practical guide to the use of correspondence analysis in marketing research*, Marketing Bulletin, 14, 2003, str. 1-15 dostupno na `http://marketing-bulletin.massey.ac.nz/v14/mb_v14_t2_bendixen.pdf` (lipanj 2015.).

[8] Ekonomija Europske unije - Wikipedia, The Free Encyclopedia, `https://en.wikipedia.org/wiki/Economy_of_the_European_Union` (lipanj 2015.).

Sažetak

Cilj ovog rada je predstaviti problem vizualizacije višedimenzionalnih podataka i prikazati rješenje problema pomoću analize korespondencije. Analiza korespondencije je statistička metoda za vizualizaciju kategorijskih podataka koju je uveo Jean-Paul Benzécri šezdesetih godina prethodnog stoljeća. U međuvremenu se analiza korespondencije nametnula kao vrlo efikasna metoda za analizu kategorijskih podataka koja se svakim danom sve više koristi u društvenim i prirodnim znanostima. U prvom poglavlju smo se najprije upoznali s biplotom kao osnovnim grafičkim alatom za vizualizaciju višedimenzionalnih podataka. Uz to, objasnili smo postupak konstrukcije različitih vrsta biplota i preko poopćene metode analize glavne komponente stigli do motivacije za analizu korespodencije. U drugom poglavlju smo detaljno opisali osnovnu analizu korespodencije, demonstrirali metodu na nekoliko primjera te uveli dvije modifikacije osnovne metode. Konačno, u trećem poglavlju se nalaze kodovi korišteni pri obradi podataka za ovaj rad.

Summary

The goal of this thesis was to present the problem of visualisation multidimensional data and provide the solution using correspondence analysis. Correspondence analysis is a statistical technique developed by Jean-Paul Benzécri during the sixties. In the meantime, correspondence analysis has imposed as very effective method for analysis of the categorical data and as such it is used in social and natural sciences more each day. In the first chapter, we introduced the biplot as the basic graphic tool for visulization of multidimensional data. In addition, we described the procedure of biplot construction, explained several types of biplot and using generalised principal component analysis developed motivation for correspondence analysis. In the second chapter, we have described, in detail, simple correspondence analysis and demonstrated method on the several examples. Later in chapter, we introduced two modifications of the simple correspondence analysis. Finaly, in the third chapter, we present the R codes used for data analysis in this thesis.

Životopis

Rođen sam 22. 11. 1991 u Splitu. Početno obrazovanje sam stekao u Osnovnoj školi Blato u Blatu na Korčuli. Po završetku osnovne škole upisao sam III. gimnaziju u Splitu. U srednjoj školi sam se aktivno zainteresirao za matematiku i sudjelovao na nekoliko državnih natjecanja iz matematike. Maturirao sam 2010. godine te sam u istoj godini upisao preddiplomski studij Matematike na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta. Neposredno nakon završetka preddiplomskog studija 2013. godine sam na istom fakultetu upisao diplomski studij Financijska i Poslovna matematika, kojeg završavam ovim radom.

Na preddiplomskom studiju sam sudjelovao u osnivanju i bio tajnik PMF ogranka neprofitne studentske udruge eSTUDENT iduće dvije godine. Posljednju godinu diplomskog studija sam proveo na praksi u Privrednoj banci Zagreb u Sektoru za upravljanje rizicima gdje sam stečena teorijska znanja primjenjivao u praksi, pri čemu mi je znanje i iskustvo koje sam stekao tijekom prakse pomoglo u izboru profesionalne karijere i dodatno me potaklo na daljnji razvoj u upravljaju rizicima.