BIPLOT

Marta Cota

16. veljače 2018.

1 Osnovna ideja

Biplot kao grafičko predočenje multivarijatnih podataka proizašao je iz potrebe za sve većim i većim količinama podataka, s mnogo varijabli. Trenutna pomama za tzv. "Big data", gdje se vrši statistička obrada velikih količina podataka, rezultirala je potrebom za grafičkim predočenjem silnih podataka na smislen način. Ipak, predočenje multivarijatnih podataka smanjenjem dimenzije problema pokazalo se kompleksnim. Recimo samo kako BIPLOT rješava kompleksan problem na, barem algebarski, relativno jednostavan način. U radu, po uzoru na autora Michaela Greenacre-a, dajemo osnovnu ideju iza biplota, primjere korištenja istog kod različitih tipova podataka te motivaciju za daljnje usavršavanje prikaza multivarijatnih podataka u 2-dimenzionalnom, odnosno 3-dimenzionalnom koordinatnom sustavu.

2 Definicija

Biplot nastaje proširenjem dijagrama raspršenja za dvije varijable početnog uzorka. Prisjetimo se, dijagram raspršenja može nam dati više informacija o koreliranosti nekih dviju varijabli. Istu ideju zagovara i Biplot. Naime, umjesto jednostavnijeg prikaza opaženih podataka pomoću dijagrama raspršenja za svake dvije varijable, biplot vizualizira vezu između vše varijabli.

3 Najjednostavniji primjer - kao na nastavi

U nastavku krećemo sa najjednostavnijim primjerom kao uvodom u biplot, pomoću faktorizacije matrice podataka Y. U nastavku rada pretpostavljamo kako su retci matrice podataka uvijek upravo opaženi uzorci, dok stupci nose promatrane karakteristike za svaki element uzorka. Neka je Y, sa pripadnim

rastavom:

$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ -1 & 1 \\ 1 & -1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 3 & 2 & -1 & -2 \\ 1 & -1 & 2 & -1 \end{pmatrix}$$

Autor M. Greenacre naziva promatrane matrice *matricom cilja, lijevom* odnosno *desnom matricom*. Iz definicije množenja matrica, znamo kako je (i, j) element matrice Y dan skalarnim produktom i-tog retka *lijeve* matrice sa j-tim stupcem *desne*. U svrhu daljnjeg pojašnjenja biplota, zapišimo gornji matrični produkt kao:

$$Y = LR^T$$

Ovdje matrice L i R^T upravo sadrže retke čije skalarne produkte računamo kako bismo došli do elemenata ciljane matrice Y. Dakle, L i R tada su matrice:

$$\begin{pmatrix} l_1^T \\ l_2^T \\ l_3^T \\ l_4^T \\ l_5^T \end{pmatrix}, \begin{pmatrix} r_1^T \\ r_2^T \\ r_3^T \\ r_4^T \end{pmatrix}$$

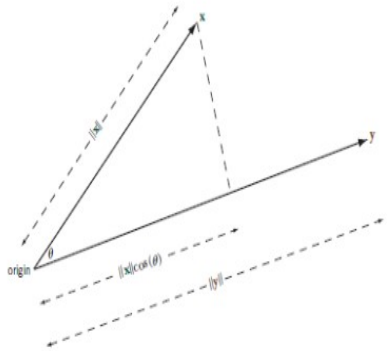
Odgovarajuće vektore dviju matrica ćemo smjestiti u graf, pri čemu vektore *lijeve* matrice konstriramo kao točke, dok elemente *desne* crtamo kao dvodimenzionalne radijvektore. Svaki od radijvektora definira jednu od *osi biplota*, na koje uvijek možemo "spustiti" projekcije točaka lijeve matrice. Kako bismo pobliže objasili vezu izračunatih porjekcija i skalarnog produkta (elemenata ciljne matrice), prisjećamo se definicije projekcije točke na vektor. Na predavanjima smo spomenuli rezultat, uz napomenu kako je θ kut između vektora x, odnosno y:

$$x^T y = \|x\|\|y\| \cos(\theta)$$

Dakle, skalarni produkt između dva vektora upravo je veličina projekcije jednog vektora na drugi, pomnožena sa normom drugog vektora. Kao što smo na nastavi i spominjali, tada iznosi za svaku od varijabli za svaki redak matrice Y postaju usporedive.

Točke lijeve matrice (retke) zvat ćemo baš *točkama*, dok ćemo stupce desne matrice crtati kao vektore u grafu, i zvat ćemo ih *vektorima*. Tako za gornji rastav matrice Y konstruiramo graf: Graf smo konstruirali u R-u pomoću naredbi:

```
Y=matrix(c(8,5,-2,2,4,2,0,-3,3,6,2,3,3,-3,-6,-6,-4,1,-1,-2),nrow=5)
L=matrix(c(2,2,1,2,-1,1,1,-1,2,-2),byrow=T,nrow=5)
Rt=matrix(c(3,1,2,-1,-1,2,-2,-1),byrow=T,nrow=4)
biplot(L,Rt,c("blue","red"),var.axes=TRUE)
```



Slika 1: Skalarni produkt

Vektori konstruirani na grafu predstavljaju osi biplota, pomoću kojih možemo dati interpretacije udaljenosti svakog od elementa uzorka do pojedinih osi. Primijetimo, u prethodnome primjeru sami smo unijeli matrice iz faktorizacije *ciljne* matrice. R omogućava, pomoću naredbe **biplot**, kreiranje biplota temeljem SVD dekompozicije ciljne matrice.

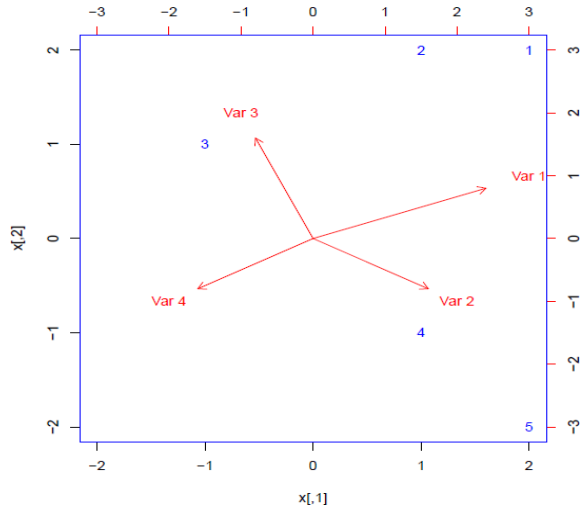
Iako se to, općenito, ne predočava na grafu, *kalibracijom* svakih od osi biplota mogli bismo dobiti ideju o veličini jedne jedinice svake od promatranih varijabli, danu invertiranjem:

$$u_i = \frac{1}{\|y_i\|}, i = 1, \ldots, 4$$

Zaista, projekcijom svake od točaka biplota na, primjerice, varijablu y_1, pomnožili smo, iz gornje formule, sve veličine sa $\|y\|$. Konačno, dobivamo "ponašanje" svakih od opaženih vrijednosti s obzirom na varijablu.

U slučaju u kojem, primjerice, dvije osi biplota leže na istom pravcu, zaključujemo vrlo jaku koreliranost varijabli poticaja za opažene uzorke. Naime, uzorci imaju vrlo slične relativne pozicije u biplotu, s obzirom na varijable - za te se uzorke dane varijable pokazuju vrlo koreliranima.

Dakako, nećemo uvijek moći dobiti ovako jednostavne rastave *ciljne* matrice. Dimenzije lijeve i desne matrice iz matrične faktorizacije kao gore ovisit će o rangu ciljne matrice. U nastavku dajemo osnove SVD metode, kao metode za aproksimaciju startne matrice matricom nižeg ranga. Sjetimo se, ipak bismo htjeli moći konstruirati biplot - htjeli bismo dimenzije 2 i 3.



Slika 2: biplot u R-u, matrica Y

4 Matrice većih dimenzija

Kod matrica opažanja većih dimenzija htjeli bismo udaljenosti među retcima matrice u visokoj dimenziji (udaljenosti mogu definirane standardnom Euklidskom udaljenošću) aproksimirati udaljenostima u nižoj dimenziji, na nekakav optimalan način. Upravo tim pristupom koristimo se kod višedimenzionalnog skaliranja (vidi [1]).

Zadavanjem ciljne funkcije pri optimizaciji dobivamo različite metode pomoću kojih možemo reducirati dimenziju problema. Naša formulacija problema za rješenje ima jedan od fundamentalnih rezultata linearne algebre, SVD dekompoziciju matrice. Prisjetimo se, *rang* r upravo je najmanji mogući broj redaka (stupaca) matrice pomoću kojih se linearnim kombinacijama postižu preostali retci (stupci) početne matrice.

Neka je Y matrica dimenzija nxm, ranga r. Tada je matrična aproksimacija matrice Y dana matricom $\hat{Y} \in M_{nm}$ ranga $p < r$, tako da "najviše sliči" početnoj matrici. "Najbližu" matricu mi ćemo upravo tražiti minimizacijom kvadratne greške, odnosno, u matričnim zapisu, minimizacijom po svim mogućim matricama ranga $p < r$:

$$\min_{\mathbf{r}(\hat{Y})=p<r} \mathbf{tr}((Y-\hat{Y})(Y-\hat{Y})^\intercal)$$

Primijetimo, funkcija koju pritom minimiziramo možemo zapisati po elementima:

$$\mathbf{tr}((Y-\hat{Y})(Y-\hat{Y})^\intercal) = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\hat{y}_{ij})^2$$

Budući da smo za funckiju cilja odabrali upravo gornju funkciju, rješenje ove minimizacije upravo je SVD dekompozicija matrice Y.

Teorem 1. *Neka je A realna matrica reda $m \times n$. Tada se A može dekomponirati kao:*

$$A = U\Sigma V^\intercal, \quad \Sigma = \begin{bmatrix} \Sigma \\ 0 \end{bmatrix}$$

pri čemu vrijedi $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$.

Teorem smo iskazali uz pretpostavku da je matrica A punog stupčanog ranga n. Prvih n vektora stupaca nazivamo *lijevi singularni vektori*, dok prvih m stupaca matrice $V^\intercal$ nazivamo *desnim singularnim vektorima*. Za matricu koja nije punog stupčanog ranga dekompozicija ide analogno: Za $\mathbf{r}(Y) = r$ U i V su ortonormalne matrice reda $n \times r$, odnosno $m \times r$. Matrica Σ je dijagonalna sa *singularnim vrijednostima* $\lambda_1, \ldots, \lambda_r$.

Dakle, ukoliko *ciljna matrica* niskog ranga (2 ili 3), dopušta faktorizaciju tipa $L \times R$ kao u početnom primjeru, koristimo se SVD dekompozicijom matrice kako bismo konstruirali dvodimenzionalni ili trodimenzionalni biplot.
Tada, za SVD dekompoziciju

$$Y = U\Sigma V^\intercal$$

će upravo odabir matrice R, odnosno L na jedan od tri načina dana u nastavku dati biplote, do na faktor slične (predavanja). Odabir faktorizacije vršimo na jedan od 3 načina:

1. $R = \Sigma V^\intercal$
2. $L = U\Sigma$
3. $L = U\Sigma^{\frac{1}{2}}, \quad R = \Sigma^{\frac{1}{2}} V^\intercal$

U praksi se nećemo susretati sa matricama opažanja ovako niskoga ranga. U tu svrhu koristimo SVD dekompoziciju za aproksimaciju početne matrice matricom niskog ranga. Tada, iz SVD dekompozicije, za aproksimaciju u p-dimenzionalnom prostoru, odabiremo prvih p stupaca matruce U iz SVD dekompozicije, Σ_p tada je dijagonalna matrica sa prvih p singularnih vrijednosti dobivenih iz dekompozicije. Također, odabiremo i prvih p stupaca matrice V.

Konačno, aproksimacija $\hat{Y}$ rješenje je gornjeg minimizacijskog problema:

$$\hat{Y} = U_p \Sigma_p V_p^\intercal$$

Aproksimacija je ponovno u formi koja nam lako daje odabir faktorizacije na lijevu i desnu matricu. Singularne vrijednosti daju nam informaciju o udaljenosti aproksimacijske matrice $\hat{Y}$ do početne matrice.

Vrijedi:

$$\mathbf{tr}(YY^\intercal) = \sum_{i=1}^{r} \lambda_i^2$$

$$\mathbf{tr}(\hat{Y}\hat{Y}^\intercal) = \sum_{j=1}^{p} \lambda_j^2.$$

Omjer gornjih vrijednosti dat će nam kvalitetu aproksimacije. Vidi([2]).
Vratimo se matrici Y iz prvog odjeljka. Računamo SVD dekompoziciju u R-u:

```
Y=matrix(c(8,5,-2,2,4,2,0,-3,3,6,2,3,3,-3,-6,-6,-4,1,-1,-2),nrow=5)
svd(Y)
```

```
> Y <- matrix(c(8,5,-2,2,4,2,0,-3,3,6,2,3,3,-3,-6,-6,-4,1,-1,-2),nrow=5)
> svd(Y)
$d
[1] 1.412505e+01 9.822577e+00 1.376116e-15 7.435554e-32

$u
           [,1]       [,2]        [,3]          [,4]
[1,] -0.6634255 -0.4574027 -0.59215653  1.121918e-16
[2,] -0.3641420 -0.4939878  0.78954203 -1.418279e-16
[3,]  0.2668543 -0.3018716 -0.06579517 -9.128709e-01
[4,] -0.2668543  0.3018716  0.06579517 -1.825742e-01
[5,] -0.5337085  0.6037432  0.13159034 -3.651484e-01

$v
           [,1]       [,2]       [,3]       [,4]
[1,] -0.7313508 -0.2551980 -0.6232141 -0.1077229
[2,] -0.4339970  0.4600507  0.4320449 -0.6429131
[3,]  0.1687853 -0.7971898  0.2209931 -0.5358750
[4,]  0.4982812  0.2961685 -0.6132728 -0.5365599
```

Dakle, iz outputa iščitavamo kako je matrica *numeričkog* ranga 2, i možemo ju faktorizirati:

$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} -0.6634 & -0.4574 \\ -0.3641 & -0.4939 \\ 0.2668 & -0.3018 \\ -0.2668 & 0.3018 \\ -0.5337 & 0.6037 \end{pmatrix} \begin{pmatrix} 14.1251 & 0 \\ 0 & 9.8226 \end{pmatrix}$$

$$\begin{pmatrix} -0.7313 & -0.4339 & 0.1687 & 0.4982 \\ -0.2551 & 0.4600 & -0.7971 & 0.2961 \end{pmatrix}$$

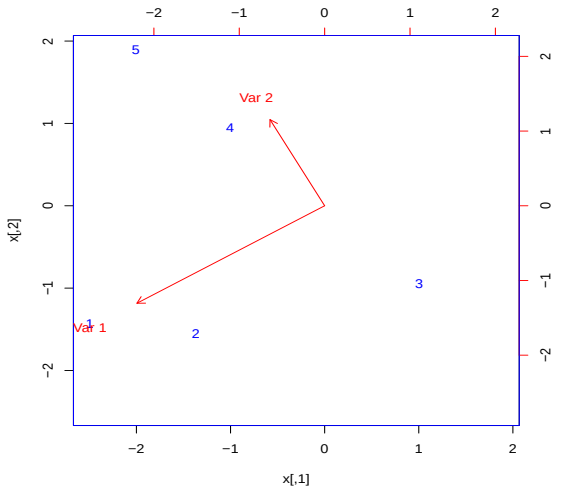
Sada, kako bismo definirali lijevu i desnu matricu faktorizacije, koristimo jedan od tri načina opisanih gore. Ostajemo konzistentni sa [1], te faktorizaciju provodimo na treći način definiran prethodno.

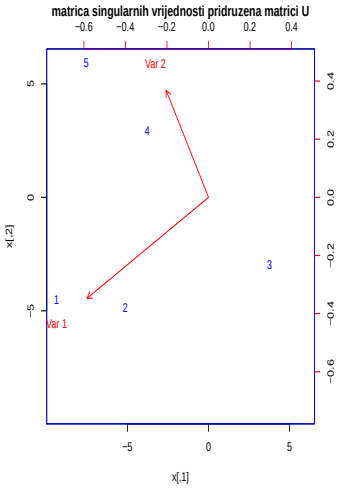
```
svd(Y)
d_p=matrix(c(14.1251,0,0,9.8226),byrow=T,nrow=2)
u_p=svd(Y)$u[,1:2]
v_p=svd(Y)$v[,1:2]
d=sqrt(d_p)
L=u_p%*%d
R=d%*%t(v_p)
```

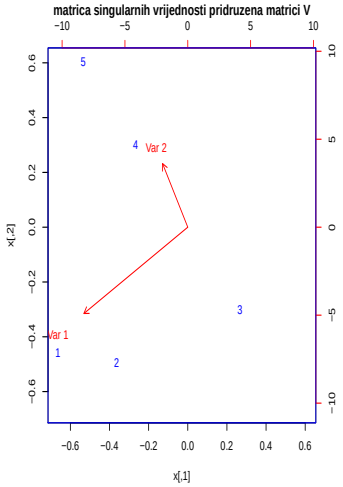
Dobivamo faktorizaciju:

$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} -2.4933 & -1.4335 \\ -1.3685 & -1.5482 \\ 1.0029 & -0.9460 \\ -1.0029 & 0.9460 \\ -2.0058 & 1.8921 \end{pmatrix} \begin{pmatrix} -2.7486 & -1.6311 & 0.6343 & 1.8727 \\ -0.7998 & 1.4418 & -2.4984 & 0.9282 \end{pmatrix}$$

Provodimo i ostale načine faktorizacije kako bismo usporedili biplote.







5 Generalizirana aproksimacija pomoću SVD-a

Osim matrice podataka Y opisane kao do sada, u praksi se pojavljuju i matrice u kojima su varijablama (stupcima matrice opažanja) ili pak elementima uzorka (retcima *ciljne matrice*) pridodane odgovarajuće težine. Primjerice, u anketama veća se težina može pridodati osobama muškog spola, ukoliko nije skupljen dostatan broj odgovora za muški dio populacije.
Također, nekim varijablama se čak po defaultu može pridodati manja težina, primjerice, zbog velike varijance vektora odziva za tu varijablu i slično. Prema tome, opisujemo metodu aproksimacije, koja će pridodati težine pri određivanju najbliže matrice $\hat{Y}$.
Uz pretpostavljenje težine $\omega_1, \omega_2, \ldots, \omega_n$ na retcima, odnosno $q_1, q_2, \ldots, q_n$ na stupcima matrice Y, minimizacija koju sada provodimo je:

$$min_{\mathbf{r}(\hat{Y})=p<r}\mathbf{tr}(D_\omega(Y-\hat{Y})D_q(Y-\hat{Y})^\intercal)$$

Gornji trag možemo ponovno zapisati po komponentama:

$$\mathbf{tr}(D_\omega(Y-\hat{Y})D_q(Y-\hat{Y})^\intercal) = \sum_{i=1}^{n}\sum_{j=1}^{m}\omega_i q_j (y_{ij}-\hat{y}_{ij})^2$$

Rješenje ovog problema tada opisujemo u tri koraka, transformacijom početne matrice Y pomoću vektora težina.

$$\begin{aligned}
&1.) \quad S := D_\omega^{\frac{1}{2}} Y D_q^{\frac{1}{2}} \\
&2.) \quad S = U D_\beta V^\intercal \\
&3.) \quad \tilde{U} := D_\omega^{\frac{-1}{2}} U, \quad \tilde{V} := D_q^{\frac{-1}{2}} V
\end{aligned}$$

Sada je matrična aproksimacija ranga p oblika:

$$\hat{Y} = \tilde{U} D_{\beta[p]} \tilde{V}.$$

Stupci matrica $\tilde{U}, \tilde{V}$ sada su ortonormalni s obzirom na skalarni produkt induciran težinama definiranim na početku. Faktorizaciju matrice aproksimacije provodimo na jedan od tri načina koji smo spomenuli na predavanjima.

5.1 Generalizirana metoda glavnih komponenti

Jedna od primjena generalizirane SVD metode upravo se javlja u generaliziranoj metodi glavnih komponenata.

Uzmimo matricu $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\intercal \\ \vdots \\ \mathbf{x}_n^\intercal \end{pmatrix}$. Neka retci matrice $\mathbf{X}$ definiraju n točaka m-dimenzionalnog prostora. Pretpotavimo da su točkama pridružene težine pomoću vektora $\mathbf{w}$, takvog da je $\mathbf{w1} = 1$. Neka je udaljenost između vektora u m-dimenzionalnom prostoru također definirana pomoću vektora težina, $\mathbf{q}$. Dakle, udaljenost među retcima matrice je:

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\intercal D_q (\mathbf{x}_i - \mathbf{x}_j).$$

Zadatak je naći pogodnu, nižedimenzionalnu aproksimaciju vektora redaka koja bi bila najbliža početnim vektorima u smislu gornje definiranih udaljenosti. Naime, uz definiciju:

$$\mathbf{Y} := (\mathbf{X} - \mathbf{1wX})$$

provodimo minimizacijski problem:

$$\min_{\mathbf{r}(\hat{Y})=p<r} \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^\intercal D_q (y_i - \hat{y}_i)$$

Dobivena aproksimacija $\hat{\mathbf{Y}}$ takva je da su njeni retci najbliži retcima matrice $\mathbf{Y}$, s obzirom na težinsku sumu kvadriranih udaljenosti po komponentama. Koordinate p-dimenzionalnih (centraliziranih) redaka sada su, po prethodno opisanoj, generaliziranoj SVD metodi upravo:

$$F := \tilde{U} D_{\beta[p]}$$

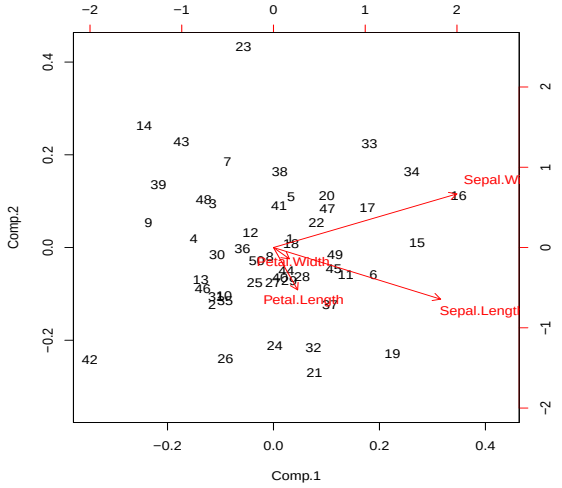
Koordinate nazivamo *glavnim komponentama* redaka. Primijetimo, tada upravo glavne komponente definiraju lijevu matricu u faktorizaciji, i čine točke biplota. Elementi desne, dani stupcima matrice $\tilde{V}$, čine osi biplota.

Što se tiče odabira načina faktorizacije matrice opažanja, razlikujemo biplote:

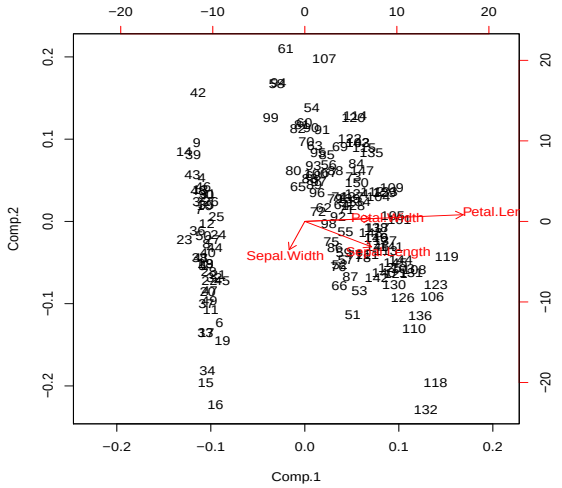
- one koordinate kojima nisu pridružene same singularne vrijednosti nazivamo *standardnim koordinatama*, dok su one koje odabiremo upravo *glavne koordinate*. U gornjem smo primjeru retke definirali kao glavne.

Za primjer biplota kod korištenja metode glavnih komponenata, uzimamo podatke dostupne u R-ovoj biblioteci, iris. Podaci uključuju mjerenja latica 150 cvjetova irisa, opisanih u R-u:

```
require(stats)
iris
irispca<-princomp(iris[1:50,-5]);
summary(irispca);
screeplot(irispca);
pdf("biplot5.pdf")
biplot(irispca,1:2,scale=1);
dev.off()
```



Prvi biplot napravljen je za prvih 50 opažanja, dok je drugi za sva opažanja (150 njih).



6 Literatura

1.Biplots in Practice, Michael Greenacre, 2010.
2.http://www.multivariatestatistics.org/biplots.html
3. R-ova dokumentacija *online*, https://stat.ethz.ch/R-manual/R-devel/library/stats/html/biplot.princomp.html
4. Predavanja prof. Singera iz Numeričke matematike, o SVD dekompoziciji: https://web.math.pmf.unizg.hr/~singer/num_anal/NA_0910/25.pdf
5. Korisni podaci za vježbu: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html
6. Bilješke s predavanja prof. Huzaka, kolegij Primijenjena statistika: http://degiorgi.math.hr/forum/viewtopic.php?t=20438
7. Biplots - The joy of SVD, Michael Greenacre, https://onlinelibrary.wiley.com/doi/full/10.1002/wics.1200