Contents

A reference for much of this material can be found in Chapter 8 of Härdle and Simar [2015]. See also Gabriel [1978] for some material on the AMMI model.

---

1 Motivation

To some extent, PCA is focused on the eigendecomposition of the matrix $\mathbf{S} = \mathbf{Y}^\top\mathbf{Y}$. The decomposition is

$$\mathbf{S} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top,$$

where $\mathbf{V}$ and $\mathbf{\Lambda}$ are $p \times p$ matrices (orthogonal and diagonal). By restricting attention to $\mathbf{S}$, we

- are investigating associations among the *columns* of $\mathbf{Y}$;

- are ignoring associations among the *rows* of $\mathbf{Y}$.

When does ignoring row associations make sense? Perhaps when there are no row associations This situation might arise if the rows are data from subjects that constitute a simple random sample from a population.

As we will see later, under the model that the rows of $\mathbf{Y}$ satisfy $\mathbf{y}_1, \ldots \mathbf{y}_n \sim$ i.i.d. $N_p(\boldsymbol{\mu}, \Sigma)$, then the elements of each principal component vector are exchangeable, and any associations among the rows of $\mathbf{Y}$ are due to noise.

However, in many cases the rows are not a simple random sample, and instead are a convenience sample, or data that were gathered under multiple conditions or from multiple populations. In such cases, understanding associations among the rows of the data matrix may be informative.

How can we describe such associations quantitatively? Based on our study of PCA, two possibilities come to mind:

- Assess row association with the PCs $\mathbf{F} = \mathbf{YV}$.
- Alternatively, just do PCA on $\mathbf{Y}^\top$.

Not surprisingly, both approaches amount to more or less the same thing, and are related to something called the *singular value decomposition* (SVD) of a matrix.

2 Ad-hoc derivation

Consider analyzing various developmental indices of the countries in the world:

```
dim(Y)

## [1] 132  20


dimnames(Y)[[1]]

##   [1] "AFG" "AGO" "ALB" "ARG" "ARM" "AUS" "AUT" "AZE" "BDI" "BEL" "BEN"
##  [12] "BFA" "BGD" "BGR" "BIH" "BLR" "BLZ" "BOL" "BRA" "BWA" "CAF" "CHE"
##  [23] "CHL" "CHN" "CMR" "COD" "COG" "COL" "CRI" "CUB" "CYP" "CZE" "DNK"
##  [34] "DOM" "DZA" "ECU" "EGY" "ERI" "ESP" "EST" "ETH" "FIN" "FRA" "GBR"
##  [45] "GEO" "GHA" "GIN" "GMB" "GRC" "GTM" "GUY" "HND" "HRV" "HUN" "IDN"
##  [56] "IND" "IRL" "IRN" "ISL" "ISR" "ITA" "JOR" "JPN" "KAZ" "KEN" "KGZ"
##  [67] "KHM" "KOR" "LBN" "LBR" "LKA" "LSO" "LTU" "LUX" "LVA" "MAR" "MDA"
##  [78] "MDG" "MEX" "MKD" "MLI" "MLT" "MNE" "MNG" "MOZ" "MRT" "MWI" "NAM"
##  [89] "NER" "NLD" "NOR" "NPL" "NZL" "OMN" "PAK" "PAN" "PER" "PHL" "POL"
## [100] "PRT" "PRY" "QAT" "ROU" "RUS" "RWA" "SAU" "SDN" "SEN" "SLV" "SRB"
## [111] "SVK" "SVN" "SWE" "SWZ" "SYR" "TCD" "TGO" "THA" "TJK" "TLS" "TUN"
## [122] "TUR" "TZA" "UGA" "UKR" "URY" "USA" "UZB" "VEN" "YEM" "ZAF" "ZWE"


substring(dimnames(Y)[[2]],first=1,last=65)

##  [1] "Adolescent fertility rate (births per 1,000 women ages 15-19)"
##  [2] "Labor force participation rate for ages 15-24, female (%) (modele"
##  [3] "Labor force participation rate for ages 15-24, male (%) (modeled "
##  [4] "Labor force participation rate for ages 15-24, total (%) (modeled"
##  [5] "Labor force participation rate, female (% of female population ag"
##  [6] "Labor force participation rate, female (% of female population ag"
##  [7] "Labor force participation rate, male (% of male population ages 1"
##  [8] "Labor force participation rate, male (% of male population ages 1"
##  [9] "Labor force participation rate, total (% of total population ages"
## [10] "Life expectancy at birth, female (years)"
## [11] "Life expectancy at birth, male (years)"
## [12] "Proportion of seats held by women in national parliaments (%)"
## [13] "Refugee population by country or territory of asylum"
## [14] "Refugee population by country or territory of origin"
## [15] "School enrollment, primary (gross), gender parity index (GPI)"
```

```
## [16] "School enrollment, primary and secondary (gross), gender parity i"
## [17] "School enrollment, secondary (gross), gender parity index (GPI)"
## [18] "School enrollment, tertiary (gross), gender parity index (GPI)"
## [19] "Unemployment, female (% of female labor force) (modeled ILO estim"
## [20] "Unemployment, male (% of male labor force) (modeled ILO estimate)"
```

I have already centered and scaled the columns of this matrix to have sample means and variances equal to zero and one, respectively. Let's do a PCA:

```
S<-crossprod(Y)
eS<-eigen(S)
V<-eS$vec
L<-eS$val

cbind(L,L/sum(L),cumsum(L)/sum(L))
```

```
##              L
##  [1,] 941.1255 0.3592082 0.359
##  [2,] 580.5128 0.2215698 0.581
##  [3,] 266.1467 0.1015827 0.682
##  [4,] 216.1152 0.0824867 0.765
##  [5,] 152.0757 0.0580442 0.823
##  [6,] 121.3734 0.0463257 0.869
##  [7,]  89.4300 0.0341336 0.903
##  [8,]  70.7231 0.0269936 0.930
##  [9,]  68.2716 0.0260579 0.956
## [10,]  31.4730 0.0120126 0.968
## [11,]  29.6506 0.0113170 0.980
## [12,]  21.2285 0.0081025 0.988
## [13,]  16.6403 0.0063512 0.994
## [14,]   6.0912 0.0023249 0.997
## [15,]   4.5663 0.0017429 0.998
## [16,]   2.0546 0.0007842 0.999
## [17,]   1.7860 0.0006817 1.000
## [18,]   0.5403 0.0002062 1.000
## [19,]   0.1658 0.0000633 1.000
```

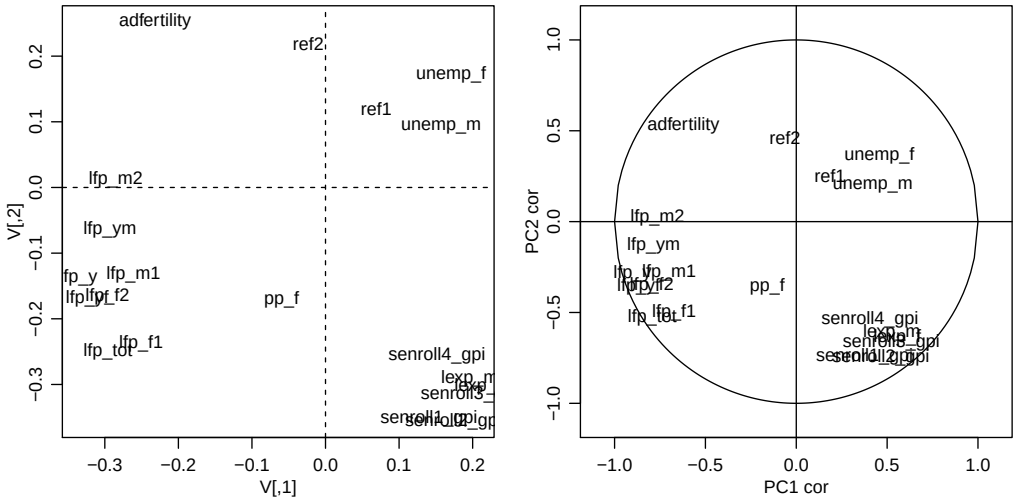
```
## [20,]   0.0297 0.0000113 1.000
```

Over half the variation in the rows can be explained as variation along the first two principal axes. We can get a sense of what these principal axes represent by looking at the entries of $\mathbf{V}$, or alternatively, the correlations between the original variables and the PCs. Recall that this latter correlation is given by

$$\sqrt{n-1}\text{Cor}[\mathbf{Y},\mathbf{F}] = \mathbf{\Psi}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}.$$

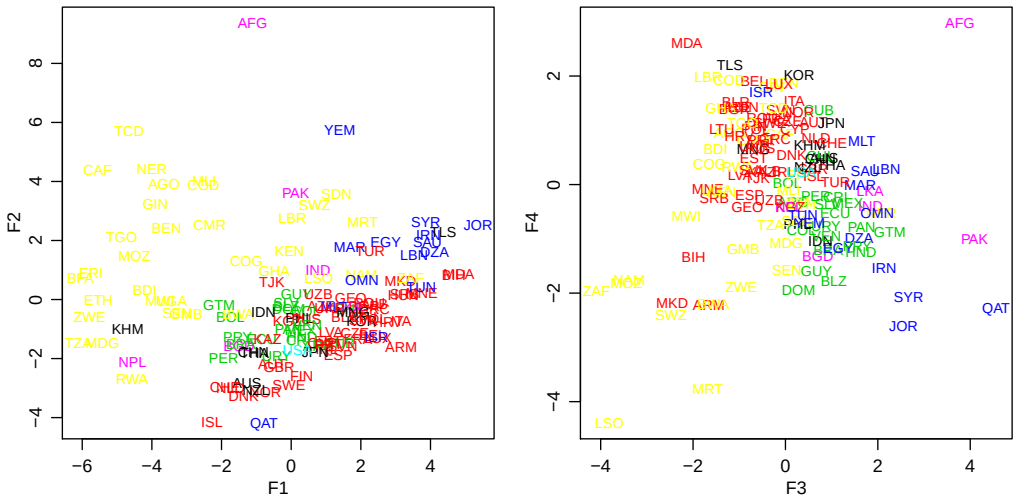
Since these data have already been scaled, $\mathbf{\Psi} = \mathbf{I}$ and so

$$\sqrt{n-1}\text{Cor}[\mathbf{Y},\mathbf{F}] = \mathbf{V}\mathbf{\Lambda}^{1/2}.$$



The first principal axis contrasts gender parity in school enrollment with labor force participation rates. The second principal axis contrasts these variables with unemployment and adolescent fertility.

Now recall that the PCs are given by $\mathbf{F} = \mathbf{YV}$. The interpretation is that $\mathbf{f}_1$ is the standardized linear combination of the columns that has the largest variability among the rows, $\mathbf{f}_2$ has the second largest, and so on. This suggests that we describe heterogeneity among the rows of $\mathbf{Y}$ with the first few PCs.



So the matrix $\mathbf{F} = \mathbf{YV}$ seems helpful in describing differences and similarities among the rows of $\mathbf{Y}$. Now recall the following:

- $\mathbf{f}_j = \mathbf{Y}\mathbf{v}_j$;
- $\mathbf{f}_j^\top \mathbf{f}_j = \lambda_j$;
- $\mathbf{f}_j^\top \mathbf{f}_k = 0$.

So the columns of $\mathbf{f}$ represent uncorrelated/orthogonal derived variables, with (sample) variances given by the λ_j's. If we standardize the columns of $\mathbf{F}$ by their magnitudes, we get a new matrix

$$\mathbf{U} = \mathbf{F}\mathbf{\Lambda}^{-1/2}$$

with jth column $\mathbf{u}_j = \mathbf{f}_j/\sqrt{\lambda_j}$.

Exercise 1. *Show that* $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_p$.

Now let's express $\mathbf{Y}$ in terms of $\mathbf{U}$, $\mathbf{\Lambda}$ and $\mathbf{V}$:

$$\begin{aligned}\mathbf{YV} &= \mathbf{F} \\ \mathbf{Y} &= \mathbf{FV}^\top \\ \mathbf{Y} &= \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}^\top = \mathbf{UDV}^\top,\end{aligned}$$

where $\mathbf{D} = \mathbf{\Lambda}^{1/2}$, a $p \times p$ diagonal matrix with nonnegative entries. Thus we have the following theorem:

Theorem 1. *Let $\mathbf{Y}$ be an $n \times p$ matrix with $p \leq n$. Then $\mathbf{Y}$ can be expressed as*

$$\mathbf{Y} = \mathbf{UDV}^\top,$$

where $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_n$, $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_p$ and $\mathbf{D}$ is a nonnegative diagonal matrix with entries $d_1 \geq \cdots \geq d_p$.

The representation $\mathbf{Y} = \mathbf{UDV}^\top$ is called the singular value decomposition (SVD) of $\mathbf{Y}$.

Now consider a different perspective: The rows of this data matrix are not in any sense a random sample, and we certainly expect there to be non-trivial sample dependencies between the rows.

Suppose we are interested in describing these covariances among the row variables of $\mathbf{Y}$. Previously, we used PCA to describe covariances of the column variables, so obviously one approach to analyzing row dependence is to transpose $\mathbf{Y}$ and use PCA. This begins with computing the eigendecomposition of the "row covariance" $\mathbf{YY}^T$. This is now easy with our new tool, the

SVD:

$$\begin{aligned}\mathbf{Y}\mathbf{Y}^\top &= \mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{V}\mathbf{D}\mathbf{U}^\top \\ &= \mathbf{U}\mathbf{D}^2\mathbf{U}^\top \\ &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top.\end{aligned}$$

This means the following:

- Each column $\mathbf{u}_j$ of $\mathbf{U}$ is an eigenvector of $\mathbf{Y}\mathbf{Y}^T$;
- The eigenvalue associated with $\mathbf{u}_j$ is λ_j;
- The principal axis of variation of *the columns* of $\mathbf{Y}$ are given by $\mathbf{U}$.
- $\mathbf{Y}\mathbf{Y}^\top$ has $n-p$ additional orthogonal unit eigenvectors with a common eigenvalue of zero.

So we see how the SVD provides both

- a description of column associations (within-row variation) via $\mathbf{V}$, and
- a description of row associations (within-column variation) via $\mathbf{U}$.

Here are some additional fun facts about the SVD:

1. The eigendecomposition of a symmetric matrix is its SVD.
2. $\mathbf{Y}\mathbf{v}_j = d_j\mathbf{u}_j$, and $\mathbf{Y}^\top\mathbf{u}_j = d_j\mathbf{v}_j$.
3. Let $\mathbf{M}_j = d_j\mathbf{u}_j\mathbf{v}_j^\top$. Then
   (a) $\mathbf{Y} = \sum_{j=1}^p \mathbf{M}_j$;

(b) $\text{tr}(\mathbf{M}_j^\top \mathbf{M}_j) = d_j^2$;

(c) $\text{tr}(\mathbf{M}_j^\top \mathbf{M}_k) = 0$ for $j \neq k$.

Exercise 2. *Prove the fun facts.*

3 Eckart-Young theorem

Recall that the eigendecomposition of $\mathbf{Y}^\top\mathbf{Y}$ played a role in obtaining low-dimensional approximations of a data matrix. Now we will show how the resulting approximations are easily expressed in terms of the SVD.

Rank-1 matrix: Let $\mathbf{a} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ and $\mathbf{b} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$. Then $\mathbf{M} = \mathbf{a}\mathbf{b}^\top \in \mathbb{R}^{n\times p}$ is a rank-1 matrix.

More formally, the rank of a matrix is the dimension of the linear space spanned by its row vectors, which is the same as the dimension of the linear space spanned by the column vectors.

Exercise 3. *Using the formal definition of rank, show that* $\text{rank}(\mathbf{a}\mathbf{b}^\top) \leq 1$.

In many applications (data compression, denoising, data description and visualization) it is useful to approximate a data matrix $\mathbf{Y}$ by a low rank matrix. Let's first consider approximation by a rank-1 matrix. In this case, we are looking for $\mathbf{a}$ and $\mathbf{b}$ so that

$$\mathbf{Y} \approx \mathbf{a}\mathbf{b}^\top$$

$$y_{i,j} \approx a_i b_j.$$

Note that a rank-1 approximation corresponds to a multiplicative model. This is in contrast to the additive models more commonly used in statistics, such as $y_{i,j} \approx a_i + b_j$.

The best rank-1 approximation by the least-squares criteria is obtained by minimizing $||\mathbf{Y} - \mathbf{a}\mathbf{b}^\top||^2$. In fact, we have already solved this problem when we studied PCA, although in a roundabout way. We will now re-solve it and see how it relates to the SVD.

Theorem 2. $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ *is minimized among rank-1 matrices* $\hat{\mathbf{Y}}$ *by*

$$\hat{\mathbf{Y}} = d_1 \mathbf{u}_1 \mathbf{v}_1^\top$$

where d_1 *is the first singular value and* $\mathbf{u}_1$ *and* $\mathbf{v}_1$ *are the first left and right singular vectors of* $\mathbf{Y}$*.*

To prove this, let $\hat{\mathbf{Y}} = \mathbf{a}\mathbf{b}^\top$ and expand out the least squares criterion:

$$\begin{aligned} ||\mathbf{Y} - \mathbf{a}\mathbf{b}^\top|| &= \mathrm{tr}((\mathbf{Y} - \mathbf{a}\mathbf{b}^\top)^\top(\mathbf{Y} - \mathbf{a}\mathbf{b}^\top)) \\ &= ||\mathbf{Y}||^2 - 2\mathrm{tr}(\mathbf{Y}^\top\mathbf{a}\mathbf{b}^\top) + \mathrm{tr}(\mathbf{b}\mathbf{a}^\top\mathbf{a}\mathbf{b}^\top) \\ &= ||\mathbf{Y}||^2 - 2\mathbf{a}^\top\mathbf{Y}\mathbf{b} + (\mathbf{a}^\top\mathbf{a})(\mathbf{b}^\top\mathbf{b}) \end{aligned}$$

Without loss of generality we can assume $\mathbf{b}^\top\mathbf{b} = 1$. Our criterion is then

$$||\mathbf{Y}^2|| - 2\mathbf{a}^\top\mathbf{Y}\mathbf{b} + \mathbf{a}^\top\mathbf{a}.$$

Let's fix $\mathbf{b}$ and minimize in $\mathbf{a}$. You should be able to see that this is then minimized by $\mathbf{a} = \mathbf{Y}\mathbf{b}$, and the criterion becomes

$$||\mathbf{Y}^2|| - 2\mathbf{b}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{b} + \mathbf{b}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{b} = ||\mathbf{Y}^2|| - \mathbf{b}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{b}.$$

Therefore minimizing $||\mathbf{Y} - \mathbf{a}\mathbf{b}^\top||$ is equivalent to maximizing $\mathbf{b}^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{b}$ among unit vectors $\mathbf{b}$. As we've discussed previously, the maximizer is the unit eigenvector $\mathbf{v}_1$ of $\mathbf{Y}^\top\mathbf{Y}$ that corresponds to the largest eigenvalue λ_1. The resulting value of $\mathbf{a}$ can be obtained from the SVD:

$$\begin{aligned} \mathbf{a} &= \mathbf{Y}\mathbf{v}_1 \\ &= \mathbf{U}\mathbf{D}\mathbf{V}^\top\mathbf{v}_1 \\ &= \mathbf{U}\mathbf{D}\mathbf{e}_1 = d_1\mathbf{u}_1. \end{aligned}$$

Thus the minimizing rank-1 matrix is given by $\hat{\mathbf{Y}} = d_1\mathbf{u}_1\mathbf{v}_1^\top$. The error in this approximation is

$$\begin{aligned} ||\mathbf{Y} - \hat{\mathbf{Y}}||^2 &= ||\mathbf{Y}||^2 - \mathbf{v}_1^\top\mathbf{Y}^\top\mathbf{Y}\mathbf{v}_1 \\ &= \sum_{j=1}^{p} \lambda - \lambda_1 \\ &= \sum_{j=2}^{p} \lambda_j = \sum_{j=2}^{p} d_j^2. \end{aligned}$$

As you can imagine, this result generalizes to higher-dimensional approximations. Using the SVD and the definition of rank, you can show that a matrix is of rank r or less if it can be expressed as the sum of r or fewer rank-1 matrices.

Rank r matrix: A rank r matrix $\mathbf{M}$ has a representation $\mathbf{M} = \sum_{k=1}^{r} \mathbf{a}_j\mathbf{b}_j^\top$.

The best rank-r approximation to a data matrix $\mathbf{Y}$ can be easily obtained from its SVD:

Theorem 3. *(Eckart-Young) Let $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ be the SVD of $\mathbf{Y}$. Then $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ is minimized among rank-r matrices by*

$$\hat{\mathbf{Y}} = \sum_{j=1}^{r} d_j\mathbf{u}_j\mathbf{v}_j^\top \equiv \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^\top.$$

The error in this approximation is

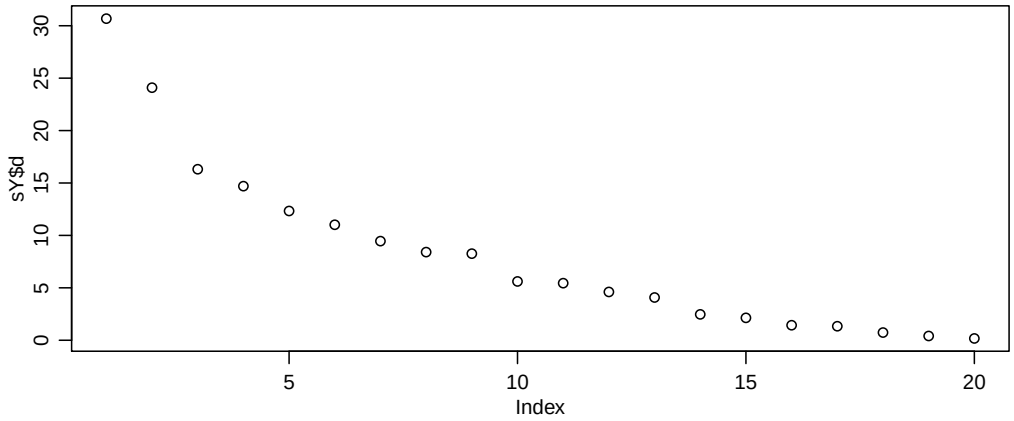
$$||\mathbf{Y} - \hat{\mathbf{Y}}||^2 = \sum_{j=1}^{p} d_j^2 - \sum_{j=1}^{r} d_j^2 = \sum_{j=r+1}^{p} d_j^2.$$

Exercise 4. *Proof Theorem 3.*

Let's illustrate the use of the SVD with the country data described earlier.

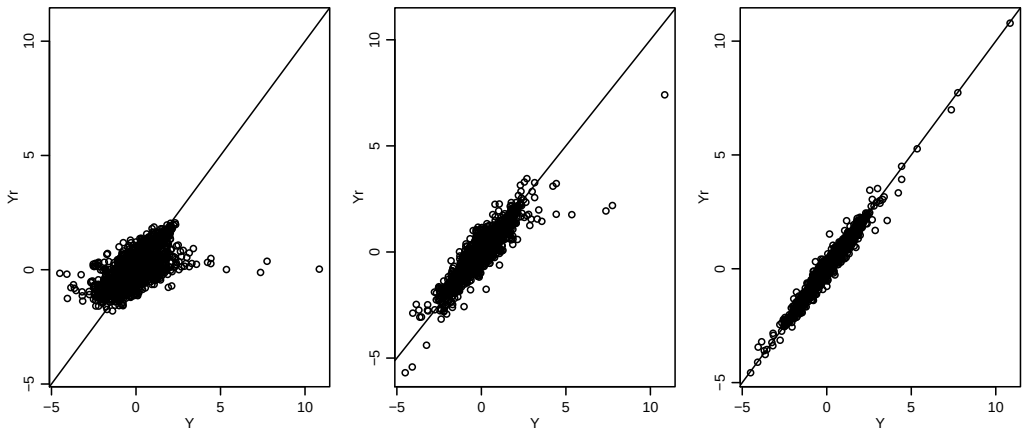
```
sY<-svd(Y)
names(sY)

## [1] "d" "u" "v"
```




```
for(r in c(1,5,10))
{
  Ur<-sY$u[,1:r,drop=FALSE]
  Vr<-sY$v[,1:r,drop=FALSE]
  Dr<-diag(sY$d[1:r],nrow=r)
  Yr<-Ur%*%Dr%*%t(Vr)
  cat(r,sum( (Y-Yr)^2 ) ,"\n" )
  plot(Y,Yr,ylim=range(c(Yr,Y))) ; abline(0,1)
}

## 1 1679
## 5 464
## 10 82.8
```




```
sum(sY$d^2) - cumsum(sY$d^2)

##  [1] 1678.87 1098.36  832.22  616.10  464.02  342.65  253.22  182.50
##  [9]  114.23   82.75   53.10   31.87   15.23    9.14    4.58    2.52
## [17]    0.74    0.20    0.03    0.00
```

Interpretation: At this point, we've discussed three decompositions of the a data matrix $\mathbf{Y}$. Mathematically, they are the same, but their interpretations are a bit different. Let's review:

Eigendecomposition of $\mathbf{Y}^\top\mathbf{Y}$:

- $\mathbf{Y}^\top\mathbf{Y}$ is roughly the "column covariance matrix."
- The first eigenvector $\mathbf{v}_1$ of $\mathbf{Y}^\top\mathbf{Y}$ gives the axis of maximum variation of the n rows of $\mathbf{Y}$ $\mathbf{y}_1, \ldots, \mathbf{y}_n$ in p-dimensional space.
- The best approximation of the rows as points along a single dimension is to project them onto $\mathbf{v}_1$, i.e. $\hat{\mathbf{y}}_i = \mathbf{v}_1\mathbf{v}_1^\top\mathbf{y}_i = d_1 u_{i,1}\mathbf{v}_1 \in \mathbb{R}^p$.

Eigendecomposition of $\mathbf{Y}\mathbf{Y}^\top$:

- $\mathbf{YY}^\top$ is roughly the "row covariance matrix."
- The first eigenvector $\mathbf{u}_1$ of $\mathbf{YY}^\top$ gives the axis of maximum variation of the p columns of $\mathbf{Y}$ $\mathbf{y}_1, \ldots, \mathbf{y}_p$ in n-dimensional space.
- The best approximation of the columns as points along a single dimension is to project them onto $\mathbf{u}_1$, i.e. $\hat{\mathbf{y}}_j = \mathbf{u}_1\mathbf{u}_1^\top\mathbf{y}_j = d_1 v_{j,1}\mathbf{u}_1 \in \mathbb{R}^n$.

SVD of $\mathbf{Y}$:

- The best rank-1 approximation of $\mathbf{Y}$ is given by $d_1\mathbf{u}_1\mathbf{v}_1^\top$, giving $\hat{y}_{i,j} = d_1 u_{i,1} v_{j,1}$.
- The best rank-r approximation of $\mathbf{Y}$ is given by $\sum_{k=1}^r d_k\mathbf{u}_k\mathbf{v}_k^\top$, giving $\hat{y}_{i,j} = \sum_1^r d_k u_{i,k} v_{j,k}$.

4 AMMI models

Yeoh et al. [2002] studied $p = 12625$ gene expression profiles of $n = 248$ biological samples taken from patients with leukemia. The patients can be divided into six different leukemia subtypes.

One way to analyze these data is with PCA, or equivalently, with an SVD of the column-demeaned data. This approach amounts to finding a representation of $\mathbf{Y}$ as

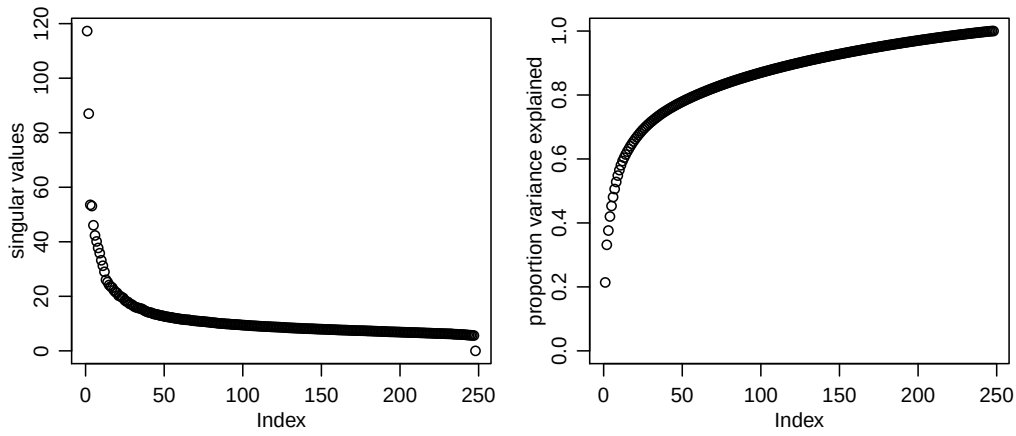
$$\mathbf{Y} \approx \mathbf{1}\boldsymbol{\mu} + \mathbf{AB}^\top.$$

where $\mathbf{A} \in \mathbb{R}^{n\times r}$ and $\mathbf{B} \in \mathbb{R}^{p\times r}$ for some $r < \min(n, p)$. As discussed above, the least squares solution is given by

$$\mathbf{Y} \approx \mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{F}_r\mathbf{V}_r^\top = \mathbf{1}\bar{\mathbf{y}}^\top + \mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^\top.$$

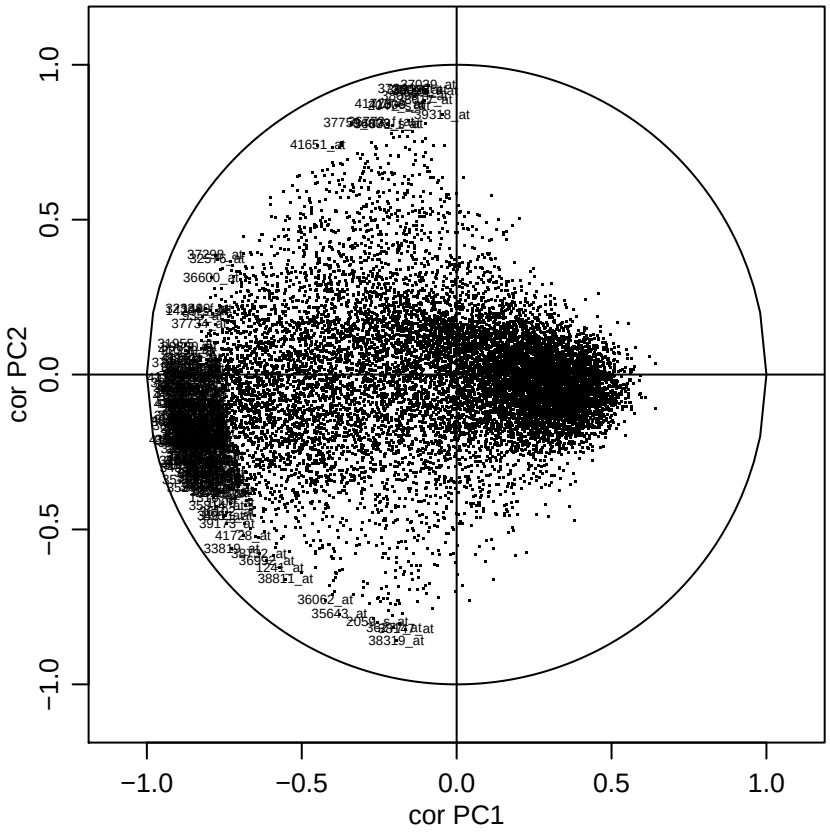
```
CY<-sweep(Y,2,apply(Y,2,mean),"-")
sCY<-svd(CY)
cbind( sCY$d^2, sCY$d^2/sum(sCY$d^2), cumsum(sCY$d^2)/sum(sCY$d^2))[1:10,]

##         [,1]    [,2]   [,3]
##  [1,] 13755 0.21392 0.2139
##  [2,]  7568 0.11770 0.3316
##  [3,]  2868 0.04461 0.3762
##  [4,]  2821 0.04387 0.4201
##  [5,]  2121 0.03299 0.4531
##  [6,]  1791 0.02786 0.4809
##  [7,]  1610 0.02504 0.5060
##  [8,]  1425 0.02215 0.5281
##  [9,]  1282 0.01994 0.5481
## [10,]  1103 0.01716 0.5652
```



So about 50% of the variation in the rows can be explained by the differences along the first seven principal axes. This is not bad - the dimension of the data is $\min(n, p) = 248$.

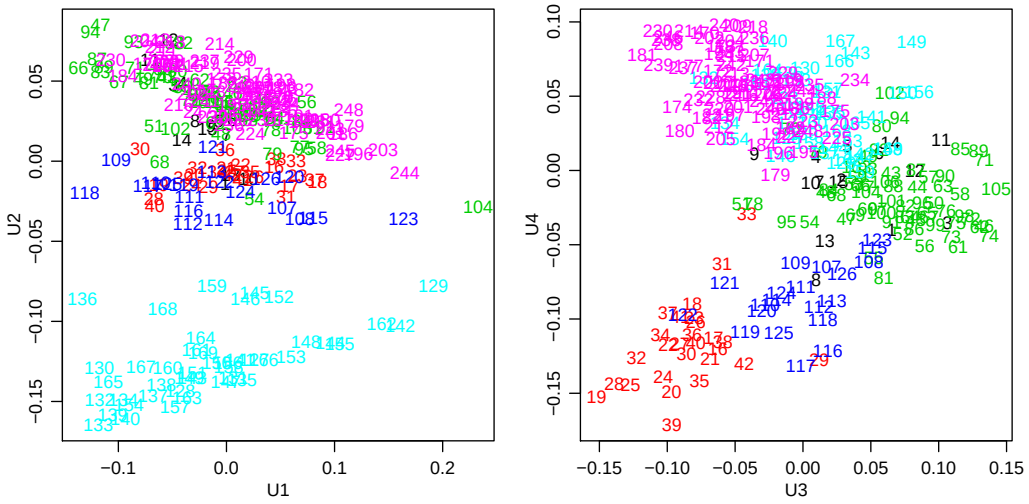
Which variables are well-explained by the first few principal components? We can investigate this with a correlation plot:



The fraction of variance explained by the first two PCs exceeds 70% for 123 of the $p = 12625$ expression profiles (indicated by text in the figure).

Next we examine variation across the samples with the first two left-singular vectors:

```
## [1] -0.3089
```



The six different leukemia subtypes are each plotted with a different color. Notice that the second, third and fourth singular vectors differentiate the groups, whereas the first does not.

That's all well and good, but notice an asymmetry in our analysis - we demeaned the columns, but not the rows. As we will discuss more later, failing to demean the rows makes sense if the rows represent a random or representative sample from some larger population. In this case, the appearance of rows with systematically high or systematically low values can and should be interpreted as positive correlation among the column variables.

However, the rows of the leukemia data do not come from a random sample. Additionally, each row is the result of a multivariate reading of a gene chip. Systematically high or low readings on a gene chip may be viewed as an experimental design artifact and not of scientific interest.

In this case, we may be interested in obtaining a different representation of

the data, of the form

$$\mathbf{Y} \approx \mu \mathbf{1}\mathbf{1}^\top + \mathbf{a}\mathbf{1}^\top + \mathbf{1}\mathbf{b}^\top + \mathbf{A}\mathbf{B}^\top. \tag{1}$$

This "model" or representation can be seen as a generalization of several models we may be familiar with:

- Two-way ANOVA decomposition/additive effects model:

$$\mathbf{Y} \approx \mu \mathbf{1}\mathbf{1}^\top + \mathbf{a}\mathbf{1}^\top + \mathbf{1}\mathbf{b}^\top$$

- Low-rank matrix approximation/multiplicative effects model:

$$\mathbf{Y} \approx \mathbf{A}\mathbf{B}^\top,\ \mathbf{A} \in \mathbb{R}^{n\times r},\ \mathbf{B} \in \mathbb{R}^{p\times r},\ r < \min(n,p)$$

- PCA-style approximation:

$$\mathbf{Y} \approx \mathbf{1}\boldsymbol{\mu} + \mathbf{A}\mathbf{B}^\top.$$

Clearly, the form of the approximation given by (1) generalizes all of these. Representations of this form are called AMMI models, which stands for "additive main effects, multiplicative interactions." This type of model has been used for many decades in in psychometrics, biology and crop science. More recently, models like this have been used for such things as social network analysis and machine learning tasks, such as building recommender systems.

First let's evaluate if there is evidence of any additive row effects. One way to do this is with the standard F-test based on the two-way additive ANOVA decomposition:

```
a<-apply(CY,1,mean)
E<-sweep(CY,1,a,"-")

SSA<-sum(a^2)*p
```

```
SSE<-sum(E^2)

dfA<-n-1
dfE<-(n-1)*(p-1)

MSA<-SSA/dfA
MSE<-SSE/dfE

MSA/MSE

## [1] 318.4


1-pf(MSA/MSE,dfA,dfE)

## [1] 0
```

The test rejects the null of no additive effects. However, keep in mind that the null assumes the errors are independent and identically distributed.

Anyway, let's proceed by finding the least-squares AMMI approximation. How can we do this? Recall our result from our discussion of PCA (using new notation):

Theorem 4. *Let* $\hat{\mathbf{Y}} = \mathbf{1}\boldsymbol{\mu}^\top + \mathbf{A}\mathbf{B}^\top$, *with* $\mathbf{A} \in \mathbb{R}^{n\times r}$ *and* $B \in \mathbb{R}^{p\times r}$. *Then* $||\mathbf{Y} - \hat{\mathbf{Y}}||^2$ *is minimized by*

$$\begin{aligned}
\boldsymbol{\mu} &= \bar{\mathbf{y}} \\
\mathbf{A} &= \mathbf{V}_r(\mathbf{v}_1, \ldots, \mathbf{v}_r), \text{ where } \mathbf{v}_j = evec_j(\mathbf{Y}^\top\mathbf{C}\mathbf{Y}) \\
\mathbf{B} &= \mathbf{C}\mathbf{Y}\mathbf{V}_r = \mathbf{U}_r\mathbf{D}_r.
\end{aligned}$$

Note that $\mathbf{U}_r\mathbf{D}_r\mathbf{V}_r^\top$ *is the best rank-r approximation to* $\mathbf{C}\mathbf{Y}$.

The result says that to find the best approximation of $\mathbf{Y}$, we can first center it

column-wise, and then find the best low-rank approximation of the resulting centered matrix $\mathbf{CY}$.

IMPORTANT: If you estimate $\mathbf{AB}^\top$ from the SVD of $\mathbf{Y}$, then estimate $\boldsymbol{\mu}$ from the resulting residuals, you will not get the least-squares estimates.

How to obtain the best AMMI representation? Intuitively, we might try the following:

1. Estimate μ by the overall average of the entries of $\mathbf{Y}$;
2. Estimate $\mathbf{a}$ by the row means of $\mathbf{Y} - \hat{\mu}\mathbf{1}\mathbf{1}^\top$;
3. Estimate $\mathbf{b}$ by the column means of $\mathbf{Y} - \hat{\mu}\mathbf{1}\mathbf{1}^\top$;
4. Estimate $\mathbf{AB}^\top$ from the SVD of $\mathbf{Y} - (\hat{\mu} + \hat{\mathbf{a}}\mathbf{1}^\top + \mathbf{1}\hat{\mathbf{b}}^\top)$.

Happily, the above procedure does in fact result in the least-squares estimates of $(\mu, \mathbf{a}, \mathbf{b}, \mathbf{AB}^\top)$. This is a result of the following more general result of Gabriel [1978]:

Theorem 5 (Gabriel, 1978)**.** *For fixed* $\mathbf{W}$*,* $\mathbf{X}$ *and* $\mathbf{Y}$*,*

$$\min_{\mathbf{F}} \min_{\mathbf{G}} \min_{\mathbf{AB}^\top} ||\mathbf{Y} - \mathbf{F}\mathbf{W}^\top - \mathbf{X}\mathbf{G}^\top - \mathbf{AB}^\top||^2 = \min_{\mathbf{AB}^\top} ||(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}(\mathbf{I} - \mathbf{P}_W) - \mathbf{AB}^\top||^2.$$

Here, $\mathbf{W}$ is a $p \times q_1$ matrix of "row covariates", $\mathbf{X}$ is an $n \times q_2$ matrix of "column regressors" and $\mathbf{P}_W$ and $\mathbf{P}_X$ are the corresponding projection matrices:

$$\mathbf{P}_X = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$
$$\mathbf{P}_W = \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top$$

The above result says that we can obtain a least-squares fit of $\mathbf{Y}$ to the model $\mathbf{F}\mathbf{W}^\top + \mathbf{X}\mathbf{G}^\top + \mathbf{AB}^\top$ by

1. obtaining the LS estimates $\hat{\mathbf{F}}, \hat{\mathbf{G}}$ from the linear model $\mathbf{F}\mathbf{W}^\top + \mathbf{X}\mathbf{G}^\top$;
2. obtaining the LS estimate $\widehat{\mathbf{A}\mathbf{B}^\top}$ from the SVD of the residual matrix.

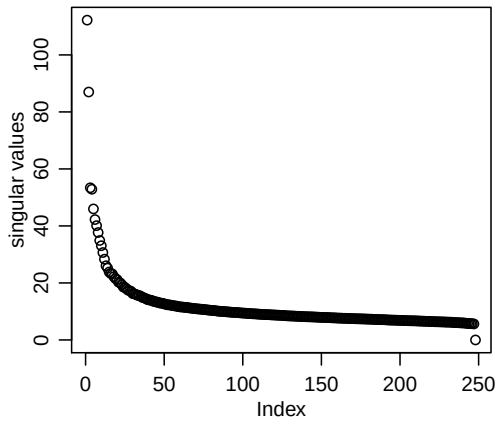
Exercise 5. *Find* $\mathbf{P}_X$ *for* $\mathbf{X} = \mathbf{1} \in \mathbb{R}^{n\times 1}$.

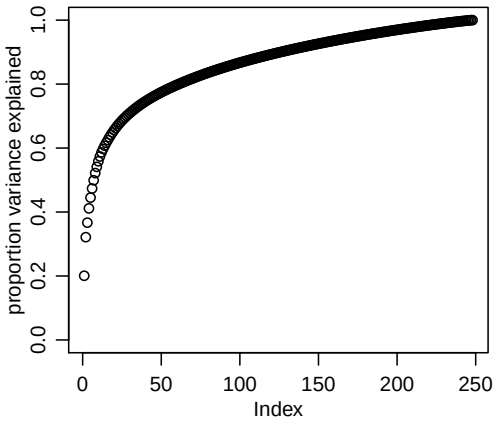
Exercise 6. *Use the above theorem to derive the method for obtaining least-squares AMMI estimates.*

Let's proceed by fitting the AMMI model to the leukemia data:

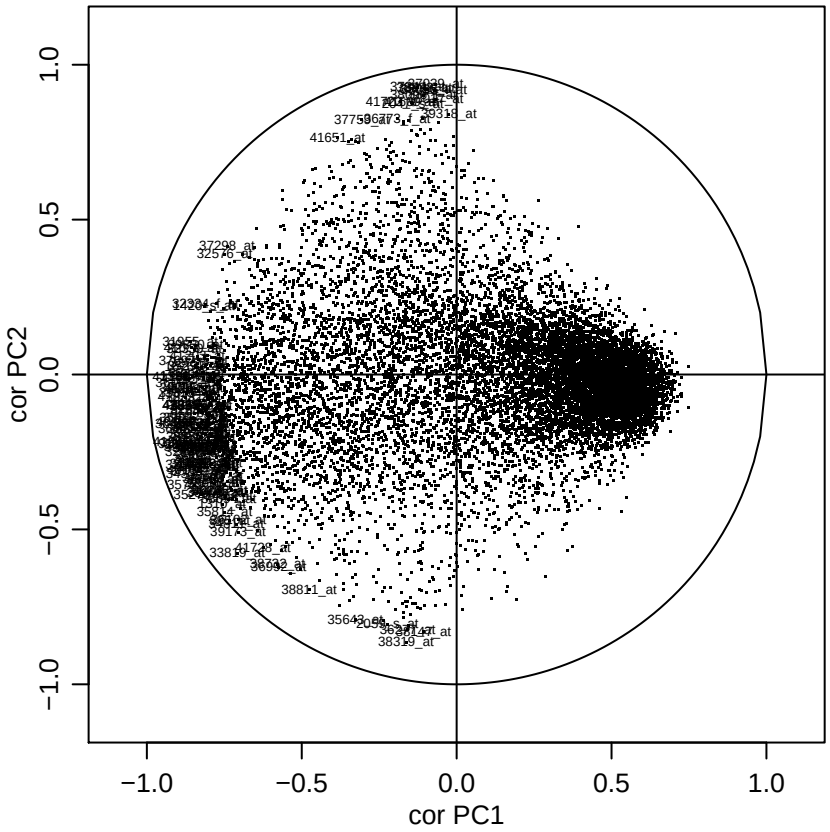
```
sE<-svd(E)
cbind( sE$d^2, sE$d^2/sum(sE$d^2), cumsum(sE$d^2)/sum(sE$d^2))[1:10,]

##         [,1]    [,2]   [,3]
##  [1,] 12590 0.20073 0.2007
##  [2,]  7568 0.12066 0.3214
##  [3,]  2851 0.04545 0.3668
##  [4,]  2795 0.04456 0.4114
##  [5,]  2114 0.03370 0.4451
##  [6,]  1784 0.02844 0.4735
##  [7,]  1608 0.02564 0.4992
##  [8,]  1421 0.02265 0.5218
##  [9,]  1218 0.01942 0.5413
## [10,]  1093 0.01743 0.5587
```

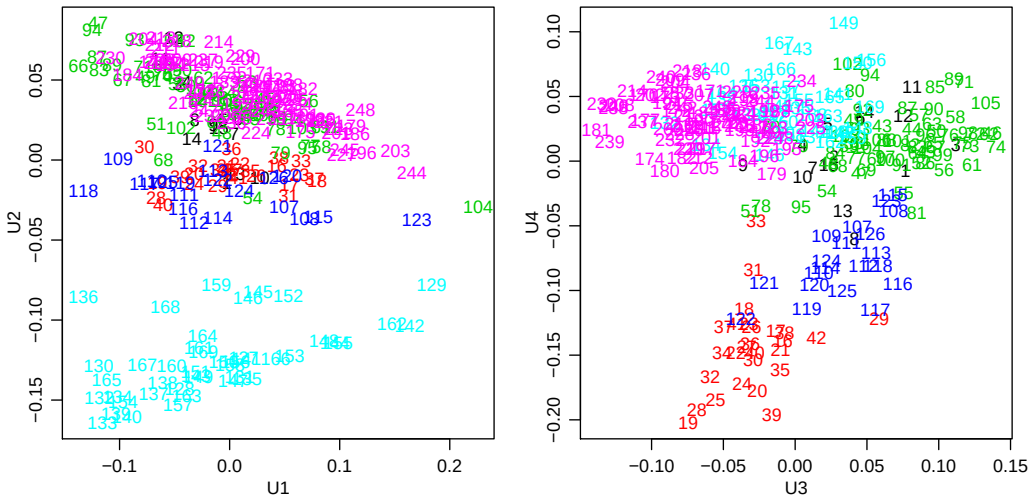




The correlation plot has changed somewhat:



The first two left-singular vectors are nearly identical, but the third and fourth are somewhat different.



References

K. R. Gabriel. Least squares approximation of matrices by additive and multiplicative models. *J. Roy. Statist. Soc. Ser. B*, 40(2):186–196, 1978. ISSN 0035-9246. URL `http://links.jstor.org/sici?sici=0035-9246(1978)40:2<186:LSAOMB>2.0.CO;2-P&origin=MSN`.

Wolfgang Karl Härdle and Léopold Simar. *Applied multivariate statistical analysis*. Springer, Heidelberg, fourth edition, 2015. ISBN 978-3-662-45170-0; 978-3-662-45171-7. doi: 10.1007/978-3-662-45171-7. URL `http://dx.doi.org/10.1007/978-3-662-45171-7`.