Week 6. Principal Component Analysis (PCA)

Principal component analysis (PCA) for dimension reduction in the variable space. PCA projection onto line: using PCA with objects/subjects ranking and application to college admission. PCA projection onto plane and 3D and application to banknote counterfeit. Quality of PCA projection as explained variance. Logistic regression with PCA loadings.

R codes: `collegePCA, swiss`

Data: `CollegeAdmData.csv, SwissBankNotes100+100.txt, iris`

The goal of PCA

The goal of PCA is to project multidimensional data onto the space of fewer dimension: line (1D), plane (2D), or space (3D) for viewing.

Data matrix is given as $\mathbf{X}^{n\times m}$ where n is the number of observations or the number of objects/subjects and m is the number of features/attributes/variables ($m < n$). In the vector representation denote $\mathbf{x}_i$ the vector of features of the ith object. Then the data is the collection of n feature vectors:

$$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\} \in R^m.$$

When $m > 3$ the data are impossible to see. We want to project the R^m data onto R^k, where $k = 1, 2$, or 3, with minimum distortion.

PCA works for **unsupervised** or **supervised** data. In the case we know the groups/classes we display the n points and color to reflect grouping.

The goal of the PCA is to derive synthetic feature, as a linear combination of m features, that most informatively reflect all features.

The first principal component reflects the most informative synthetic feature, as a linear combination of m features, the second principal component reflects the next most informative synthetic feature, etc.

Projection on the line and its application to ranking

To be specific, we refer to the ith row of $\mathbf{X}$ as the ith subject with m attributes $x_{i1}, x_{i2}, ..., x_{im}$ combined into vector $\mathbf{x}_i$. We want to project the data on line using vector $\mathbf{p}^{m\times 1} = (p_1, p_2, ..., p_m)'$ as a vector of coefficients. Sometimes, coefficients p_j are called loadings. In other words,

$$y_i = \sum_{j=1}^{m} p_j x_{ij}, \quad i = 1, 2, ..., n.$$

We treat vector $\mathbf{y} = (y_1, ..., y_n)'$ as a vector which represents n subjects with a single synthetic attribute (the linear combination of the original attributes). In vector/matrix form

$$\mathbf{y}^{n\times 1} = \mathbf{X}\mathbf{p}^{m\times 1}.$$

There are several ways to arrive at the PCA solution: (a) maximum variance, (b) optimal projection on the line.

(a) Justification of PCA as the maximum variance. First, we derive the PCA solution in the probabilistic framework and then apply it to the data matrix $\mathbf{X}$. Let us assume that $\mathbf{x}$ is a random m-dimensional vector with the $m \times m$ covariance matrix $\mathbf{W}$. We aim to find a one-dimensional random variable Y as a linear combination of $\mathbf{x}$ in the form

$$Y = \mathbf{x}'\mathbf{p} = \sum_{j=1}^{m} p_j x_j,$$

where $p_1, p_2, ..., p_m$ are the coefficients to be determined. Among all $\mathbf{p}$ we want to find such vector that makes Y as scattered as possible, i.e. having maximum variance, because otherwise values will be difficult to distinguish and the linear combination is non-informative.

Layman language: given a random vector with known covariance matrix find a linear combination with minimum variance.

Since the variance is unbounded if the norm of $\mathbf{p}$ is not limited we may assume that $\mathbf{p}$ has unit length. One may say that we put Y on the unit scale. In mathematical language we want to solve the following optimization problem:

$$\max_{\|\mathbf{p}\|^2=1} var(Y) = \max_{\|\mathbf{p}\|^2=1} var(\mathbf{x}'\mathbf{p}).$$

Now we use a well known fact that

$$var(\mathbf{x}'\mathbf{p}) = \mathbf{p}'\mathbf{W}\mathbf{p}.$$

Then the problem simplifies to

$$\max_{\|\mathbf{p}\|^2=1} \mathbf{p}'\mathbf{W}\mathbf{p}.$$

As we know from linear algebra $\mathbf{p}'\mathbf{W}\mathbf{p}$ reaches its maximum at $\mathbf{p}$ as the maximum eigenvector

$$\mathbf{p} = \max \text{ eigenvector } \mathbf{W}.$$

This can also be seen from Lagrange multiplier function

$$\mathcal{L}(\mathbf{p}; \lambda) = \mathbf{p}'\boldsymbol{\Omega}\mathbf{p} - \lambda(\|\mathbf{p}\|^2 - 1).$$

Indeed,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}} = 2\boldsymbol{\Omega}\mathbf{p} - 2\lambda\mathbf{p} = \mathbf{0}$$

and

$$\boldsymbol{\Omega}\mathbf{p} = \lambda\mathbf{p}.$$

Since

$$\mathbf{p}'\boldsymbol{\Omega}\mathbf{p} = \lambda \|\mathbf{p}\|^2 = \lambda$$

we pick $\mathbf{p} = \mathbf{p}_{\max}$ the maximum eigenvector with $\lambda = \lambda_{\max}$ as the maximum eigenvalue.

In summary, the maximum eigenvector of matrix $\mathbf{W}$ is the optimal vector of coefficients.

Now we turn our attention to matrix $\mathbf{X}$ treating rows $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ as observation vectors with the same population covariance matrix $\boldsymbol{\Omega}$. An estimator of this matrix (discrete version of the covariance matrix) is

$$\mathbf{W} =\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'.$$

Solution: the first principal component of matrix $\mathbf{X}$ is the linear combination of vector columns $\mathbf{Xp}_{\max}$ = synthetic feature:

$$\mathbf{X} \rightarrow \mathbf{Xp}_{\max},$$

where

$$\begin{aligned}\mathbf{p}_{\max} &= \text{maximum eigenvector of the covariance matrix } \mathbf{W},\ \|\mathbf{p}_{\max}\| = 1 \\ \lambda_{\max} &= \max_{\|\mathbf{p}\|^2=1} \mathbf{p}'\mathbf{W}\mathbf{p} >0\end{aligned}$$

The larger $\lambda_{\max}$ the better.

Computation in R:

1. `W=var(X)`
2. `eigenW=eigen(W,sym=T)`
3. `p.max=eigenW$vectors[,1]`
4. `lambda.max=eigenW$values[1]`
5. `proj1=(X-rep(1,nrow(X)))%*%t(colMeans(X))%*%p.max`

(b) Justification of PCA as the optimal projection on the line. We want to project points $\{\mathbf{x}_i \in R^m, i = 1, 2, ..., n\}$ onto the straight line in R^m. The straight line is defined as $\{\mathbf{a}+\lambda\mathbf{p}, -\infty < \lambda < \infty\}$, where $\mathbf{a} \in R^m$ is the translation vector and $\mathbf{p} \in R^m$ is the line direction vector. Without loss of generality, we can assume that $\|\mathbf{p}\|^2 = 1$. We want to find **a, p**, and projection coordinates λ_i such that the line represents vectors $\{\mathbf{x}_i, i = 1, 2, ..., n\}$ in the closest way meaning that

$$\min_{\lambda_i,\mathbf{a},\|\mathbf{p}\|^2=1} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{a} - \lambda_i\mathbf{p}\|^2 .$$

First we eliminate $\mathbf{a}$ when $\lambda_i\mathbf{p}$ is held fixed:

$$\mathbf{a} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i - \lambda\mathbf{p} =\overline{\mathbf{x}} - \lambda\mathbf{p}.$$

This gives

$$\sum_{i=1}^{n} \|(\mathbf{x}_i - \overline{\mathbf{x}}) - \lambda_i\mathbf{p}\|^2 .$$

Now we find optimal λ_i by differentiation:

$$-2((\mathbf{x}_i - \overline{\mathbf{x}}) - \lambda_i\mathbf{p})'\mathbf{p} = 0,$$

which gives

$$\lambda_i = (\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{p}.$$

Plugging this back into the criterion function we obtain

$$\|(\mathbf{x}_i - \overline{\mathbf{x}}) - \lambda_i \mathbf{p}\|^2 = (\mathbf{x}_i - \mathbf{a})'(\mathbf{I} - \mathbf{p}\mathbf{p}')(\mathbf{x}_i - \mathbf{a}) = \|(\mathbf{x_i} - \overline{\mathbf{x}})\|^{\mathbf{2}} - ((\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{p})^2.$$

The minimum attains when

$$\max_{\|\mathbf{p}\|^2=1} ((\mathbf{x}_i - \overline{\mathbf{x}})'\mathbf{p})^2 = \max_{\|\mathbf{p}\|^2=1} \mathbf{p}' \left(\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})' \right) \mathbf{p}$$

This means that $\mathbf{p}$ is the maximum eigenvector of matrix

$$\sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'.$$

Sometimes $\{p_j, j = 1, ..., m\}$ are called the loadings.

Conclusion 1 *Since the eigenvectors are defined up to ± you can get ranks in ascending or descending order (from worst to best or other way around). Be careful!*

A way to check if the direction of the found maximum eigenvector is correct is to normalize

$$\text{synthetic score} = p_1 x_1 + p_2 x_2 + ... + p_m x_m.$$

Synthetic score must match common sense (total rank), see below.

Application of PCA to subjects/objects ranking

Trivial ranking based on the total rank: sum the ranks over the features: `rank(y)` or `rank(-y)`

Example:

```
> y=c(3.3,1.3,4.1,2.9)
> rank(y)
[1] 3 1 4 2 #1=min,4=max, from min to max
> rank(-y) #1=max,4=min, from max to min
[1] 2 4 1 3
```

Alternative: 1st principal component PCA: projection onto line

Example 2 *College admission. Five thousand students applied for a prestigious college, but only one thousand can be admitted. Select students for admission based on the data presented in CollegeAdmData.csv.*

Solution. Seven score criteria are used to rank students:

1. HSC=High school curriculum

2. SAT=SAT score
3. CI=College interview
4. OSA=Out of school activity
5. SR=Sport & research programs
6. ES=Essay
7. LR=Letters of recommendation

1st principal component PCA:

$$\text{synthetic score} = \mathrm{p}_1 \times \mathrm{HSC}+\mathrm{p}_2 \times \mathrm{SAT}+\mathrm{p}_3 \times \mathrm{CI}+\mathrm{p}_4 \times \mathrm{OSA}+\mathrm{p}_5 \times \mathrm{SR}+\mathrm{p}_6 \times \mathrm{ES}+\mathrm{p}_7 \times \mathrm{LR}$$

```
collegePCA(job=1)
collegePCA(job=2)
collegePCA(job=3)
```

Wrong direction of the max eigenvector, turn by 180 degree

Projection onto plane

Centerize given data matrix $\mathbf{X}$ as

$$\mathbf{Z} = \mathbf{X} - \mathbf{1}\overline{\mathbf{x}}',$$

that is, subtract the means from the columns:

```
X.means=colMeans(X)
Z=X-rep(1,nrow(X))%*%t(X.means)
```

$$[\mathbf{p}_1, \mathbf{p}_2] = \text{first two max eigenvectors } \mathbf{Z}'\mathbf{Z}$$

Projection

$$\mathbf{y}^{n\times 2} = \mathbf{Z}[\mathbf{p}_1, \mathbf{p}_2]$$

Plot $\{(y_{1i}, y_{2i}), i = 1, ..., n\}$

Alternatively, use `var`:

Computation in R:

1. `W=var(X)`
2. `eigenW=eigen(W,sym=T)`
3. `p.max12=eigenW$vectors[,1:2]`
4. `lambda.max12=eigenW$values[1:2]`
5. `Z=X-rep(1:nrow(X))%*%t(colMeans(X))`
6. `proj12=Z%*%p.max12`

Normalized PCA: PCA on the correlation matrix:

$$h_{ij} = \frac{x_{ij} - \overline{x}_j}{SD_j}$$

where

$$SD_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2},$$

the SD of the jth column (row-vectors are iid).

Computation in R for normalized PCA:

1. `R=cor(X)`
2. `eigenW=eigen(R,sym=T)`
3. `p.max12=eigenW$vectors[,1:2]`
4. `lambda.max12=eigenW$values[1:2]`
5. `SD=sqrt(diag(var(X)))`
6. `proj12=((X-rep(1,nrow(X)))%*%t(X.means))/SD)%*%p.max12`

Example: PCA against counterfeit (R function `swiss`) The first half of these measurements are from genuine bank notes, the other half are from counterfeit bank notes.

X1 = length of the bill
X2 = height of the bill (left)
X3 = height of the bill (right)
X4 = distance of the inner frame to the lower border
X5 = distance of the inner frame to the upper border
X6 = length of the diagonal of the central picture.

Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics, A practical Approach*, Cambridge, University Press.

Forged bank notes.

Dataset: SwissBankNotes100+100.txt, R function `swiss`

swiss(job=1)

The maximum eigenvector:

-0.6033324 -0.1339827 -0.1330675 0.5343121 0.5274588 -0.1913882

The projection line in R^6 is defined

X1=-0.6033324×λ

X2=-0.1339827×λ

...

X6=-0.1913882×λ

```
swiss(job=1)
swiss(job=1.1)
```

PCA2 logistic regression and LDA

See `swiss(job=2):` How to plot the line on the plane and define the rule for classification?

PCA2 logistic regression:

If the membership information is known we can construct linear rule to classify the vectors into two classes. Let $y_i = 0$ if the vector of measurements comes from genuine banknote and $y_i = 1$ otherwise.

Run

```
oLOG.R=glm(y~proj,family=binomial)
a=coef(oLOG.R)
```

Predictive model

$$\Pr(Y = 1) = \frac{\exp(a_1 + a_2 proj_1 + a_3 proj_2)}{1 + \exp(a_1 + a_2 proj_1 + a_3 proj_2)}$$

If

$$a_1 + a_2 proj_1 + a_3 proj_2 < 0$$

prob<0.5. If

$$a_1 + a_2 proj_1 + a_3 proj_2 > 0$$

prob>0.5

Thus the separation line is defined as

$$a_1 + a_2 proj_1 + a_3 proj_2 = 0.$$

How to plot?

Linear predictor

$$a_1 + a_2 x + a_2 y = 0, \quad y = -(a_1 + a_2 x)/a_3.$$

The separation line: `y=-(a[1]+a[2]*x)/a[3]`

PCA2 LDA:

```
X=proj
X0=X[y==0,];n0=nrow(X0)
OM0=var(X0)
mu0=colMeans(X0)

X1=X[y==1,];n1=nrow(X1)
OM1=var(X1)
mu1=colMeans(X1)

OM=((n0-1)*OM0+(n1-1)*OM1)/(n0+n1-2)
iOM=solve(OM)
a=iOM%*%(mu1-mu0)
mu=colMeans(X)
yLDA=mu[2]-(x-mu[1])*a[1]/a[2]
```

Quality of projection

Definition 3 *Let $\mathbf{Y}$ be an $m \times 1$ random vector with the $m \times m$ covariance matrix $\mathbf{\Omega}$. The total variance of $\mathbf{Y}$ is the sum of component variances,*

$$\sum_{j=1}^{m} var(Y_j) = \sum_{j=1}^{m} \Omega_{jj} = tr(\mathbf{\Omega}).$$

Lemma 4 *For any $m \times m$ symmetric matrix $\mathbf{M}$ with eigenvalues λ_j we have*

$$tr(\mathbf{M}) = \sum_{j=1}^{m} \lambda_j.$$

Proof. Since $\mathbf{M}$ is symmetric we can apply spectral Jordan decomposition

$$\mathbf{M} = \mathbf{P\Lambda P}' = \sum_{j=1}^{m} \lambda_j \mathbf{p}_j \mathbf{p}_j'$$

where λ_j is the eigenvalue with the corresponding eigenvector $\mathbf{p}_j$, and

$$\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_m], \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_m \end{bmatrix}$$

Note that λ_j and $\mathbf{p}_j$ can be arranged is any order, random, ascending, or descending.

Lemma 5 *Let $\mathbf{Y}_k$ be the PCA projection of the $m \times 1$ random vector $\mathbf{Y}$ onto R^k where $k < m$ such as $k = 1, 2,$ or $3,$ that is,*

$$\mathbf{Y}_k = \mathbf{P}_k \mathbf{Y},$$

where

$$\mathbf{P}_k = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_k].$$

Then the total variance of $\mathbf{Y}_k$ is

$$\sum_{j=1}^{k} \lambda_j$$

where the eigenvalues and eigenvectors are arranged is descending order: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$.

Definition 6 *The variance explained by PCA projection of R^m onto R^k is defined as*

$$\frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{m} \lambda_j},$$

assuming that eigenvalues and eigenvectors are arranged in descending order: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k$.

When projecting on the line the variance explained is

$$\frac{\lambda_{\max}}{\sum_{j=1}^m \lambda_j}.$$

For discrete PCA the covariance matrix is estimated as

$$\mathbf{M}^{m\times m}=\frac{1}{n}\mathbf{Z}'\mathbf{Z}.$$

Then for projection onto line $\mathbf{y} = \mathbf{Z}\mathbf{p}_{\max}$ we have

$$\text{Variance explained}_1 = \frac{var(\mathbf{y})}{tr(\mathbf{Z}'\mathbf{Z})} = \frac{\mathbf{p}'_{\max}\mathbf{Z}'\mathbf{Z}\mathbf{p}_{\max}}{tr(\mathbf{Z}'\mathbf{Z})} = \frac{\lambda_{\max}}{\sum_{j=1}^m \lambda_j},$$

the quality of the projection onto line, the proportion of total variance captured by the 1st component. The quantity

$$\frac{\sum_{j=2}^m \lambda_j}{\sum_{j=1}^m \lambda_j}$$

is the proportion of the variance leftover (unexplained by PCA).

For projection onto plane we have

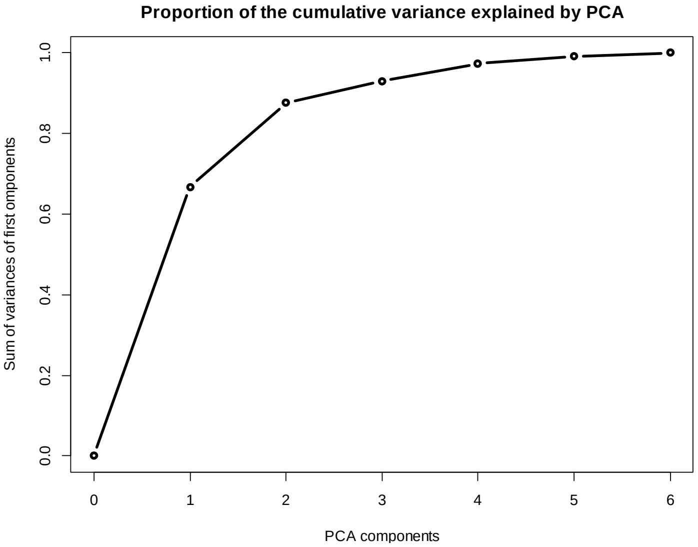
$$\text{Variance explained}_2 = \frac{\lambda_1+\lambda_2}{\sum_{j=1}^m \lambda_j}$$

as the proportion of the variance quality of the projection onto plane.

In general case

$$\text{Variance explained by projection on the first } k \text{ components} = \frac{\sum_{j=1}^k \lambda_j}{\sum_{j=1}^m \lambda_j}$$

```
swiss(3)
```



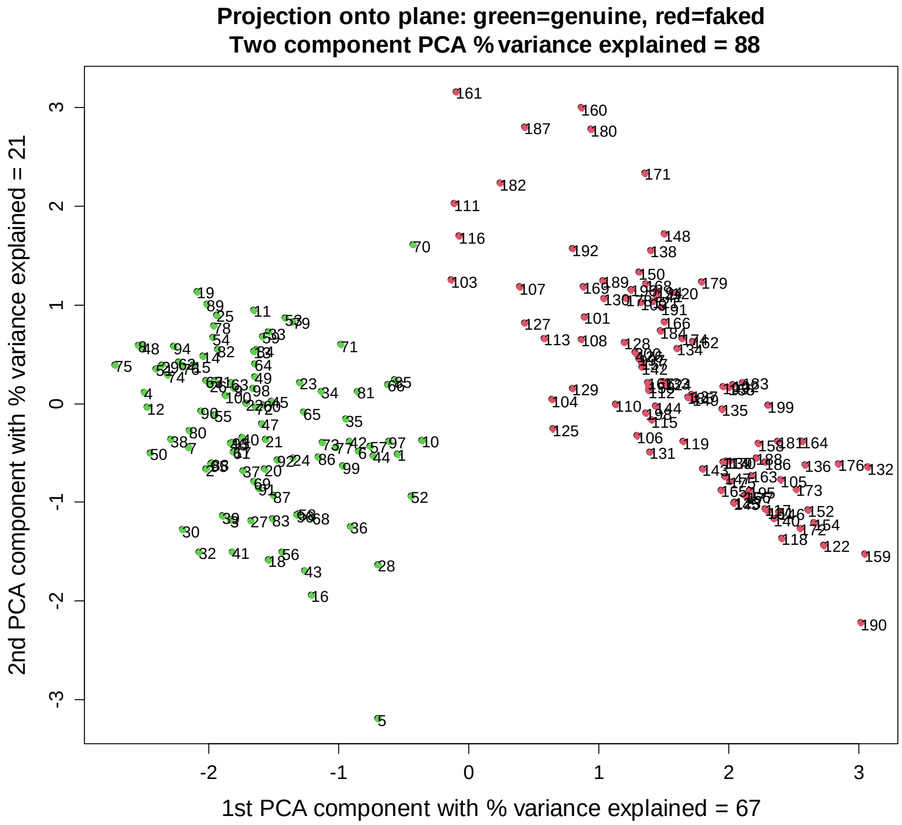
It is a good practice to show % variance explained by the two PCA components as labels at the x- and y-axis separately, that is,

$$\frac{\lambda_1}{\sum_{j=1}^m \lambda_j} \times 100\% \text{ and } \frac{\lambda_2}{\sum_{j=1}^m \lambda_j} \times 100\%,$$

and on the top the total variance explained by the two components, that is

$$\frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{m} \lambda_j} \times 100\%.$$

```
swiss(job=4)
```



Homework 6

1. (20 points). (a) Apply PCA to project the `iris` data onto the plane, display and color each point. (b) Compute and display the proportion of variance explained by individual components and two components together on the top of the graph. (c) Repeat the same tasks for the 3D projection (use `theta=30` and `phi=30`).

2. (20 points). (a) Project `Goldman.imputed.csv` data onto plane. (b) Use `red` to color male and `green` to color female. (c) Compute and display the non-optimized/standard PCA logistic regression separation line. (d) Compute and display the optimized PCA logistic regression line with the threshold that minimizes the total misclassification error. (e) Display the two ROC curves along with the respective AUCs. (f) Explain why the PCA logistic regression yields a worse result than just using one RFEB variable.