CS724: Topics in Algorithms
Principal Component Analysis

Prof. Dan A. Simovici

UMASS BOSTON

Principal Components Analysis (PCA) is used to transform the values of variables of a data sample matrix into values of new uncorrelated variables which explain the variability of data. In general, this is accompanied by a reduction of data dimensionality.

Definition

Let D be a centered data sample matrix. The *principal directions* of D are the eigenvectors of the covariance matrix $cov(D)$.

UMASS BOSTON

If $R \in \mathbb{R}^{n \times n}$ is an orthogonal matrix $R = (\mathbf{r}_1, \ldots, \mathbf{r}_n)$ that diagonalizes the matrix $cov(D)$, then the principal directions of D are the columns of R because the equality

$$R' cov(D) R = \text{diag}(d_1, \ldots, d_n)$$

is equivalent to $cov(D)R = R\text{diag}(d_1, \ldots, d_n)$, that is, $cov(D)\mathbf{r}_j = d_j \mathbf{r}_j$ for $1 \leqslant j \leqslant n$.
We shall assume from now on that

$$d_1 \geqslant d_2 \geqslant \cdots \geqslant d_n.$$

Note that the covariance matrix $cov(D)$ is a scalar multiple of the Gram matrix $D'D$ of the columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of the centered data matrix D.

Definition

Let D be a centered data matrix. The first eigenvector $\mathbf{r}_1$ of $cov(D)$ that corresponds to the largest eigenvalue d_1 of $cov(D)$ is the *first principal direction* of D.
In general, the k^{th} eigenvector (that is, the eigenvector that corresponds to the k^{th} eigenvalue of $cov(D)$) is the k^{th} *principal direction* of D.

The principal directions of a centered data matrix D are linked to the singular value decomposition of D. Namely, if

$$D\mathbf{s} = \sigma\mathbf{r} \text{ and } D'\mathbf{r} = \sigma\mathbf{s},$$

then $\mathbf{r}$ is a *principal direction* of D and $\mathbf{s}$ is the *vector of principal components of D* that corresponds to $\mathbf{r}$.

If the centered data matrix is $D = \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix}$, we have the equalities

$$\mathbf{u}_1'\mathbf{r} = \sigma s_1, \ldots, \mathbf{u}_m'\mathbf{r} = \sigma s_m,$$

which show that the principal components are the projection of the centered data points on the principal directions.

Theorem

Let $D \in \mathbb{R}^{m \times n}$ be a centered data matrix and let R be an orthogonal matrix that diagonalizes $cov(D)$, that is, $R' cov(D) R = diag(d_1, \ldots, d_n)$, where $d_1 \geqslant d_2 \geqslant \cdots \geqslant d_n$.
Let $Q \in \mathbb{R}^{n \times \ell}$ be a matrix with orthogonal columns and let $X = DQ \in \mathbb{R}^{m \times \ell}$. Then, $trace(cov(X))$ is maximized when Q consists of the first ℓ columns of R and is minimized when Q consists of the last ℓ columns of R.

Proof.

This statement are a consequence of Ky Fan's Theorem applied to the symmetric covariance matrix of the transformed data set. □

If $D = U\text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)V' \in \mathbb{R}^{m \times n}$ is the thin SVD decomposition of the centered data matrix D, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ are matrices having orthogonal columns. For the covariance matrix we have

$$\begin{aligned} cov(D) &= \frac{1}{m-1} D'D = \frac{1}{m-1} V\text{diag}(\sigma_1, \ldots, \sigma_r)U'U\text{diag}(\sigma_1, \ldots, \sigma_r)V' \\ &= \frac{1}{m-1} V\text{diag}(\sigma_1^2, \ldots, \sigma_r^2)V', \end{aligned}$$

because the columns of U are orthogonal. The matrix V is known as the *matrix of loadings*.

The matrix $S = U\text{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{m \times r}$ is *matrix of scores*. Note that $D = SV$, where S is the scores matrix and V is the loadings matrix. Since the columns of V are orthogonal, we also have $S = DV$.
The SVD of D can be written as

$$D = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i'.$$

This implies $D'D\mathbf{v}_i = \sigma_i^2 \mathbf{v}_i$. Since $\mathbf{u}_i' D = \sigma_i \mathbf{v}_i'$, it follows that $\mathbf{v}_i'$ is a weighted sum of the rows of D. Similarly, $\mathbf{u}_i$ is a weighted sum of the columns of D.

We discuss the function `PCA` of the package `FactoMineR` that performs principal component analysis in **R** .

- The result of the application of this function consists of a set of eigenvalues, a table with the *scores* of principal components and a table of *loadings* (or correlations between variables and principal components).
- The scores provide information about the structure of the observations; the loadings offer information about relationships between variables and about variable associations with the principal components.

We use the data set `USArrests` of the package `stats`. The initial part of this data frame is

```
> head(USArrests)
           Murder Assault UrbanPop Rape
Alabama      13.2     236       58 21.2
Alaska       10.0     263       48 44.5
Arizona       8.1     294       80 31.0
Arkansas      8.8     190       50 19.5
California    9.0     276       91 40.6
Colorado      7.9     204       78 38.7
```

A principal component object can be obtained using the function `PCA` of the package `FactoMineR`. Alternative choices are the functions `prcomp` and `princomp` of the package `stats`; the `PCA` function is preferable because it provides more detailed results.

The function `PCA` returns a list that consists of the following components:

- `eig`: a matrix containing all the eigenvalues, the percentage of variance, and the cumulative percentage of variance;
- `var`: a list of matrices containing all the results for the active variables (coordinates, correlation between variables and axes, square cosine, contributions);
- `ind`: a list of matrices containing all the results for the active individuals (coordinates, square cosine, contributions);

UMASS BOSTON

```
> pca <- PCA(USArrests,graph=FALSE)

> pca$eig
       eigenvalue percentage of variance
comp 1  2.4802416             62.006039
comp 2  0.9897652             24.744129
comp 3  0.3565632              8.914080
comp 4  0.1734301              4.335752

cumulative percentage of variance
62.00604
86.75017
95.66425
100.00000
```

```
> pca$var

$coord
              Dim.1      Dim.2      Dim.3       Dim.4
Murder    0.8439764 -0.4160354  0.2037600  0.27037052
Assault   0.9184432 -0.1870211  0.1601192 -0.30959159
UrbanPop  0.4381168  0.8683282  0.2257242  0.05575330
Rape      0.8558394  0.1664602 -0.4883190  0.03707412
```

```
$cor
              Dim.1      Dim.2      Dim.3       Dim.4
Murder    0.8439764 -0.4160354  0.2037600  0.27037052
Assault   0.9184432 -0.1870211  0.1601192 -0.30959159
UrbanPop  0.4381168  0.8683282  0.2257242  0.05575330
Rape      0.8558394  0.1664602 -0.4883190  0.03707412
```

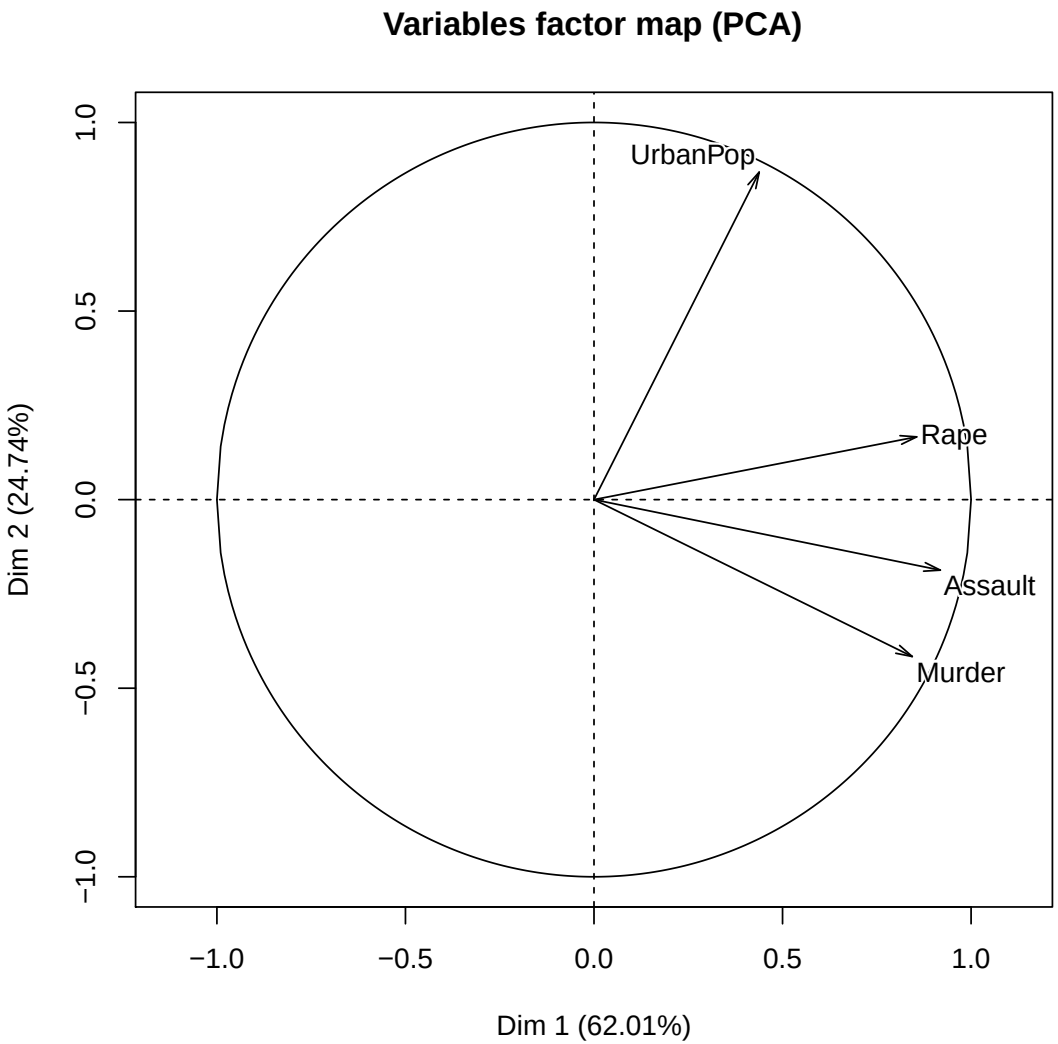
```
$cos2
              Dim.1     Dim.2      Dim.3       Dim.4
Murder    0.7122962 0.1730854 0.04151814 0.073100217
Assault   0.8435380 0.0349769 0.02563817 0.095846950
UrbanPop  0.1919463 0.7539938 0.05095143 0.003108430
Rape      0.7324611 0.0277090 0.23845544 0.001374491

$contrib
              Dim.1     Dim.2     Dim.3     Dim.4
Murder    28.718825 17.487524 11.643977 42.149674
Assault   34.010315  3.533859  7.190358 55.265468
UrbanPop   7.739016 76.179065 14.289594  1.792325
Rape      29.531844  2.799553 66.876071  0.792533
```

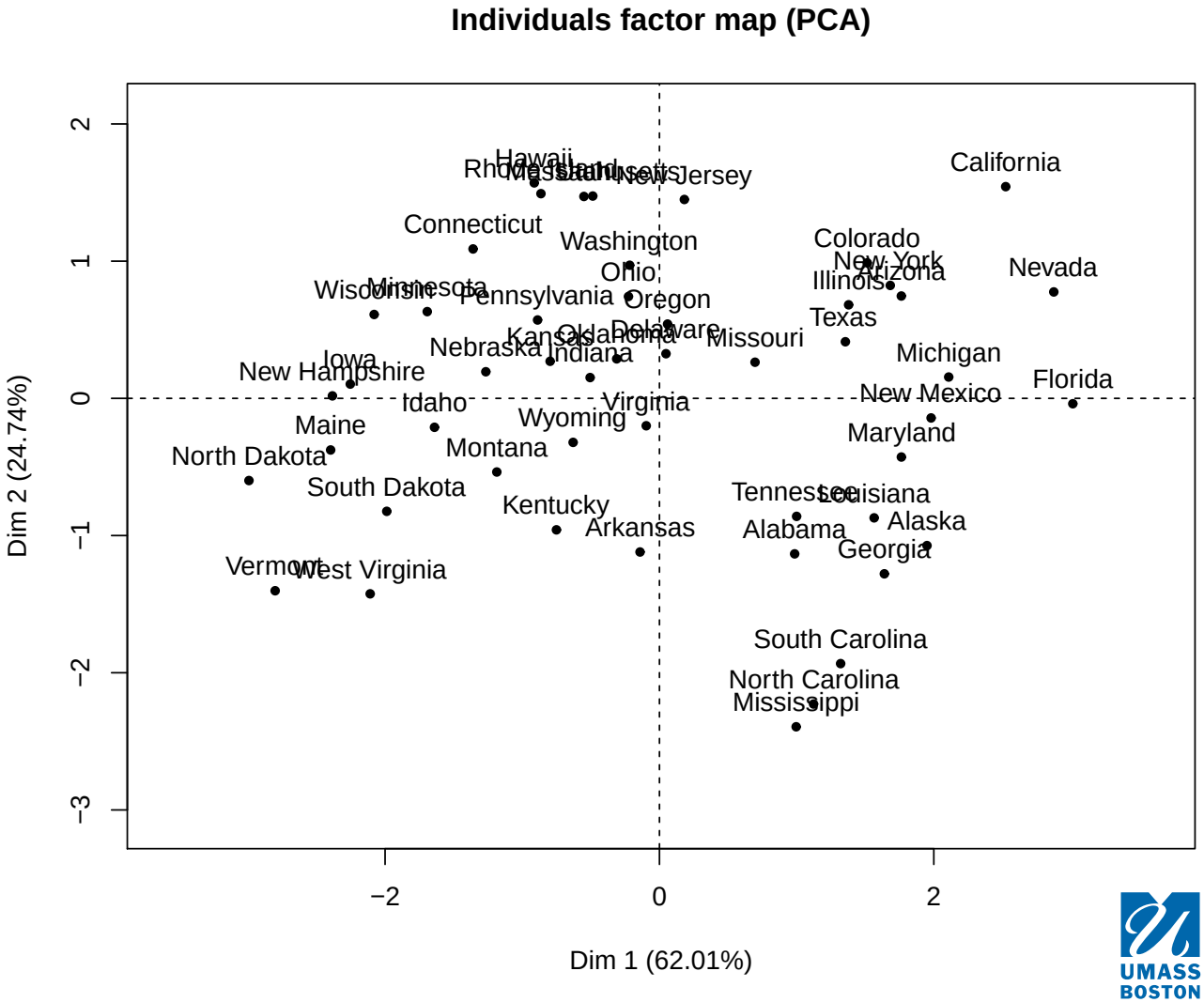
```
> head(pca$ind$coord)
                Dim.1      Dim.2       Dim.3        Dim.4
Alabama     0.9855659 -1.1333924  0.44426879  0.156267145
Alaska      1.9501378 -1.0732133 -2.04000333 -0.438583440
Arizona     1.7631635  0.7459568 -0.05478082 -0.834652924
Arkansas   -0.1414203 -1.1197968 -0.11457369 -0.182810896
California  2.5239801  1.5429340 -0.59855680 -0.341996478
Colorado    1.5145629  0.9875551 -1.09500699  0.001464887
```

The variables factor map shown next presents a view of the observed variables projected into the plane spanned by the first two principal components. This shows the structural relationship between the variables and the components, and helps us name the components. The projection of a variable vector onto the component axis allows us to directly read the correlation between the variable and the component.

Variables factor map in PCA



The individuals factor map shown next is a plot of the principal component scores for individuals on the first two principal components.



Principal component loadings can be calculated directly using the following piece of code:

```
> V <- eigen(cor(USArrests))$vectors[,1:2]
> V
           [,1]       [,2]
[1,] -0.5358995  0.4181809
[2,] -0.5831836  0.1879856
[3,] -0.2781909 -0.8728062
[4,] -0.5434321 -0.1673186

> D <- eigen(cor(USArrests))$values[1:2]
> D
[1] 2.4802416 0.9897652
> D <- diag(D)
> D
         [,1]      [,2]
[1,] 2.480242 0.0000000
[2,] 0.000000 0.9897652
```

```
> D.half <- sqrt(D)
> D.half
         [,1]      [,2]
[1,] 1.574878 0.0000000
[2,] 0.000000 0.9948694

> F <- V %*% D.half
> F
            [,1]       [,2]
[1,] -0.8439764  0.4160354
[2,] -0.9184432  0.1870211
[3,] -0.4381168 -0.8683282
[4,] -0.8558394 -0.1664602
```

```
> rownames(F) <- colnames(USArrests)
> F
                [,1]       [,2]
Murder   -0.8439764  0.4160354
Assault  -0.9184432  0.1870211
UrbanPop -0.4381168 -0.8683282
Rape     -0.8558394 -0.1664602
> F <- -F
> F
                [,1]       [,2]
Murder    0.8439764 -0.4160354
Assault   0.9184432 -0.1870211
UrbanPop  0.4381168  0.8683282
Rape      0.8558394  0.1664602
```