CS724: Topics in Algorithms
Data Sample Matrices
Slide Set 9

Prof. Dan A. Simovici

UMASS BOSTON

Matrices are natural tools for organizing data sets.
Let such a data set consist of a sequence $\mathcal{E}$ of m vectors of $\mathbb{R}^n$, $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$. The j^{th} components $(\boldsymbol{u}_i)_j$ of these vectors correspond to the values of a random variable $\mathcal{V}_j$, where $1 \leq j \leq n$.
This data series will be represented as a matrix having m rows $\boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_m$ and n columns $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$. We refer to matrices obtained in this manner as *sample matrices*. The number m is the *size* of the sample.

Each row vector $\boldsymbol{u}_i'$ corresponds to an experiment E_i in the series of experiments $\mathcal{E} = (E_1, \ldots, E_m)$; the experiment E_i consists of measuring the n components of $\boldsymbol{u}_i' = (x_{i1}, \ldots, x_{in})$, as shown below.

	$\boldsymbol{v}_1$	$\cdots$	$\boldsymbol{v}_n$
$\boldsymbol{u}_1'$	x_{11}	$\cdots$	x_{1n}
$\boldsymbol{u}_2'$	x_{21}	$\cdots$	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{u}_m'$	x_{m1}	$\cdots$	x_{mn}

The column vector

$$\boldsymbol{v}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{mj} \end{pmatrix}$$

represents the measurements of the j^{th} variable $\mathcal{V}_j$ of the experiment, for $1 \leq j \leq n$, as shown below. These variables are usually referred to as *attributes* or *features* of the series $\mathcal{E}$.

Definition

The *sample matrix* of $\mathcal{E}$ is the matrix $X \in \mathbb{C}^{m \times n}$ given by

$$X = \begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_m' \end{pmatrix} = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n).$$

Clearly, we have $(\boldsymbol{v}_j)_i = (\boldsymbol{u}_i')_j = x_{ij}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. If $\mathcal{E}$ is clear from context, the subscript $\mathcal{E}$ is omitted. We will use both representations of the sample matrix and will write

$$X = \begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_m' \end{pmatrix} = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n),$$

when we are interested in the vectors that represent results of experiments and

$$X = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n),$$

when we need to work with vectors that represent the values of variables.

Pairwise distances between the row vectors of the sample matrix $X \in \mathbb{R}^{m \times n}$ can be computed with the MATLAB function `pdist(X)`. This form of the function returns a vector $\boldsymbol{D}$ having $\frac{m(m-1)}{2}$ components corresponding to $\binom{m}{2}$ pairs of observations arranged in the order $d_2(\boldsymbol{u}_2', \boldsymbol{u}_1'), d_2(\boldsymbol{u}_3', \boldsymbol{u}_1'), d_2(\boldsymbol{u}_3', \boldsymbol{u}_2'), \ldots$, that is the order of the lower triangle of the distance matrix.

Example

Let X be the data matrix

$$X = \begin{pmatrix} 1 & 4 & 5 \\ 2 & 3 & 7 \\ 5 & 1 & 4 \\ 6 & 2 & 4 \end{pmatrix}$$

The function call `D = pdist(X)` returns

```
D =
    2.4495    6.0000    5.4772    7.3485    5.0990    5.0990
```

Example

Equivalently, a distance matrix can be obtained using the auxiliary function `squareform`, by writing `E = squareform(D)`, which yields

```
E =
         0    2.4495    6.0000    5.4772
    2.4495         0    7.3485    5.0990
    6.0000    7.3485         0    5.0990
    5.4772    5.0990    5.0990         0
```

There are versions of `pdist` that can return other distances by using a second string parameter. For instance, `pdist(X,'cityblock')` computes $d_1(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and `pdist(X,'cebyshev')` computes $d_\infty(\boldsymbol{x}_i, \boldsymbol{x}_j)$. In general, the Minkowski's distance d_p can be computed using
`D = pdist(X,'minkowski',p)`.

A *linear data mapping* for a data sequence $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m) \in \mathbf{Seq}_m(\mathbb{R}^n)$ is the morphism $\mathrm{r} : \mathbb{R}^n \longrightarrow \mathbb{R}^q$. If $R \in \mathbb{R}^{n \times q}$ is the matrix that represents this mapping, then $\mathrm{r}(\boldsymbol{u}_i) = R\boldsymbol{u}_i$ for $1 \leq i \leq m$.
If $q < n$, we refer to r as a *linear dimensionality-reduction mapping*. The *reduced data matrix* is given by

$$\mathrm{r}(X_{\mathcal{E}}) = \begin{pmatrix} \mathrm{r}(\boldsymbol{u}_1)' \\ \vdots \\ \mathrm{r}(\boldsymbol{u}_m)' \end{pmatrix} = \begin{pmatrix} (R\boldsymbol{u}_1)' \\ \vdots \\ (R\boldsymbol{u}_m)' \end{pmatrix} = X_{\mathcal{E}} R \in \mathbb{R}^{m \times q}$$

The reduced data set $\mathsf{r}(X_{\mathcal{E}})$ has new variables $\mathcal{Y}_1, \ldots, \mathcal{Y}_q$. We denote this by writing

$$(\mathcal{Y}_1, \ldots, \mathcal{Y}_q) = \mathsf{r}(\mathcal{V}_1, \ldots, \mathcal{V}_n).$$

The mapping r is a *linear feature selection mapping* if $R \in \{0,1\}^{q \times n}$ is a 0/1-matrix having exactly one unit in every row and at most one unit in every column.

Definition

Let $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ be a series of observations in $\mathbb{R}^n$. The *sample mean* of this sequence is the vector

$$\tilde{\boldsymbol{u}} = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{u}_i \in \mathbb{R}^n.$$

The series is *centered* if $\tilde{\boldsymbol{u}} = \boldsymbol{0}_n$.

Note that the series $(\boldsymbol{u}_1 - \tilde{\boldsymbol{u}}, \ldots, \boldsymbol{u}_m - \tilde{\boldsymbol{u}})$ is always centered.
If $n = 1$, the series of observation is reduced to a vector $\boldsymbol{v} \in \mathbb{R}^m$.

UMASS BOSTON

Definition

The *standard deviation of a vector* $\boldsymbol{v} \in \mathbb{R}^m$ is the number

$$s_{\boldsymbol{v}} = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(v_i - v)^2},$$

where v is the mean of the components of $\boldsymbol{v}$.
The *standard deviation of sample matrix* $X \in \mathbb{R}^{m\times n}$, where $X = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n)$ is the row $\boldsymbol{s} = (s_{\boldsymbol{v}_1}, \ldots, s_{\boldsymbol{v}_n})$.

If the measurement scale for the variables $\mathcal{V}_1, \ldots, \mathcal{V}_n$ involved in the experiment are very different due to different measurement units, some variables may inappropriately influence the analysis process. Therefore, the columns of the data sample matrix need to be *scaled* in order to make. To scale a matrix we need to replace each column $\boldsymbol{v}_i$ by $\frac{1}{s_{\boldsymbol{v}_i}}\boldsymbol{v}_i$. This will yield a matrix having the standard deviation of each column equal to 1.

Next, we examine the effect of centering on a sample matrix.

Theorem

Let $X \in \mathbb{R}^{m \times n}$ is a sample matrix

$$X = \begin{pmatrix} \boldsymbol{u}_1' \\ \vdots \\ \boldsymbol{u}_m' \end{pmatrix}.$$

The sample matrix that corresponds to the centered sequence is

$$\hat{X} = (I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}_m')X.$$

Proof

The matrix that corresponds to the centered sequence is

$$\hat{X} = \begin{pmatrix} \boldsymbol{u}_1' - \tilde{\boldsymbol{u}}' \\ \vdots \\ \boldsymbol{u}_m' - \tilde{\boldsymbol{u}}' \end{pmatrix} = X - \mathbf{1}_m \tilde{\boldsymbol{u}}'.$$

It follows that

$$\hat{X} = X - \mathbf{1}_m \tilde{\boldsymbol{u}}' = X - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m' X = \left(I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}_m' \right) X,$$

which yields the desired equality.

By the theorem on Slide 12 to center a data matrix $X \in \mathbb{R}^{m \times n}$ we need to multiply it at the left by the *centering matrix*

$$H_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m \in \mathbb{R}^{m \times m},$$

that is, $\hat{X} = H_m X$. Note that $H_m = I_m - \frac{1}{m}J_m$. It is easy to see that H_m is both symmetric and idempotent. Since

$$H_m\mathbf{1}_m = \mathbf{1}_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m\mathbf{1}_m = \mathbf{0},$$

it follows that H_m has the eigenvalue 0.

If $X \in \mathbb{R}^{m \times n}$ is a matrix the standard deviations are computed in MATLAB using the function `std(X)`, which returns an n-dimensional row $\boldsymbol{s}$ containing the square roots of the sample variances of the columns of U, that is, their standard deviations. The means of the columns of X is computed in MATLAB using the function `mean(X)`.
The MATLAB function `Z = zscore(X)` computes a centered and scaled version of a data sample matrix having the same format as X.
If X is a matrix, then z-scores are computed using the mean and standard deviation along each column of X. The columns of Z have sample mean zero and sample standard deviation one (unless a column of X is constant, in which case that column of Z is constant at 0). If we use the format

```
[Z,mu,sigma] = zscore(X)
```

also returns the mean vector is returned to `mu` and the vector of standard deviations to `sigma`.

Example

Let X be the matrix

```
X =
     1     12     77
     3     15     80
     2     15     75
     5     18     98
```

The means and the standard deviations of the columns of X are obtained as follows.

```
>> m = mean(X)

m =
    2.7500   15.0000   82.5000

>> s=std(X)

s =
    1.7078    2.4495   10.5357
```

Example

Finally, to compute together the mean, the standard deviation, and the matrix Z, we write

```
>> [Z,m,s]=zscore(A)

Z =
   -1.0247   -1.2247   -0.5220
    0.1464         0   -0.2373
   -0.4392         0   -0.7119
    1.3175    1.2247    1.4712


m =
    2.7500   15.0000   82.5000


s =
    1.7078    2.4495   10.5357
```

Definition

Let $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$ be a sequence of vectors in $\mathbb{R}^n$. The *inertia* of this sequence relative to a vector $\boldsymbol{z} \in \mathbb{R}^n$ is the number

$$I_{\boldsymbol{z}}(\boldsymbol{u}) = \sum_{j=1}^{m} \| \boldsymbol{u}_j - \boldsymbol{z} \|_2^2 .$$

Theorem

(Huygens' Inertia Theorem) *Let* $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m) \in \mathbf{Seq}_m(\mathbb{R}^n)$. *We have*

$$I_{\boldsymbol{z}}(\boldsymbol{u}) - I_{\tilde{\boldsymbol{u}}}(\boldsymbol{u}) = m \parallel \tilde{\boldsymbol{u}} - \boldsymbol{z} \parallel_2^2,$$

for every $\boldsymbol{z} \in \mathbb{R}^n$.

UMASS BOSTON

Proof

The inertia of $\boldsymbol{u}$ relative to $\tilde{\boldsymbol{u}}$ is

$$\begin{aligned} I_{\tilde{\boldsymbol{u}}}(\boldsymbol{u}) &= \sum_{j=1}^{m} \| \boldsymbol{u}_j - \tilde{\boldsymbol{u}} \|_2^2 \\ &= \sum_{j=1}^{m} (\boldsymbol{u}_j - \tilde{\boldsymbol{u}})'(\boldsymbol{u}_j - \tilde{\boldsymbol{u}}) \\ &= \sum_{j=1}^{m} (\boldsymbol{u}_j'\boldsymbol{u}_j - \tilde{\boldsymbol{u}}'\boldsymbol{u}_j - \boldsymbol{u}_j'\tilde{\boldsymbol{u}} + \tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}}). \end{aligned}$$

Similarly, we have

$$I_{\boldsymbol{z}}(\boldsymbol{u}) = \sum_{j=1}^{m} (\boldsymbol{u}_j'\boldsymbol{u}_j - \boldsymbol{z}'\boldsymbol{u}_j - \boldsymbol{u}_j'\boldsymbol{z} + \boldsymbol{z}'\boldsymbol{z}).$$

This allows us to write

$$I_{\boldsymbol{z}}(\boldsymbol{u}) - I_{\tilde{\boldsymbol{u}}}(\boldsymbol{u}) = \sum_{j=1}^{m} (\tilde{\boldsymbol{u}} - \boldsymbol{z})'\boldsymbol{u}_j + \sum_{j=1}^{m} \boldsymbol{u}_j'(\tilde{\boldsymbol{u}} - \boldsymbol{z}) + \boldsymbol{z}'\boldsymbol{z} - \tilde{\boldsymbol{u}}'\tilde{\boldsymbol{u}}$$

Corollary

Let $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m) \in \mathbf{Seq}_m(\mathbb{R}^n)$. *The minimal value of the inertia* $I_{\boldsymbol{z}}(\boldsymbol{u})$ *is achieved for* $\boldsymbol{z} = \tilde{\boldsymbol{u}}$.

Let $\boldsymbol{u}$ and $\boldsymbol{w}$ be two vectors in $\mathbb{R}^m$, where $m > 1$, having the means u and w, and the standard deviations s_u and s_v, respectively.

Definition

The *covariance coefficient* of $\boldsymbol{u}$ and $\boldsymbol{w}$ is the number

$$cov(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{m-1} \sum_{i=1}^{m-1} (u_i - u)(w_i - w)$$

The *correlation coefficient* of $\boldsymbol{u}$ and $\boldsymbol{w}$ is the number

$$\rho(\boldsymbol{u}, \boldsymbol{w}) = \frac{cov(\boldsymbol{u}, \boldsymbol{w})}{s_u s_w}$$

By Cauchy-Schwarz Inequality, we have

$$\left|\sum_{i=1}^{m}(u_i - u)(w_i - w)\right| \leq \sqrt{\sum_{i=1}^{m}(u_i - u)^2} \cdot \sqrt{\sum_{i=1}^{m}(w_i - w)^2},$$

which implies

$$-1 \leq \rho(\boldsymbol{u}, \boldsymbol{w}) \leq 1.$$

Definition

Let $X \in \mathbb{R}^{m \times n}$ be a sample matrix and let $\hat{X}$ be the centered sample matrix corresponding to X. The *sample covariance matrix* is the matrix

$$cov(X) = \frac{1}{m-1}\hat{X}'\hat{X} \in \mathbb{R}^{n \times n}.$$

Note that if X is centered, $cov(X) = \frac{1}{m-1}X'X$.
If $n = 1$ the matrix is reduced to one column $X = (\boldsymbol{v})$ and

$$cov(\boldsymbol{v}) = \frac{1}{m-1}\boldsymbol{v}'\boldsymbol{v} \in \mathbb{R}.$$

In this case we refer to $cov(\boldsymbol{v})$ as the *variance* of $\boldsymbol{v}$; this number is denoted by $var(\boldsymbol{v})$.

If $X = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n)$, then $(cov(X))_{ij} = cov(\boldsymbol{v}_i, \boldsymbol{v}_j)$ for $1 \leq i, j \leq n$. The covariance matrix can be written also as

$$cov(X) = \frac{1}{m-1} X' H_m H_m X = \frac{1}{m-1} X' H_m X.$$

The *sample correlation matrix* is the matrix $corr(X)$ given by $(corr(X))_{ij} = \rho(\boldsymbol{v}_i, \boldsymbol{v}_j)$ for $1 \leq i, j \leq n$.
If X is centered, then $cov(X) = \frac{1}{m-1} X'X$. Clearly, the covariance matrix is a symmetric, positive semidefinite matrix. Furthermore, the rank of $cov(X)$ is the same as the rank of $\hat{X}$ and, since m, the size of the sample is usually much larger than n we are often justified in assuming that $\text{rank}(cov(X)) = n$.

Let $X = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n) \in \mathbb{R}^{m \times n}$ be a sample matrix. Note that

$$H_m \boldsymbol{v}_p = (I_m - \frac{1}{m}\mathbf{1}_m \mathbf{1}_m')\boldsymbol{v}_p = \boldsymbol{v}_p - \frac{1}{m}\mathbf{1}_m \mathbf{1}_m' \boldsymbol{v}_p = \boldsymbol{v}_p - a_p \mathbf{1}_m,$$

because $\frac{1}{m}\mathbf{1}_m' \boldsymbol{v}_p = a_p$ for $1 \leq p \leq n$, where $\tilde{\boldsymbol{u}}' = (a_1, \ldots, a_n)$.
The covariance matrix can be written as

$$\begin{aligned} cov(X) &= \frac{1}{m-1}(\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n)' H_m' H_m (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_n) \\ &= \frac{1}{m-1}(H_m \boldsymbol{v}_1 \ \cdots \ H_m \boldsymbol{v}_n)'(H_m \boldsymbol{v}_1 \ \cdots \ H_m \boldsymbol{v}_n), \end{aligned}$$

which implies that the (p, q)-entry of this matrix is

$$cov(X)_{pq} = \frac{1}{m-1}(H_m \boldsymbol{v}_p)'(H_m \boldsymbol{v}_q) = \frac{1}{m-1}(\boldsymbol{v}_p - a_p \mathbf{1}_m)'(\boldsymbol{v}_q - a_q \mathbf{1}_m).$$

For a diagonal element we have

$$cov(X)_{pp} = \frac{1}{m-1}\sum_{i=1}^{m}(\boldsymbol{v}_q - a_q\mathbf{1}_m)_i^2,$$

which shows that $cov(X)_{pp}$ measures the scattering of the values of the p^{th} variable around the corresponding component a_i of the mean sample. This quantity is known as the p^{th} *variance* and is denoted by σ_p^2 for $1 \leq p \leq n$. The *total variance* tvar(X) of X is trace$(cov(X))$.

For $p \neq q$ the element c_{pq} of the matrix $C = cov(X)$ is referred to as the *(p, q)-covariance*. We have:

$$\begin{aligned}
(cov(X))_{pq} &= \frac{1}{m-1}(\boldsymbol{v}_p - a_p \mathbf{1}_m)'(\boldsymbol{v}_q - a_q \mathbf{1}_m) \\
&= \frac{1}{m}\left(\boldsymbol{v}_p' \boldsymbol{v}_q - a_p \mathbf{1}_m' \boldsymbol{v}_q - a_q \boldsymbol{v}_p' \mathbf{1}_m + m a_p a_q\right) \\
&= \boldsymbol{v}_p' \boldsymbol{v}_q - a_p a_q.
\end{aligned}$$

If $cov(X)_{pq} = 0$, then we say that the variables $\mathcal{V}_p$ and $\mathcal{V}_q$ are *uncorrelated*.

The behavior of the covariance matrix with respect to multiplication by orthogonal matrices is discussed next.

Theorem

Let

$$X = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix}$$

be a centered sample matrix and let $R \in \mathbb{R}^{n \times n}$ *be an orthogonal matrix. If* $Z \in \mathbb{R}^{m \times n}$ *is a matrix such that* $Z = XR$*, then* Z *is centered,* $cov(Z) = R' cov(X) R$ *and* $tvar(Z) = tvar(X)$.

Proof

By writing explicitly the rows of the matrix Z,

$$Z = \begin{pmatrix} \boldsymbol{z}_1 \\ \vdots \\ \boldsymbol{z}_m \end{pmatrix},$$

we have $\boldsymbol{z}_i = \boldsymbol{x}_i R$ for $1 \leq i \leq m$ because $Z = XR$. Note that the sample mean of Z is

$$\tilde{Z} = \frac{1}{m}\mathbf{1}'_m Z = \frac{1}{m}\mathbf{1}'_m XR = \tilde{X}R,$$

where $\tilde{X}$ is the sample mean of X. Since X is centered, we have $\tilde{Z} = \tilde{X} = \mathbf{0}'_n$, so Z is centered as well.

Proof cont'd

The covariance matrix of Z is

$$cov(Z) = \frac{1}{m-1} Z'Z = \frac{1}{m-1} R'X'XR = R'cov(X)R.$$

Since the trace of two similar matrices are equal and $cov(Z)$ is similar to $cov(X)$, the total variance of Z equals the total variance of X, that is,

$$\text{tvar}(Z) = \text{trace}(cov(Z)) = \text{trace}(cov(X)) = \text{tvar}(X).$$

Since the covariance matrix of a centered matrix X, $cov(X) = \frac{1}{m-1}X'X \in \mathbb{R}^{n\times n}$ is symmetric, $cov(X)$ is orthonormally diagonalizable, so there exists an orthogonal matrix $R \in \mathbb{R}^{n\times n}$ such that $R'cov(X)R = D$, which corresponds to a sample matrix $Z = XR$. Let $cov(Z) = D = \text{diag}(d_1, \ldots, d_n)$. The number d_p is the sample variance of the p^{th} variable of the data matrix, and the covariances of the form $cov(Z)_{pq}$ with $p \neq q$ are 0. From a statistical point of view, this means that the components p and q are uncorrelated. Without loss of generality we can assume that $d_1 \geq \cdots \geq d_n$.
The columns of the matrix Z correspond to the new variables $\mathcal{Z}_1, \ldots, \mathcal{Z}_n$.

Often the variables of a data sample matrix are not expressed using different units. In this case, the components of the covariance have no meaning because variables that have large numerical values have a disproportionate influence compared to variables that have small numerical value. For example, if a spatial variable is measured in millimeters, its values are three order of magnitude larger than the values of a variable expressed in meters.

UMASS BOSTON

Biplots offer a succinct and powerful way of representing graphically the elements of a matrix using two sets of vectors (hence, the term *biplot*). Let $A \in \mathbb{R}^{m \times n}$ be a matrix that can be written as a product, $A = LR$, where $L \in \mathbb{R}^{m \times r}$, $R \in \mathbb{R}^{r \times n}$ are the *left* and the *right* factors, respectively. Suppose that

$$L = \begin{pmatrix} \boldsymbol{l}'_1 \\ \vdots \\ \boldsymbol{l}'_m \end{pmatrix} \text{ and } R = (\boldsymbol{r}_1 \ \cdots \ \boldsymbol{r}_n)$$

where $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_m, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ are $m + n$ vectors in $\mathbb{R}^r$. Then, each element a_{ij} of A can be regarded as a inner product of two vectors in $\mathbb{R}^r$:

$$a_{ij} = \boldsymbol{l}'_i \boldsymbol{r}_j \tag{1}$$

for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Such matrix factorizations are common in linear algebra and we have already discussed a number of factorization techniques (full-rank decompositions, QR decompositions, etc.) Starting from the factorization $A = LR$ new factorizations of A can be built as $A = (LK')(R'K^{-1})'$ for every invertible matrix $K \in \mathbb{R}^{r \times r}$. Therefore, the above representation for A is not unique in general.

Thus, to use a biplot for a representation of the relations between the rows $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ of A one could choose R such that $RR' = I_r$, which yields

$$AA' = LRR'L' = LL'.$$

This implies $\boldsymbol{w}_i'\boldsymbol{w}_j = \boldsymbol{l}_i'\boldsymbol{l}_j$ for $1 \leq i, j \leq n$. Taking $i = j$ we have $\| \boldsymbol{w}_i' \| = \| \boldsymbol{l}_i' \|$, which, in turn, implies

$$\angle(\boldsymbol{w}_i', \boldsymbol{w}_j') = \angle(\boldsymbol{l}_i', \boldsymbol{l}_j').$$

A similar choice can be made for the columns of A by imposing the requirement $L'L = I_r$, which implies $A'A = R'R$.

The case when the rank r of the matrix A is 2 is especially interesting because we can draw the vectors $\boldsymbol{l}_1, \ldots, \boldsymbol{l}_m, \boldsymbol{r}_1, \ldots, \boldsymbol{r}_n$ to obtain an exact two-dimensional representation of A, as we show in the next example.

Example

Let

$$A = \begin{pmatrix} 18 & 8 & 20 \\ -4 & 20 & 1 \\ 25 & 8 & 27 \\ 9 & 4 & 10 \end{pmatrix}$$

be a matrix of rank 2 in $\mathbb{R}^{4\times 3}$ that can be written as $A = LR$, where

$$L = \begin{pmatrix} 2 & 4 \\ -2 & 3 \\ 3 & 5 \\ 1 & 2 \end{pmatrix} \text{ and } R = \begin{pmatrix} 5 & -4 & 4 \\ 2 & 4 & 3 \end{pmatrix}.$$

Each vector $\boldsymbol{l}_i$ corresponds to a row of A and each vector $\boldsymbol{r}_j$ to a column of A.

Example

The vectors that help us with the representation of A are

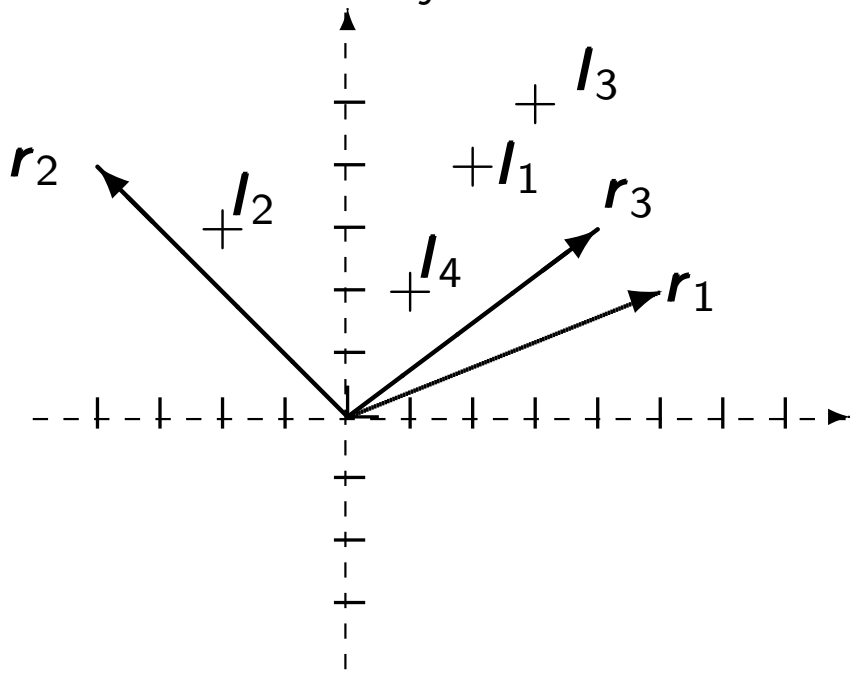
$$\boldsymbol{l}_1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, \boldsymbol{l}_2 = \begin{pmatrix} -2 \\ 3 \end{pmatrix}, \boldsymbol{l}_3 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}, \boldsymbol{l}_4 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

and

$$\boldsymbol{r}_1 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, \boldsymbol{r}_2 = \begin{pmatrix} -4 \\ 4 \end{pmatrix}, \boldsymbol{r}_3 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}.$$

UMASS BOSTON

Representation of the vectors l_i and r_j



Each vector l_i corresponds to an observation and each vector r_j to a variable.

- When we can factor a sample data matrix X as $X = LR$ a column of the right factor $\boldsymbol{r}_j$ is referred to as the *biplot axis* and corresponds to a variable $\mathcal{V}_j$.
- Each vector $\boldsymbol{l}'_i$ represents an *observation* in the sample matrix.
- The magnitude of projection of $\boldsymbol{l}_i$ on the biplot axis $\boldsymbol{r}_j$ is

$$\| \boldsymbol{l}_i \|_2 \cos \angle(\boldsymbol{l}_i, \boldsymbol{r}_j) = \frac{\boldsymbol{l}'_i \boldsymbol{r}_j}{\| \boldsymbol{r}_j \|_2} = \frac{a_{ij}}{\| \boldsymbol{r}_j \|_2}.$$

Therefore, if we choose the unit of measure on the axis $\boldsymbol{r}_j$ the number $\frac{1}{\|\boldsymbol{r}_j\|_2}$ we can read the values of the entries a_{ij} directly on the axis $\boldsymbol{r}_j$.

For instance, the unit along the biplot axis is $\frac{1}{\|\boldsymbol{r}_3\|_2} = 0.2$. It is also clear that if two axis of the biplot point roughly in the same direction, the corresponding variables will show a strong correlation.

In general, the rank of the data matrix A is larger than 2. In this case, approximative representations of A can be obtained by using the thin singular value decomposition of matrices.
Let A be a matrix of rank r and let

$$A = UDV' = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i',$$

be the thin SVD, where $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ are matrices of rank r (and, therefore, full-rank matrices) having orthonormal sets of columns. Here $U = (\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_r)$ and $V = (\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_r)$.

The matrix D containing singular values can be split between U and V by defining $L = U\sqrt{D}$ and $R = \sqrt{D}V'$. The usefulness of the SVD for biplots is based on the Eckhart-Young Theorem which stipulates that the best approximation of A in the sense of the matrix norm $\|\cdot\|_2$ in the class of matrix of rank k is the matrix defined by

$$B(k) = \sum_{i=1}^{k} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i'.$$

The same matrix $B(k)$ is the best approximation of A in the sense of Frobenius norm. The extent of the deficiency of this approximation is measured by $\| A - B(k) \|_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_r^2$. Since $\| A \|_F^2 = \sigma_1^2 + \cdots + \sigma_r^2$, an absolute measure of the quality of the approximation of A by $B(k)$ is

$$q_k = 1 - \frac{\| A - B(k) \|_F^2}{\| A \|_F^2} = \frac{\sigma_1^2 + \cdots + \sigma_k^2}{\sigma_1^2 + \cdots + \sigma_r^2}$$

In the special case, $k = 2$, the quality of the approximation is

$$q_2 = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2 + \cdots + \sigma_r^2}$$

and it is desirable that this number is as close as one as possible. The rank-2 approximation of A is useful because we can apply biplots to the visualization of A.

Example

Let $A \in \mathbb{R}^{5\times 3}$ be the matrix defined by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is easy to see that the rank of this matrix is 3 and, using MATLAB , a singular value decomposition can be obtained as shown on the next silde.

Example cont'd

Example

```
U =
    0.2787   -0.2176   -0.7071   -0.2996   -0.5341
    0.2787   -0.2176    0.7071   -0.2996   -0.5341
    0.7138    0.3398   -0.0000   -0.4037    0.4605
    0.5573   -0.4352   -0.0000    0.7033    0.0736
    0.1565    0.7749    0.0000    0.4037   -0.4605
S =
    2.3583         0         0
         0    1.1994         0
         0         0    1.0000
         0         0         0
         0         0         0
V =
    0.6572   -0.2610   -0.7071
    0.6572   -0.2610    0.7071
    0.3690    0.9294    0.0000
```

The rank-2 approximation of this matrix is

$$B(2) = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^{\mathsf{H}} + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^{\mathsf{H}},$$

and is computed in MATLAB using

```
>> B2 = 2.3583* U(:,1) * V(:,1)' + 1.1994 * U(:,2) * V(:,2)'

B2 =
    0.5000    0.5000   -0.0000
    0.5000    0.5000   -0.0000
    1.0000    1.0000    1.0000
    1.0000    1.0000   -0.0000
   -0.0000   -0.0000    1.0000
```

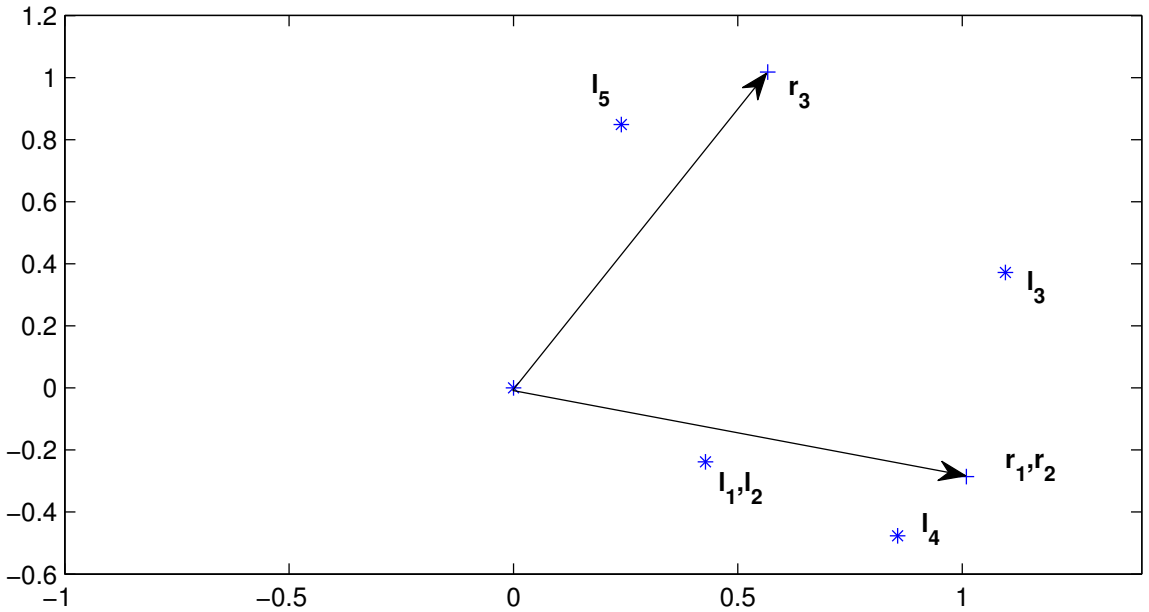
Since

$$B(2) = (\sqrt{\sigma_1}\boldsymbol{u}_1)(\sqrt{\sigma_1}\boldsymbol{v}_1)^{\mathsf{H}} + (\sqrt{\sigma_2}\boldsymbol{u}_2)(\sqrt{\sigma_2}\boldsymbol{v}_2)^{\mathsf{H}},$$

$B(2)$ can be written as

$$B(2) = \begin{pmatrix} 0.4280 & -0.2383 \\ 0.4280 & -0.2383 \\ 1.0962 & 0.3721 \\ 0.8559 & -0.4766 \\ 0.2403 & 0.8487 \end{pmatrix} \begin{pmatrix} 1.0092 & 1.0092 & 0.5667 \\ -0.2858 & -0.2858 & 1.0179 \end{pmatrix}.$$

The biplot that represents matrix A is shown in next



The quality of the approximation of A is

$$q_2 = \frac{2.3583^2 + 1.1994^2}{2.3583^2 + 1.1994^2 + 1} = 0.875$$

The "allocation" of singular values among the columns of the matrices U and V may lead to biplots that have distinct properties. For example, we could write

$$B(2) = (\sigma_1 \boldsymbol{u}_1)\boldsymbol{v}_1^{\mathsf{H}} + (\sigma_2 \boldsymbol{u}_2)\boldsymbol{v}_2^{\mathsf{H}},$$

or

$$B(2) = \boldsymbol{u}_1(\sigma_1 \boldsymbol{v}_1)^{\mathsf{H}} + \boldsymbol{u}_2(\sigma_2 \boldsymbol{v}_2)^{\mathsf{H}}.$$

The first allocation leads to the factorization $B(2) = LR$, where

$$L = \begin{pmatrix} 0.6572 & -0.2610 \\ 0.6572 & -0.2610 \\ 1.6834 & 0.4075 \\ 1.3144 & -0.5219 \\ 0.3690 & 0.9294 \end{pmatrix}$$

and

$$R = \begin{pmatrix} 0.6572 & 0.6572 & 0.3690 \\ -0.2610 & -0.2610 & 0.9294 \end{pmatrix}$$

The second yields the factors

$$L = \begin{pmatrix} 0.2787 & -0.2176 & 0.2787 & -0.2176 & 0.7138 & 0.3398 & 0.5573 & -0.4352 & 0.1565 & 0.7749 \end{pmatrix}$$

and

$$R = \begin{pmatrix} 1.5499 & 1.5499 & 0.8703 \\ -0.3130 & -0.3130 & 1.1147 \end{pmatrix}$$

The first variant leads to a representation, where the distances between the vectors $\boldsymbol{l}_i$ approximate the Euclidean distances between rows, while for the second variant, the cosine of angles between the vectors $\boldsymbol{r}_j$ approximate the correlations between variables.