CS724: Topics in Algorithms
Data Sample Characteristics

Prof. Dan A. Simovici

UMASS BOSTON

We present algebraic properties of vectors and matrices associated with data sets.

In general, vectors in $\mathbb{R}^k$ are written as column vectors. If $\mathbf{r}$ is a vector, its transpose (which is a row vector) is denoted as $\mathbf{r}'$.

Definition

Let $X = \{x_1, \ldots, x_n\}$ be a set of n numbers, where $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n$. The *median of* X, $\mu(X)$ is defined as

$$\mu(X) = \begin{cases} x_{\frac{n+1}{2}} & \text{if } n \text{ is odd,} \\ x_{\lfloor \frac{n+1}{2} \rfloor} & \text{if } n \text{ is even.} \end{cases}$$

Some authors refer to $x_{\lfloor \frac{n+1}{2} \rfloor}$ as the *lower median* and to $x_{\lceil \frac{n+1}{2} \rceil}$ as the *upper median* of U, respectively, when n is even.
When compared to the mean the median of a finite set of numbers is less sensitive to the existence of outlier values.

Example

For $X = \{1, \ldots, 9\}$ the median is 5: the numbers $1, \ldots, 4$ are smaller than the median, while $6, \ldots, 9$ are larger.
For the set $V = \{1, 2, \ldots, 10\}$ which has an even number of members the lower median is 5, while its upper median is 6.

Theorem

Let $X = \{x_1, \ldots, x_n\}$ be a set of n numbers, where $x_1 \leqslant x_2 \leqslant x_n$. The sum $f_X(a) = \sum_{i=1}^{n} |x_i - a|$ is minimal when a is the median of X (when n is odd) and the lower and the upper median, when n is even; finding this minimum can be done in $O(n \log n)$ time.

Proof

Let $u, v \in \mathbb{R}$, $u \leqslant v$. Since $v - u = |v - u| \leqslant |u - a| + |v - a|$ if follows that the minimum of $f_{u,v}(a) = |u - a| + |v - a|$ is $v - u$ which is attained when $a \in [u, v]$. This shows that in this case, the minimum is attained when a is the lower median (u) or the upper median v.
Note that

$$f_X(a) = f_{x_1,x_n}(a) + f_{x_2,x_{n-1}}(a) + \cdots + f_{x_{\lfloor \frac{n+1}{2} \rfloor}, x_{\lceil \frac{n+1}{2} \rceil}}(a),$$

and all terms of the sum in the right member are non-negative. Therefore, $f_X(a)$ is minimal when

$$a \in [x_1, x_n], a \in [x_2, x_{n-1}], \ldots, a \in [x_{\lfloor \frac{n+1}{2} \rfloor}, x_{\lceil \frac{n+1}{2} \rceil}],$$

which implies $a \in [x_{\lfloor \frac{n+1}{2} \rfloor}, x_{\lceil \frac{n+1}{2} \rceil}]$.

The notion of median can be extended to finite sets of vectors in $\mathbb{R}^m$ by replacing the absolute values used for $\mathbb{R}$ by the $\| \cdot \|_1$.

Definition

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n vectors in $\mathbb{R}^m$ and let $f_X(\mathbf{z}) = \sum_{i=1}^{n} \| \mathbf{z} - \mathbf{x}_i \|$. A *median* of X is a vector $\mathbf{x}_k \in X$ if

$$f_X(\mathbf{x}_k) = \min_{1 \leqslant p \leqslant n} \sum_{i=1}^{n} \| \mathbf{x}_p - \mathbf{x}_i \|_1 .$$

An analogous extension involves the notion of *medoid*, using the norm $\| \cdot \|_2$ on $\mathbb{R}^m$.

Definition

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n vectors in $\mathbb{R}^m$ and let $g_X(\mathbf{z}) = \sum_{i=1}^{n} \| \mathbf{z} - \mathbf{x}_i \|_2^2$.
A *medoid* of X is a vector $\mathbf{x}_k \in X$ if

$$g_X(\mathbf{x}_k) = \min_{1 \leqslant p \leqslant n} \sum_{i=1}^{n} \| \mathbf{x}_p - \mathbf{x}_i \|_2^2 .$$

Definition

The *centroid* of a finite set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $X \subseteq \mathbb{R}^m$ is the point $\mathbf{c}_X$ of $\mathbb{R}^m$ given by

$$\mathbf{c}_X = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i.$$

The centroid of a set of points is a generalization to $\mathbb{R}^m$ of the notion of mean of a set of numbers. Note that the centroid of U is not necessarily a member of X.

The notion of *inertia of* a finite subset U of $\mathbb{R}^m$ relative to a vector $\mathbf{z}$ is a notion that originates in solid mechanics.

Definition

Let $X = \{\mathbf{u}_1, \ldots, \mathbf{u}_n\}$ be a sequence of vectors in $\mathbb{R}^m$. The *inertia relative to a vector* $\mathbf{z} \in \mathbb{R}^m$ is the number

$$I_{\mathbf{z}}(X) = \sum_{j=1}^{n} \| \mathbf{x}_j - \mathbf{z} \|_2^2 .$$

The special case of the inertia of X relative to the vector $\mathbf{c}_X$ is referred to as the *sum of square errors* of X. We denote $I_{\mathbf{c}_X}(X)$ by $sse(X)$. The *mean square error* of the set X is the number $r(X)$ defined by

$$r(X) = \frac{sse(X)}{|X|}.$$

Theorem

(Huygens' Inertia Theorem) *Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a finite set of vectors in $\mathbb{R}^m$. We have*

$$I_{\mathbf{z}}(X) - I_{\mathbf{c}_X}(X) = n \parallel \mathbf{c}_X - \mathbf{z} \parallel_2^2,$$

for every $\mathbf{z} \in \mathbb{R}^m$.

Proof

The inertia of X relative to $\mathbf{c}_X$ is

$$\begin{aligned} I_{\mathbf{c}_X}(X) &= \sum_{j=1}^{n} \| \mathbf{x}_j - \mathbf{c}_X \|_2^2 = \sum_{j=1}^{n} (\mathbf{x}_j - \mathbf{c}_X)'(\mathbf{x}_j - \mathbf{c}_X) \\ &= \sum_{j=1}^{n} (\mathbf{x}_j'\mathbf{x}_j - \mathbf{c}_X'\mathbf{x}_j - \mathbf{x}_j'\mathbf{c}_X + \mathbf{c}_X'\mathbf{c}_X). \end{aligned}$$

Similarly, we have $I_{\mathbf{z}}(X) = \sum_{j=1}^{n} (\mathbf{x}_j'\mathbf{x}_j - \mathbf{z}'\mathbf{x}_j - \mathbf{x}_j'\mathbf{z} + \mathbf{z}'\mathbf{z})$, hence

$$\begin{aligned} I_{\mathbf{z}}(X) - I_{\mathbf{c}_X}(X) &= \sum_{j=1}^{n} (\mathbf{c}_X - \mathbf{z})'\mathbf{x}_j + \sum_{j=1}^{n} \mathbf{x}_j'(\mathbf{c}_X - \mathbf{z}) + \mathbf{z}'\mathbf{z} - \mathbf{c}_X'\mathbf{c}_X \\ &= (\mathbf{c}_X - \mathbf{z})' \sum_{i=1}^{n} \mathbf{x}_j + \left(\sum_{j=1}^{n} \mathbf{x}_j \right)' (\mathbf{c}_X - \mathbf{z}) + n(\mathbf{z}'\mathbf{z} - \mathbf{c}_X'\mathbf{c}_X) \\ &= n(\mathbf{c}_X - \mathbf{z})'\mathbf{c}_X + n\mathbf{c}_X'(\mathbf{c}_X - \mathbf{z}) + n(\mathbf{z}'\mathbf{z} - \mathbf{c}_X'\mathbf{c}_X) \\ &= n \| \mathbf{c}_X - \mathbf{z} \|_2^2 . \end{aligned}$$

Corollary

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ *be a set of vectors in* $\mathbb{R}^m$. *The minimal value of the inertia* $I_{\mathbf{z}}(X)$ *is achieved for* $\mathbf{z} = \mathbf{c}_X$.

UMASS BOSTON

Corollary

The sum of all squared distances between the members of a set divided by its cardinality equals the sum of the square errors of that set.

Proof

By Huygens' Theorem, the inertia of X relative to one of its members $\mathbf{x}_k$ is

$$\sum_{i=1}^{n} \| \mathbf{x}_i - \mathbf{x}_k \|^2 = I_{\mathbf{x}_k}(X) = I_{\mathbf{c}_X} + n \| \mathbf{c}_X - \mathbf{x}_k \|_2^2 .$$

Therefore,

$$\begin{aligned} \sum_{k=1}^{n} \sum_{i=1}^{n} \| \mathbf{x}_i - \mathbf{x}_k \|^2 &= 2 \sum \{ \| \mathbf{x}_i - \mathbf{x}_k \|^2 \mid 1 \leqslant k < i \leqslant n \} \\ &= n I_{\mathbf{c}_X} + n \sum_{k=1}^{n} \| \mathbf{c}_X - \mathbf{x}_k \|_2^2 = 2n I_{\mathbf{c}_X}, \end{aligned}$$

which implies the statement of the corollary.

Definition

For a set X and a partition $\pi = \{U_1, \ldots, U_k\}$ of X, the *sum of the squared errors* of π is the number

$$sse(\pi) = \sum_{i=1}^{k} sse(U_i) = \sum_{i=1}^{k} \sum \{ \| \mathbf{x} - \mathbf{c}_{U_i} \|^2 \, | \mathbf{x} \in U_i \}. \tag{1}$$

Corollary

The sum of square errors of a partition $\pi = \{U_1, \ldots, U_k\}$ *of a finite subset* U *of* $\mathbb{R}^m$ *equals the sum over all blocks of mean square errors,* $\sum_{j=1}^{k} r(U_j)$.

Lemma

Let W be a subset of $\mathbb{R}^m$ and let $\sigma = \{U, V\}$ be a bipartition of W. We have:

$$sse(W) = sse(U) + sse(V) + \frac{|U|\,|V|}{|W|} \parallel \mathbf{c}_U - \mathbf{c}_V \parallel^2 .$$

UMASS BOSTON

Proof

By applying the definition of the sum of square errors we have:

$$\begin{aligned} & sse(W) - sse(U) - sse(V) \\ = & \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 \, | \mathbf{x} \in U \cap V\} \\ & - \sum\{\| \mathbf{x} - \mathbf{c}_U \|^2 \, | \mathbf{x} \in U\} - \sum\{\| \mathbf{x} - \mathbf{c}_V \|^2 \, | \mathbf{x} \in V\}. \end{aligned}$$

The centroid of W is given by:

$$\mathbf{c}_W = \frac{1}{|W|} \sum\{\mathbf{x} | \mathbf{x} \in W\} = \frac{|U|}{|W|} \mathbf{c}_U + \frac{|V|}{|W|} \mathbf{c}_V.$$

This allows us to evaluate the variation of the sum of squared errors:

$$\begin{aligned} & sse(W) - sse(U) - sse(V) \\ = & \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 \, | \, \mathbf{x} \in U \cup V\} \\ & - \sum\{\| \mathbf{x} - \mathbf{c}_U \|^2 \, | \, \mathbf{x} \in U\} - \sum\{\| \mathbf{x} - \mathbf{c}_V \|^2 \, | \, \mathbf{x} \in V\} \\ = & \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_U \|^2 \, | \, \mathbf{x} \in U\} \\ & + \sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_V \|^2 \, | \, \mathbf{x} \in V\}. \end{aligned}$$

Proof (cont'd)

Observe that:

$$
\begin{aligned}
&\sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_U \|^2 \mid \mathbf{x} \in U\} \\
&= \sum_{\mathbf{x} \in U} \left((\mathbf{x} - \mathbf{c}_W)'(\mathbf{x} - \mathbf{c}_W) - (\mathbf{x} - \mathbf{c}_U)'(\mathbf{x} - \mathbf{c}_U) \right) \\
&= |U|(\mathbf{c}_W'\mathbf{c}_W - \mathbf{c}_U'\mathbf{c}_U) + 2(\mathbf{c}_U' - \mathbf{c}_W') \sum_{\mathbf{x} \in U} \mathbf{x} \\
&= |U|(\mathbf{c}_W'\mathbf{c}_W - \mathbf{c}_U'\mathbf{c}_U) + 2|U|(\mathbf{c}_U' - \mathbf{c}_W')\mathbf{c}_U \\
&= |U|(\| \mathbf{c}_W \|^2 - \| c_U \|^2 + 2 \| \mathbf{c}_U \|^2 - 2\mathbf{c}_W'\mathbf{c}_U) \\
&= |U|(\| \mathbf{c}_W \|^2 + \| \mathbf{c}_U \|^2 - 2\mathbf{c}_W'\mathbf{c}_U) \\
&= |U| \| \mathbf{c}_W - \mathbf{c}_U \|^2 .
\end{aligned}
$$

Proof (cont'd)

Using the equality

$$\mathbf{c}_W - \mathbf{c}_U = \frac{|U|}{|W|}\mathbf{c}_U + \frac{|V|}{|W|}\mathbf{c}_V - \mathbf{c}_U = \frac{|V|}{|W|}\left(\mathbf{c}_V - \mathbf{c}_U\right), \tag{2}$$

we obtain

$$\sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_U \|^2 \mid \mathbf{x} \in U\} = \frac{|U||V|^2}{|W|^2} \| \mathbf{c}_V - \mathbf{c}_U \|^2 .$$

Proof (cont'd)

In a similar manner we have:

$$\sum\{\| \mathbf{x} - \mathbf{c}_W \|^2 - \| \mathbf{x} - \mathbf{c}_V \|^2 \mid \mathbf{x} \in V\} = \frac{|U|^2|V|}{|W|^2} \| \mathbf{c}_V - \mathbf{c}_U \|^2,$$

so,

$$sse(W) - sse(U) - sse(V) = \frac{|U||V|}{|W|} \| \mathbf{c}_V - \mathbf{c}_U \|^2,$$

which is the equality we needed to prove.

Theorem

Let X be a finite set. The function $sse : PART(X) \longrightarrow \mathbb{R}_{\geqslant 0}$ between the posets $(PART(X), \leqslant)$ and $(\mathbb{R}_{\geqslant 0}, \leqslant)$ is monotonic.

Proof

It suffices to show that if $\pi \prec \pi'$, then $sse(\pi) \leqslant sse(\pi')$. If two blocks U and V of a partition π are fused into a new block W to yield a new partition π' that covers π then, by a previous Lemma the variation of the sum of squared errors is given by

$$sse(\pi') - sse(\pi) = sse(W) - sse(U) - sse(V) = \frac{|U|\,|V|}{|W|} \parallel \mathbf{c}_U - \mathbf{c}_V \parallel^2 \geqslant 0.$$

Theorem

Let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ *be a finite subset of* $\mathbb{R}^m$. *The centroid of* X *is the point* $\mathbf{c}$ *that minimizes* $\sum_{\mathbf{x} \in X} \| \mathbf{x} - \mathbf{c} \|_2^2$.

UMASS BOSTON

Proof

Since the function $f(\mathbf{c}) = \sum_{i=1}^{n} \| \mathbf{x}_i - \mathbf{c} \|_2^2$ is differentiable, the necessary conditions for the minimum are $\frac{\partial f}{\partial c_p}(\mathbf{c}) = 0$ for $1 \leqslant p \leqslant m$, and these amount to

$$2nc_p - 2\sum_{i=1}^{n}(\mathbf{x}_i)_p = 0,$$

for $1 \leqslant p \leqslant m$, hence $c_p = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i)_p$. Thus, $\mathbf{c} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$.

Let a data set consist of a sequence $\mathcal{E}$ of m vectors of $\mathbb{R}^n$, $(\mathbf{u}_1, \ldots, \mathbf{u}_m)$.

- The j^{th} components $(\mathbf{u}_i)_j$ 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 $\mathbf{u}'_1, \ldots, \mathbf{u}'_m$ and n columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$. We refer to matrices obtained in this manner as *sample matrices*. The number m is the *size* of the sample.

Each row vector $\mathbf{u}_i'$ 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 $\mathbf{u}_i' = (x_{i1}, \ldots, x_{in})$, as shown below.

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

The column vector

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

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 *mean* of D is the vector $\tilde{D} = \frac{1}{m} D' \mathbf{1}_m \in \mathbb{R}^n$. D is *centered* if $\tilde{D} = \mathbf{0}_n$.

In particular, if $D = \mathbf{v} \in \mathbb{R}^m$, then $\tilde{D} = \frac{1}{m} \mathbf{v}' \mathbf{1}_m$.

The following data matrix records the weights and heigths of five individuals:

weight (lbs)	height (m)
180	1.78
150	1.64
210	1.90
140	1.50
170	1.89

This data matrix D belongs to $\mathbb{R}^{5\times 2}$ and has two features: weight and height, and five observations:

$$\begin{pmatrix} 180 \\ 1.78 \end{pmatrix}, \begin{pmatrix} 150 \\ 1.64 \end{pmatrix}, \begin{pmatrix} 210 \\ 1.90 \end{pmatrix}, \begin{pmatrix} 140 \\ 1.50 \end{pmatrix}, \begin{pmatrix} 170 \\ 1.89 \end{pmatrix}$$

We have $D \in \mathbb{R}^{5\times 2}$, so $\tilde{D} \in \mathbb{R}^2$ is given by

$$\begin{aligned}
\tilde{D} &= \frac{1}{5} D' \mathbf{1}_5 \\
&= \frac{1}{5} \begin{pmatrix} 180 & 150 & 210 & 140 & 170 \\ 1.78 & 1.64 & 1.90 & 1.50 & 1.89 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \\
&= \begin{pmatrix} 170 \\ 1.74 \end{pmatrix}
\end{aligned}$$

Theorem

Let $D \in \mathbb{R}^{m \times n}$ be a data matrix and let

$$H_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m = I_m - \frac{1}{m}J_{m,m}.$$

Then $H_m D$ is a centered data matrix.

Proof

We have

$$\begin{aligned}\widetilde{(H_m D)} &= \frac{1}{m} D' H'_m \mathbf{1}_m = \frac{1}{m} D' \left(I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}'_m \right)' \mathbf{1}_m \\ &= \frac{1}{m} D' \left(I_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}'_m \right) \mathbf{1}_m = \frac{1}{m} D' \left(\mathbf{1}_m - \frac{1}{m} \mathbf{1}_m \mathbf{1}'_m \mathbf{1}_m \right) = \mathbf{0}_n.\end{aligned}$$

Theorem

The centering matrix $H_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m$ *is both symmetric and idempotent; further,* $H_m\mathbf{1}_m = \mathbf{0}_m$.

Proof.

Note that

$$\begin{aligned} H_m^2 &= (I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m)(I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m) \\ &= I_m - 2\frac{1}{m}\mathbf{1}_m\mathbf{1}'_m + \frac{1}{m^2}\mathbf{1}_m\mathbf{1}'_m\mathbf{1}_m\mathbf{1}'_m. \end{aligned}$$

Since $\mathbf{1}'_m\mathbf{1}_m = m$, it follows that $H_m^2 = H_m$, hence H_m is indeed idempotent.

The symmetry of H_m is immediate. □

For $D \in \mathbb{R}^{5\times 2}$ the centering matrix is

$$H_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} - \frac{1}{5} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} (1\ 1\ 1\ 1\ 1),$$

so

$$H = \begin{pmatrix} 0.8000 & -0.2000 & -0.2000 & -0.2000 & -0.2000 \\ -0.2000 & 0.8000 & -0.2000 & -0.2000 & -0.2000 \\ -0.2000 & -0.2000 & 0.8000 & -0.2000 & -0.2000 \\ -0.2000 & -0.2000 & -0.2000 & 0.8000 & -0.2000 \\ -0.2000 & -0.2000 & -0.2000 & -0.2000 & 0.8000 \end{pmatrix}.$$

Thus, the centered matrix is

$$H_5 A = \begin{pmatrix} 10.0000 & -0.1220 \\ -20.0000 & -0.0620 \\ 40.0000 & 0.1980 \\ -30.0000 & -0.2020 \\ 0 & 0.1880 \end{pmatrix}$$

For a data matrix

$$D = \begin{pmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_m' \end{pmatrix} = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n) \in \mathbb{R}^{m \times n}$$

and $\mathbf{z} \in \mathbb{R}^n$. The minimal value of the inertia $I_{\mathbf{z}}(\mathbf{u})$ is achieved for $\mathbf{z} = \tilde{D}$.

Definition

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

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

where $\tilde{v} = \frac{1}{m}\sum_{i=1}^{m} v_i$ is the mean of the components of $\mathbf{v}$.
The *variance* is the number $var(\mathbf{v})$ given by

$$var(\mathbf{v}) = s_{\mathbf{v}}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(v_i - \tilde{v})^2.$$

Definition

The *standard deviation of a data matrix* $D \in \mathbb{R}^{m \times n}$, where $D = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)$ is the row

$$\mathbf{s} = (s_{\mathbf{v}_1}, \ldots, s_{\mathbf{v}_n}) \in \mathbb{R}^n.$$

The *standard deviation of a data matrix* $D \in \mathbb{R}^{m \times n}$, where $D = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)$ is the row $\mathbf{s} = (s_{\mathbf{v}_1}, \ldots, s_{\mathbf{v}_n}) \in \mathbb{R}^n$.

UMASS BOSTON

Let $\mathbf{u}$ and $\mathbf{w}$ be two vectors in $\mathbb{R}^m$, where $m > 1$, having the means $\tilde{u}$ and $\tilde{w}$, and the standard deviations s_u and s_v, respectively.
The *covariance coefficient* of $\mathbf{u}$ and $\mathbf{w}$ is the number

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

Definition

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

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

The *covariance matrix* of a data matrix $D \in \mathbb{R}^{m \times n}$ is

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

The *total variance* $\text{TVAR}(D)$ of D is the number

$$\text{TVAR}(D) = trace(cov(X)).$$

Theorem

For $\mathbf{v} \in \mathbb{R}^m$ *we have:*

$$var(\mathbf{v}) = \frac{1}{m-1}\left(\| \mathbf{v} \|^2 - m\tilde{v}^2\right).$$

Proof

Since

$$var(\mathbf{v}) = \frac{1}{m-1}\sum_{i=1}^{m}(v_i - \tilde{v})^2,$$

we have

$$\begin{aligned} var(\mathbf{v}) &= \frac{1}{m-1}\left(\sum_{i=1}^{m} v_i^2 - 2\tilde{v}\sum_{i=1}^{m} v_i + m\tilde{v}^2\right) \\ &= \frac{1}{m-1}\left(\sum_{i=1}^{m} v_i^2 - 2m\tilde{v}^2 + m\tilde{v}^2\right) \\ &= \frac{1}{m-1}\left(\sum_{i=1}^{m} v_i^2 - m\tilde{v}^2\right), \end{aligned}$$

which is the desired equality.

Theorem

Let $D \in \mathbb{R}^{m \times n}$ be a data matrix, where

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

The mean square distance between column vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is equal to twice the sum of row variances, $\sum_{i=1}^{m} var(\mathbf{u}_i)$.

Proof

The mean square distance between the columns of D is

$$\begin{aligned}
&\frac{2}{n(n-1)} \sum_{i<j} \| \mathbf{v}_i - \mathbf{v}_j \|^2 \\
&= \frac{2}{n(n-1)} \left(\sum_{j=1}^{n} \| \mathbf{v}_j \|^2 - 2 \sum_{i<j} \mathbf{v}_i' \mathbf{v}_j \right) \\
&= \frac{2}{n(n-1)} \left((n-1) \| D \|_F^2 + \| D \|_F^2 - \mathbf{1}_n' D D' \mathbf{1}_n \right) \\
&= \frac{2}{n(n-1)} \left(n \| D \|_F^2 - \mathbf{1}_n' D D' \mathbf{1}_n \right).
\end{aligned}$$

Since each vector $\mathbf{u}_k$ belongs to $\mathbb{R}^n$, the the sum of row variances is

$$\sum_{k=1}^{m} var(\mathbf{u}_k) = \sum_{k=1}^{m} \frac{1}{n-1} \left(\| \mathbf{u}_k \|^2 - n\tilde{u}_k^2 \right) = \frac{1}{n-1} \| D \|_F^2 - \frac{n}{n-1} \sum_{k=1}^{n} \tilde{u}_k^2.$$

Taking into account that

$$\tilde{u}_k = \frac{1}{n}\mathbf{1}_n' D' \mathbf{e}_k = \frac{1}{n} e_k' D \mathbf{1}_n,$$

we have $\tilde{u}_k^2 = \frac{1}{n^2}\mathbf{1}_n' D' \mathbf{e}_k e_k' D \mathbf{1}_n$, which implies

$$\sum_{k=1}^{n} \tilde{u}_k^2 = \frac{1}{n^2}\mathbf{1}_n' D' \left(\sum_{k=1}^{n} \mathbf{e}_k \mathbf{e}_k' \right) D \mathbf{1}_n = \frac{1}{n^2}\mathbf{1}_n' D' D \mathbf{1}_n,$$

because $\sum_{k=1}^{n} \mathbf{e}_k \mathbf{e}_k' = I_m$. The desired equality follows immediately.

Theorem

For any vectors $\mathbf{u}, \mathbf{w} \in \mathbb{R}^m$ *we have* $-1 \leq \rho(\mathbf{u}, \mathbf{w}) \leq 1$.

Proof

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

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

Theorem

Let

$$D \in \mathbb{R}^{m \times n} = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)$$

be a centered data matrix and let $R \in \mathbb{R}^{n \times n}$ *be an orthogonal matrix. The matrix* $Z = DR \in \mathbb{C}^{m \times n}$ *is centered and* $cov(DR) = R' cov(D) R$.

UMASS BOSTON

Proof

By writing explicitly the rows of the matrix Z,

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

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

$$\tilde{Z}' = \frac{1}{m}\mathbf{1}'_m Z = \frac{1}{m}\mathbf{1}'_m DR = \tilde{D}' R.$$

Since D is centered, we have $\tilde{D}' = \mathbf{0}'_n$, so Z is centered as well.

Proof (cont'd)

The covariance matrix of Z is

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

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

$$\text{TVAR}(Z) = trace(cov(Z)) = trace(cov(D)) = \text{TVAR}(D).$$

Since the covariance matrix of a centered matrix D, $cov(D) = \frac{1}{m-1} D'D \in \mathbb{R}^{n \times n}$ is symmetric $cov(X)$ is orthonormally diagonalizable, so there exists an orthogonal matrix $R \in \mathbb{R}^{n \times n}$ such that $R'cov(D)R = D$, which corresponds to a matrix $Z = DR$.

Proof (cont'd)

Let $cov(Z) = D = \text{diag}(d_1, \ldots, d_n)$.
d_p is the 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.