Linear Methods in Data Mining

Dan A. Simovici

Why Linear Methods?

- linear methods are well understood, simple and elegant;
- algorithms based on linear methods are widespread: data mining, computer vision, graphics, pattern recognition;
- excellent general software available for experimental work.

Software available

- the JAMA java package available free on Internet;
- MATLAB, excellent, but expensive;
- SCILAB (free)
- OCTAVE (free)

Least Squares for Linear Regression

Main Problem: Given observations of p variables in n experiments $\mathbf{a}_1, \ldots, \mathbf{a}_p$ in $\mathbb{R}^n$ and the vector of the outcomes of the experiments $\mathbf{b} \in \mathbb{R}^n$ determine $b_0, b_1, \ldots, b_p$ to express the outcome of the experiment $\mathbf{b}$ as

$$\mathbf{b} = \alpha_0 + \sum_{i=1}^{p} \mathbf{a}_i \alpha_i.$$

Data Sets as Tables

Series of n experiments measuring p variables and an outcome:

$\mathbf{a}_1$	$\cdots$	$\mathbf{a}_p$	$\mathbf{b}$
a_{11}	$\cdots$	a_{1p}	b_1
$\vdots$	$\cdots$	$\vdots$	$\vdots$
a_{n1}	$\cdots$	a_{np}	b_n

Least Squares for Linear Regression

In Matrix Form: Determine $\alpha_0, \alpha_1, \ldots, \alpha_p$ such that

$$(\mathbf{1}\ \mathbf{a}_1\ \cdots\ \mathbf{a}_p) \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \mathbf{b}.$$

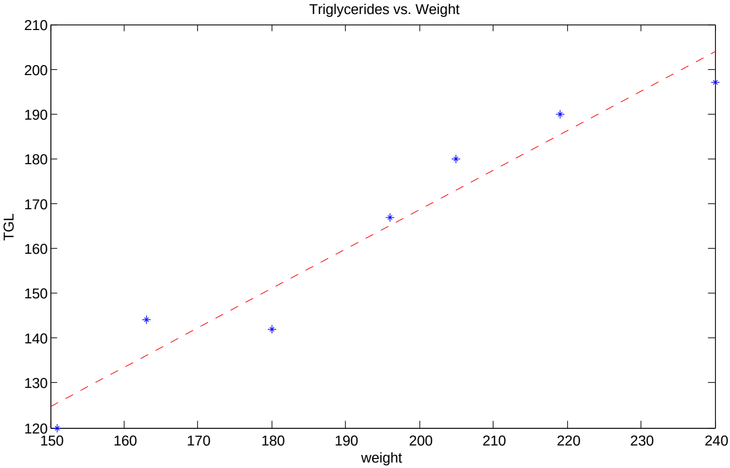
This system consists of n equations and $p+1$ unknowns $\alpha_0, \ldots, \alpha_p$ and is overdetermined.

Example

Data on relationship between weight (in lbs) and blood triglycerides:

weight	tgl
151	120
163	144
180	142
196	167
205	180
219	190
240	197

Example cont'd



Regression Quest

Find α_0 and α_1 such that $\text{tgl} = \alpha_0 + \alpha_1 * \text{weight}$.

$$
\begin{aligned}
\alpha_0 + 151\alpha_1 &= 120 \\
\alpha_0 + 163\alpha_1 &= 144 \\
\alpha_0 + 180\alpha_1 &= 142 \\
\alpha_0 + 196\alpha_1 &= 167 \\
\alpha_0 + 205\alpha_1 &= 180 \\
\alpha_0 + 219\alpha_1 &= 190 \\
\alpha_0 + 240\alpha_1 &= 197
\end{aligned}
$$

2 unknowns and 7 equations!

In matrix form:

$$\begin{pmatrix} 1 & 151 \\ 1 & 163 \\ 1 & 180 \\ 1 & 196 \\ 1 & 205 \\ 1 & 219 \\ 1 & 240 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 120 \\ 144 \\ 142 \\ 167 \\ 180 \\ 190 \\ 197 \end{pmatrix}$$

or $A\boldsymbol{\alpha} = \mathbf{b}$.

Finding parameters allows building a model and making predictions for values of tgl based on weight.

The Least Square Method (LSM)

- LSM is used in data mining as a method of estimating the parameters of a model.
- Estimation process is known as regression.
- Several types of regression exist depending on the nature of the assummed model of dependency.

Making the Best of Overdetermined Systems

- If $A\mathbf{x} = \mathbf{b}$ has no solution, the "next best thing" is finding $\mathbf{c} \in \mathbb{R}^n$ such that

$$\| A\mathbf{c} - \mathbf{b} \|_2 \leq \| A\mathbf{x} - \mathbf{b} \|_2$$

  for every $\mathbf{x} \in \mathbb{R}^n$.
- $A\mathbf{x} \in \text{range}(A)$ for any $\mathbf{x} \in \mathbb{R}^n$.
- Find a $\mathbf{u} \in \text{range}(A)$ such that $A\mathbf{u}$ is as close to $\mathbf{b}$ as possible.

- $A \in \mathbb{R}^{m \times n}$ be a full-rank matrix ($\text{rank}(A) = n$) such that $m > n$.
- The symmetric square matrix $A'A \in \mathbb{R}^{n \times n}$ has the same rank n as the matrix A
- $(A'A)\mathbf{x} = A'\mathbf{b}$ has a unique solution.
- $A'A$ is positive definite because $\mathbf{x}'A'A\mathbf{x} = (A\mathbf{x})'A\mathbf{x} = \parallel A\mathbf{x} \parallel_2^2 > 0$ if $\mathbf{x} \neq \mathbf{0}$.

Linear Regression Theorem

Let $A \in \mathbb{R}^{m \times n}$ be a full-rank matrix such that $m > n$ and let $\mathbf{b} \in \mathbb{R}^m$. The unique solution of the system $(A'A)\mathbf{x} = A'\mathbf{b}$ equals the projection of the vector $\mathbf{b}$ on the subspace range(A).

$$(A'A)\mathbf{x} = A'\mathbf{b}$$

is known as the system of normal equations of A and $\mathbf{b}$.

Example in MATLAB

$C = A' * A$

$$C = \begin{pmatrix} 7 & 1354 \\ 1354 & 267772 \end{pmatrix}$$

$x = C \setminus (A' * b)$

$$x = \begin{pmatrix} -7.3626 \\ 0.8800 \end{pmatrix}$$

Regression line: $\text{tgl} = 0.8800 * \text{weight} - 7.3626$

Danger: Choosing the Simplest Way!

- Condition number of $A' * A$ is the square of condition number of A.
- Forming $A' * A$ leads to loss of information.

The Thin QR Decompositions

Let $A \in \mathbb{R}^{m \times n}$ be a matrix such that $m \geq n$ and $\text{rank}(A) = n$ (full-rank matrix).
A can be factored as

$$A = Q \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix},$$

where $Q \in \mathbb{R}^{m \times m}$ and $R \in \mathbb{R}^{n \times n}$ such that

i. Q is an orthonormal matrix, and
ii. $R = (r_{ij})$ is an upper triangular matrix.

LSQ Approximation and QR Decomposition

$$
\begin{aligned}
A\mathbf{u} - \mathbf{b} &= Q \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \mathbf{u} - \mathbf{b} \\
&= Q \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \mathbf{u} - QQ'\mathbf{b} \\
&\quad \text{(because } Q \text{ is orthonormal and therefore } QQ' = I_m) \\
&= Q \left(\begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \mathbf{u} - Q'\mathbf{b} \right).
\end{aligned}
$$

LSQ Approximation and QR Decomposition (cont'd)

$$\| A\mathbf{u} - \mathbf{b} \|_2^2 = \left\| \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \mathbf{u} - Q'\mathbf{b} \right\|_2^2$$

If we write $Q = (L_1 \; L_2)$, where $L_1 \in \mathbb{R}^{m\times n}$ and $L_2 \in \mathbb{R}^{m\times(m-n)}$, then

$$\begin{aligned} \| A\mathbf{u} - \mathbf{b} \|_2^2 &= \left\| \begin{pmatrix} R \\ O_{m-n,n} \end{pmatrix} \mathbf{u} - \begin{pmatrix} L_1'\mathbf{b} \\ L_2'\mathbf{b} \end{pmatrix} \right\|_2^2 \\ &= \left\| \begin{pmatrix} R\mathbf{u} - L_1'\mathbf{b} \\ -L_2'\mathbf{b} \end{pmatrix} \right\|_2^2 \\ &= \| R\mathbf{u} - L_1'\mathbf{b} \|_2^2 + \| L_2'\mathbf{b} \|_2^2 . \end{aligned}$$

LSQ Approximation and QR Decomposition (cont'd)

The system $R\mathbf{u} = L_1'\mathbf{b}$ can be solved and its solution minimizes $\| A\mathbf{u} - \mathbf{b} \|_2$.

Example

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{pmatrix} \text{ and } b = \begin{pmatrix} 20 \\ 18 \\ 25 \\ 28 \\ 27 \\ 30 \end{pmatrix}$$

$$[Q, R] = \mathrm{qr}(A, 0)$$

$$Q = \begin{pmatrix} -0.4082 & -0.5976 \\ -0.4082 & -0.3586 \\ -0.4082 & -0.1195 \\ -0.4082 & 0.1195 \\ -0.4082 & 0.3586 \\ -0.4082 & 0.5976 \end{pmatrix} \text{ and } R = \begin{pmatrix} -2.4495 & -8.5732 \\ 0 & 4.1833 \end{pmatrix}$$

$x = R \setminus (Q' * b)$

$$x = (16.66672.2857)$$

$\mathrm{cond}(A) = 9.3594$

Sample Data Sets

$$
\begin{array}{cc|ccccc|}
\cline{3-7}
 & & V_1 & \cdots & V_j & \cdots & V_p \\
\cline{3-7}
E_1 & \mathbf{x}_1 & x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\
E_i & \mathbf{x}_i & x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\
E_n & \mathbf{x}_n & x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \\
\cline{3-7}
\end{array}
$$

Sample Matrix

$$X_{\mathcal{E}} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \in \mathbb{R}^{n \times p}.$$

Sample Mean: $\tilde{\mathbf{x}} = \frac{1}{n}(\mathbf{x}_1 + \cdots + \mathbf{x}_n) = \frac{1}{n}\mathbf{1}'X.$

Centered Sample Matrix

$$\hat{X} = \begin{pmatrix} \mathbf{x}_1 - \tilde{\mathbf{x}} \\ \vdots \\ \mathbf{x}_n - \tilde{\mathbf{x}} \end{pmatrix}$$

Covariance Matrix: $\mathrm{cov}(X) = \frac{1}{n-1}\hat{X}'\hat{X} \in \mathbb{R}^{p \times p}$.

- $$\mathrm{cov}(X)_{ii} = \frac{1}{n-1}\sum_{k=1}^{n}((\mathbf{v}_i)_k)^2 = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ik} - \tilde{x}_i)^2,$$

c_{ii} is the i^{th} variance and c_{ij} is the (i,j)-covariance.

Centered Sample Matrix

$$\hat{X} = \begin{pmatrix} \mathbf{x}_1 - \tilde{\mathbf{x}} \\ \vdots \\ \mathbf{x}_n - \tilde{\mathbf{x}} \end{pmatrix}$$

Covariance Matrix: $\mathrm{cov}(X) = \frac{1}{n-1} \hat{X}' \hat{X} \in \mathbb{R}^{p \times p}$.

- $$\mathrm{cov}(X)_{ii} = \frac{1}{n-1} \sum_{k=1}^{n} ((\mathbf{v}_i)_k)^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ik} - \tilde{x}_i)^2,$$
- For $i \neq j$,
$$(\mathrm{cov}(X))_{ij} = \frac{n}{n-1} \left(\frac{1}{n} \sum_{k=1}^{n} x_{ik} x_{jk} - \tilde{x}_i \tilde{x}_j \right)$$

c_{ii} is the i^{th} variance and c_{ij} is the (i, j)-covariance.

Total Variance

tvar(X) of X is trace(C).
Let

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} \in \mathbb{R}^{n \times p}$$

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

Properties of the Covariance matrix

$\mathrm{cov}(X) = \frac{1}{n-1} X'X$ is symmetric, so is orthonormally diagonalizable:
There exists an orthonormal matrix R such that $R'\mathrm{cov}(X)R = D$, which corresponds to a sample matrix $Z = XR$. Since

$$Z = \begin{pmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_n \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} R,$$

New data form:

$$\mathbf{z}_i = \mathbf{x}_i R$$

Properties of Diagonalized Covariance Matrix

$\mathrm{cov}(Z) = D = \mathrm{diag}(d_1, \ldots, d_p)$, then d_i is the variance of the i^{th} component; the covariances of the form $\mathrm{cov}(i, j)$ are 0. From a statistical point of view, this means that the components i and j are uncorrelated.

The *principal components* of the sample matrix X are the eigenvectors of the matrix $\mathrm{cov}(X)$: the column of R that diagonalizes $\mathrm{cov}(X)$.

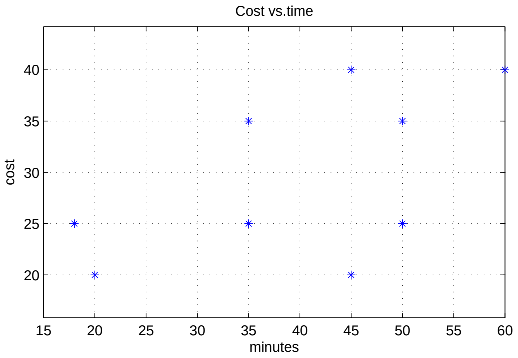
Variance Explanation

The sum of the elements of D's main diagonal is equal to the total variance $\text{tvar}(X)$.
The principal components "explain" the sources of the total variance: sample vectors grouped around $\mathbf{p}_1$ explain the largest portion of the variance; sample vectors grouped around $\mathbf{p}_2$ explain the second largest portion of the variance, etc.

Example

		minutes	cost
E_1	$\mathbf{x}_1$	15	20
E_2	$\mathbf{x}_2$	18	25
E_3	$\mathbf{x}_3$	20	20
E_4	$\mathbf{x}_4$	35	25
E_5	$\mathbf{x}_5$	35	35
E_6	$\mathbf{x}_6$	45	20
E_7	$\mathbf{x}_7$	45	40
E_8	$\mathbf{x}_8$	50	25
E_9	$\mathbf{x}_9$	50	35
E_{10}	$\mathbf{x}_{10}$	60	40

Price vs. repair time



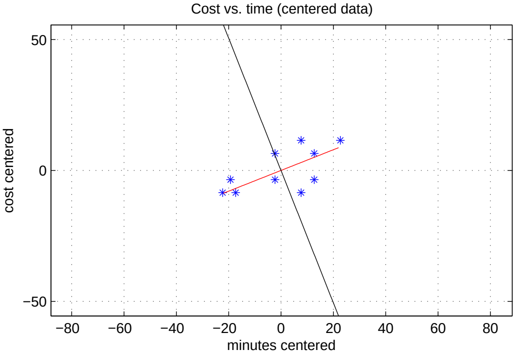
The principal components of X:

$$\mathbf{r}_1 = \begin{pmatrix} 0.37 \\ -0.93 \end{pmatrix} \text{ and } \mathbf{r}_2 = \begin{pmatrix} -0.93 \\ -0.37 \end{pmatrix},$$

corresponding to the eigenvalues

$$\lambda_1 = 35.31 \text{ and } \lambda_2 = 268.98.$$

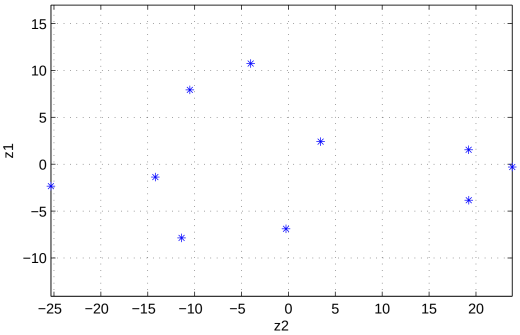
Centered Data and Principal Components



The New Variables

$$
\begin{aligned}
z_1 &= 0.37\,(\texttt{minutes} - 37.30) - 0.93\,(\texttt{price} - 28.50) \\
z_2 &= -0.93\,(\texttt{minutes} - 37.30) - 0.37\,(\texttt{price} - 28.50),
\end{aligned}
$$

The New Variables



The Optimality Theorem of PCA

Let $X \in \mathbb{R}^{n \times p}$ be a centered sample matrix and let $R \in \mathbb{R}^{p \times p}$ be the orthonormal matrix such that $(p-1)R'\text{cov}(X)R$ is a diagonal matrix $D \in \mathbb{R}^{p \times p}$, where $d_{11} \geq \cdots \geq d_{pp}$.
Let $Q \in \mathbb{R}^{p \times \ell}$ be a matrix whose set of columns is orthonormal and $1 \leq \ell \leq p$, and let $W = XQ \in \mathbb{R}^{n \times \ell}$. Then, $\text{trace}(\text{cov}(W))$ 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.

Singular values and vectors

Let $A \in \mathbb{C}^{m \times n}$ be a matrix. A number $\sigma \in \mathbb{R}_{>0}$ is a *singular value* of A if there exists a pair of vectors $(\mathbf{u}, \mathbf{v}) \in \mathbb{C}^n \times \mathbb{C}^m$ such that

$$A\mathbf{v} = \sigma\mathbf{u} \text{ and } A^{\mathsf{H}}\mathbf{u} = \sigma\mathbf{v}.$$

The vector $\mathbf{u}$ is the *left singular vector* and $\mathbf{v}$ is the *right singular vector* associated to the singular value σ.

SVD Facts

- The matrices $A^{\mathsf{H}}A$ and AA^{H} have the same non-zero eigenvalues.
- If σ is a singular value of A, then σ^2 is an eigenvalue of both $A^{\mathsf{H}}A$ and AA^{H}.
- Any right singular vector $\mathbf{v}$ is an eigenvector of $A^{\mathsf{H}}A$ and any left singular vector $\mathbf{u}$ is an eigenvector of AA^{H}.

SVD Facts (cont'd)

- If λ be an eigenvalue of $A^{\mathsf{H}}A$ and AA^{H}, λ is a real non-negative number.
- There is $\mathbf{v} \in \mathbb{C}^n$ such that $A^{\mathsf{H}}A\mathbf{v} = \lambda\mathbf{v}$. Let $\mathbf{u} \in \mathbb{C}^m$ be the vector defined by $A\mathbf{v} = \sqrt{\lambda}\mathbf{u}$.
- We have $A^{\mathsf{H}}\mathbf{u} = \sqrt{\lambda}\mathbf{v}$, so $\sqrt{\lambda}$ is a singular value of A and $(\mathbf{u}, \mathbf{v})$ is a pair of singular vectors associate with the singular value $\sqrt{\lambda}$.
- If $A \in \mathbb{C}^{n \times n}$ is invertible and σ is a singular value of A, then $\frac{1}{\sigma}$ is a singular value of the matrix A^{-1}.

Example: SVD of a Vector

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

be a non-zero vector is $\mathbb{C}^n$, which can also be regarded as a matrix in $\mathbb{C}^{n \times 1}$. The square of a singular value of A is an eigenvalue of the matrix

$$A^{\mathsf{H}} A = \begin{pmatrix} \bar{a}_1 a_1 & \cdots & \bar{a}_n a_1 \\ \bar{a}_1 a_2 & \cdots & \bar{a}_n a_2 \\ \vdots & \cdots & \vdots \\ \bar{a}_1 a_n & \cdots & \bar{a}_n a_n \end{pmatrix}$$

The unique non-zero eigenvalue of this matrix is $\| a \|_2^2$, so the unique singular value of $\mathbf{a}$ is $\| a \|_2$.

A Decomposition of Square Matrices

Let $A \in \mathbb{C}^{n \times n}$ be a unitarily diagonalizable matrix. There exists a diagonal matrix $D \in \mathbb{C}^{n \times n}$ and a unitary matrix $U \in \mathbb{C}^{n \times n}$ such that $A = U^{\mathsf{H}} D U$; equivalently, we have $A = V D V^{\mathsf{H}}$, where $V = U^{\mathsf{H}}$.
If $V = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n)$, then $A\mathbf{v}_i = d_i \mathbf{v}_i$, where $D = \mathrm{diag}(d_1, \ldots, d_n)$, so $\mathbf{v}_i$ is a unit eigenvector of A that corresponds to the eigenvalue d_i.

$$A = (\mathbf{v}_1 \ \cdots \ \mathbf{v}_n) \begin{pmatrix} d_1 \mathbf{v}_1^{\mathsf{H}} \\ \vdots \\ d_n \mathbf{v}_n^{\mathsf{H}} \end{pmatrix}$$

implies

$$A = d_1 \mathbf{v}_1 \mathbf{v}_1^{\mathsf{H}} + \cdots + d_n \mathbf{v}_n \mathbf{v}_n^{\mathsf{H}},$$

Decomposition of Rectangular Matrices

Theorem (SVD Decomposition Theorem)

Let $A \in \mathbb{C}^{m \times n}$ be a matrix with singular values $\sigma_1, \sigma_2, \ldots, \sigma_p$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$ and $p \leq \min\{m, n\}$. There exist two unitary matrices $U \in \mathbb{C}^{m \times m}$ and $V \in \mathbb{C}^{n \times n}$ such that

$$A = U diag(\sigma_1, \ldots, \sigma_p) V^H, \tag{1}$$

where $diag(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}$.

SVD Properties

Let $A = U\mathrm{diag}(\sigma_1, \ldots, \sigma_p)V^{\mathsf{H}}$ be the SVD decomposition of the matrix A, where $\sigma_1, \ldots, \sigma_p$ are the singular values of A. The rank of A equals p, the number of the non-zero elements located on the diagonal of D.

The Thin SVD Decomposition

Let $A \in \mathbb{C}^{m \times n}$ be a matrix with singular values $\sigma_1, \sigma_2, \ldots, \sigma_p$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$ and $p \leq \min\{m, n\}$. Then, A can be factored as $A = UDV^{\mathsf{H}}$, where $U \in \mathbb{C}^{m \times p}$ and $V \in \mathbb{C}^{n \times p}$ are matrices having orthonormal columns and D is the diagonal matrix

$$D = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \cdots & \sigma_p \end{pmatrix}.$$

SVD Facts

- The rank-1 matrices of the form $\mathbf{u}_i\mathbf{v}_i^{\mathsf{H}}$, where $1 \leq i \leq p$ are pairwise orthogonal.
- $\parallel \mathbf{u}_i\mathbf{v}_i^{\mathsf{H}} \parallel_F = 1$ for $1 \leq i \leq p$.

Eckhart-Young Theorem

Let $A \in \mathbb{C}^{m\times n}$ be a matrix whose sequence of non-zero singular values is $\sigma_1 \geq \cdots \geq \sigma_p > 0$. A can be written as

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{\mathsf{H}} + \cdots + \sigma_p \mathbf{u}_p \mathbf{v}_p^{\mathsf{H}}.$$

Let $B(k) \in \mathbb{C}^{m\times n}$ be

$$B(k) = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathsf{H}}.$$

If $r_k = \inf\{\|A - X\|_2 \mid X \in \mathbb{C}^{m\times n} \text{ and } \operatorname{rank}(X) \leq k\}$, then

$$\|A - B(k)\|_2 = r_k = \sigma_{k+1},$$

for $1 \leq k \leq p$, where $\sigma_{p+1} = 0$.

Central Issue in IR

The computation of sets of documents that contain terms specified by queries submitted by users.

Challenges:

- a concept can be expressed by many equivalent words (*synonimy*);
- and the same word may mean different things in various contexts (*polysemy*);
- this can lead the retrieval technique to return documents that are irrelevant to the query (*false positive*) or to omit documents that may be relevant (*false negatives*).

Documents, Corpora, Terms

A corpus is a pair $\mathcal{K} = (T, \mathcal{D})$, where $T = \{t_1, \ldots, t_m\}$ is a finite set whose elements are referred to as terms, and $\mathcal{D} = (D_1, \ldots, D_n)$ is a set of documents. Each document D_i is a finite sequence of terms, $D_j = (t_{j1}, \ldots, t_{jk}, \ldots, t_{j\ell_j})$.

Documents, Corpora, Terms (cont'd)

Each term t_i generates a row vector $(a_{i1}, a_{i2}, \ldots, a_{in})$ referred to as a *term vector* and each document d_j generates a column vector

$$\mathbf{d}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}.$$

A *query* is a sequence of terms $q \in \mathbf{Seq}(T)$ and it is also represented as a vector

$$\mathbf{q} = \begin{pmatrix} q_1 \\ \vdots \\ q_m \end{pmatrix},$$

where $q_i = 1$ if the term t_i occurs in q, and 0 otherwise.

Retrieval

When a query q is applied to a corpus $\mathcal{K}$, an IR system that is based on the vector model computes the similarity between the query and the documents of the corpus by evaluating the cosine of the angle between the query vector $\mathbf{q}$ and the vectors of the documents of the corpus. For the angle α_j between $\mathbf{q}$ and $\mathbf{d}_j$ we have

$$\cos \alpha_j = \frac{(\mathbf{q}, \mathbf{d}_j)}{\| \mathbf{q} \|_2 \| \mathbf{d}_j \|_2}.$$

The IR system returns those documents D_j for which this angle is small, that is $\cos \alpha_j \geq t$, where t is a parameter provided by the user.

LSI

The LSI methods aims to capture relationships between documents motivated by the underlying structure of the documents. This structure is obscured by synonimy, polysemy, the use of insignificant syntactic-sugar words, and plain noise, which is caused my misspelled words or counting errors.

Assumptions:

i. $A \in \mathbb{R}^{m \times n}$ is the matrix of a corpus $\mathcal{K}$;
ii. $A = UDV^{\mathsf{H}}$ is an SVD of A;
iii. $\text{rank}(A) = p$

The first p columns of U form an orthonormal basis for range(A), the subspace generated by the vector documents of $\mathcal{K}$;
The last $n - p$ columns of V constitute an orthonormal basis for nullsp(A).
The first p transposed columns of V form an orthonormal basis for the subspace of $\mathbb{R}^n$ generated by the term vectors of $\mathcal{K}$.

Example

$\mathcal{K} = (T, \mathcal{D})$, where $T = \{t_1, \ldots, t_5\}$ and $\mathcal{D} = \{D_1, D_2, D_3\}$.

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

D_1 and D_2 are fairly similar (they contain two common terms, t_3 and t_4).

Example (cont'd)

Suppose that t_1 and t_2 are synonyms.

$$\mathbf{q} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ returns } D_1.$$

The matrix representation can not directly account for the equivalence of the terms t_1 and t_2. However, this is an acceptable assumption because these terms appear in the common context $\{t_3, t_4\}$ in D_1 and D_2.

An SVD of A

$$A = \begin{pmatrix} -0.2787 & 0.2176 & 0.7071 & -0.5314 & 0.3044 \\ -0.2787 & 0.2176 & -0.7071 & -0.5314 & 0.3044 \\ -0.7138 & -0.3398 & -0.0000 & -0.1099 & -0.6024 \\ -0.5573 & 0.4352 & 0.0000 & 0.6412 & 0.2980 \\ -0.1565 & -0.7749 & -0.0000 & 0.1099 & 0.6024 \end{pmatrix}$$

$$\cdot \begin{pmatrix} 2.3583 & 0 & 0 \\ 0 & 1.1994 & 0 \\ 0 & 0 & 1.0000 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -0.6572 & -0.6572 & -0.3690 \\ 0.2610 & 0.2610 & -0.9294 \\ 0.7071 & -0.7071 & -0.0000 \end{pmatrix}.$$

Successive Approximations of A:

$$
\begin{aligned}
B(1) &= \sigma_1 * \mathbf{u}_1 * \mathbf{v}_1^{\mathrm{H}} \\
&= \begin{pmatrix} 0.4319 & 0.4319 & 0.2425 \\ 0.4319 & 0.4319 & 0.2425 \\ 1.1063 & 1.1063 & 0.6213 \\ 0.8638 & 0.8638 & 0.4851 \\ 0.2425 & 0.2425 & 0.1362 \end{pmatrix}, \\
B(2) &= \sigma_1 * \mathbf{u}_1 * \mathbf{v}_1^{\mathrm{H}} + \sigma_2 * \mathbf{u}_2 * \mathbf{v}_2^{\mathrm{H}} \\
&= \begin{pmatrix} 0.5000 & 0.5000 & 0.0000 \\ 0.5000 & 0.5000 & -0.0000 \\ 1.0000 & 1.0000 & 1.0000 \\ 1.0000 & 1.0000 & 0.0000 \\ 0.0000 & -0.0000 & 1.0000 \end{pmatrix}.
\end{aligned}
$$

A and $B(1)$

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad B(1) = \begin{pmatrix} 0.4319 & 0.4319 & 0.2425 \\ 0.4319 & 0.4319 & 0.2425 \\ 1.1063 & 1.1063 & 0.6213 \\ 0.8638 & 0.8638 & 0.4851 \\ 0.2425 & 0.2425 & 0.1362 \end{pmatrix}$$

A and $B(2)$

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad B(2) = \begin{pmatrix} 0.5000 & 0.5000 & 0.0000 \\ 0.5000 & 0.5000 & -0.0000 \\ 1.0000 & 1.0000 & 1.0000 \\ 1.0000 & 1.0000 & 0.0000 \\ 0.0000 & -0.0000 & 1.0000 \end{pmatrix}$$

Basic Properties of SVD Approximation:

1. the rank-1 matrices $\mathbf{u}_i\mathbf{v}_i^H$ are pairwise orthogonal;
2. their Frobenius norms are all equal to 1;
3. noise as distributed with relative uniformity with respect to the p orthonormal components of the SVD.

By omitting several such components that correspond to relatively small singular values we eliminate a substantial part of the noise and we obtain a matrix that better reflects the underlying hidden structure of the corpus.

Example

Let $\mathbf{q}$ be a query whose vector is

$$\mathbf{q} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

The similarity between $\mathbf{q}$ and the document vectors $\mathbf{d}_i$, $1 \leq i \leq 3$, that constitute the columns of $\mathcal{A}$ is

$$\cos(\mathbf{q}, \mathbf{d}_1) = 0.8165, \cos(\mathbf{q}, \mathbf{d}_2) = 0.482, \text{ and } \cos(\mathbf{q}, \mathbf{d}_3) = 0,$$

suggesting that d_1 is by far the most relevant document for q.

Example (cont'd)

If we compute the same value of cosine for q and the columns $\mathbf{b}_1, \mathbf{b}_2$ and $\mathbf{b}_3$ of the matrix $B(2)$ we have

$$\cos(\mathbf{q}, \mathbf{b}_1) = \cos(\mathbf{q}, \mathbf{b}_2) = 0.6708, \text{ and } \cos(\mathbf{q}, \mathbf{b}_3) = 0.$$

This approximation of A uncovers the hidden similarity of d_1 and d_2, a fact that is quite apparent from the structure of the matrix $B(2)$.

Recognizing Chinese Characters - N. Wang

你好

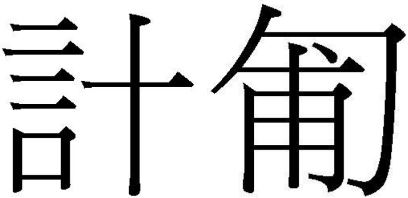
Characteristics of Chinese Characters

- square in shape;
- more than 5000 characters;
- set of characters indexed by the number of strokes which varies from 1 to 32.

Difficulties with the Stroke Counts

- length of stroke is variable;
- number of strokes does not capture topology of characters;
- strokes are not line segments; rather, they are calygraphic units.

Two Characters with 9 Strokes Each



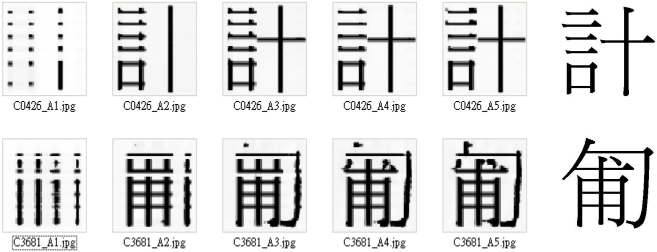
Characters are Digitized (black and white)

Matrix of Chinese characters

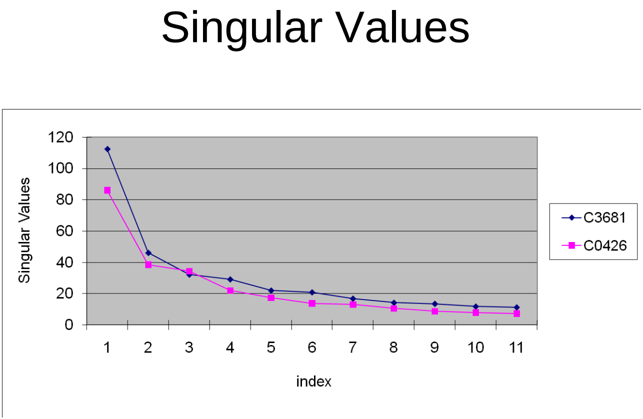
SVD for Chinese characters

$$A_k = \sum_{i=1}^{k} u_i \sigma_i v_i^T$$

The best rank k approximation to A



Scree Diagram: Variation of Singular Values



Several Important Sources for Linear Methods

- G. H. Golub and C. F. Van Loan: Matrix Computations, The Johns Hopkins University Press, Baltimore, 1989
- I. T.Jolliffe: Principal Component Analysis, Springer, New York, 2002
- R. A. Horn and C. R. Johnson, Matrix Analysis, Cambridge University Press, Cambridge, UK, 1996
- L. Elden: Matrix Methods in Data Mining and Pattern Recognition, SIAM, Philadelpha, 2007