Principal Components Analysis

18th-23th of March 2013

The main Theorem I

Theorem (SVD decomposition)

Let A be a $m \times n$ matrix with rank $r \leq \min(m, n)$. Then, there exist U isometry on $\mathbb{R}^n$, V isometry on $\mathbb{R}^m$ and D a $m \times n$ matrix with

$$A = VDU, D = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & 0 & \dots \\ \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & s_r & 0 \cdots \\ 0 & \cdots & 0 & 0 \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} s_1 \geq s_2 \geq s_3 \cdots \geq s_r > 0.$$

$s_1, s_2, s_3, \cdots, s_r$ are called the singular value of A.

The main Theorem II

$$A = VDU$$

set

$$U = \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{pmatrix} \qquad V = \begin{pmatrix} v_1 v_2 \dots v_m \end{pmatrix}$$

u_j are vectors of $\mathbb{R}^n$ and v_j are vectors of $\mathbb{R}^m$. Moreover,

$$Au_j = s_j v_j \text{ and } A^T v_j = s_j u_j.$$

Real life is nice I

- Let A be a real life matrix measuring m quantitative variables on n individuals

Real life is nice I

- Let A be a real life matrix measuring m quantitative variables on n individuals
- Let, $\overline{A}$ is the $m \times n$ matrix containing constant lines. The constant is the mean

Real life is nice I

- Let A be a real life matrix measuring m quantitative variables on n individuals
- Let, $\overline{A}$ is the $m \times n$ matrix containing constant lines. The constant is the mean
- Then, $A - \overline{A}$ has very few big singular values !!!!!

Real life is nice II

Example

$A^T =$

	MAT	FIS	FRAN	INGL
Juan	6.00	6.00	5.00	5.50
Alan	8.00	8.00	8.00	8.00
Ana	6.00	7.00	11.00	9.50
Monica	14.50	14.50	15.50	15.00
Didier	14.00	14.00	12.00	12.50
Andres	11.00	10.00	5.50	7.00
Pedro	5.50	7.00	14.00	11.50
Brigita	13.00	12.50	8.50	9.50
Maria	9.00	9.50	12.50	12.00

Real life is nice III

Example

$$A^T - \overline{A}^T = \begin{pmatrix} -3.6666667 & -3.8333333 & -5.2222222 & -4.5555556 \\ -1.6666667 & -1.8333333 & -2.2222222 & -2.0555556 \\ -3.6666667 & -2.8333333 & 0.7777778 & -0.5555556 \\ 4.8333333 & 4.6666667 & 5.2777778 & 4.9444444 \\ 4.3333333 & 4.1666667 & 1.7777778 & 2.4444444 \\ 1.3333333 & 0.1666667 & -4.7222222 & -3.0555556 \\ -4.1666667 & -2.8333333 & 3.7777778 & 1.4444444 \\ 3.3333333 & 2.6666667 & -1.7222222 & -0.5555556 \\ -0.6666667 & -0.3333333 & 2.2777778 & 1.9444444 \end{pmatrix}$$

Singular values: $15.940949 \geq 10.405523 \geq 0.54193 \geq 0.3087624$

Real life is nice IV

Approximation by only two factors (2 biggest sv(s))

Real life is nice IV

Approximation by only two factors (2 biggest sv(s))

absolute error

$$\begin{pmatrix}
-0.0314396 & 0.0150058 & -0.0551034 & 0.0737196 \\
0.0459759 & -0.0508687 & 0.0177687 & -0.0136492 \\
-0.0389639 & 0.0854931 & 0.0769781 & -0.1263953 \\
0.0988983 & -0.1116579 & 0.0333692 & -0.0220862 \\
-0.1309320 & 0.1547786 & -0.0290764 & 0.0066033 \\
0.0974981 & -0.1207342 & 0.0097540 & 0.0129175 \\
-0.0120300 & 0.0548056 & 0.0854607 & -0.1315031 \\
-0.0614005 & 0.1156365 & 0.0798577 & -0.1370491 \\
0.0323937 & -0.1424589 & -0.2190086 & 0.3374427
\end{pmatrix}$$

PCA picture

Cloud approximation by projecting on only two biggest factors

PCA picture

Cloud approximation by projecting on only two biggest factors