Singular Value Decomposition (SVD) and Principal Component Analysis (PCA)

Rodrigo B. Platte

RTG Seminar, 08/29/2016
Announcement

ICERM Workshop (Brown University)
Women in Data Science and Mathematics Research Collaboration Workshop (WiSDM)

July 17-21, 2017

Click here!
Motivation for this talk on SVD/PCA

- SVD is a standard tool in Theoretical, Applied and Computational Mathematics as well as Statistics.
- Students might have learned about SVD in a linear algebra class as a tool to study linear operators, but might not have seen it as a tool for data analysis.
- SVD might not have received much emphasis in your undergraduate LA course.
- Applications: Data compression (or dimension reduction), data analysis, inversion/regularization of operators.

My bias is to see matrices as operators and not as data.
The goal of PCA is to find the direction(s) in which the variance in the data is maximum.
Orthogonal (unitary) matrices

Let \( Q \) be a unitary matrix. Then

\[
Q^* Q = QQ^* = I
\]

\[
\|Qx\|_2 = \|x\|_2
\]

\[
\langle Qx, Qy \rangle = \langle x, y \rangle
\]

- Unitary matrices are ideal for many applications. Their ”condition number” is 1 and hence optimal.
- If \( Q \) is a data matrix, on the other hand, PCA will not help in reducing the dimensionality of the problem.
Singular Value Decomposition

\[ A_{m \times n} = U_{m \times m} S_{m \times n} V_{n \times n}^*, \]

where \( U \) and \( V \) are unitary (orthogonal), and \( S \) is (sorta) diagonal,

\[
S = \begin{bmatrix}
\sigma_1 & 0 & \ldots & 0 \\
0 & \sigma_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \sigma_n \\
0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 0
\end{bmatrix}
\]

or

\[
S = \begin{bmatrix}
\sigma_1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & \sigma_2 & \ldots & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \sigma_m & 0 & \ldots & 0
\end{bmatrix}.
\]

The diagonal entries are always real and positive,

\[ \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r, \]

where

\[ r = \min\{m, n\}, \]

are the singular values.
SVD 2x2 illustration by Bruno Welfert

SVD = rotation + scaling + rotation

\[ V^T \text{ rotates } \{v_1, v_2\} \text{ into } \{e_1, e_2\} \]

\[ \Sigma \text{ stretches } \{e_1, e_2\} \text{ into } \{\sigma_1 e_1, \sigma_2 e_2\} \]

\[ U \text{ rotates } \{\sigma_1 e_1, \sigma_2 e_2\} \text{ into } \{Av_1, Av_2\} \]

\[ A = U\Sigma V^T \]
Singular Value Decomposition

The columns of $U$ and $V$ are called the left and right singular vectors,

$$U = [u_1 \ u_2 \ \ldots \ u_m]$$
$$V = [v_1 \ v_2 \ \ldots \ v_n].$$

The relation

$$A = USV^*$$

gives

$$Av_k = \sigma_k u_k.$$  

From

$$AV = US$$

we get

$$A^* U = VS^*,$$

and then

$$A^* u_k = \sigma_k v_k.$$  

Put it all together, and what do you get?

$$A^* Av_k = \sigma_k^2 v_k.$$
Singular Value Decomposition

Put it all together, and what do you get?

\[ A^* A v_k = \sigma_k^2 v_k. \]

Thus \( \sigma_k^2 \) and \( v_k \) are an eigenpair of \( A^* A \). This is not a good way of computing \( \sigma_k \) since \( A^* A \) is usually more ill-conditioned than \( A \) itself.

Remarks:

- \( A^* A \) is self-adjoint (symmetric), hence it has real eigenvalues.
- \( A^* A x = \sigma_k^2 x \), then \( x^* A^* A x = \sigma_k^2 x^* x \), and \( \sigma_k^2 = \| A x \|_2^2 / \| x \|_2^2 \geq 0 \).
- The number of nonzero singular values \( (\sigma_k) \) is the rank of \( A \).
- If \( A \) is full rank, \( \sigma_k > 0 \) for all \( k \).
**Theorem:** \( \|A\|_2 = \max_k \sigma_k = \sigma_1. \)

**Proof:** Recall that

\[
\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=1} \|Ax\|_2
\]

Now let (here \( \| \| = \| \|_2 \))

\( A = USV^* \), then \( Ax = USV^*x \). Define \( z = V^*x \), then \( \|z\| = \|x\| \).
The 2-norm of A

**Theorem:** \( \|A\|_2 = \max_k \sigma_k = \sigma_1. \)

Proof: Recall that

\[
\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2=1} \|Ax\|_2
\]

Now let (here \( \| \| = \| \|_2 \))

\( A = USV^* \), then \( Ax = USV^*x \). Define \( z = V^*x \), then \( \|z\| = \|x\| \).

\[
\|Ax\| = \|USz\| = \|Sz\|
\]
Theorem: \( \|A\|_2 = \max_k \sigma_k = \sigma_1 \).

Proof: Recall that
\[
\|A\|_2 = \max_x \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2 = 1} \|Ax\|_2
\]

Now let (here \( \| \| = \| \|_2 \))
\( A = USV^* \), then \( Ax = USV^*x \). Define \( z = V^*x \), then \( \|z\| = \|x\| \).

\[
\|Ax\| = \|USz\| = \|Sz\|
\]

\[
\frac{\|Ax\|}{\|x\|} = \frac{\|Sz\|}{\|z\|}, \quad \max_x \frac{\|Ax\|_2}{\|x\|_2} = \max_x \frac{\|Sz\|}{\|z\|} = \|S\|
\]

Because \( S \) is a diagonal matrix, \( \|S\| = \max_k \sigma_k \). (HW2 problem)
SVD truncation

Let

\[ A_k = \sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^* + \cdots + \sigma_k u_k v_k^*. \]

We have

\[ A = USV^* = [u_1 \ u_2 \ \cdots \ u_m] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} v_1^* \\ v_2^* \\ \vdots \\ v_n^* \end{bmatrix} \]

\[ A = [\sigma_1 u_1 v_1^* + \sigma_2 u_2 v_2^* + \cdots + \sigma_n u_n v_n^*]; \]

so

\[ A_n - A_k = \sigma_{k+1} u_{k+1} v_{k+1}^* + \cdots + \sigma_n u_n v_n^* \]
Then we have
\[ \|A - A_k\|_2 = \sigma_{k+1}, \quad k > 1. \]

**Proof.**

Let
\[ E = A - A_k = \sum_{j=k+1}^{n} \sigma_j u_j v_j^* \]

\[ \|E\|_2 = \sigma_{k+1} \]

is the largest singular value of \( E \). \qed
Image compression

load mandrill
image(X), colormap(gray(220)), axis image, shg
[U,S,V] = svd(X);
plot(diag(S))
X1 = U(:,1:40)*S(1:40,1:40)*V(:,1:40)';
size(X1)
image([X X1]), colormap(gray(220)), axis image, shg
480*500
size(U)
40*480+40+500*40
ans/(480*500)
norm(X-X1,2)
S(41,41)
S(41,41)/S(1,1)
X1 = U(:,1:20)*S(1:20,1:20)*V(:,1:20)';
image([X X1]), colormap(gray(220)), axis image,
Recall that given a sequence of numbers \( \{x_i\}_{i=1}^m \), its mean and variance are given by

\[
\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i \quad \text{and} \quad \sigma(x) = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})^2,
\]

and the covariance of two sequences \( \{x_i\}_{i=1}^m \) and \( \{y_i\}_{i=1}^m \) is

\[
cov(x, y) = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y}).
\]

The standard deviation is \( \sqrt{\sigma(x)} \).

Why \( m - 1 \)? See [https://en.wikipedia.org/wiki/Bessel’s_correction](https://en.wikipedia.org/wiki/Bessel’s_correction)
Moreover, in the case of zero mean sequences

\[ \text{cov}(x, y) = \frac{1}{m-1} \sum_{i=1}^{m} x_i y_i = \frac{1}{m-1} y^T x \quad \text{and} \]

\[ \sigma(x) = \text{cov}(x, x) = \frac{1}{m-1} \|x\|^2, \]

where \( x \) and \( y \) are vectors with elements \( x_i \) and \( y_i \) respectively.

Hence, the matrix \( A^T A \) is called the covariance matrix.
Data matrix

Let $A$ be an $m \times n$ data matrix, where $n$ is the data dimension and $m$ is the number of samples.
Example: (2D data with zero column mean)

Nsamples = 200;
A = zeros(Nsamples,2);
for k = 1:Nsamples
    A(k,:) = [1 1]*5*randn+[-1 1]*randn;
end
mu = mean(A);
A = A - repmat(mu,Nsamples,1);
plot(A(:,1),A(:,2),’*’)
The goal of PCA is to find the direction(s) in which the variance in the data is maximum.
Principal component analysis

Our goal is to find a new orthogonal coordinate system (change of basis) such that

\[ T_{m \times n} = A_{m \times n}Q_{n \times n}, \quad \text{or} \quad QT^* = IA^* \]

\[
\begin{bmatrix}
q_1 & q_2 & \ldots
\end{bmatrix}
\begin{bmatrix}
\alpha \\
\beta \\
\vdots
\end{bmatrix}
= \begin{bmatrix}
e_1 & e_2 & \ldots
\end{bmatrix}
\begin{bmatrix}
x \\
y \\
\vdots
\end{bmatrix}
\]

and \( Q^*Q = I \). Notice that

\[
t_1 = Aq_1, \ t_2 = Aq_2, \ldots, t_n = Aq_n.
\]

The vectors \( q_1, q_2, \ldots \) are called the principal directions and \( t_1, t_2, \ldots \) scores.

Notice that if the columns \( A \) have zero mean, then the columns of \( T \) also have zero mean:

\[
\sum_{i=1}^{m} a_{i,j} = 0 \Rightarrow \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j} q_{j,k} = \sum_{j=1}^{n} q_{j,k} \sum_{i=1}^{m} a_{i,j} = 0
\]
Principal component analysis

Since $q_1$ is the principal direction, the variance in that direction should be maximum. That is,

$$q_1 = \arg\max_{\|q\|=1} \frac{1}{m} t^* t = \arg\max_{\|q\|=1} q^* A^* A q$$

Theorem

Let $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_n^2 \geq 0$ be the eigenvalues of $A^* A$ with corresponding eigenvectors $v_1, v_2, \ldots, v_n$ then

$$\sigma_1^2 = \max_{\|q\|=1} q^* A^* A q \quad \text{and} \quad v_1 = \arg\max_{\|q\|=1} q^* A^* A q;$$

$$\sigma_2^2 = \max_{\|q\|=1, \ q'v_1=0} q^* A^* A q \quad \text{and} \quad v_2 = \arg\max_{\|q\|=1, \ q'v_1=0} q^* A^* A q,$$

and so on.

Proof.

HW (hint: $A^* A = U D U^*$ $\Rightarrow$ $q^* U D U^* q = z^* D z$).
Principal component analysis

- The principal directions are the eigenvectors of $A^*A$. The eigenvalues are the variances of the data along the principal directions (multiplied by $m - 1$).
- The principal directions are the singular vectors of $A$. The singular values are the standard deviations of the data along the principal directions (multiplied by $\sqrt{m - 1}$).
Back to the example

\[
\begin{align*}
\text{[Q,T,Var]} &= \text{pca}(X); \\
Q & \quad \text{Var} \\
[V,D] &= \text{eig}(X'X); \\
V & \quad D/(N\text{samples}-1) \\
T(:,1)'*T(:,1)/(N\text{samples}-1) & \quad Q = \\
& \begin{pmatrix}
0.693529494948996 & 0.720428233508231 \\
0.720428233508231 & -0.693529494948996
\end{pmatrix}
\end{align*}
\]

\[
\text{Var} = \\
\begin{pmatrix}
48.174626337324085 & 0 \\
1.645003805276812 & 0
\end{pmatrix}
\]

\[
\text{V} = \\
\begin{pmatrix}
-0.720428233508231 & 0.693529494948996 \\
0.693529494948996 & 0.720428233508231
\end{pmatrix}
\]

\[
\text{ans} = \\
\begin{pmatrix}
1.645003805276811 & 0 \\
0 & 48.174626337324078
\end{pmatrix}
\]

\[
\text{ans} = \\
48.174626337324085
\]
Back to the example

principal directions scaled by the corresponding standard deviations

data represented in the new basis $T = AQ$
Another example

PCA3D.m