Singular Value Decomposition and its Applications in Computer Vision

Subhashis Banerjee

Department of Computer Science and Engineering
IIT Delhi

Graphs and Geometry Workshop, NIT Warangal
October 24, 2013
Overview

- Linear algebra basics
- Singular value decomposition
- Linear equations and least squares
- Principal component analysis
- Latent semantics and topic discovery
- Clustering?
Linear systems

- $m$ equations in $n$ unknowns. $A\mathbf{x} = \mathbf{b}$. $A \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{b} \in \mathbb{R}^m$.

- Two reasons usually offered for importance of linearity:
  
  **Superposition:** If $f_1$ produces $a_1$ and $f_2$ produces $a_2$, then a combined force $f_1 + \alpha f_2$ produces $a_1 + \alpha a_2$.

  **Pragmatics:**
  - $f(x, y) = 0$ and $g(x, y) = 0$ yields $F(x) = 0$ by elimination.
  - Degree of $F = \text{degree of } f \times \text{degree of } g$.
  - A system of $m$ quadratic equation gives a polynomial of degree $2^m$.
  - The only case in which the exponential is harmless is when the base is 1 (linear).
Linear (in)dependence

- Given vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ and scalars $x_1, \ldots, x_n$, the vector

$$\mathbf{b} = \sum_{j=1}^{n} x_j \mathbf{a}_j$$

is a *linear combination* of the vectors.

- The vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ are *linearly dependent* iff at least one of them is a linear combination of the others (ones that precedes it).

- A set of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$ is a *basis* for a set $B$ of vectors if they are linearly independent and every vector in $B$ can be expressed as a linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_n$.

- Two different bases for the same vector space $B$ have the same number of vectors (*dimension*).
Inner product and orthogonality

- **2-norm:**

\[
\|b\|^2 = b_1^2 + \sum_{j=2}^{m} b_j e_j^2 = \sum_{j=1}^{m} b_j^2 = b^T b
\]

- **inner product:** \(b^T c = \|b\| \|c\| \cos \theta\)
- **orthogonal:** \(b^T c = 0\)
- **projection of \(b\) onto \(c\):**

\[
\frac{cc^T}{c^T c} b
\]
Orthogonal subspaces and rank

- Any basis $\mathbf{a}_1, \ldots, \mathbf{a}_n$ for a subspace $A$ of $\mathbb{R}^m$ can be extended to a basis for $\mathbb{R}^m$ by adding $m - n$ vectors $\mathbf{a}_{n+1}, \ldots, \mathbf{a}_m$.

- If vector space $A$ is a subspace of $\mathbb{R}^m$ for some $m$, then the **orthogonal complement** ($A^\perp$) of $A$ is the set of all vectors in $\mathbb{R}^m$ that are orthogonal to all the vectors in $A$.

  - $\dim(A) + \dim(A^\perp) = m$

- $\text{null}(A) = \{ \mathbf{x} : A\mathbf{x} = \mathbf{0} \}$. $\dim(\text{null}(A)) = h$ (nullity).

- $\text{range}(A) = \{ \mathbf{b} : A\mathbf{x} = \mathbf{b} \text{ for some } \mathbf{x} \}$. $\dim(\text{range}(A)) = r$ (rank).

  - $r = n - h$.

- Number of linearly independent rows of $A$ is equal to its number of linearly independent columns.
Solutions of a linear system: $Ax = b$

- $\text{range}(A)$; dimension $r = \text{rank}(A)$
- $\text{null}(A)$; dimension $h = \text{nullity}(A)$
- $\text{range}(A)^\perp$; dimension $m - r$
- $\text{null}(A)^\perp$; dimension $n - h$

$$
\begin{align*}
\text{null}(A)^\perp &= \text{range}(A^T) \\
\text{range}(A)^\perp &= \text{null}(A^T)
\end{align*}
$$

- $b \notin \text{range}(A) \implies$ no solutions
- $b \in \text{range}(A)$
  - $r = n = m$. Invertible. Unique solution.
  - $r = n$, $m > n$. Redundant. Unique solution.
  - $r < n$. Under determined. $\infty^{n-r}$ solutions.
Orthogonal matrices

- A set of vectors $V$ is *orthogonal* if its elements are pairwise orthogonal. *Orthonormal*, if in addition for each $x \in V$, $\|x\| = 1$.

- Vectors in an orthonormal set are linearly independent.

- $V = [v_1, \ldots, v_n]$ is an *orthogonal matrix*.

- $V^{-1}V = V^T V = V^{-1}V = VV^T = I$.

- The norm of a vector $x$ is not changed by multiplication by an orthogonal matrix:

\[
\|Vx\|^2 = x^T V^T V x = x^T x = \|x\|^2
\]
Vector norms

\[ \| x \|_1 = \sum_{i=1}^{m} |x_i|, \]
\[ \| x \|_2 = \left( \sum_{i=1}^{m} |x_i|^2 \right)^{1/2} = \sqrt{x^*x}, \]
\[ \| x \|_\infty = \max_{1 \leq i \leq m} |x_i|, \]
\[ \| x \|_p = \left( \sum_{i=1}^{m} |x_i|^p \right)^{1/p} \quad (1 \leq p < \infty). \]
Matrix norms

\[ A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix} \]
Overview

- Linear algebra basics
- **Singular value decomposition** (Golub and Van Loan, 1996, Golub and Kahan, 1965)
- Linear equations and least squares
- Principal component analysis
- Latent semantics and topic discovery
- Clustering?
Singular value decomposition

Geometric view: An $m \times n$ matrix $A$ of rank $r$ maps the $r$-dimensional unit hypersphere in rowspace($A$) into an $r$-dimensional hyperellipse in range($A$).

Algebraic view: If $A$ is a real $m \times n$ matrix then there exists orthogonal matrices

$$
U = \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix} \in \mathbb{R}^{m \times m}
$$
$$
V = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \in \mathbb{R}^{n \times n}
$$

such that

$$
U^T AV = \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_p) \in \mathbb{R}^{m \times n}
$$

where $p = \min(m, n)$, and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$

Equivalently,

$$
A = U\Sigma V^T
$$
Proof (sketch):

- Consider all vectors of the form $Ax = b$ for $x$ on the unit hypersphere $\|x\| = 1$. Consider the scalar function $\|Ax\|$. Let $v_1$ be a vector on the unit sphere in $\mathbb{R}^n$ where the scalar function is maximised.

- Let $\sigma_1 u_1$ be the corresponding vector with $\sigma_1 u_1 = Av_1$ and $\|u_1\| = 1$. Let $u_1$ and $v_1$ be extended to orthonormal bases for $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. Let the corresponding matrices be $U_1$ and $V_1$.

- We have $U_1^T AV_1 = S_1 = \begin{bmatrix} \sigma_1 & w^T \\ 0 & A_1 \end{bmatrix}$

- Consider the length of the vector

$$\frac{1}{\sqrt{\sigma_1^2 + w^T w}} S_1 \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} = \frac{1}{\sqrt{\sigma_1^2 + w^T w}} \begin{bmatrix} \sigma_1^2 + w^T w \\ A_1 w \end{bmatrix}$$

- Conclude $w = 0$ and induct.
SVD geometry:

1. $\xi = V^T x$, where $V = [v_1 \ v_2]\n
2. $\eta = \Sigma \xi$, where $\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}\n
3. Finally, $b = U \eta$. 
SVD: structure of a matrix

- Suppose $\sigma_1 \geq \ldots \geq \sigma_r > \sigma_{r+1} = 0$. Then,

\[
\begin{align*}
\text{rank}(A) &= r \\
\text{null}(A) &= \text{span}\{v_{r+1}, \ldots, v_n\} \\
\text{range}(A) &= \text{span}\{u_1, \ldots, u_r\}
\end{align*}
\]

- Setting $U_r = U(:,1:r)$, $\Sigma_r = \Sigma(1:r,1:r)$, and $V_r = V(:,1:r)$, we have $A = U_r \Sigma_r V_r = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

\[
\begin{align*}
\|A\|_F &= \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sigma_1^2 + \ldots + \sigma_p^2 \\
\|A\|_2 &= \sigma_1
\end{align*}
\]
SVD: low rank approximation

For any $\nu$ with $0 \leq \nu \leq r$, define $A_\nu = \sum_{i=1}^{\nu} \sigma_i u_i v_i^T$. If $\nu = p = \min(m, n)$, define $\sigma_{\nu+1} = 0$. Then,

$$\|A - A_\nu\|_2 = \inf_{B \in \mathbb{R}^{m \times n}, \text{rank}(B) \leq \nu} \|A - B\|_2 = \sigma_{\nu+1}$$

Proof (sketch):

- $Aw$ is maximised by that $w$ which is closest in direction to most of the rows of $A$.
- The projections of the rows of $A$ onto $v_1$ is given by $Av_1v_1^T$. This is indeed the best rank 1 approximation:

$$\|A - Av_1v_1^T\|_2 = \|A - \sigma_1 u_1 v_1^T\|_2$$

is the smallest over $\|A - B\|_2$ where $B$ is any rank 1 matrix.
Overview

- Linear algebra basics
- Singular value decomposition
- Linear equations and least squares
- Principal component analysis
- Latent semantics and topic discovery
- Clustering?
Least squares

The minimum-norm least squares solution to a linear system $Ax = b$, that is, the shortest vector $x$ that achieves

$$\min_x \|Ax - b\|$$

is unique and is given by

$$x = V\Sigma^\dagger U^T b$$

where $\Sigma^\dagger = \text{diag}(1/\sigma_1, \ldots, 1/\sigma_r, 0)$ is a $n \times m$ diagonal matrix. The matrix

$$A^\dagger = V\Sigma^\dagger U^T$$

is called the pseudoinverse of $A$. 
Pseudoinverse proof (sketch):

\[
\begin{align*}
\min_x \|Ax - b\| &= \min_x \|U\Sigma V^T x - b\| = \min_x \|U(\Sigma V^T x - U^T b)\| \\
&= \min_x \|\Sigma V^T x - U^T b\|
\end{align*}
\]

Setting \( y = V^T x \) and \( c = U^T b \), we have

\[
\min_x \|Ax - b\| = \min_y \|\Sigma y - c\|
\]
Least squares for homogenous systems

The solution to
\[
\min_{\|x\|=1} \|Ax\|
\]
is given by \(v_n\), the last column of \(V\).

Proof:

\[
\min_{\|x\|=1} \|Ax\| = \min_{\|x\|=1} \|U\Sigma V^T x\| = \min_{\|x\|=1} \|\Sigma V^T x\| = \min_{\|y\|=1} \|\Sigma y\|
\]
where \(y = V^T x\).

Clearly this is minimised by the vector \(y = [0, \ldots, 0, 1]^T\).
A couple of other least squares problems

- Given an $m \times n$ matrix $A$ with $m \geq n$, find the vector $x$ that minimises $\|Ax\|$ subject to $\|x\| = 1$ and $Cx = 0$.
- Given an $m \times n$ matrix $A$ with $m \geq n$, find the vector $x$ that minimises $\|Ax\|$ subject to $\|x\| = 1$ and $x \in \text{range}(G)$. 

Graphs and Geometry Workshop, NIT Warangal
Overview

- Linear algebra basics
- Singular value decomposition
- Linear equations and least squares
- **Principal component analysis** (Pearson, 1901, Schlens 2003)
- Latent semantics and topic discovery
- Clustering?
PCA: A toy problem

\[
X(t) = [x_A(t) \ y_A(t) \ x_B(t) \ y_B(t) \ x_C(t) \ y_C(t)]^T,
X = [X(1) \ X(2) \ \cdots \ X(n)]^T.
\]

Is there another basis, which is a linear combination of the original basis, that \textit{best} expresses our data set?

\[
PX = Y
\]
PCA Issues: noise and redundancy

Noise

\[ SNR = \frac{\sigma_{signal}^2}{\sigma_{noise}^2} \gg 1 \]

Redundancy

Graphs and Geometry Workshop, NIT Warangal | SVD and its Applications in Computer Vision
Consider zero mean vectors $\mathbf{a} = [a_1 \ a_2 \ \ldots \ a_n]$ and $\mathbf{b} = [b_1 \ b_2 \ \ldots \ b_n]$. 

- Variance: $\sigma^2_a = \langle a_i a_i \rangle_i$ and $\sigma^2_b = \langle b_i b_i \rangle_i$.
- Covariance: $\sigma^2_{ab} = \langle a_i b_i \rangle_i = \frac{1}{n-1} \mathbf{a} \mathbf{b}^T$.

If $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_m]^T$ ($m \times n$) then the covariance matrix is:

$$S_X = \frac{1}{n-1} \mathbf{X} \mathbf{X}^T$$

- $ij^{th}$ value of $S_X$ is obtained by substituting $\mathbf{x}_i$ for $\mathbf{a}$ and $\mathbf{x}_j$ for $\mathbf{b}$.
- $S_X$ is square, symmetric, $m \times m$.
- Diagonal entries of $S_X$ are the variance of particular measurement types.
- The off-diagonal entries of $S_X$ are the covariance between measurement types.
Solving PCA

\[ S_Y = \frac{1}{n-1} YY^T = \frac{1}{n-1} (PX)(PX)^T = \frac{1}{n-1} PXX^T P^T \]

- Writing \( X = U\Sigma V^T \), we have
  \[ XX^T = U\Sigma U^T \]

- Setting \( P = U^T \), we have
  \[ S_Y = \frac{1}{n-1} \Sigma \]

- Data is maximally uncorrelated.
- Effective rank \( r \) of \( \Sigma \) gives dimensionality reduction.
PCA: tacit assumptions

- Linearity.
- Mean and variance are sufficient statistics $\rightarrow$ Gaussian distribution.
- Large variances have important dynamics.
- The principal components are orthogonal.
Application: eigenfaces (Turk and Pentland, 1991)

- Obtain a set $S$ of $M$ face images:
  
  $$S = \{\Gamma_1, \ldots, \Gamma_M\}$$

- Obtain the mean image $\Psi$:
  
  $$\Psi = \frac{1}{M} \sum_{j=1}^{M} \Gamma_j$$
Application: eigenfaces (Turk and Pentland, 1991)

- Compute centered images
  \[ \Phi_i = \Gamma_i - \Psi \]

- The covariance matrix is
  \[ C = \frac{1}{M} \sum_{j=1}^{M} \Phi_j \Phi_j^T = AA^T \]

  Size is \( N^2 \times N^2 \). Intractable.

- If \( \mathbf{v}_i \) is an eigenvector of \( A^T A \) (\( M \times M \)), then \( A \mathbf{v}_i \) an eigenvector of \( AA^T \).

  \[ A^T A \mathbf{v}_i = \mu_i \mathbf{v}_i \iff AA^T A \mathbf{v}_i = \mu_i A \mathbf{v}_i \]
Application: eigenfaces (Turk and Pentland, 1991)

Recognition:
- $\omega_k = u_k(\Gamma - \Psi)$
- Compute minimum distance to database of faces
Overview

- Linear algebra basics
- Singular value decomposition
- Linear equations and least squares
- Principal component analysis
- **Latent semantics and topic discovery** (Scott et. al. 1990, Papadimitriou et. al. 1998)
- Clustering?
Consider a $m \times n$ matrix $A$ where the $ij^{th}$ entry denotes the marks obtained by the $i^{th}$ student in the $j^{th}$ test (Naveen Garg, Abhiram Ranade).

- Are the marks obtained by the $i^{th}$ student in various tests correlated?
- What are the capabilities of the $i^{th}$ student?
- What does the $j^{th}$ test evaluate?
- What is the expected rank of $A$?
Suppose there are really only three abilities (topics) that determine a student’s marks in tests: verbal, logical and quantitative.
Suppose \( v_i, l_i \) and \( q_i \) characterise these abilities of the \( i^{th} \) student; let \( V_j, L_j \) and \( Q_j \) characterise the extent to which the \( j^{th} \) test evaluates these abilities.
A generative model for the \( ij^{th} \) entry of \( A \) may be given as

\[
v_i V_j + l_i L_j + q_i Q_j
\]
A new \( m \times 1 \) term vector \( t \) can be projected into the LSI space as:

\[
\hat{t} = t^T U_k \Sigma_k^{-1}
\]

A new \( 1 \times n \) document vector \( d \) can be projected into the LSI space as:

\[
\hat{d} = d V_k \Sigma_k^{-1}
\]
Topic discovery example

- An example with more than 2000 images and with 12 topics (LDA)

Graphs and Geometry Workshop, NIT Warangal
SVD and its Applications in Computer Vision
Overview

- Linear algebra basics
- Singular value decomposition
- Linear equations and least squares
- Principal component analysis
- Latent semantics and topic discovery
- **Clustering** (Drineas et. al. 1999)
Clustering

- Partition rows of a matrix so that “similar” rows (points in $n$ dimensional space) are clustered together.

- Given points $a_1, \ldots, a_m \in \mathbb{R}^m$, find $c_1, \ldots, c_k \in \mathbb{R}^m$ so as to minimize

$$\sum_i d(a_i, \{c_1, \ldots, c_k\})^2$$

where $d(a, S)$ is the smallest distance from a point $a$ to any of the points in $S$. (*k-means*)

- $k$ is a constant. Consider $k = 2$ for simplicity. Even then the problem is NP-complete for arbitrary $n$.

- We have $k$ centres. If $n = k$ then the problem can be solved in polynomial time.
The points belonging to the two clusters can be separated by the perpendicular bisector of the line joining the two centres.

The centre selected for a group must be its centroid.

There are only a polynomial number of lines to consider (Each set of cluster centres define a Voronoi diagram. Each cell is a polyhedron and the total number of faces in $k$ cells is no more than $\binom{k}{2}$. Enumerate all sets of hyperplanes (faces) each of which contains $k$ independent points of $A$ such that they define exactly $k$ cells. Assign each point of $A$ lying on a hyperplane to one of the sides.)

The best $k$ dimensional subspace can be found using SVD.

Gives a 2-approximation.
Other applications

- High dimensional matching
- Graph partitioning
- Metric embedding
- Image compression
- ... Learn SVD well

Learn SVD well