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 $\mathbf{f}_1$ produces $\mathbf{a}_1$ and $\mathbf{f}_2$ produces $\mathbf{a}_2$, then a combined force $\mathbf{f}_1 + \alpha\mathbf{f}_2$ produces $\mathbf{a}_1 + \alpha\mathbf{a}_2$.

  Pragmatics:
  - $f(x, y) = 0$ and $g(x, y) = 0$ yields $F(x) = 0$ by elimination.
  - Degree of F = degree of f $\times$ 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*).

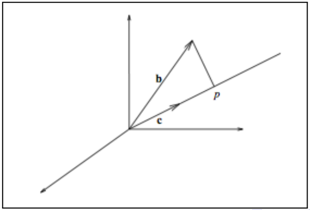
Inner product and orthogonality

- 2-*norm*:

$$\|\mathbf{b}\|^2 = b_1^2 + \|\sum_{j=2}^{m} b_j \mathbf{e}_j\|^2 = \sum_{j=1}^{m} b_j^2 = \mathbf{b}^T\mathbf{b}$$

- *inner product:* $\mathbf{b}^T\mathbf{c} = \|\mathbf{b}\|\|\mathbf{c}\| \cos\theta$
- *orthogonal:* $\mathbf{b}^T\mathbf{c} = 0$
- *projection of* $\mathbf{b}$ *onto* $\mathbf{c}$*:*

$$\frac{\mathbf{c}\mathbf{c}^T}{\mathbf{c}^T\mathbf{c}}\mathbf{b}$$



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: $A\mathbf{x} = \mathbf{b}$

- range(A); dimension $r = \text{rank}(A)$
- null(A); dimension $h = \text{nullity}(A)$
- range$(A)^\perp$; dimension $m - r$
- null$(A)^\perp$; dimension $n - h$
- $$\begin{aligned} \text{null}(A)^\perp &= \text{range}(A^T) \\ \text{range}(A)^\perp &= \text{null}(A^T) \end{aligned}$$
- $\mathbf{b} \notin \text{range}(A) \implies$ no solutions
- $\mathbf{b} \in \text{range}(A)$
  - $r = n = m$. Invertible. Unique solution.
  - $r = n$, $m > n$. Redundant. Unique solution.
  - $r < n$. Under determined. ∞^{n-r} solutions.

Orthogonal matrices

- A set of vectors V is *orthogonal* if its elements are pairwise orthogonal. *Orthonormal*, if in addition for each $\mathbf{x} \in V$, $\|\mathbf{x}\| = 1$.
- Vectors in an orthonormal set are linearly independent.
- $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ is an *orthogonal matrix*.
- $V^{-1}V = V^T V = V^{-1}V = VV^T = \mathbf{I}$.
- The norm of a vector $\mathbf{x}$ is not changed by multiplication by an orthogonal matrix:

$$\|V\mathbf{x}\|^2 = \mathbf{x}^T V^T V\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|^2$$

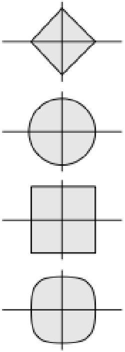
Vector norms

$$\|x\|_1 = \sum_{i=1}^{m} |x_i|,$$

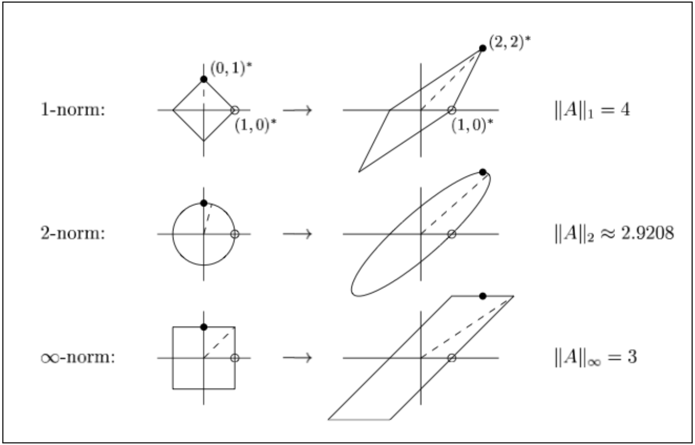
$$\|x\|_2 = \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2} = \sqrt{x^* x},$$

$$\|x\|_\infty = \max_{1 \le i \le m} |x_i|,$$

$$\|x\|_p = \left(\sum_{i=1}^{m} |x_i|^p \right)^{1/p} \quad (1 \le p < \infty).$$



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

$$\begin{aligned} U &= [\mathbf{u}_1 \cdots \mathbf{u}_m] \in \mathbb{R}^{m \times m} \\ V &= [\mathbf{v}_1 \cdots \mathbf{v}_n] \in \mathbb{R}^{n \times n} \end{aligned}$$

such that

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

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

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

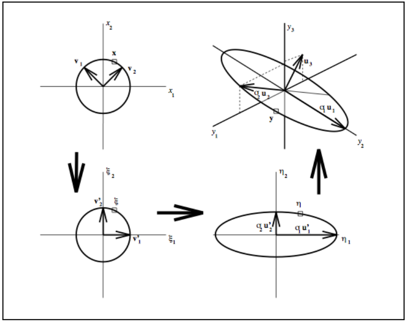
Proof (sketch):

- Consider all vectors of the form $A\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ on the unit hypersphere $\|\mathbf{x}\| = 1$. Consider the scalar function $\|A\mathbf{x}\|$. Let $\mathbf{v}_1$ be a vector on the unit sphere in $\mathbb{R}^n$ where the scalar function is maximised.
- Let $\sigma_1\mathbf{u}_1$ be the corresponding vector with $\sigma_1\mathbf{u}_1 = A\mathbf{v}_1$ and $\|\mathbf{u}_1\| = 1$. Let $\mathbf{u}_1$ and $\mathbf{v}_1$ 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 $U_1^T A V_1 = S_1 = \begin{bmatrix} \sigma_1 & \mathbf{w}^T \\ \mathbf{0} & A_1 \end{bmatrix}$
- Consider the length of the vector

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

- Conclude $\mathbf{w} = \mathbf{0}$ and induct.

SVD geometry:



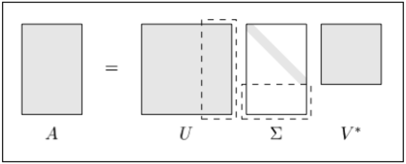
1. $\xi = V^T\mathbf{x}$, where $V = [\mathbf{v}_1 \;\; \mathbf{v}_2]$
2. $\eta = \Sigma\xi$, where $\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{bmatrix}$
3. Finally, $\mathbf{b} = U\eta$.

SVD: structure of a matrix

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

$$\begin{aligned} \text{rank}(A) &= r \\ \text{null}(A) &= \text{span}\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\} \\ \text{range}(A) &= \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\} \end{aligned}$$

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



- $\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sigma_1^2 + \ldots + \sigma_p^2$
- $\|A\|_2 = \sigma_1$

SVD: low rank approximation

For any ν with $0 \leqslant \nu \leqslant r$, define $A_\nu = \sum_{i=1}^{\nu} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. If $\nu = p = \min(m, n)$, define $\sigma_{\nu+1} = 0$. Then,

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

Proof (sketch):

- $A\mathbf{w}$ is maximised by that $\mathbf{w}$ which is closest in direction to most of the rows of A.
- The projections of the rows of A onto $\mathbf{v}_1$ is given by $A\mathbf{v}_1\mathbf{v}_1^T$. This is indeed the best rank 1 approximation:

$$\|A - A\mathbf{v}_1\mathbf{v}_1^T\|_2 = \|A - \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T\|_2$$

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 $A\mathbf{x} = \mathbf{b}$, that is, the shortest vector $\mathbf{x}$ that achieves

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|$$

is unique and is given by

$$\mathbf{x} = V\Sigma^{\dagger}U^{T}\mathbf{b}$$

where $\Sigma^{\dagger} = \text{diag}(1/\sigma_1, \ldots, 1/\sigma_r, \mathbf{0})$ 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):

- $$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\| = \min_{\mathbf{x}} \|U\Sigma V^T x - \mathbf{b}\| = \min_{\mathbf{x}} \|U(\Sigma V^T \mathbf{x} - U^T \mathbf{b})\|$$
  $$= \min_{\mathbf{x}} \|\Sigma V^T \mathbf{x} - U^T \mathbf{b}\|$$
- Setting $\mathbf{y} = V^T\mathbf{x}$ and $\mathbf{c} = U^T\mathbf{b}$, we have
  $$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\| = \min_{\mathbf{y}} \|\Sigma\mathbf{y} - \mathbf{c}\|$$
- $$\begin{bmatrix} \sigma_1 & 0 & & \cdots & & 0 \\ 0 & \ddots & & \cdots & & 0 \\ & & \sigma_r & & & \\ \vdots & & & 0 & & \vdots \\ & & & & \ddots & \\ 0 & & & & & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} c_1 \\ \vdots \\ c_r \\ c_{r+1} \\ \vdots \\ c_m \end{bmatrix}$$

Least squares for homogenous systems

The solution to

$$\min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

is given by $\mathbf{v}_n$, the last column of V.

Proof:

$$\min_{\|\mathbf{x}\|=1} \|A\mathbf{x}\| = \min_{\|\mathbf{x}\|=1} \|U\Sigma V^T\mathbf{x}\| = \min_{\|\mathbf{x}\|=1} \|\Sigma V^T\mathbf{x}\| = \min_{\|\mathbf{y}\|=1} \|\Sigma\mathbf{y}\|$$

where $\mathbf{y} = V^T\mathbf{x}$.

Clearly this is minimised by the vector $\mathbf{y} = [0, \ldots, 0, 1]^T$.

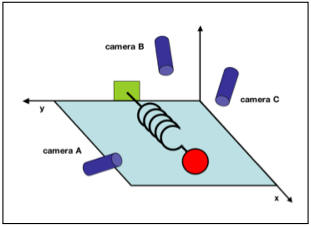
A couple of other least squares problems

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

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



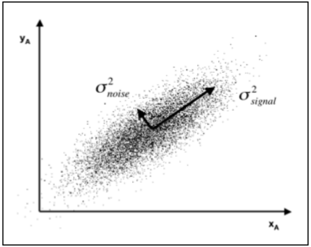
$$X(t) = [x_A(t)\ y_A(t)\ x_B(t)\ y_B(t)\ x_C(t)\ y_C(t)]^T,$$
$$X = [X(1)\ X(2)\ \cdots\ X(n)]^T.$$

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

$$PX = Y$$

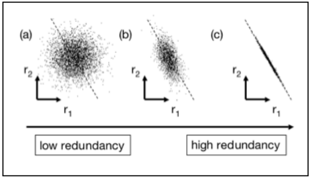
PCA Issues: noise and redundancy

Noise



$$SNR = \frac{\sigma^2 signal}{\sigma^2 noise} \gg 1$$

Redundancy



Covariance

- Consider zero mean vectors $\mathbf{a} = [a_1 \ a_2 \ \ldots \ a_n]$ and $\mathbf{b} = [b_1 \ b_2 \ \ldots \ b_n]$.
- Variance: $\sigma_{\mathbf{a}}^2 = \langle a_i a_i \rangle_i$ and $\sigma_{\mathbf{b}}^2 = \langle b_i b_i \rangle_i$
- Covariance: $\sigma_{\mathbf{ab}}^2 = \langle a_i b_i \rangle_i = \frac{1}{n-1}\mathbf{ab}^T$.
- If $X = [\mathbf{x}_1 \ \mathbf{x}_2 \ \ldots \ \mathbf{x}_m]^T$ $(m \times n)$ then the *covariance matrix* is:

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

- 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 Σ gives dimensionality reduction.

PCA: tacit assumptions

- Linearity.
- Mean and variance are sufficient statistics $\Longrightarrow$ 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 Ψ:

$$\Psi = \frac{1}{M} \sum_{j=1}^{M} \Gamma_j$$

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.

$$A^T A\mathbf{v}_i = \mu_i \mathbf{v}_i \Leftrightarrow AA^T A\mathbf{v}_i = \mu_i A\mathbf{v}_i$$

Application: eigenfaces (Turk and Pentland, 1991)

Recognition:

- $\omega_k = \mathbf{u}_k(\Gamma - \Psi)$
- 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?

Latent semantics and topic discovery

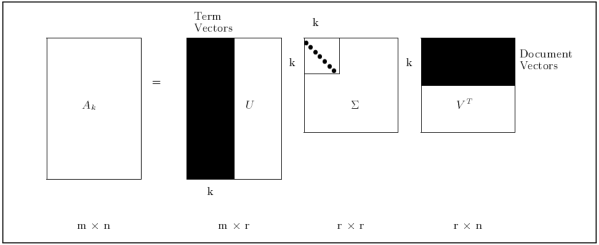
- Consider a $m \times n$ matrix A where the ij^{th} entry denotes the marks obtained by the i^{th} student in the j^{th} test (Naveen Garg, Abhiram Ranade).
- Are the marks obtained by the i^{th} student in various tests correlated?
- What are the capabilities of the i^{th} student?
- What does the j^{th} test evaluate?
- What is the expected rank of A?

Latent semantics and topic discovery

- 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

$$v_i V_j + l_i L_j + q_i Q_j$$

Latent semantics and topic discovery



- A new $m \times 1$ term vector t can be projected in to the LSI space as:

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

- A new $1 \times n$ document vector d can be projected in to 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)

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 $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^m$, find $c_1, \ldots, c_k \in \mathbb{R}^m$ so as to minimize
$$\sum_i d(\mathbf{a}_i, \{c_1, \ldots, c_k\})^2$$
where $d(\mathbf{a}, S)$ is the smallest distance from a point $\mathbf{a}$ 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.

Clustering

- 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