Lecture 6: Singular Value Decomposition - I

November 6, 2017

Introduction

- n data points in d space - each a row of *Data Matrix A* ($n \times d$ matrix).

Introduction

- n data points in d space - each a row of *Data Matrix A* ($n \times d$ matrix).
- Singular Value Decomposition (SVD) consists of *best fit k* dimensional subspace for A, for every k, $k = 1, 2, \ldots$rank(A).

Introduction

- n data points in d space - each a row of *Data Matrix A* ($n \times d$ matrix).
- Singular Value Decomposition (SVD) consists of *best fit k* dimensional subspace for A, for every k, $k = 1, 2, \ldots$rank(A).
- Best Fit in the sense of minimum sum of squared (perpendicular) distances of data points to subspace. Will see best fit for every k simultaneously.

Introduction

- n data points in d space - each a row of *Data Matrix A* ($n \times d$ matrix).
- Singular Value Decomposition (SVD) consists of *best fit k* dimensional subspace for A, for every k, $k = 1, 2, \ldots$rank(A).
- Best Fit in the sense of minimum sum of squared (perpendicular) distances of data points to subspace. Will see best fit for every k simultaneously.
- Equivalently, maximum sum of squares of the lengths of projection of data points into subspace.

Minimize distances $\equiv$ Maximize projections

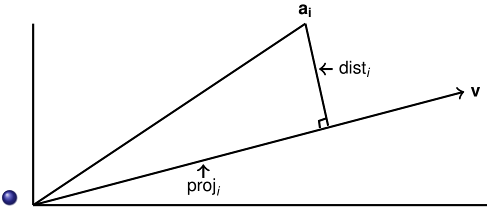

Figure: The projection of the point $\mathbf{a_i}$ onto the line through the origin in the direction of $\mathbf{v}$.

Minimize distances $\equiv$ Maximize projections

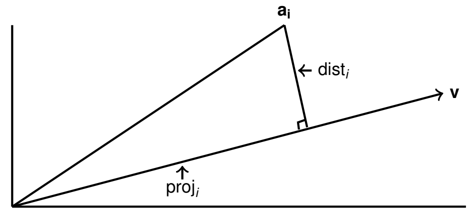

Figure: The projection of the point $\mathbf{a_i}$ onto the line through the origin in the direction of $\mathbf{v}$.

- Min $\sum_i \text{dist}_i^2 \equiv$ Max $\sum_i \text{proj}_i^2$ courtesy Pythogorus

Minimize distances $\equiv$ Maximize projections

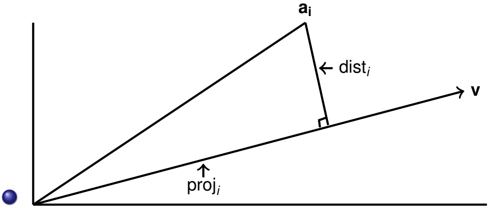

Figure: The projection of the point $\mathbf{a_i}$ onto the line through the origin in the direction of $\mathbf{v}$.

- Min $\sum_i \text{dist}_i^2 \equiv$ Max $\sum_i \text{proj}_i^2$ courtesy Pythogorus
- Contrast: "Least-Squares Fit": Given $(x_i, y_i), i = 1, 2, \ldots, n$ Min$_{a,b}(ax_i + b - y_i)^2$. Dist.s "vertical", not perp to line. [PICTURE]

Minimize distances $\equiv$ Maximize projections

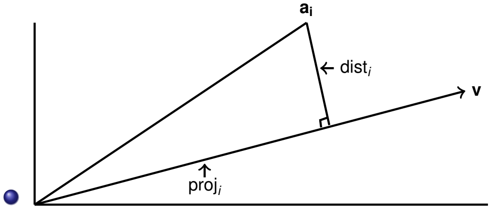

Figure: The projection of the point $\mathbf{a_i}$ onto the line through the origin in the direction of $\mathbf{v}$.

- Min $\sum_i \text{dist}_i^2 \equiv$ Max $\sum_i \text{proj}_i^2$ courtesy Pythogorus
- Contrast: "Least-Squares Fit": Given $(x_i, y_i), i = 1, 2, \ldots, n$ $\text{Min}_{a,b}(ax_i + b - y_i)^2$. Dist.s "vertical", not perp to line. [PICTURE]
- Least Squares- Not nec. through **0**. But SVD : subspace, so has **0**. See later: best-fit affine subspace passes through centroid of data. Can translate to make centroid $=$ **0**.

Will Show: Greedy works

- First find the best-fit 1-dimensional subspace to data: line (through the origin).

Will Show: Greedy works

- First find the best-fit 1-dimensional subspace to data: line (through the origin).
- Then, find best-fit line (through $\mathbf{0}$) perpendicular to first line.

Will Show: Greedy works

- First find the best-fit 1-dimensional subspace to data: line (through the origin).
- Then, find best-fit line (through $\mathbf{0}$) perpendicular to first line.
- At the i th step, find best fit line perp to $i-1$ lines found so far. Until: rank(A).

Will Show: Greedy works

- First find the best-fit 1-dimensional subspace to data: line (through the origin).
- Then, find best-fit line (through $\mathbf{0}$) perpendicular to first line.
- At the i th step, find best fit line perp to $i-1$ lines found so far. Until: rank(A).
- Will Show: When done, we can write $A = UDV^T$, where, columns of V are unit vectors along lines found above; D is a diagonal matrix with positive entries and columns of U, V are orthonormal. [$A = UDV^T$ is called SVD.] Now focus on the just the best-fit lines, not on the matrix factorization yet.

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.
- If $\mathbf{v}$ is the unit vector along the best-fit line, $\mathbf{v}$ minimizes among all unit length vectors $\mathbf{x}$ the quantity: $\sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 - (\mathbf{a_i} \cdot \mathbf{x})^2 \right) = \sum_{i=1}^{n} |\mathbf{a_i}|^2 - |A\mathbf{x}|^2$.

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.
- If **v** is the unit vector along the best-fit line, **v** minimizes among all unit length vectors **x** the quantity:
  $\sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 - (\mathbf{a_i} \cdot \mathbf{x})^2 \right) = \sum_{i=1}^{n} |\mathbf{a_i}|^2 - |A\mathbf{x}|^2$.
- Now, $\sum_i |\mathbf{a_i}|^2 = ||A||_F^2$ does not depend on **x**. So (as we said earlier), equivalent to maximizing $|A\mathbf{x}|^2$.

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.
- If $\mathbf{v}$ is the unit vector along the best-fit line, $\mathbf{v}$ minimizes among all unit length vectors $\mathbf{x}$ the quantity: $\sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 - (\mathbf{a_i} \cdot \mathbf{x})^2 \right) = \sum_{i=1}^{n} |\mathbf{a_i}|^2 - |A\mathbf{x}|^2$.
- Now, $\sum_i |\mathbf{a_i}|^2 = ||A||_F^2$ does not depend on $\mathbf{x}$. So (as we said earlier), equivalent to maximizing $|A\mathbf{x}|^2$.
- Define the *first singular vector of A* by:

  $\mathbf{v_1} = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|$

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.
- If $\mathbf{v}$ is the unit vector along the best-fit line, $\mathbf{v}$ minimizes among all unit length vectors $\mathbf{x}$ the quantity:
  $\sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 - (\mathbf{a_i} \cdot \mathbf{x})^2\right) = \sum_{i=1}^{n} |\mathbf{a_i}|^2 - |A\mathbf{x}|^2$.
- Now, $\sum_i |\mathbf{a_i}|^2 = ||A||_F^2$ does not depend on $\mathbf{x}$. So (as we said earlier), equivalent to maximizing $|A\mathbf{x}|^2$.
- Define the *first singular vector of A* by:
  $\mathbf{v_1} = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|$
- There can be ties. [What is an obvious tie for $\mathbf{v_1}$?] Break them arbitrarily. Will use ArgMax for arb. broken ties.

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.
- If $\mathbf{v}$ is the unit vector along the best-fit line, $\mathbf{v}$ minimizes among all unit length vectors $\mathbf{x}$ the quantity:
  $\sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 - (\mathbf{a_i} \cdot \mathbf{x})^2\right) = \sum_{i=1}^{n} |\mathbf{a_i}|^2 - |A\mathbf{x}|^2$.
- Now, $\sum_i |\mathbf{a_i}|^2 = ||A||_F^2$ does not depend on $\mathbf{x}$. So (as we said earlier), equivalent to maximizing $|A\mathbf{x}|^2$.
- Define the *first singular vector of A* by:
  $\mathbf{v_1} = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|$
- There can be ties. [What is an obvious tie for $\mathbf{v_1}$?] Break them arbitrarily. Will use ArgMax for arb. broken ties.
- Value $\sigma_1(A) = |A\mathbf{v_1}|$ is called the *first singular value* of A.

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.
- If $\mathbf{v}$ is the unit vector along the best-fit line, $\mathbf{v}$ minimizes among all unit length vectors $\mathbf{x}$ the quantity: $\sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 - (\mathbf{a_i} \cdot \mathbf{x})^2\right) = \sum_{i=1}^{n} |\mathbf{a_i}|^2 - |A\mathbf{x}|^2$.
- Now, $\sum_i |\mathbf{a_i}|^2 = ||A||_F^2$ does not depend on $\mathbf{x}$. So (as we said earlier), equivalent to maximizing $|A\mathbf{x}|^2$.
- Define the *first singular vector of A* by: $\mathbf{v_1} = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|$
- There can be ties. [What is an obvious tie for $\mathbf{v_1}$?] Break them arbitrarily. Will use ArgMax for arb. broken ties.
- Value $\sigma_1(A) = |A\mathbf{v_1}|$ is called the *first singular value* of A.
- If all data points lie on a line through the origin, the line on which projections are maximized is precisely the same line, so it is the first singular vectors.

First Singular Vector

- Notation: A $n \times d$ data matrix; each row is a data point.
- If $\mathbf{v}$ is the unit vector along the best-fit line, $\mathbf{v}$ minimizes among all unit length vectors $\mathbf{x}$ the quantity: $\sum_{i=1}^{n} \left(|\mathbf{a_i}|^2 - (\mathbf{a_i} \cdot \mathbf{x})^2\right) = \sum_{i=1}^{n} |\mathbf{a_i}|^2 - |A\mathbf{x}|^2$.
- Now, $\sum_i |\mathbf{a_i}|^2 = ||A||_F^2$ does not depend on $\mathbf{x}$. So (as we said earlier), equivalent to maximizing $|A\mathbf{x}|^2$.
- Define the *first singular vector of A* by: $\mathbf{v_1} = \arg\max_{|\mathbf{v}|=1} |A\mathbf{v}|$
- There can be ties. [What is an obvious tie for $\mathbf{v_1}$?] Break them arbitrarily. Will use ArgMax for arb. broken ties.
- Value $\sigma_1(A) = |A\mathbf{v_1}|$ is called the *first singular value* of A.
- If all data points lie on a line through the origin, the line on which projections are maximized is precisely the same line, so it is the first singular vectors.
- Further singular vectors. Think - what if data points are coplanar? Would like to get two perpendicular vectors spanning the plane.

Greedy and Definition of further singular vectors

- How to define further singular vectors? Again, think coplanar data points. Would like the 2-dim subspace maximizing sum of squared projections.

Greedy and Definition of further singular vectors

- How to define further singular vectors? Again, think coplanar data points. Would like the 2-dim subspace maximizing sum of squared projections.
- Try Greedy: Define the second singular vector $\mathbf{v_2}$ as the one (break ties arbitrarily) maximizing the sum of projections squared subject to being perpendicular to first. Algebra: same as:

Greedy and Definition of further singular vectors

- How to define further singular vectors? Again, think coplanar data points. Would like the 2-dim subspace maximizing sum of squared projections.
- Try Greedy: Define the second singular vector $\mathbf{v_2}$ as the one (break ties arbitrarily) maximizing the sum of projections squared subject to being perpendicular to first. Algebra: same as:
- $\mathbf{v_2} = \arg\max_{\substack{\mathbf{v} \perp \mathbf{v_1} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.

Greedy and Definition of further singular vectors

- How to define further singular vectors? Again, think coplanar data points. Would like the 2-dim subspace maximizing sum of squared projections.
- Try Greedy: Define the second singular vector $\mathbf{v_2}$ as the one (break ties arbitrarily) maximizing the sum of projections squared subject to being perpendicular to first. Algebra: same as:
- $\mathbf{v_2} = \arg\max_{\substack{\mathbf{v} \perp \mathbf{v_1} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.
- $\mathbf{v_3} = \arg\max_{\substack{\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.

Greedy and Definition of further singular vectors

- How to define further singular vectors? Again, think coplanar data points. Would like the 2-dim subspace maximizing sum of squared projections.
- Try Greedy: Define the second singular vector $\mathbf{v_2}$ as the one (break ties arbitrarily) maximizing the sum of projections squared subject to being perpendicular to first. Algebra: same as:
- $\mathbf{v_2} = \arg\max_{\substack{\mathbf{v} \perp \mathbf{v_1} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.
- $\mathbf{v_3} = \arg\max_{\substack{\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.
- Define $\sigma_2(A) = |A\mathbf{v_2}|$; $\sigma_3(A) = |A\mathbf{v_3}|$... as the singular values of A.

Greedy and Definition of further singular vectors

- How to define further singular vectors? Again, think coplanar data points. Would like the 2-dim subspace maximizing sum of squared projections.
- Try Greedy: Define the second singular vector $\mathbf{v_2}$ as the one (break ties arbitrarily) maximizing the sum of projections squared subject to being perpendicular to first. Algebra: same as:
- $\mathbf{v_2} = \arg\max_{\substack{\mathbf{v} \perp \mathbf{v_1} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.
- $\mathbf{v_3} = \arg\max_{\substack{\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.
- Define $\sigma_2(A) = |A\mathbf{v_2}|$; $\sigma_3(A) = |A\mathbf{v_3}|$... as the singular values of A.
- Stop when $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ have been found and $\max_{\substack{\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r} \\ |\mathbf{v}|=1}} |A\mathbf{v}| = 0$.

Greedy and Definition of further singular vectors

- How to define further singular vectors? Again, think coplanar data points. Would like the 2-dim subspace maximizing sum of squared projections.
- Try Greedy: Define the second singular vector $\mathbf{v_2}$ as the one (break ties arbitrarily) maximizing the sum of projections squared subject to being perpendicular to first. Algebra: same as:
- $\mathbf{v_2} = \arg\max_{\substack{\mathbf{v}\perp\mathbf{v_1} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.
- $\mathbf{v_3} = \arg\max_{\substack{\mathbf{v}\perp\mathbf{v_1},\mathbf{v_2} \\ |\mathbf{v}|=1}} |A\mathbf{v}|$.
- Define $\sigma_2(A) = |A\mathbf{v_2}|$; $\sigma_3(A) = |A\mathbf{v_3}|$... as the singular values of A.
- Stop when $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ have been found and $\max_{\substack{\mathbf{v}\perp\mathbf{v_1},\mathbf{v_2},\ldots,\mathbf{v_r} \\ |\mathbf{v}|=1}} |A\mathbf{v}| = 0$.
- Will prove: $r = \text{rank}(A)$ and even if there are ties, the singular values $\sigma_1(A), \sigma_2(A), \ldots$ are unique.

Greedy Works

- Define *best-fit k dim subspace for A* as the one maximizing the sum of squared projection lengths of data points into subspace over all k dim subspaces.

Greedy Works

- Define *best-fit k dim subspace for A* as the one maximizing the sum of squared projection lengths of data points into subspace over all k dim subspaces.
- **Theorem (The Greedy Algorithm Works)**
  Let A be an $n \times d$ matrix with singular vectors $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$. For $1 \leq k \leq r$, let V_k be the subspace spanned by $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_k}$. For each k, V_k is the best-fit k-dimensional subspace for A.

Proof

- Proof by induction on k. Statement obvious for $k = 1$.

Proof

- Proof by induction on k. Statement obvious for $k = 1$.
- Lets do $k = 2$. Assume W is the best-fit 2-d subspace.

Proof

- Proof by induction on k. Statement obvious for $k = 1$.
- Lets do $k = 2$. Assume W is the best-fit 2-d subspace.
- **Claim** There is a $\mathbf{w_2} \in W$ s.t. $|\mathbf{w_2}| = 1$, $\mathbf{w_2} \cdot \mathbf{v_1} = 0$.

Proof

- Proof by induction on k. Statement obvious for $k = 1$.
- Lets do $k = 2$. Assume W is the best-fit 2-d subspace.
- **Claim** There is a $\mathbf{w_2} \in W$ s.t. $|\mathbf{w_2}| = 1$, $\mathbf{w_2} \cdot \mathbf{v_1} = 0$.
- Because the projection of $\mathbf{v_1}$ onto W spans a (at most) one dim subspace W' of W. Take $\mathbf{w_2} \in W$ perp to W'.

Proof

- Proof by induction on k. Statement obvious for $k = 1$.
- Lets do $k = 2$. Assume W is the best-fit 2-d subspace.
- **Claim** There is a $\mathbf{w_2} \in W$ s.t. $|\mathbf{w_2}| = 1$, $\mathbf{w_2} \cdot \mathbf{v_1} = 0$.
- Because the projection of $\mathbf{v_1}$ onto W spans a (at most) one dim subspace W' of W. Take $\mathbf{w_2} \in W$ perp to W'.
- Choose $\mathbf{w_1} \in W$ of unit length perpendicular to $\mathbf{w_2}$. $\mathbf{w_1}, \mathbf{w_2}$ form a (orthonormal) basis for W. [Convention: Basis means orthonormal..]

Proof

- Proof by induction on k. Statement obvious for $k = 1$.
- Lets do $k = 2$. Assume W is the best-fit 2-d subspace.
- **Claim** There is a $\mathbf{w_2} \in W$ s.t. $|\mathbf{w_2}| = 1$, $\mathbf{w_2} \cdot \mathbf{v_1} = 0$.
- Because the projection of $\mathbf{v_1}$ onto W spans a (at most) one dim subspace W' of W. Take $\mathbf{w_2} \in W$ perp to W'.
- Choose $\mathbf{w_1} \in W$ of unit length perpendicular to $\mathbf{w_2}$. $\mathbf{w_1}, \mathbf{w_2}$ form a (orthonormal) basis for W. [Convention: Basis means orthonormal..]
- Sum of squared projections of data points into W equals $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2$ - Why ? Algebra..

Proof

- Proof by induction on k. Statement obvious for $k = 1$.
- Lets do $k = 2$. Assume W is the best-fit 2-d subspace.
- **Claim** There is a $\mathbf{w_2} \in W$ s.t. $|\mathbf{w_2}| = 1$, $\mathbf{w_2} \cdot \mathbf{v_1} = 0$.
- Because the projection of $\mathbf{v_1}$ onto W spans a (at most) one dim subspace W' of W. Take $\mathbf{w_2} \in W$ perp to W'.
- Choose $\mathbf{w_1} \in W$ of unit length perpendicular to $\mathbf{w_2}$. $\mathbf{w_1}, \mathbf{w_2}$ form a (orthonormal) basis for W. [Convention: Basis means orthonormal..]
- Sum of squared projections of data points into W equals $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2$ - Why ? Algebra..
- $|A\mathbf{w_1}|^2 \leq |A\mathbf{v_1}|^2$ (Why?)

Proof

- Proof by induction on k. Statement obvious for $k = 1$.
- Lets do $k = 2$. Assume W is the best-fit 2-d subspace.
- **Claim** There is a $\mathbf{w_2} \in W$ s.t. $|\mathbf{w_2}| = 1, \mathbf{w_2} \cdot \mathbf{v_1} = 0$.
- Because the projection of $\mathbf{v_1}$ onto W spans a (at most) one dim subspace W' of W. Take $\mathbf{w_2} \in W$ perp to W'.
- Choose $\mathbf{w_1} \in W$ of unit length perpendicular to $\mathbf{w_2}$. $\mathbf{w_1}, \mathbf{w_2}$ form a (orthonormal) basis for W. [Convention: Basis means orthonormal..]
- Sum of squared projections of data points into W equals $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2$ - Why ? Algebra..
- $|A\mathbf{w_1}|^2 \leq |A\mathbf{v_1}|^2$ (Why?)
- $|A\mathbf{w_2}|^2 \leq |A\mathbf{v_2}|^2$ (why?) Add to get: V_2 as good as W.

Proof for general $k > 2$

- Inductive hypothesis implies: V_{k-1} is best-fit $k-1$ dim subspace.

Proof for general $k > 2$

- Inductive hypothesis implies: V_{k-1} is best-fit $k-1$ dim subspace.
- Suppose W is best-fit k dim subspace. **Claim** There is a unit length vector $\mathbf{w_k}$ in W perpendicular to V_{k-1} because: projections of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$ onto W span a (at most) $k-1$ dimensional subspace of W, so there is a $\mathbf{w_k}$ perpendicular to this in W.

Proof for general $k > 2$

- Inductive hypothesis implies: V_{k-1} is best-fit $k-1$ dim subspace.
- Suppose W is best-fit k dim subspace. **Claim** There is a unit length vector $\mathbf{w_k}$ in W perpendicular to V_{k-1} because: projections of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$ onto W span a (at most) $k-1$ dimensional subspace of W, so there is a $\mathbf{w_k}$ perpendicular to this in W.
- Choose a basis $\mathbf{w_1}, \mathbf{w_2}, \ldots \mathbf{w_k}$ of W.

Proof for general $k > 2$

- Inductive hypothesis implies: V_{k-1} is best-fit $k - 1$ dim subspace.
- Suppose W is best-fit k dim subspace. **Claim** There is a unit length vector $\mathbf{w_k}$ in W perpendicular to V_{k-1} because: projections of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$ onto W span a (at most) $k - 1$ dimensional subspace of W, so there is a $\mathbf{w_k}$ perpendicular to this in W.
- Choose a basis $\mathbf{w_1}, \mathbf{w_2}, \ldots \mathbf{w_k}$ of W.
- $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 + \cdots + |A\mathbf{w_{k-1}}|^2 \leq |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 + \cdots + |A\mathbf{v_{k-1}}|^2$ - Why?

Proof for general $k > 2$

- Inductive hypothesis implies: V_{k-1} is best-fit $k-1$ dim subspace.
- Suppose W is best-fit k dim subspace. **Claim** There is a unit length vector $\mathbf{w_k}$ in W perpendicular to V_{k-1} because: projections of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$ onto W span a (at most) $k-1$ dimensional subspace of W, so there is a $\mathbf{w_k}$ perpendicular to this in W.
- Choose a basis $\mathbf{w_1}, \mathbf{w_2}, \ldots \mathbf{w_k}$ of W.
- $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 + \cdots + |A\mathbf{w_{k-1}}|^2 \leq |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 + \cdots + |A\mathbf{v_{k-1}}|^2$ - Why?
  - Induction.

Proof for general $k > 2$

- Inductive hypothesis implies: V_{k-1} is best-fit $k-1$ dim subspace.
- Suppose W is best-fit k dim subspace. **Claim** There is a unit length vector $\mathbf{w_k}$ in W perpendicular to V_{k-1} because: projections of $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_{k-1}}$ onto W span a (at most) $k-1$ dimensional subspace of W, so there is a $\mathbf{w_k}$ perpendicular to this in W.
- Choose a basis $\mathbf{w_1}, \mathbf{w_2}, \ldots \mathbf{w_k}$ of W.
- $|A\mathbf{w_1}|^2 + |A\mathbf{w_2}|^2 + \cdots + |A\mathbf{w_{k-1}}|^2 \leq |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 + \cdots + |A\mathbf{v_{k-1}}|^2$ - Why?
  - Induction.
- $|A\mathbf{w_k}|^2 \leq |A\mathbf{v_k}|^2$. Why? Add to get V_k as good as W. QED

Consequences

- Theorem Proves: $\sigma_1(A), \sigma_2(A), \ldots$ are unique even if there were ties for the singular vectors.

Consequences

- Theorem Proves: $\sigma_1(A), \sigma_2(A), \ldots$ are unique even if there were ties for the singular vectors.
- Proof: $\sigma_1(A)$ obviously (first) exists (closed, bounded etc.) and is unique.

Consequences

- Theorem Proves: $\sigma_1(A), \sigma_2(A), \ldots$ are unique even if there were ties for the singular vectors.
- Proof: $\sigma_1(A)$ obviously (first) exists (closed, bounded etc.) and is unique.
- Now, there is a unique value of the maximum over all 2-d subspaces of sum of projections squared onto subspace, because the set of 2-d subspaces in closed etc.; call this value μ_2. Theorem says: $\sigma_1(A)^2 + \sigma_2^2(A) = |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 = \mu_2$. So, $\sigma_2^2(A) = \mu_2 - \sigma_1^2(A)$ is unique.

Consequences

- Theorem Proves: $\sigma_1(A), \sigma_2(A), \ldots$ are unique even if there were ties for the singular vectors.
- Proof: $\sigma_1(A)$ obviously (first) exists (closed, bounded etc.) and is unique.
- Now, there is a unique value of the maximum over all 2-d subspaces of sum of projections squared onto subspace, because the set of 2-d subspaces in closed etc.; call this value μ_2. Theorem says: $\sigma_1(A)^2 + \sigma_2^2(A) = |A\mathbf{v_1}|^2 + |A\mathbf{v_2}|^2 = \mu_2$. So, $\sigma_2^2(A) = \mu_2 - \sigma_1^2(A)$ is unique.
- General k: assume $\sigma_1(A), \sigma_2(A), \ldots, \sigma_{k-1}(A)$ are unique. Let μ_k be the maximum over all $k-$d subspaces of the sum of squared projections onto the subspace. Then, theorem implies that $\sigma_1^2(A) + \sigma_2^2(A) + \cdots + \sigma_k(A)^2 = \mu_k$. Using inductive hypothesis, now, $\sigma_k(A)$ is unique. Provided μ_k exists - Prove.

Aside: Convergence of Subspaces

- Suppose $V_1, V_2, \ldots$ are an infinite sequence of k dimensional subspaces of $\mathbf{R}^d$. There is a convergent sub-sequence. [Caution: Subspaces seem to be unbounded objects.]

Aside: Convergence of Subspaces

- Suppose $V_1, V_2, \ldots$ are an infinite sequence of k dimensional subspaces of $\mathbf{R}^d$. There is a convergent sub-sequence. [Caution: Subspaces seem to be unbounded objects.]
- Choose a basis for each V_i. First take a subsequence of $\{V_i\}$ in which the first basis vector converges.

Aside: Convergence of Subspaces

- Suppose $V_1, V_2, \ldots$ are an infinite sequence of k dimensional subspaces of $\mathbf{R}^d$. There is a convergent sub-sequence. [Caution: Subspaces seem to be unbounded objects.]
- Choose a basis for each V_i. First take a subsequence of $\{V_i\}$ in which the first basis vector converges.
- Then take subsequence of the subsequence where the second basis vector converges. Repeat.

Aside: Convergence of Subspaces

- Suppose $V_1, V_2, \ldots$ are an infinite sequence of k dimensional subspaces of $\mathbf{R}^d$. There is a convergent sub-sequence. [Caution: Subspaces seem to be unbounded objects.]
- Choose a basis for each V_i. First take a subsequence of $\{V_i\}$ in which the first basis vector converges.
- Then take subsequence of the subsequence where the second basis vector converges. Repeat.
- Finally, get a subsequence with each basis vector converging. Prove: in the limit, each "basis vector" is of length 1 and they are orthonormal. [Just convergent sequence of reals.]

Singular Values and Norm

- $A\mathbf{v_1}$ is list of lengths (with signs) of projections of rows of A onto $\mathbf{v_1}$.

Singular Values and Norm

- $A\mathbf{v_1}$ is list of lengths (with signs) of projections of rows of A onto $\mathbf{v_1}$.
- So, $\sigma_1(A) = |A\mathbf{v_1}|$ may be viewed as "component" of A along $\mathbf{v_1}$.

Singular Values and Norm

- $A\mathbf{v_1}$ is list of lengths (with signs) of projections of rows of A onto $\mathbf{v_1}$.
- So, $\sigma_1(A) = |A\mathbf{v_1}|$ may be viewed as "component" of A along $\mathbf{v_1}$.
- $\sigma_2(A)$ is "component" of A along $\mathbf{v_2}$.... Analogous to decomposing a vector into its components along the basis vectors.

Singular Values and Norm

- $A\mathbf{v_1}$ is list of lengths (with signs) of projections of rows of A onto $\mathbf{v_1}$.
- So, $\sigma_1(A) = |A\mathbf{v_1}|$ may be viewed as "component" of A along $\mathbf{v_1}$.
- $\sigma_2(A)$ is "component" of A along $\mathbf{v_2}$.... Analogous to decomposing a vector into its components along the basis vectors.
- Better have sum of squares of components $=$ whole to complete the analogy.

Singular Values and Norm

- $A\mathbf{v_1}$ is list of lengths (with signs) of projections of rows of A onto $\mathbf{v_1}$.
- So, $\sigma_1(A) = |A\mathbf{v_1}|$ may be viewed as "component" of A along $\mathbf{v_1}$.
- $\sigma_2(A)$ is "component" of A along $\mathbf{v_2}$.... Analogous to decomposing a vector into its components along the basis vectors.
- Better have sum of squares of components $=$ whole to complete the analogy.
- Since $\mathbf{a_i} \cdot \mathbf{v} = 0$ for all $\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ (why?), we have: $\sum_{t=1}^{r} (\mathbf{a_i} \cdot \mathbf{v_t})^2 = |\mathbf{a_i}|^2$.

Singular Values and Norm

- $A\mathbf{v_1}$ is list of lengths (with signs) of projections of rows of A onto $\mathbf{v_1}$.
- So, $\sigma_1(A) = |A\mathbf{v_1}|$ may be viewed as "component" of A along $\mathbf{v_1}$.
- $\sigma_2(A)$ is "component" of A along $\mathbf{v_2}$.... Analogous to decomposing a vector into its components along the basis vectors.
- Better have sum of squares of components $=$ whole to complete the analogy.
- Since $\mathbf{a_i} \cdot \mathbf{v} = 0$ for all $\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ (why?), we have: $\sum_{t=1}^{r} (\mathbf{a_i} \cdot \mathbf{v_t})^2 = |\mathbf{a_i}|^2$.
- $\sum_{j=1}^{n} |\mathbf{a_j}|^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} |A\mathbf{v_i}|^2 = \sum_{i=1}^{r} \sigma_i^2(A)$.

Singular Values and Norm

- $A\mathbf{v_1}$ is list of lengths (with signs) of projections of rows of A onto $\mathbf{v_1}$.
- So, $\sigma_1(A) = |A\mathbf{v_1}|$ may be viewed as "component" of A along $\mathbf{v_1}$.
- $\sigma_2(A)$ is "component" of A along $\mathbf{v_2}$.... Analogous to decomposing a vector into its components along the basis vectors.
- Better have sum of squares of components $=$ whole to complete the analogy.
- Since $\mathbf{a_i} \cdot \mathbf{v} = 0$ for all $\mathbf{v} \perp \mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ (why?), we have: $\sum_{t=1}^{r} (\mathbf{a_i} \cdot \mathbf{v_t})^2 = |\mathbf{a_i}|^2$.
- $\sum_{j=1}^{n} |\mathbf{a_j}|^2 = \sum_{j=1}^{n} \sum_{i=1}^{r} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n} (\mathbf{a_j} \cdot \mathbf{v_i})^2 = \sum_{i=1}^{r} |A\mathbf{v_i}|^2 = \sum_{i=1}^{r} \sigma_i^2(A)$.
- **Lemma** $\sum_{t=1}^{r} \sigma_t^2(A) = ||A||_F^2$.

Right and left singular vectors

- $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are called the *right-singular vectors*.

Right and left singular vectors

- $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are called the *right-singular vectors*.
- The vectors $A\mathbf{v_i}$ form a fundamental set of vectors. Normalize to length 1: $\mathbf{u_i} = \frac{1}{\sigma_i(A)} A\mathbf{v_i}$.

Right and left singular vectors

- $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are called the *right-singular vectors*.
- The vectors $A\mathbf{v_i}$ form a fundamental set of vectors. Normalize to length 1: $\mathbf{u_i} = \frac{1}{\sigma_i(A)} A\mathbf{v_i}$.
- Will show later $\mathbf{u}_i$ similarly maximizes $|\mathbf{u}^T A|$ over all $\mathbf{u}$ perpendicular to $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$.

Right and left singular vectors

- $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are called the *right-singular vectors*.
- The vectors $A\mathbf{v_i}$ form a fundamental set of vectors. Normalize to length 1: $\mathbf{u_i} = \frac{1}{\sigma_i(A)} A\mathbf{v_i}$.
- Will show later $\mathbf{u}_i$ similarly maximizes $|\mathbf{u}^T A|$ over all $\mathbf{u}$ perpendicular to $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$.
- $\mathbf{u}_i$ are called the *left-singular vectors*.

Right and left singular vectors

- $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are called the *right-singular vectors*.
- The vectors $A\mathbf{v_i}$ form a fundamental set of vectors. Normalize to length 1: $\mathbf{u_i} = \frac{1}{\sigma_i(A)} A\mathbf{v_i}$.
- Will show later $\mathbf{u}_i$ similarly maximizes $|\mathbf{u}^T A|$ over all $\mathbf{u}$ perpendicular to $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$.
- $\mathbf{u}_i$ are called the *left-singular vectors*.
- By definition, the right-singular vectors are orthogonal.

Right and left singular vectors

- $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}$ are called the *right-singular vectors*.
- The vectors $A\mathbf{v_i}$ form a fundamental set of vectors. Normalize to length 1: $\mathbf{u_i} = \frac{1}{\sigma_i(A)} A\mathbf{v_i}$.
- Will show later $\mathbf{u}_i$ similarly maximizes $|\mathbf{u}^T A|$ over all $\mathbf{u}$ perpendicular to $\mathbf{u}_1, \ldots, \mathbf{u}_{i-1}$.
- $\mathbf{u}_i$ are called the *left-singular vectors*.
- By definition, the right-singular vectors are orthogonal.
- Will show later that the left-singular vectors are also orthogonal.

Singular Value Decomposition

- A any matrix, $\mathbf{v_t}, t = 1, 2, \ldots, r$, $\mathbf{u_t}, t = 1, 2, \ldots, r$, $\sigma_t, t = 1, 2, \ldots, r$, its right singular vectors, left singular vectors and singular values respy.

Singular Value Decomposition

- A any matrix, $\mathbf{v_t}, t = 1, 2, \ldots, r$, $\mathbf{u_t}, t = 1, 2, \ldots, r$, $\sigma_t, t = 1, 2, \ldots, r$, its right singular vectors, left singular vectors and singular values respy.
- **Theorem (Singular Value Decomposition** $A = \sum_{t=1}^{r} \sigma_t \mathbf{u_t} \mathbf{v_t}^T$**.** Sum of r outer products.

Singular Value Decomposition

- A any matrix, $\mathbf{v_t}, t = 1, 2, \ldots, r$, $\mathbf{u_t}, t = 1, 2, \ldots, r$, $\sigma_t, t = 1, 2, \ldots, r$, its right singular vectors, left singular vectors and singular values respy.
- **Theorem (Singular Value Decomposition** $A = \sum_{t=1}^{r} \sigma_t \mathbf{u_t} \mathbf{v_t}^T$**.** Sum of r outer products.
- **Claim** Matrices A, B are identical iff for all $\mathbf{v}$, we have $A\mathbf{v} = B\mathbf{v}$. If $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$, this holds for each unit vector e_j and so j th col of A, B are the same for all j.

Singular Value Decomposition

- A any matrix, $\mathbf{v_t}, t = 1, 2, \ldots, r$, $\mathbf{u_t}, t = 1, 2, \ldots, r$, $\sigma_t, t = 1, 2, \ldots, r$, its right singular vectors, left singular vectors and singular values respy.
- **Theorem (Singular Value Decomposition** $A = \sum_{t=1}^{r} \sigma_t \mathbf{u_t} \mathbf{v_t}^T$. Sum of r outer products.
- **Claim** Matrices A, B are identical iff for all $\mathbf{v}$, we have $A\mathbf{v} = B\mathbf{v}$. If $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$, this holds for each unit vector e_j and so j th col of A, B are the same for all j.
- Let $B = \sum_{t=1}^{r} \sigma_t \mathbf{u_t} \mathbf{v_t}^T$. Want to show $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$. Enough to show for a set of $\mathbf{v}$ forming a basis of space. Take a convienient basis: $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}, \mathbf{v_{r+1}}, \ldots, \mathbf{v_d}$, containing the r singular vectors of A. [Such a basis exists. Why?]

Singular Value Decomposition

- A any matrix, $\mathbf{v_t}, t = 1, 2, \ldots, r$, $\mathbf{u_t}, t = 1, 2, \ldots, r$, $\sigma_t, t = 1, 2, \ldots, r$, its right singular vectors, left singular vectors and singular values respy.
- **Theorem (Singular Value Decomposition** $A = \sum_{t=1}^{r} \sigma_t \mathbf{u_t} \mathbf{v_t}^T$. Sum of r outer products.
- **Claim** Matrices A, B are identical iff for all $\mathbf{v}$, we have $A\mathbf{v} = B\mathbf{v}$. If $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$, this holds for each unit vector e_j and so j th col of A, B are the same for all j.
- Let $B = \sum_{t=1}^{r} \sigma_t \mathbf{u_t} \mathbf{v_t}^T$. Want to show $A\mathbf{v} = B\mathbf{v}$ for all $\mathbf{v}$. Enough to show for a set of $\mathbf{v}$ forming a basis of space. Take a convienient basis: $\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_r}, \mathbf{v_{r+1}}, \ldots, \mathbf{v_d}$, containing the r singular vectors of A. [Such a basis exists. Why?]
- For $t = 1, 2, \ldots, r$: $A\mathbf{v_t} = \sigma_t \mathbf{u_t}$ and $B\mathbf{v_t} = \sigma_t \mathbf{v_t}$ too by the ortho nomrality of $\mathbf{v_1}, \ldots, \mathbf{v_r}$.
- For $t \geq r + 1$, $A\mathbf{v_t} = \mathbf{0}$ (Why?) and so is $B\mathbf{v_t}$. QED