SVD and Low-rank Approximation

Lecture 19

Nov 1, 2022

Matrix Rank

Given $m \times n$ matrix A the *column rank* of A is the maximum number of linearly independent columns of A. The *row rank* is the maximum number of linearly independent rows of A.

Non-obvious fact: column rank = to row rank = $\text{rank}(A)$

Fact: A has rank r iff A can be written as sum of k rank 1 matrices

$$A = \sum_{i=1}^{r} y_i z_i^T = YZ^T$$

where Y is $m \times r$ matrix and Z is $r \times n$ matrix.

Singular Value Decomposition (SVD)

Let A be a $m \times n$ real-valued matrix

- a_i denotes vector corresponding to row i
- m rows. think of each row as a data point in $\mathbb{R}^n$
- Data applications: $m \gg n$
- Other notation: A is a $n \times d$ matrix.

Singular Value Decomposition (SVD)

Let A be a $m \times n$ real-valued matrix

- a_i denotes vector corresponding to row i
- m rows. think of each row as a data point in $\mathbb{R}^n$
- Data applications: $m \gg n$
- Other notation: A is a $n \times d$ matrix.

SVD theorem: A can be written as UDV^T where

- V is a $n \times n$ orthonormal matrix
- D is a $m \times n$ diagonal matrix with $\leq \min\{m, n\}$ non-zeroes called the singular values of A
- U is a $m \times m$ orthonormal matrix

SVD

Let $d = \min\{m, n\}$.

- $u_1, u_2, \ldots, u_m$ columns of U, left singular vectors of A
- $v_1, v_2, \ldots, v_n$ columns of V (rows of V^T) right singular vectors of A
- $\sigma_1 \geq \sigma_2 \geq \ldots, \geq \sigma_d \geq 0$ are non-negative singular values where $d = \min\{m, n\}$. And $\sigma_i = D_{i,i}$

$$A = \sum_{i=1}^{d} \sigma_i u_i v_i^T$$

SVD

Let $d = \min\{m, n\}$.

- $u_1, u_2, \ldots, u_m$ columns of U, left singular vectors of A
- $v_1, v_2, \ldots, v_n$ columns of V (rows of V^T) right singular vectors of A
- $\sigma_1 \geq \sigma_2 \geq \ldots, \geq \sigma_d \geq 0$ are non-negative singular values where $d = \min\{m, n\}$. And $\sigma_i = D_{i,i}$

$$A = \sum_{i=1}^{d} \sigma_i u_i v_i^T$$

We can in fact restrict attention to r the rank of A.

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

SVD

Interpreting A as a linear operator $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$

- Columns of V is an orthonormal basis and hence $V^T x$ for $x \in \mathbb{R}^n$ expresses x in the V basis. Note that $V^T x$ is a rigid transformation (does not change length of x).
- Let $y = V^T x$. D is a diagonal matrix which only stretches y along the coordinate axes. Also adjusts dimension to go from n to m with right number of zeroes.
- Let $z = Dy$. Then Uz is a rigid transformation that expresses z in the basis corresponding to rows of U.

Thus any linear operator can be split into a sequence of three simpler/basic type of transformations

Low rank approximation property of SVD

Question: Given $A \in \mathbb{R}^{m \times n}$ and integer k find a matrix B of rank at most k such that $\|A - B\|$ is minimized

Low rank approximation property of SVD

Question: Given $A \in \mathbb{R}^{m \times n}$ and integer k find a matrix B of rank at most k such that $\|A - B\|$ is minimized

Fact: For Frobenius norm *and* spectral norm optimum for all k is captured by SVD.

That is, $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ is the best rank k approximation to A

$$\|A - A_k\|_F = \min_{B:\text{rank}(B) \leq k} \|A - B\|_F = \sqrt{\sum_{i>k} \sigma_i^2}$$

$$\|A - A_k\|_2 = \min_{B:\text{rank}(B) \leq k} \|A - B\|_2 = \sigma_{k+1}$$

Low rank approximation property of SVD

Question: Given $A \in \mathbb{R}^{m \times n}$ and integer k find a matrix B of rank at most k such that $\|A - B\|$ is minimized

Fact: For Frobenius norm *and* spectral norm optimum for all k is captured by SVD.

That is, $A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ is the best rank k approximation to A

$$\|A - A_k\|_F = \min_{B:\text{rank}(B) \leq k} \|A - B\|_F = \sqrt{\sum_{i>k} \sigma_i^2}$$

$$\|A - A_k\|_2 = \min_{B:\text{rank}(B) \leq k} \|A - B\|_2 = \sigma_{k+1}$$

Why this magic? Frobenius norm and basic properties of vector

Geometric meaning

What is the best rank 1 matrix B that minimizes $\|A - B\|_F$

Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$

Without loss of generality v is a unit vector

Geometric meaning

What is the best rank 1 matrix B that minimizes $\|A - B\|_F$
Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Without loss of generality v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

Geometric meaning

What is the best rank 1 matrix B that minimizes $\|A - B\|_F$
Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Without loss of generality v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

If we know v then best u to minimize above is determined. Why?

Geometric meaning

What is the best rank 1 matrix B that minimizes $\|A - B\|_F$
Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$
Without loss of generality v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

If we know v then best u to minimize above is determined. Why?
For fixed v, $u(i) = \langle a_i, v \rangle$

Geometric meaning

What is the best rank 1 matrix B that minimizes $\|A - B\|_F$

Since B is rank 1, $B = uv^T$ where $v \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$

Without loss of generality v is a unit vector

$$\|A - uv^T\|_F^2 = \sum_{i=1}^{m} ||a_i - u(i)v||^2$$

If we know v then best u to minimize above is determined. Why?

For fixed v, $u(i) = \langle a_i, v \rangle$

$\|a_i - \langle a_i, v \rangle v\|_2$ is distance of a_i from line described by v.

Geometric meaning

What is the best rank 1 matrix B that minimizes $\|A - B\|_F$
It is to find unit vector/direction v to minimize

$$\sum_{i=1}^{m} ||a_i - \langle a_i, v \rangle v||^2$$

which is same as finding unit vector v to maximize

$$\sum_{i=1}^{m} \langle a_i, v \rangle^2$$

Geometric meaning

What is the best rank 1 matrix B that minimizes $\|A - B\|_F$
It is to find unit vector/direction v to minimize

$$\sum_{i=1}^{m} ||a_i - \langle a_i, v\rangle v||^2$$

which is same as finding unit vector v to maximize

$$\sum_{i=1}^{m} \langle a_i, v\rangle^2$$

Note: Maximum value is $\|A\|_2^2$, the spectral norm square! How to find best v? Not obvious: we will come to it a bit later

Best rank two approximation

Consider $k = 2$. What is the best rank 2 matrix B that minimizes $\|A - B\|_F$

Since B has rank 2 we can assume without loss of generality that $B = u_1 v_1^T + u_2 v_2^T$ where v_1, v_2 are orthogonal unit vectors (span a space of dimension 2)

Best rank two approximation

Consider $k = 2$. What is the best rank 2 matrix B that minimizes $\|A - B\|_F$

Since B has rank 2 we can assume without loss of generality that $B = u_1 v_1^T + u_2 v_2^T$ where v_1, v_2 are orthogonal unit vectors (span a space of dimension 2)

Minimizing $\|A - B\|_F^2$ is same as finding orthogonal vectors v_1, v_2 to maximize

$$\sum_{i=1}^{m} (\langle a_i, v_1 \rangle^2 + \langle a_i, v_2 \rangle^2)$$

in other words the best fit 2-dimensional space

Greedy algorithm

- Find v_1 as the best rank 1 approximation. That is $v_1 = \arg\max_{v, \|v\|_2=1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$
- For v_2 solve $\arg\max_{v \perp v_1, \|v\|_2=1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$.

  Alternatively: let $a_i' = a_i - \langle a_i, v_1 \rangle v_1$. Let $v_2 = \arg\max_{v, \|v\|_2=1} \sum_{i=1}^{m} \langle a_i', v \rangle^2$

Greedy algorithm

- Find $\mathbf{v}_1$ as the best rank 1 approximation. That is $\mathbf{v}_1 = \arg\max_{\mathbf{v}, \|\mathbf{v}\|_2 = 1} \sum_{\mathbf{i}=1}^{\mathbf{m}} \langle \mathbf{a}_{\mathbf{i}}, \mathbf{v} \rangle^2$
- For $\mathbf{v}_2$ solve $\arg\max_{\mathbf{v} \perp \mathbf{v}_1, \|\mathbf{v}\|_2 = 1} \sum_{\mathbf{i}=1}^{\mathbf{m}} \langle \mathbf{a}_{\mathbf{i}}, \mathbf{v} \rangle^2$.

  Alternatively: let $\mathbf{a}'_{\mathbf{i}} = \mathbf{a}_{\mathbf{i}} - \langle \mathbf{a}_{\mathbf{i}}, \mathbf{v}_1 \rangle \mathbf{v}_1$. Let $\mathbf{v}_2 = \arg\max_{\mathbf{v}, \|\mathbf{v}\|_2 = 1} \sum_{\mathbf{i}=1}^{\mathbf{m}} \langle \mathbf{a}'_{\mathbf{i}}, \mathbf{v} \rangle^2$

Greedy algorithm works!

Greedy algorithm correctness

Proof that Greedy works for $k = 2$.

Suppose w_1, w_2 are orthogonal unit vectors that form the best fit 2-d space. Let H be the space spanned by w_1, w_2.

Claim: Any two orthogonal unit vectors in H will yield same value.

Suffices to prove that

$$\sum_{i=1}^{m}(\langle a_i, v_1\rangle^2 + \langle a_i, v_2\rangle^2) \geq \sum_{i=1}^{m}(\langle a_i, w_1\rangle^2 + \langle a_i, w_2\rangle^2)$$

Greedy algorithm correctness

Case 1: $v_1 \in H$ then done because we can assume wlog that $w_1 = v_1$ and v_2 is at least as good as w_2.

Case 2: $v_1 \notin H$. Let v_1' be projection of v_1 onto H and $v_1'' = v_1 - v_1'$ be the component of v_1 orthogonal to H.

Greedy algorithm correctness

Case 1: $\boldsymbol{v}_1 \in \boldsymbol{H}$ then done because we can assume wlog that $\boldsymbol{w}_1 = \boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is at least as good as $\boldsymbol{w}_2$.

Case 2: $\boldsymbol{v}_1 \notin \boldsymbol{H}$. Let $\boldsymbol{v}_1'$ be projection of $\boldsymbol{v}_1$ onto $\boldsymbol{H}$ and $\boldsymbol{v}_1'' = \boldsymbol{v}_1 - \boldsymbol{v}_1'$ be the component of $\boldsymbol{v}_1$ orthogonal to $\boldsymbol{H}$. Note that $\|\boldsymbol{v}_1'\|^2 + \|\boldsymbol{v}_1''\|_2^2 = \|\boldsymbol{v}_1\|_2^2 = 1$.

Wlog we can assume by rotation that $\boldsymbol{w}_1 = \frac{1}{\|\boldsymbol{v}_1'\|_2}\boldsymbol{v}_1'$ and $\boldsymbol{w}_2$ is orthogonal to $\boldsymbol{v}_1'$. Hence $\boldsymbol{w}_2$ is orthogonal to $\boldsymbol{v}_1$.

Greedy algorithm correctness

Case 1: $\mathbf{v}_1 \in \mathbf{H}$ then done because we can assume wlog that $\mathbf{w}_1 = \mathbf{v}_1$ and $\mathbf{v}_2$ is at least as good as $\mathbf{w}_2$.

Case 2: $\mathbf{v}_1 \not\in \mathbf{H}$. Let $\mathbf{v}_1'$ be projection of $\mathbf{v}_1$ onto $\mathbf{H}$ and $\mathbf{v}_1'' = \mathbf{v}_1 - \mathbf{v}_1'$ be the component of $\mathbf{v}_1$ orthogonal to $\mathbf{H}$. Note that $\|\mathbf{v}_1'\|^2 + \|\mathbf{v}_1''\|_2^2 = \|\mathbf{v}_1\|_2^2 = 1$.

Wlog we can assume by rotation that $\mathbf{w}_1 = \frac{1}{\|\mathbf{v}_1'\|_2}\mathbf{v}_1'$ and $\mathbf{w}_2$ is orthogonal to $\mathbf{v}_1'$. Hence $\mathbf{w}_2$ is orthogonal to $\mathbf{v}_1$.

Therefore $\mathbf{v}_2$ is at least as good as $\mathbf{w}_2$, and $\mathbf{v}_1$ is at least as good as $\mathbf{w}_1$ which implies the desired claim.

Greedy algorithm for general k

- Find v_1 as the best rank 1 approximation. That is $v_1 = \arg\max_{v, \|v\|_2 = 1} \sum_{i=1}^{m} \langle a_i, v \rangle^2$
- For v_k solve $\arg\max_{v \perp v_1, v_2, \ldots, v_{k-1}, \|v\|_2 = 1} \sum_{i=1}^{k} \langle a_i, v \rangle^2$ which is same as solving $k = 1$ with vectors $a'_1, a'_2, \ldots, a'_m$ that are residuals. That is $a'_i = a_i - \sum_{j=1}^{k-1} \langle a_i, v_j \rangle v_j$

Proof of correctness is via induction and is a straight forward generalization of the proof for $k = 2$

Summarizing

$\sigma_j^2 = \sum_{i=1}^{m} \langle a_i, v_j \rangle^2$

By greedy contruction $\sigma_1 \geq \sigma_2 \geq \ldots,$

Let r be the (row) rank of A. $v_1, v_2, \ldots, v_r$ span the row space of A and $\sigma_j = 0$ for $j > r$. Can choose $v_{r+1}, \ldots, v_n$ to ensure orthonormal basis of R^n

u_1 determined by v_1 and u_2 determined by v_1, v_2 and so on. Can show that they are orthogonal.

$$A = \sum_{i=1}^{n} \sigma_i u_i v_i^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

Power method

Thus SVD relies on being able to solve $k = 1$ case

Given m vectors $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ solve

$$\max_{v \in \mathbb{R}^n, \|v\|_2 = 1} \langle a_i, v \rangle^2$$

How do we solve the above problem?

Let $B = A^T A$ Then

$$\begin{aligned} B &= \left(\sum_{i=1}^{m} \sigma_i v_i u_i^T\right)\left(\sum_{i=1}^{r} \sigma_i u_i v_i^T\right) \\ &= \sum_{i=1}^{r} \sigma_i^2 v_i v_i^T \end{aligned}$$

Power method continued

Let $B = A^T A$ Then

$$\begin{aligned} B^2 &= (\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T)(\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T) \\ &= \sum_{i=1}^{r} \sigma_i^4 v_i v_i^T. \end{aligned}$$

More generally

$$B^k = \sum_{i=1}^{r} \sigma_i^{2k} v_i v_i^T$$

Power method continued

Let $B = A^T A$ Then

$$\begin{aligned} B^2 &= (\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T)(\sum_{i=1}^{r} \sigma_i^2 v_i v_i^T) \\ &= \sum_{i=1}^{r} \sigma_i^4 v_i v_i^T. \end{aligned}$$

More generally

$$B^k = \sum_{i=1}^{r} \sigma_i^{2k} v_i v_i^T$$

If $\sigma_1 > \sigma_2$ then B^k converges to $\sigma_1^{2k} v_1 v_1^T$ and we can identify v_1 from B^k. But expensive to compute B^k

Power method continued

Pick a random (unit) vector $\mathbf{x} \in \mathbb{R}^n$. Then $\mathbf{x} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i$ since $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ is a basis for $\mathbb{R}^n$.

$$B^k x = (\sum_{i=1}^{r} \sigma_i^{2k} v_i v_i^T)(\sum_{i=1}^{d} \lambda_i v_i) \rightarrow \sigma_1^{2k} \lambda_1 v_1$$

Can obtain $\mathbf{v}_1$ by normalizing $B^k x$ to a unit vector.
Computing $B^k x$ is easier via a series of matrix vector multiplications

Power method continued

Pick a random (unit) vector $\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$. Then $\boldsymbol{x} = \sum_{i=1}^{n} \boldsymbol{\lambda_i v_i}$ since $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v_n}$ is a basis for $\mathbb{R}^{\boldsymbol{n}}$.

$$B^k x = (\sum_{i=1}^{r} \sigma_i^{2k} v_i v_i^T)(\sum_{i=1}^{d} \lambda_i v_i) \to \sigma_1^{2k} \lambda_1 v_1$$

Can obtain $\boldsymbol{v}_1$ by normalizing $\boldsymbol{B^k x}$ to a unit vector.
Computing $\boldsymbol{B^k x}$ is easier via a series of matrix vector multiplications

Why random $\boldsymbol{x}$? So as to ensure $\boldsymbol{\lambda}_1 > 0$ with good probability.

Theorem

Suppose $\boldsymbol{\sigma}_1 > \boldsymbol{\sigma}_2$. Then with probability $(1-\delta)$, power method converges to a vector $\boldsymbol{v}$ such that $\langle \boldsymbol{v}, \boldsymbol{v}_1 \rangle \geq (1-\epsilon)$ after $O(\frac{\log n + \log(1/\epsilon) + \log(1/\delta)}{\log(\sigma_1/\sigma_2)})$ iterations.

Power method continued

Pick a random (unit) vector $x \in \mathbb{R}^n$. Then $x = \sum_{i=1}^{n} \lambda_i v_i$ since $v_1, v_2, \ldots, v_n$ is a basis for $\mathbb{R}^n$.

$$B^k x = (\sum_{i=1}^{r} \sigma_i^k v_i v_i^T)(\sum_{i=1}^{d} \lambda_i v_i) \to \sigma_1^{2k} \lambda_1 v_1$$

Convergence dependes on σ_1/σ_2. What if $\sigma_1 \simeq \sigma_2$? Power method may not converge to v_1 but output will be some vector in the space spanned by $v_1, v_2, \ldots, v_h$ where σ_h is the largest h such that $\sigma_1 \simeq \sigma_h$. This is good enough in various applications. See references.

Principal Component Analysis

Consider $\boldsymbol{A}$ a $\boldsymbol{m} \times \boldsymbol{n}$ matrix where rows $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_{\boldsymbol{m}}$ are data points in $\mathbb{R}^{\boldsymbol{n}}$

$\boldsymbol{B} = \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A}$ is a symmetrix positive definite matrix and has real non-negative eigenvalues

Via SVD $\boldsymbol{B} = (\boldsymbol{UDV}^{\boldsymbol{T}})^{\boldsymbol{T}}(\boldsymbol{UDV}^{\boldsymbol{T}}) = (\boldsymbol{VD}^{\boldsymbol{T}}\boldsymbol{DV}^{\boldsymbol{T}})$

Can check that $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{\boldsymbol{r}}$ are eigen vectors of $\boldsymbol{B}$ with eigen values $\boldsymbol{\sigma}_1^2, \boldsymbol{\sigma}_2^2, \ldots, \boldsymbol{\sigma}_{\boldsymbol{r}}^2$

Principal Component Analysis

Consider A a $m \times n$ matrix where rows $a_1, a_2, \ldots, a_m$ are data points in $\mathbb{R}^n$

- Compute eigenvectors of $B = A^T A$ or singular vectors $v_1, v_2, \ldots, v_n$ which are also called the principal directions
- Approximate each a_i by its projection onto the first k singular vectors for some small k. That is $a'_i = \sum_{j=1}^{k} \langle a_i, v_j \rangle v_j$.
- Thus $a'_1, a'_2, \ldots, a'_m$, a kind of *dimensionality reduction* along first k principal directions. Different from JL and is motivated by different applications (mainly statistical analysis)

PCA and Covariance Matrix

Covariance of two real-valued random variables $\boldsymbol{X}, \boldsymbol{Y}$ is defined as

$$\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) := \mathrm{E}[(\boldsymbol{X} - \mathrm{E}[\boldsymbol{X}])(\boldsymbol{Y} - \mathrm{E}[\boldsymbol{Y}])]$$

Note that $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X}) = \mathrm{Var}(\boldsymbol{X})$. If $\boldsymbol{X}, \boldsymbol{Y}$ independent then $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = 0$ but converse is not necessarily true. There is also a related normalized measure (value in $[-1, 1]$)

$$\mathrm{Correlation}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})}{\boldsymbol{\sigma_X \sigma_Y}}$$

The sign of $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})$ is an "indication" of positive vs negative correlation. Non-linear relationships between $\boldsymbol{X}, \boldsymbol{Y}$ are not necessarily captured by covariance but still useful in many situations.

PCA and Covariance Matrix

Suppose $X = (X_1, X_2, \ldots, X_n)$ is a n-dimensional random variable. Thus, each X_i is a random variable and may be correlated with the other variables.

Given X we can define a covariance matrix C where $C_{i,j} = \text{Cov}(X_i, X_j)$. Note that if the X_i are independent then C will be a diagonal matrix. Similarly one can also define a correlation matrix where the entries are the correlation coefficients instead of covariances.

PCA of C reveals useful information if X is in fact obtained via a linear transformation from another random variable Y that lives in a lower dimension. Typically X will be a noisy version of Y and hence will not be a pure low rank matrix but a low rank approximation gives the important directions.

PCA and Covariance Matrix

Suppose $X = (X_1, X_2, \ldots, X_n)$ is a n-dimensional random variable.

Suppose we have m data points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ drawn independently from the distribution of X. We create a $m \times n$ matrix A where a_i is i'th row. Given the empirical data matrix A we would like to estimate the covariance matrix C of X.

Assuming we know for each i, $\mu_i = \mathrm{E}[X_i]$ we can estimate $\mathrm{Cov}(X_i, X_j)$ from the m data samples as $\frac{1}{m}\sum_{\ell=1}^{m}(a_\ell(i) - \mu_i)(a_\ell(j) - \mu_j)$.

By setting $a'_\ell = a_\ell - \mu$ where μ is the vector of expectations we see that $C = \frac{1}{m}(A')^T A'$ is the desired estimated covariance matrix. Thus PCA on $(A')^T A'$ helps identify important features in the underlying distribution X

PCA and Covariance Matrix

Suppose $X = (X_1, X_2, \ldots, X_n)$ is a n-dimensional random variable. Suppose we have m data points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$ drawn independently from the distribution of X. We create a $m \times n$ matrix A where a_i is i'th row. Given the empirical data matrix A we would like to estimate the covariance matrix C of X.

Suppose we do not know the means $\mu_i = \mathrm{E}[X_i]$. We can compute an empirical estimate from the data itself as $\frac{1}{m}\sum_{\ell=1}^{m} a_\ell(i)$ and then the empirical mean vector it to "center" the data to compute an estimated covariance matrix as in the previous slide. Sometimes data is already assumed to be centered in which case we simply work with $A^T A$.

Linear least square/Regression and SVD

Linear least squares: Given $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{n}}$ and $\boldsymbol{b} \in \mathbb{R}^{\boldsymbol{m}}$ find $\boldsymbol{x}$ to minimize $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$.

Interesting when $\boldsymbol{m} > \boldsymbol{n}$ the over constrained case when there is no solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and want to find best fit.

Linear least square/Regression and SVD

Linear least squares: Given $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{n}}$ and $\boldsymbol{b} \in \mathbb{R}^{\boldsymbol{m}}$ find $\boldsymbol{x}$ to minimize $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$.

Interesting when $\boldsymbol{m} > \boldsymbol{n}$ the over constrained case when there is no solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and want to find best fit.

Geometrically $\boldsymbol{A}\boldsymbol{x}$ is a linear combination of columns of $\boldsymbol{A}$. Hence we are asking what is the vector $\boldsymbol{z}$ in the column space of $\boldsymbol{A}$ that is closest to vector $\boldsymbol{b}$ in ℓ_2 norm.

Linear least square/Regression and SVD

Linear least squares: Given $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{m} \times \boldsymbol{n}}$ and $\boldsymbol{b} \in \mathbb{R}^{\boldsymbol{m}}$ find $\boldsymbol{x}$ to minimize $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$.

Interesting when $\boldsymbol{m} > \boldsymbol{n}$ the over constrained case when there is no solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ and want to find best fit.

Geometrically $\boldsymbol{A}\boldsymbol{x}$ is a linear combination of columns of $\boldsymbol{A}$. Hence we are asking what is the vector $\boldsymbol{z}$ in the column space of $\boldsymbol{A}$ that is closest to vector $\boldsymbol{b}$ in ℓ_2 norm.

Closest vector to $\boldsymbol{b}$ is the projection of $\boldsymbol{b}$ into the column space of $\boldsymbol{A}$ so it is "obvious" geometrically. How do we find it?

Linear least square/Regression and SVD

Linear least squares: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ find x to minimize $\|Ax - b\|_2$.

Interesting when $m > n$ the over constrained case when there is no solution to $Ax = b$ and want to find best fit.

Geometrically Ax is a linear combination of columns of A. Hence we are asking what is the vector z in the column space of A that is closest to vector b in ℓ_2 norm.

Closest vector to b is the projection of b into the column space of A so it is "obvious" geometrically. How do we find it? Find an orthonormal basis $z_1, z_2, \ldots, z_r$ for the columns of A. Compute projection b' as $b' = \sum_{j=1}^{r} \langle b, z_j \rangle z_j$ and output answer as $\|b - b'\|_2$.

Linear least square/Regression and SVD

Linear least squares: Given $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ find x to minimize $\|Ax - b\|_2$.

Closest vector to b is the projection of b into the column space of A so it is "obvious" geometrically. Find an orthonormal basis $z_1, z_2, \ldots, z_r$ for the columns of A. Compute projection b' as $b' = \sum_{j=1}^{r} \langle b, z_j \rangle z_j$ and output answer as $\|b - b'\|_2$.

Finding the basis is the expensive part. Recall SVD gives $v_1, v_2, \ldots, v_r$ which form a basis for the *row* space of A but then $u_1^T, u_2^T, \ldots, u_m^T$ form a basis for the *column* space of A. Hence SVD gives us all the information to find b'. In fact we have

$$\min_x \|Ax - b\|_2^2 = \sum_{i=r+1}^{m} \langle u_i^T, b \rangle^2$$