Singular Value Decomposition (1)

- Data reduction
- Data-driven generalization of Fourier transform (FFT)
- Tailored to specific problem.

- Solve $Ax=b$ for non square A → regression
- Basis PCA → correlation

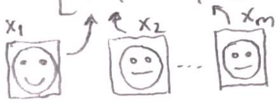
Simple/interpretable linear algebra.

Scalable.

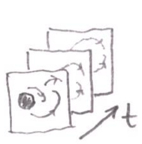
(2)

$$\underline{X} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & & | \end{bmatrix} = U \Sigma V^T = \underbrace{\begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix}}_{\text{eigen faces}} \underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \\ & & 0 & \end{bmatrix}}_{\text{ordered by importance}} \underbrace{\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_m \\ | & | & & | \end{bmatrix}^T}_{\text{eigen time series}} = U \Sigma V^T$$

singular values (Σ); U: left singular vectors; V: right singular vectors



$x_k \in \mathbb{R}^n$



$$\underline{X} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & & | \end{bmatrix} \xrightarrow{t}$$

- orthogonal
- U, V unitary

$$UU^T = U^TU = I_{n \times n}$$

$$VV^T = V^TV = I_{m \times m}$$

Σ diagonal $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_m \geq 0$

$$V^T = \begin{bmatrix} - & v_1^T & - \\ - & v_2^T & - \\ & \vdots & \\ - & v_m^T & - \end{bmatrix}$$

first column → ★ mixture of u's to make x_1

★ Guaranteed to exist

★ Unique

$$\underline{X} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & & | \end{bmatrix} = U \Sigma V^T = \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_m \\ & & 0 & \end{bmatrix} \begin{bmatrix} - & v_1^T & - \\ - & v_2^T & - \\ & \vdots & \\ - & v_m^T & - \end{bmatrix} = \hat{U} \hat{\Sigma} V^T$$

$\hat{U}$ → m columns of U; $\hat{\Sigma}$ → $m \times m$ matrix. } economy SVD

Matrix Approximation (SVD) (3)

$$\underline{X} = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_m u_m v_m^T + 0$$

$$= \underbrace{\underbrace{\sigma_1 u_1 v_1^T}_{\text{rank-1 matrix}} + \underbrace{\sigma_2 u_2 v_2^T}_{\text{rank-1 matrix}}}_{\text{Ex: Best rank-2 matrix}} + \cdots + \sigma_m u_m v_m^T$$

(truncate at rank r) $\approx \tilde{U} \tilde{\Sigma} \tilde{V}^T$: rank r approximation.

Eckard-Young thm [1936]

$$\underset{\tilde{X} \text{ s.t. } \operatorname{rank}(\tilde{X}) = r}{\operatorname{argmin}} \| \underline{X} - \tilde{\underline{X}} \|_F = \tilde{U} \tilde{\Sigma} \tilde{V}^T$$

After truncating $\tilde{U}^T \tilde{U} = I_{r \times r}$

$$\tilde{U} \tilde{U}^T \neq I$$

Dominant Correlations (SVD) (4)

$$\underbrace{X^T X}_{m\times m} = \underbrace{\begin{bmatrix} — x_1^T — \\ — x_2^T — \\ \vdots \\ — x_m^T — \end{bmatrix}}_{m\times n} \underbrace{\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & & | \end{bmatrix}}_{n\times m} = \underbrace{\begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \cdots & x_1^T x_m \\ x_2^T x_1 & x_2^T x_2 & \cdots & x_2^T x_m \\ \vdots & \vdots & & \vdots \\ x_m^T x_1 & x_m^T x_2 & \cdots & x_m^T x_m \end{bmatrix}}_{\text{Correlation Matrix (column-wise)}} \rightarrow \text{PSD}$$

$n \geq m$

$$x_i^T x_j = \langle x_i, x_j \rangle$$

If $X = \hat{U}\hat{\Sigma}V^T$

$X^T = V\hat{\Sigma}\hat{U}^T$

$$X^T X = V\hat{\Sigma}\underbrace{\hat{U}^T\hat{U}}_{I}\hat{\Sigma}V^T = V\hat{\Sigma}^2 V^T$$

$$\underbrace{X^T X}\, V = V\hat{\Sigma}^2$$

(column-wise correlation matrix)

$(m\times n)(n\times m) = m\times m$

V: eigenvectors of correlation matrix; $\hat{\Sigma}^2$: eigen values of correlation matrix.

Similarly,

$$X X^T = \hat{U}\hat{\Sigma}\underbrace{V^T V}_{I}\hat{\Sigma}\hat{U}^T = \hat{U}\hat{\Sigma}^2\hat{U}^T$$

$$X X^T \hat{U} = \hat{U}\hat{\Sigma}^2$$

(row-wise correlation matrix)

$(n\times m)(m\times n) = n\times n$ — bigger than column-wise correlation matrix.

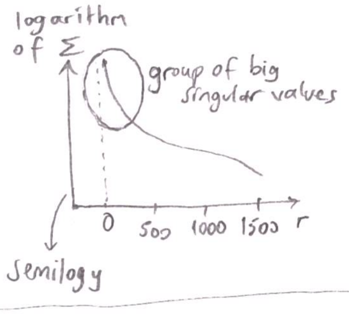
$\hat{U}$: eigenvectors, $\hat{\Sigma}^2$: eigenvalues of the correlation matrix

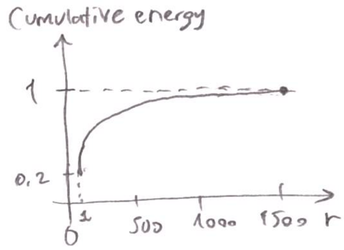
---

(5)

$$\left(\frac{\text{Cumulative sum of first } r \text{ singular values}}{\text{sum of all singular values}}\right) \Rightarrow$$

How much energy/information is in the first r modes compared to all modes.





The Frobenius Norm of Matrices (7)

$$\underset{\tilde{X}\text{ s.t. } \operatorname{rank}(\tilde{X})=r}{\operatorname{argmin}} \|X - \tilde{X}\|_F = \tilde{U}\tilde{\Sigma}\tilde{V}^T$$

$$\underbrace{\|A\|_F}_{\text{frobenius norm}} = \sqrt{\sum_{i,j} |A_{i,j}|^2}$$

(A) matrix to a big vector.

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}$$

SVD Method of Snapshots

(8)

Sirovich [1987]

If the matrix is too large

$$\underbrace{X^T X}_{\substack{\text{small} \\ m\times m}} \longrightarrow \underbrace{\boxed{V, \hat{\Sigma}}}_{\substack{\text{first find} \\ V,\hat{\Sigma}}} \qquad X = \hat{U}\hat{\Sigma}V^T \qquad \boxed{\hat{U} = X V \hat{\Sigma}^{-1}}$$

second find $\hat{U}$

- It is not recommend to compute SVD using correlation matrices but if the matrix is too big that you cannot load it to memory:

  Load first column and take dot product of itself.
  Load first column and second column and dot product.
  ⋮
  Only loading two vectors, the result $m\times m$ small correlation matrix can be obtained.

$$\underbrace{\begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \cdots \\ \vdots & & \\ & & \end{bmatrix}}_{\text{correlation matrix}}$$

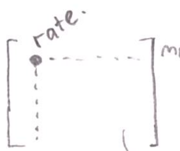
$x_i^T x_j = \langle x_i, x_j \rangle$ dot product.

Matrix Completion and the Netflix Price

(9)

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_m \\ | & | & & | \end{bmatrix} \begin{array}{l} \to \text{movie 1} \\ \to \text{movie 2} \\ \vdots \end{array}$$

x_1 → person 1, x_2 → p2, x_m → pm



empty areas.
missing values.
(how to fill them?)

$\underbrace{R}_{\text{Robust}}\underbrace{PCA}_{\text{principal component analysis.}}$

Unitary Transformations   Geometry

(10)

$$X = \underbrace{U\Sigma V^T}_{\text{SVD}} = \underbrace{\hat{U}\hat{\Sigma}V^T}_{\text{econ SVD}}$$

$$UU^T = U^T U = I$$
$$VV^T = V^T V = I$$

(U, V unitary matrices)

over the real number field U, V are called orthogonal

$$Q^T Q = QQ^T = I$$
$$Q^T = Q^{-1}$$

Unitary transformations preserve angles and lengths of vectors.

Just rotate vectors.

$$\langle x, y\rangle = \langle Ux, Uy\rangle \quad \forall x, y \in \mathbb{R}^n$$

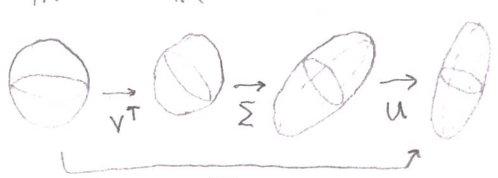
If X complex $\in \mathbb{C}^{n\times m}$

then X^* is complex conjugate transpose.

(11)(12)

SVD: Rotate + Scale + Rotate

$$\mathbb{R}^m \xrightarrow{\text{Rotate}} \mathbb{R}^n \xrightarrow{\text{Scale}} \mathbb{R}^n \xrightarrow{\text{Rotate}} \mathbb{R}^m$$



Linear Systems of Equations, Least Square Regression, Pseudoinverse (13)

$\underline{\underline{A}}\, x = \underline{b}$ linear system of equations

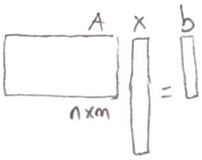
$\underline{\underline{A}}$ known, x ↓ solve, $\underline{b}$ known

SVD allow us to generalize to non-square A. *

- Underdetermined, $n<m$ (short fat A)

∞ many solution x given b.

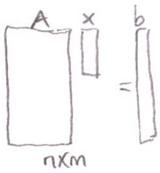
(infinite solutions) *



- Overdetermined, $n>m$ (tall skinny A)

zero solutions x for generic b.

(no solution) *



$$A = U\Sigma V^T \Rightarrow A^\dagger$$ (economy SVD) pseudo-inverse

$$\boxed{A^\dagger = V\Sigma^{-1}U^T}$$ * moore-penrose left pseudo inverse

$Ax=b$ ↓ left

$$Ax=b$$

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

$$\underbrace{V\Sigma^{-1}\underbrace{U^TU}\Sigma V^T}_{I} x = \underbrace{V\Sigma^{-1}U^T}_{A^\dagger} b$$

$$\tilde{x} = V\Sigma^{-1}U^Tb$$

$$\tilde{x} = A^\dagger b$$ †dagger

$\tilde{x}$ ↓ best x

- Underdetermined

$\min \|\tilde{x}\|_2$ s.t. $A\tilde{x}=b$ minimum 2-norm x.

(minimum norm solution)

↳ cheapest way of multiplying A with x to get b. let's say x is energy.

- Overdetermined.

$\min \|A\tilde{x}-b\|_2$ ← 2 norm error

(least square solution)

Least Square Regression and the SVD (14)

$$A\tilde{x} = b$$

$$U\Sigma V^T \quad V\Sigma^{-1}U^Tb$$

$$\overbrace{U\Sigma \underbrace{\underbrace{V^TV}_{I}\Sigma^{-1}}_{I} U^Tb}^{A\tilde{x}}$$

$\hat{U}\hat{U}^T \neq I$ for econ SVD.

$$\overbrace{\underbrace{\hat{U}\hat{U}^T b}}^{A\tilde{x}}$$

projection of b onto span$(\hat{U})$ = span(A)

Linear system of equations, $Ax=b$ (15)

* Solution of $Ax=b$ only exists if b is in column space of A.

$$\underbrace{A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}}$$

no solution even though $Ax=b$ underdetermined $(n<m)$

$$\underbrace{A = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix} \quad b = \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} \quad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}}$$

there might be a solution even though $Ax=b$ overdetermined $(n>m)$

- col(A) = col$(\hat{U})$ range
- ker(A^T) orthogonal complement

↓ kernel

- row(A) = col(V)
- ker(A) null space

Set of all vec. x_{null} s.t. $Ax_{null}=0$

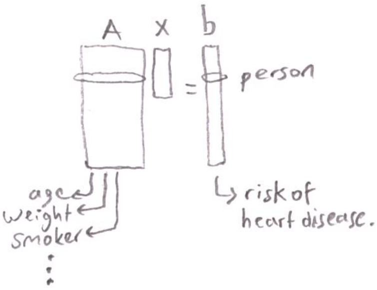
* Also, dim(ker(A)) $\neq 0$ (there are some vectors that map to zero) then ∞ many solutions.

$$A(x+x_{null}) = b$$

Linear Regression

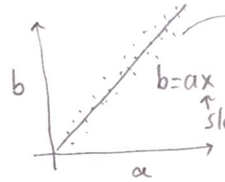
$Ax = b$ linear system of equations $\Rightarrow \boxed{\tilde{x} = A^{+} b}$ $\quad A = \hat{U}\hat{\Sigma}V^{T} \Rightarrow A^{+} = V\Sigma^{-1}U^{T}$

Linear regression model from data.



best x model

1- Dimensional Case



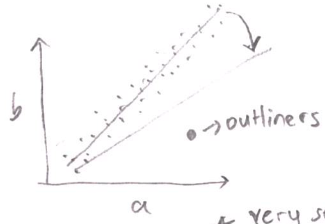
$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix} \quad \begin{bmatrix} b_1 \\ b_2 \\ \vdots \end{bmatrix}$$

$$\left.\begin{aligned} U &= \frac{a}{\|a\|_2} \\ \Sigma &= \|a\|_2 \\ V &= 1 \end{aligned}\right\}$$

best fit slope

$$\tilde{x} = \frac{a^{T} b}{\|a\|_2^{2}}$$

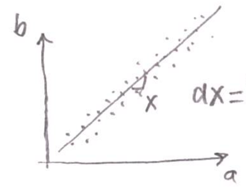
* Add L1 penalty term to discount outliners.



* very sensitive to the outliners.

If the data is clean or only has gaussian distribution, it is optimal to fit data (SVD)

Linear Regression



$x = 3$

$a = [-2, 2]$ $\Delta a = 0.1$ → + noise

$b = ax + \text{noise}$

both a & b have noisy data.

} Example of Data.

$$\left.\begin{aligned} Ax &= b \\ A &= U\Sigma V^{T} \end{aligned}\right\} \boxed{\tilde{x} = V\Sigma^{-1}U^{T} b}$$

SVD $\Rightarrow [U, \Sigma, V] = \text{svd}(a, \text{"econ"})$

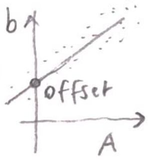
$$\tilde{x} = \underbrace{V\Sigma^{-1}U^{T}}_{\text{pseudo-inverse of matrix } A} b$$

* Note: If there is offset

Add ones for the offset

$$\begin{bmatrix} 1 & \\ 1 & \\ 1 & A \\ 1 & \\ 1 & \end{bmatrix} \begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} b \end{bmatrix}$$

Add a x for the offset.



$ax + \text{offset} = b$

$\boxed{\tilde{x} = V\Sigma^{-1}U^{T} b}$

$$A = \begin{bmatrix} 7 & 26 & 6 & 60 \\ 10 & 59 & 7 & 14 \\ & \vdots & & \\ & & & \end{bmatrix} \begin{matrix} \leftarrow \text{experiments} \\ \leftarrow \; 13 \\ \leftarrow \\ \\ \end{matrix} \qquad b = \begin{bmatrix} \\ \\ \\ \end{bmatrix} \rightarrow \text{heat}$$

↑ ↑ ↑ ↑ 13x4

4 ingredients

(overdetermined)

Ex:

- Use 10 data to create model
- Use 3 data to validate / how accurate

avoid overfitting.

A → parameters b → house value.

↳ area

↳ close to school

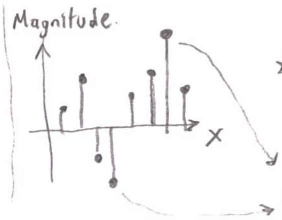
↳ ⋮

19, 20, 21 python

500 data → 250 data : model
→ 250 data : validate

maybe data collected by one neighbourhood and another neighbourhood.

* Instead of taking first half, shuffle the data.



Dominant parameter for higher house price.

Dominant parameter for lower house price.

Principal Component Analysis (PCA) [dimentionality reduction]

(22)

Statistical Interpretation of the SVD.

Hierarchical coordinate system (based on data)

$$\mathbb{X} = \begin{bmatrix} — x_1 — \\ — x_2 — \\ \vdots \end{bmatrix}$$

→ measurements from a single experiment. (row instead of column)

1. Compute mean row

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

$$\bar{\mathbb{X}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}[\bar{x}]$$

2. Subtract mean $B = \mathbb{X} - \bar{\mathbb{X}}$

3. Covariance matrix of the rows of B.

$$C = B^T B$$

4. Compute eigs of C.

$$V_1^T B^T B V_1$$

↳ biggest eigenvector

$$CV = VD$$

↳ eigenvalues.

↓ eigenvectors.

$$\lambda = \sigma^2$$

mean subtracted data

$$T = BV$$

↳ loadings.

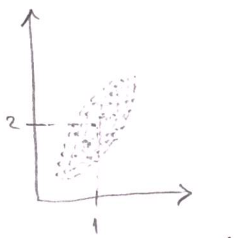
↓ principal components

$$B = U\Sigma V^T \Rightarrow T = U\Sigma$$

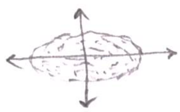
$$\frac{\sum_{k=1}^{r} \lambda_k}{\sum_{k=1}^{n} \lambda_k}$$

Principal Component Analysis

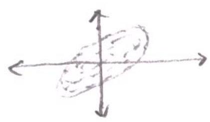
(23)



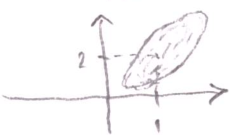
Principal axes [2;0.5]



Rotate by $\pi/3$



Add offset



1. mean ⇒ $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$

2. $B = \mathbb{X} - \bar{\mathbb{X}}$ mean centered matrix B.

3. Compute SVD of normalized B

B/sqrt(number of Points)

↓

Principal axes and rotation captured from data.

Σ should capture $\begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}$

U should capture R by $\frac{\pi}{3}$

↓

first column of U → first direction of maximal variance

second column of U → second direction of maximal variance

★ SVD can be used to compute the principal component analysis which can give you some information about the distribution/statistics how data is distributed, what are the principal directions of variance and the directions that have the least variance, the most variance and you can visulize very high dimensional data.