SVD and Least Squares Problems

Note Title

LECTURE 17

☆ LS via SVD

Recall the LS solution via QR factorization:

(1) Compute reduced QR of A.
(2) Compute $y = \hat{Q}^T b$.
(3) Solve $\hat{R}x = y$ — (*)

If A: full rank, then $\hat{R}_{ii} \neq 0$, $1 \le i \le n$, and the triangular system (*) has a unique LS solution.

Now using the reduced SVD of A, i.e., $A = \hat{U}\hat{\Sigma}V^T$, we can also solve the normal eqn:

$$A^TAx = A^Tb$$
$$\Leftrightarrow (\hat{U}\hat{\Sigma}V^T)^T(\hat{U}\hat{\Sigma}V^T)x = (\hat{U}\hat{\Sigma}V^T)^Tb$$
$$\Leftrightarrow V\hat{\Sigma}^T\hat{U}^T\hat{U}\hat{\Sigma}V^Tx = V\hat{\Sigma}\hat{U}^Tb$$
$$\Leftrightarrow V\hat{\Sigma}^T\hat{\Sigma}V^Tx = V\hat{\Sigma}^T\hat{U}^Tb$$
$$\Leftrightarrow \hat{\Sigma}^T\hat{\Sigma}V^Tx = \hat{\Sigma}^T\hat{U}^Tb \quad \text{since } V\text{: ortho.}$$
$$\Leftrightarrow \hat{\Sigma}V^Tx = \hat{U}^Tb$$

if A: full rank, i.e., $\sigma_j > 0$, $1 \le j \le n$

This can be solved easily.

(1) Compute reduced SVD of A.
(2) Compute $y = \hat{U}^Tb$.
(3) Solve $\hat{\Sigma}w = y$. —— (**)
(4) Set $x = Vw$.

Note: (**) is a diagonal system, easier to solve than (*) !!

☆ Pseudoinverse and SVD

Recall that if $A \in \mathbb{R}^{m\times n}$ is full rank,

$m > n$: $A^\dagger = (A^T A)^{-1} A^T$

$m = n$: $A^\dagger = A^{-1}$

$m < n$: $A^\dagger = A^T (A A^T)^{-1}$

However, we can define the pseudo inv. using SVD even if A is not full rank!

$$A = U \Sigma V^T, \qquad \Sigma = \begin{array}{c} \begin{array}{cc} \overbrace{\qquad\qquad}^{r} & \overbrace{\quad}^{n-r} \end{array} \\ \left[\begin{array}{c|c} \begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_r \end{matrix} & 0 \\ \hline 0 & 0 \end{array}\right] \begin{array}{l} \}\, r \\ \}\, m-r \end{array} \end{array}$$

Define

$$A^\dagger := V \Sigma^\dagger U^T, \qquad \Sigma^\dagger := \begin{array}{c} \begin{array}{cc} \overbrace{\qquad\qquad}^{r} & \overbrace{\quad}^{m-r} \end{array} \\ \left[\begin{array}{c|c} \begin{matrix} 1/\sigma_1 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_r \end{matrix} & 0 \\ \hline 0 & 0 \end{array}\right] \begin{array}{l} \}\, r \\ \}\, n-r \end{array} \end{array}$$

As we discussed before, $A^\dagger$ satisfies the following Moore-Penrose conditions:

(i) $AXA = A$; (ii) $XAX = X$

(iii) $(AX)^T = AX$; (iv) $(XA)^T = XA$.

Such X is uniquely determined and $X = A^\dagger$!!

★ Pseudoinverse & Orthogonal Projectors

Thm $AA^\dagger$ is an ortho. proj. onto range(A)
and $AA^\dagger = U_r U_r^T$
$A^\dagger A$ is an ortho. proj. onto range(A^T)
and $A^\dagger A = V_r V_r^T$
where $U_r \in \mathbb{R}^{m\times r}$, $V_r \in \mathbb{R}^{n\times r}$ consist of the first r columns of U, V, respectively.
$r = \text{rank}(A)$.

(Proof) Let $P_A := AA^\dagger$, $P_{A^T} := A^\dagger A$.

Now, $P_A = U\Sigma V^T V \Sigma^\dagger U^T$

$= U\Sigma\Sigma^\dagger U^T = U \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U^T$

$= U_r U_r^T$ ✓

$P_A^2 = U_r \underbrace{U_r^T U_r}_{= I_r} U_r^T = U_r U_r^T = P_A$ ✓
so it's a proj.!

$P_A^T = (U_r U_r^T)^T = (U_r^T)^T U_r^T = U_r U_r^T = P_A$ ✓
So it's an ortho. proj.!

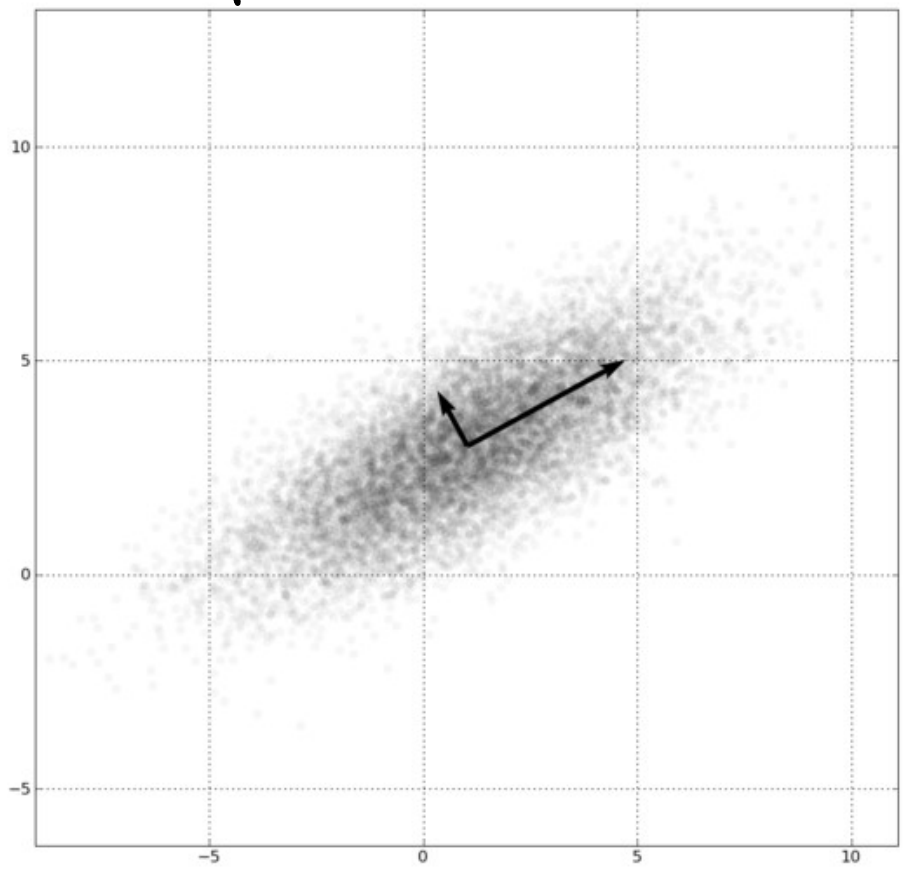
Finally, it's also clear that
P_A maps onto range(A) since
range$(A) = \langle u_1, \cdots, u_r \rangle$. ✓
You can do similarly for P_{A^T} ///

Note: Consider any $x \in$ range(A).
Then $\exists\, y \in \mathbb{R}^n$ s.t. $x = Ay$.
Now $P_A x = AA^\dagger x = \underbrace{AA^\dagger A}_{\text{"}A \text{ via Moore-Penrose (i)}} y$
$= Ay = x$.

☆ Principal Component Analysis (PCA)

(a.k.a. Karhunen-Loève Transform) is a data analysis technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables called "principal components."

2D example (from Wikipedia)



One can understand PCA using SVD! But before doing so, we need a bit of statistics.

Suppose we are given a set of vectors (observations)

$x_1, x_2, \cdots, x_n$

often these are viewed as n realizations of some stochastic process.

and each $x_j \in \mathbb{R}^d$. d: could be huge (ex. a face image database).

Let $X := [x_1 \; x_2 \; \cdots \; x_n] \in \mathbb{R}^{d \times n}$

You know the mean (or average) of this data set

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

And define the centered data matrix

$$\tilde{X} := [x_1 - \bar{x} \;\; x_2 - \bar{x} \;\; \cdots \;\; x_n - \bar{x}]$$

Note: $\tilde{X} = X\left(I_n - \frac{1}{n} \mathbb{1}_n \mathbb{1}_n^T\right)$

↳ Good exercise!

Now the sample covariance matrix S is defined as

$$S := \frac{1}{n} \tilde{X} \tilde{X}^T \in \mathbb{R}^{d \times d}$$

S_{ij} indicates the covariance or mutual correlation between the ith and jth entries of data vectors.

PCA is nothing but an eigenvalue decomposition of S, i.e.,

$$S = \Phi \Lambda \Phi^T, \quad \Lambda = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$$

Let's sort λ_i's as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.
Because $S^T = S$, and $S = \frac{1}{n}\tilde{X}\tilde{X}^T$, we can show that $\lambda_i \geq 0$, $1 \leq i \leq d$.
$\Phi = [\phi_1 \cdots \phi_d] \in \mathbb{R}^{d \times d}$
is a matrix containing the eigenvectors.
Also thanks to $S^T = S$, Φ is an orthogonal matrix whose columns form an ONB of $\mathbb{R}^d$.
The change of the bases from $[e_1 \cdots e_d]$ to $[\phi_1 \cdots \phi_d]$ is achieved simply by $\Phi^T \tilde{X}$.

$\phi_j^T \tilde{X}$ is called the jth principal components of X.

PCA was known for a long time, e.g., since the time of Pearson (1901) and Hotelling (1933).
Those days, the measurement dimension d was much smaller than the number of samples n, i.e. $d \ll n$
This is called the "classical" setting.
Ex. 5 exam scores of 2000 students
$d = 5$, $n = 2000$.
Due to the advent of computers and sensor technology, now we often have $d \gg n$, the "neo-classical" setting.
Ex. The face database: $d = 128^2$, $n = 143$.