SVD and Least Squares Problems

* LS via SVD

Recall the LS solution via QR factorization:

\[
\begin{align*}
(1) & \text{ Compute reduced QR of } A. \\
(2) & \text{ Compute } y = \hat{Q}^T lb. \\
(3) & \text{ Solve } \hat{R} x = y - (\star) \\
\end{align*}
\]

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

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

\[
\begin{align*}
A^T A x &= A^T lb \\
\iff (\hat{U} \hat{\Sigma} \hat{V}^T)^T (\hat{U} \hat{\Sigma} \hat{V}^T) x &= (\hat{U} \hat{\Sigma} \hat{V}^T)^T lb \\
\iff \hat{V} \hat{\Sigma}^T \hat{U}^T (\hat{U} \hat{\Sigma} \hat{V}^T) x &= \hat{V} \hat{\Sigma}^T \hat{U}^T lb \\
\iff \hat{V} \hat{\Sigma}^T \hat{V}^T x &= \hat{V} \hat{\Sigma}^T \hat{U}^T lb \\
\iff \hat{\Sigma}^T \hat{V}^T x &= \hat{U}^T lb \\
\end{align*}
\]

This can be solved easily:

\[
\begin{align*}
(1) & \text{ Compute reduced SVD of } A. \\
(2) & \text{ Compute } y = \hat{U}^T lb. \\
(3) & \text{ Solve } \hat{\Sigma} w = y - (\star\star) \\
(4) & \text{ Set } x = \hat{V} w. \\
\end{align*}
\]

Note: (\( \star\star \)) is a diagonal system, easier to solve than (\( \star \))!!
Pseudo-inverse and SVD

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

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

$m = n$ : $A^+ = A^{-1}$

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

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

$A = U \Sigma V^T$,  \[ \Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_r & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} r \\ \vdots \\ \vdots \\ m-r \end{pmatrix} \begin{pmatrix} r \\ \vdots \\ \vdots \\ m-r \end{pmatrix} \]

Define

$A^+ := V \Sigma^+ U^T$,  \[ \Sigma^+ := \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_r} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} r \\ \vdots \\ \vdots \\ n-r \end{pmatrix} \begin{pmatrix} r \\ \vdots \\ \vdots \\ n-r \end{pmatrix} \]

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

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

(iii) $(A X)^T = A X$;  
(iv) $(X A)^T = X A$

Such $X$ is uniquely determined and $X = A^+$ !!
Pseudoinverse & Orthogonal Projectors

Thm. \( AA^+ \) is an ortho. proj. onto range (A)
and \( AA^+ = U_r U_r^T \)

\( A^+ A \) is an ortho. proj. onto range \( (A^T) \)
and \( A^+ A = V_r V_r^T \)

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 \( P_A := AA^+ \), \( P_A^T := A^+ A \).

Now, \( P_A = U \Sigma V^T V \Sigma^+ U^T \)

\[ = U \Sigma \Sigma^+ U^T = U \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} U^T \]
\[ = U_r U_r^T \checkmark \]

\( P_A^2 = U_r U_r^T U_r U_r^T = U_r U_r^T = P_A \checkmark \)

\( = I_r \) so it's a proj. !

\( P_A^T = (U_r U_r^T)^T = (U_r^T)^T U_r^T = U_r U_r^T = P_A \checkmark \)

So it's an ortho. proj. !

Finally, it's also clear that

\( P_A \) maps onto range \( (A) \) since

\( \text{range}(A) = \langle u_1, \ldots, u_r \rangle \).

You can do similarly for \( P_A^T \)

Note: Consider any \( x \in \text{range}(A) \).

Then \( \exists y \in \mathbb{R}^n \) s.t. \( x = Ay \).

Now \( P_A x = AA^+ x = AA^+ A y \)

\[ = A y = x. \quad \text{"A via Moore-Penrose (i)"} \]
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) 

\[ \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \]

and each \( \mathbf{X}_j \in \mathbb{R}^d \). \( d \) could be huge (e.g., a face image database).

Let \( \mathbf{X} := [\mathbf{X}_1 \mathbf{X}_2 \cdots \mathbf{X}_n] \in \mathbb{R}^{d \times n} \)

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

\[ \overline{\mathbf{X}} := \frac{1}{n} \sum_{j=1}^{n} \mathbf{X}_j \]

And define the centered data matrix

\[ \widetilde{\mathbf{X}} := [\mathbf{X}_1 - \overline{\mathbf{X}} \mathbf{X}_2 - \overline{\mathbf{X}} \cdots \mathbf{X}_n - \overline{\mathbf{X}}] \]

Note: \( \widetilde{\mathbf{X}} = \mathbf{X} (\mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T) \)

Good exercise!

Now the sample covariance matrix \( \mathbf{S} \) is defined as

\[ \mathbf{S} := \frac{1}{n} \widetilde{\mathbf{X}} \widetilde{\mathbf{X}}^T \in \mathbb{R}^{d \times d} \]

\( \mathbf{S}_{ij} \) indicates the covariance or mutual correlation between the \( i \)th and \( j \)th entries of data vectors.

PCA is nothing but an eigenvalue decomposition of \( \mathbf{S} \), i.e.,

\[ \mathbf{S} = \mathbf{\Psi} \Lambda \mathbf{\Psi}^T, \quad \Lambda = \text{diag} (\lambda_1, \cdots, \lambda_g) \]
Let’s sort $\lambda_i$‘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 $S^T = S$, $\Phi$ 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 $\tilde{\Phi}^T \tilde{X}$.

$\tilde{\Phi}_j^T \tilde{X}$ is called the $j$th 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: $d = 128^2$, $n = 143$. 