1 Principle Component Analysis

Principle Component Analysis (PCA) is short for PCA. It is one of the most important dimension reduction approaches. Let us have a set of (zero centered) points in the high dimension space

\[ \{x_n\}_{n=1}^N \]

where all \( x_n \)'s are in \( \mathbb{R}^D \) (\( D \) is usually very large). A very important task is to visualize all points in the low dimensional space. So the idea of PCA is to apply a linear mapping \( L \in \mathbb{R}^{d \times D} \) to map any \( x_n \) to \( y_n \) in a low dimensional space \( y_n = Lx_n \in \mathbb{R}^d \). Apparently, it does not make sense if choose an arbitrary linear mapping \( L \). So we need a criteria to decide how to choose \( L \). PCA basically chooses to minimize the reconstruction error to decide the optimal mapping \( L \in \mathbb{R}^{d \times D} \)

\[
\min_{L \in \mathbb{R}^{d \times D}} \sum_{n=1}^N \| x_m - \underbrace{L^T Lx_n}_{\text{reconstruction}} \|^2.
\]

You can imagine \( y_n = Lx_n \in \mathbb{R}^d \) is to reduce the dimension and \( L^\top Lx_n = L^\top y_n \in \mathbb{R}^D \) is to reconstruct the \( x_n \) from \( y_n \). If we define the data matrix

\[ X = [x_1, x_2, \ldots, x_N] \]

then the objective can be cast into an equivalent from

\[
\min_{L \in \mathbb{R}^{d \times D}} \| X - L^\top LX \|_F^2,
\]

where \( \| \cdot \|_F \) is the Frobenius norm. This problem is not convex, but it has a closed form solution. Let the (compact) singular value decomposition (SVD) of \( X \) be

\[ X = U \Sigma V^\top \]

where \( U = [u_1, u_2, \ldots, u_r] \in \mathbb{R}^{D \times r} \) and \( V = [v_1, v_2, \ldots, v_r] \in \mathbb{R}^{d \times r} \) have orthogonal columns, that is, \( U^\top U = V^\top V = I \), and \( \Sigma \in \mathbb{R}^{r \times r} \) is a positive diagonal matrix

\[
\Sigma = \begin{bmatrix}
\sigma_1 & 0 & \cdots & 0 \\
0 & \sigma_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \sigma_r
\end{bmatrix}
\]

\(^1\)It means that the average of all points is 0, that is, \( \frac{1}{N} \sum x_n = 0 \). In practice, if they are not centered you should shift all points along the same direction such that the mean is 0, that is, \( x_n = x_n - \frac{1}{N} \sum_{n=1}^N x_n \).
with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Note that $r$ is the rank of matrix $X$. You can use existing functions to compute the (compact) SVD from Matlab, R, Python, or any other programming languages.

Now we are ready to define the optimal (or closed form of) $L$ for PCA:

$$L = [u_1, u_2, \cdots, u_d]^\top.$$

Note that $d$ should be no more than $r$. So the optimal low dimension representation $y_n \in \mathbb{R}^d$ can be obtained accordingly

$$[y_1, \cdots, y_N] = LX = \Sigma_{1:d,1:d} [v_1, v_2, \cdots, v_d]^\top = [\sigma_1 v_1, \sigma_2 v_2, \cdots, \sigma_d v_d]^\top \in \mathbb{R}^{d \times N},$$

where $\Sigma_{1:d,1:d}$ denotes the sub matrix of $\Sigma$ with the first $d$ rows and $d$ columns.

So the typical procedure of applying PCA is to

- make the center of all data points $x_n$’s to be 0 by shifting all $x_n$ along the same direction;
- apply (1) to obtain the representation $y_n$ in low dimension space.