Lecture 10: Principal Component Analysis

Recall our overarching idea of decomposing data/signals/images etc. as linear combinations of "possibly" simpler pieces.

$$f(t) = \sum_k c_k \psi_k(t)$$

The principal Component Analysis (PCA) is yet another approach to this, but aims to find the c_k the coeffs. & ψ_k at the same time.

At the same time PCA can be viewed as a dimensionality reduction (think compression) technique as well. Allowing us to represent high dimensional data using a few coefficients (low dimensional representation).

10.1 Theory of PCA & its Connection to SVD

Let us consider the following setting:

We are given a dataset

$$X = \{\underline{x}_0, \cdots, \underline{x}_{N-1}\}, \quad \underline{x}_j \in \mathbb{R}^d$$

*don't confuse notation with signal processing. each $\underline{x}_j$ here is an instance of data (an image, audio, etc).

Naturally we can further overload our notation & write

data set → $X = \left[\left| \right| \cdots \left| \right| \right] \in \mathbb{R}^{d \times N}$ (columns $x_0, \ldots, x_N$)

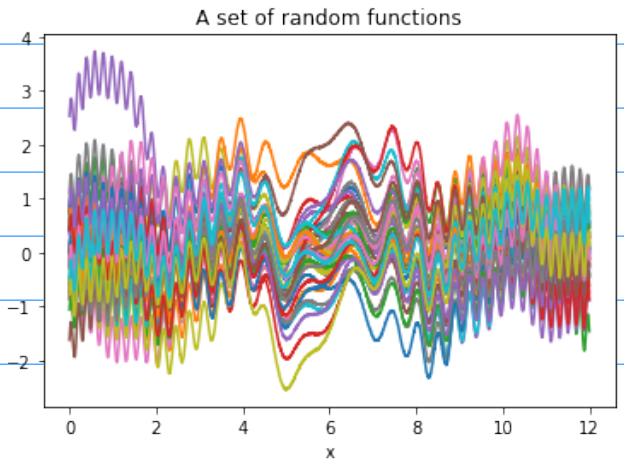
Then here is a one sentence description of PCA

> The left Singular vectors of X are the Principal components of the data x_j and the optimal (in some sense) basis in which the x_j can be represented.

*So, from the linear algebra view point PCA is just an application of SVD. (Of course there is a lot more to it as we will see soon).

Let's look at a quick demo:

Here we are given a dataset consisting of 40 randomly generated functions.



The question now is what is a good basis for representing these functions?

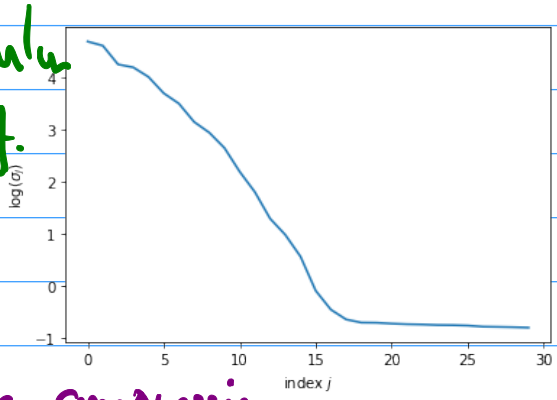
Alternatively, can we find a low dimensional/compressed representation?

$$X = \begin{bmatrix} f_0^{(1)} & f_0^{(2)} & \cdots & f_0^{(N-1)} \\ \vdots & \vdots & & \\ f_d^{(1)} & f_d^{(2)} & & f_d^{(N-1)} \end{bmatrix}$$

$\tilde{X}$ ← Center data by subtracting the mean.

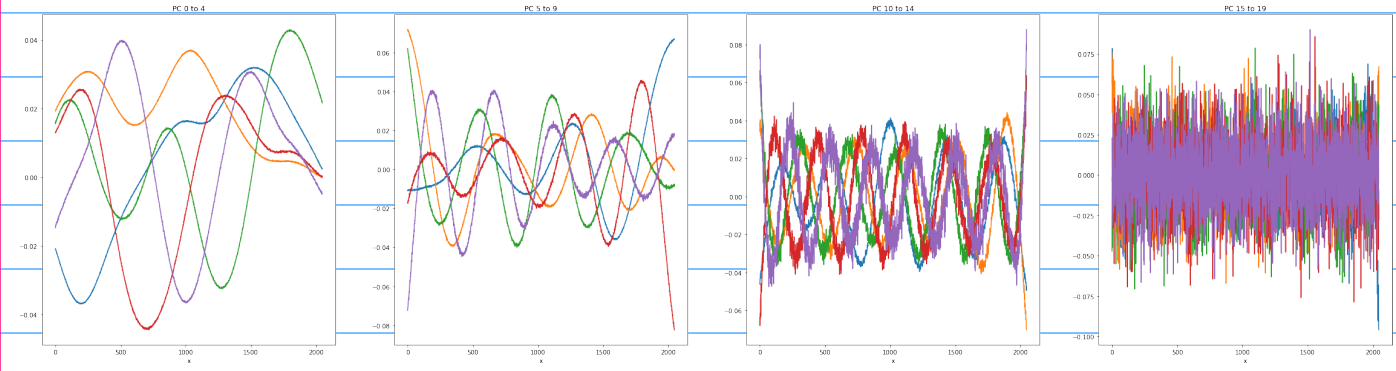
First, lets plot the singular values of $\tilde{X}$:

observe that the first 15-20 singular values are clearly dominant.



This tells us that $\tilde{X}$ is close to a rank(15-20) matrix! in otherwords. There is a lot of room for compression.

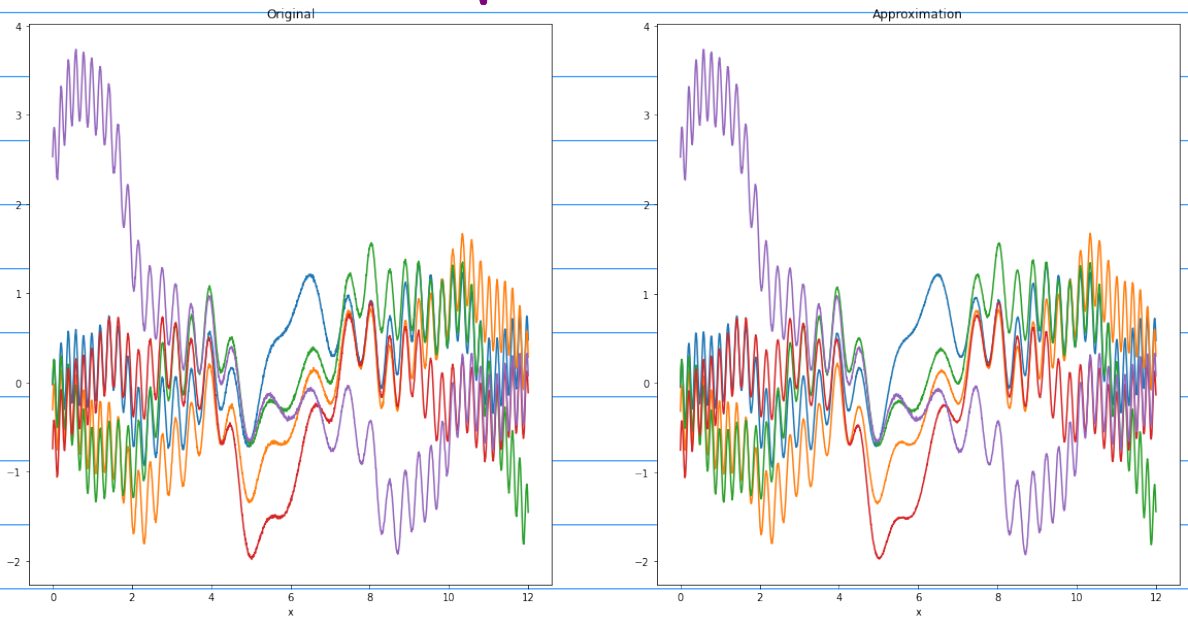
Next we look at the left singular vecs. of $\tilde{X}$ to see if they tell us anything about the variations in the data & a basis to represent it.



low freq. Mid freq high freq.

Obsan similarities to trigonometric functions, ie, Fourier bases!

We can now approximate $\tilde{X}$, & therefore X simply by omitting the singular vectors corresponding to the smaller singular values. This is nothing but a low-rank approx. to X.



So, lets take a closer look at what is going on.

Recall that we defined the covariance matrix of an $\mathbb{R}^d$ valued centered random variable $\underline{x}$ as

$$\mathrm{Cov}(X) = \mathbb{E}\, \underline{X}\underline{X}^T$$

$$\underline{x}_j \text{ iid} \longrightarrow \approx \frac{1}{N} \sum_{j=0}^{N-1} \underline{x}_j \underline{x}_j^T = \tilde{C} \quad \text{where } \tilde{C}_{ik} = \frac{1}{N} \sum_{j=0}^{N-1} x_{ij} x_{kj}$$

$$\approx \frac{1}{N-1} \sum_{j=0}^{N-1} \underline{x}_j \underline{x}_j^T \quad \leftarrow \text{unbiased estimator}$$

if we arrange the $\underline{x}_j$ in a matrix $X = \left[\underline{x}_0 \,\middle|\, \cdots \,\middle|\, \underline{x}_{N-1} \right]$ then

$$\frac{1}{N-1} X X^T = \tilde{C}$$

so $\frac{1}{N-1} XX^T$ is simply the empirical approx. to Cov(X) which henceforth we denote as C_X.

C_X is non-negative definite & symmetric (NDS)
Hence it has an eigen decomposition

$$C_X = Q \Lambda Q^T$$

the eigenvectors of C_X are called the Principal Components or the Karhunen-Loève modes or PCA modes of C_X.
At the same time write $X = U \Sigma V^T$ then,

$$C_X = \frac{1}{N-1} XX^T = \frac{1}{N-1} U \Sigma V^T V \Sigma^T U^T$$
$$= \frac{1}{N-1} U \Sigma^2 U^T$$

Thus, columns of U, the left singular vectors of X are precisely the Principal Components of C_X.

But why are the PCA modes useful? are they optimal in some sense? Yes!

A random variable $\underline{x}: \Omega \to \mathbb{R}^d$ is called Gaussian if its PDF has the form,

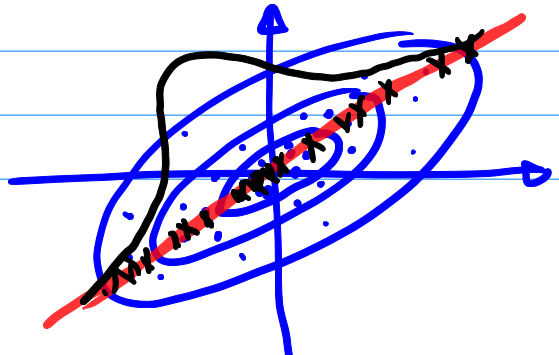
$$P(\underline{x}) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{m})^T C^{-1}(\underline{x}-\underline{m})\right)$$

where $\underline{m} \in \mathbb{R}^d$ is the mean & $C \in \mathbb{R}^{d \times d}$ is the covariance matrix (strictly pos. def.)

Gaussian random variables are completely identified by their mean $\underline{m}$ & Covariance matrix C. Hence, the notation $\underline{x} \sim \mathcal{N}(\underline{m}, C)$. The particular case $\mathcal{N}(0, Id)$ is called the standard normal dist.

Now suppose our dataset is centered (mean zero). Then we can ask, what is the closest Gaussian distribution that resembles our data? Since a mean-0 Gaussian is completely identified by its covariance matrix we simply take $\mathcal{N}(0, C_X)$.

We can push this idea further by asking, what is the closest Gaussian that approx. X whose Covariance matrix has rank at most $r < N$?



Then it is natural for us to formulate the following optimization problem

$$\underset{\substack{M \in \mathbb{R}^{d\times d} \\ \text{rank}(M)\leq r}}{\text{minimize}} \; \|M - C_X\|_F^2$$

Then the low-rank approx to C_X obtained via SVD truncation is precisely the solution to this problem!

In summary, the low rank/compressed approx. obtained via PCA/SVD is based in approximation of our data via a degenerate (low rank) Gaussian distribution.

But what is the significance of the PCA modes?

Lemma Suppose $\underline{x} \sim \mathcal{N}(\underline{m}, C)$, $\underline{m} \in \mathbb{R}^d$, $C \in \mathbb{R}^{d\times d}$. Let $\underline{b} \in \mathbb{R}^n$ & $A \in \mathbb{R}^{n\times d}$ then $\underline{z} = A\underline{x} + \underline{b}$ is also Gaussian & $\underline{z} \sim \mathcal{N}(b + A\underline{m}, ACA^T)$.

Now suppose we have a centered Gaussian $\underline{X} \sim N(0, C)$ with $C = U\Sigma^2 U^T$. By previous lemma we have

in distribution

$$\underline{X} \overset{d}{=} U\Sigma\underline{\xi} \quad \text{where} \quad \underline{\xi} \sim N(0, I).$$

$$\overset{d}{=} \sum_{j=0}^{d-1} \xi_j \sigma_j \underline{u}_j, \quad \xi_j \overset{iid}{\sim} N(0,1)$$

Thus, any centered Gaussian r.v. can be written as a linear combination of iid standard normals. This is called the Karhunen-Loève theorem.

Furthermore, larger σ_j represent larger variance in the direction of the vector $\underline{u}_j$! Hence, the PCA modes corresponding to the larger σ_j are indeed more significant as they represent the directions of maximum variance in the data!