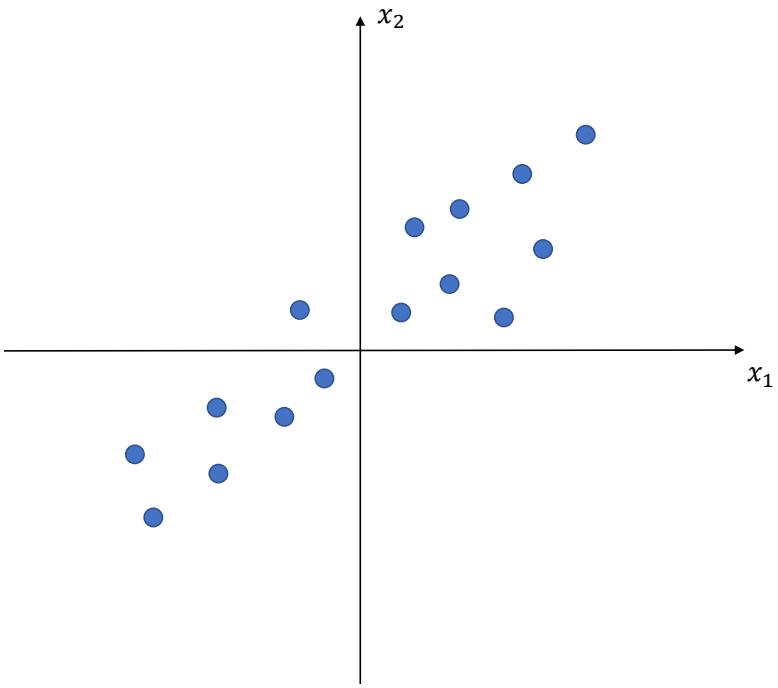
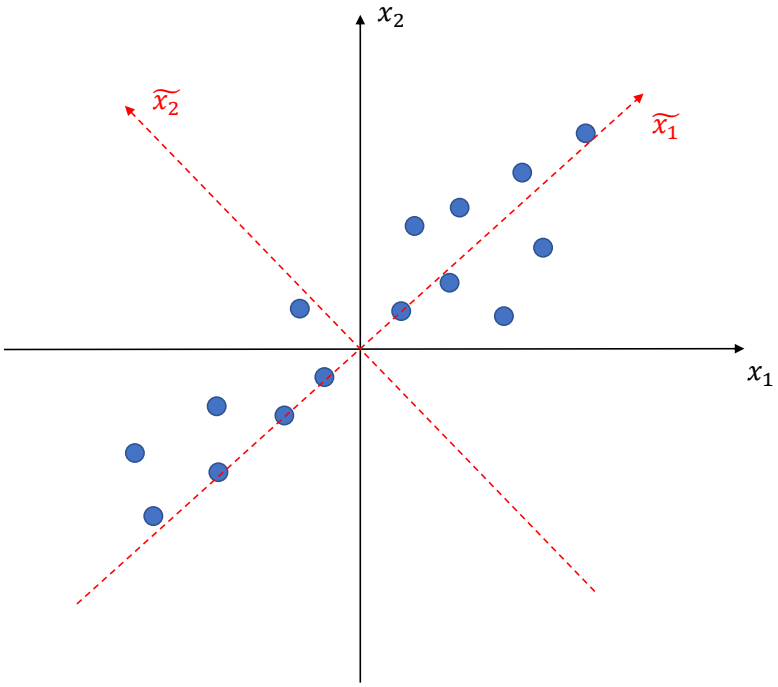
Lecture 2

1 Principal Component Analysis - basics

When we are given a large data set that is multi-dimensional (dimensionality is $d > 1$), it may be the case that it is in fact embedded in a low-dimensional linear subspace. As a very simple example, let $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \ldots, \mathbf{x}^N\}$ be a two-dimensional data set of $N \in \mathbb{N}$ points $\mathbf{x}^i \in \mathbb{R}^2$, $i = 1, 2, \ldots, N$. Each point has $d = 2$ coordinates, $\mathbf{x}^i = [x_1^i, x_2^i]^T$, $x_1^i, x_2^i \in \mathbb{R}$. One can plot all the points in a two-dimensional Cartesian graph. You may obtain the following figure:



We see that the data has some sort of "structure". But what exactly do we mean by saying that? The data is really 1-dimensional (modulo some "noise"). This would become more apparent if we expressed the same data in a new set of co-ordinate axes $\widetilde{x_1}$ and $\widetilde{x_2}$, obtained by rotating our original axes about the origin:



Most of the data structure is captured by axis $\widetilde{x_1}$. Projections of our points onto the second axis $\widetilde{x_2}$ represent "noise". In this sense we can say that our data set is one-dimensional, aligned along $\widetilde{x_1}$ with a certain amount of noise aligned along $\widetilde{x_2}$. In other words, the axis $\widetilde{x_1}$ is special - it preserves most of the variability in $\mathcal{D}$. So **when projecting our data onto a lower dimensional subspace, we need to do it "intelligently" - i.e. by picking a subspace that contains most of the variability (and structure) of the original data.**

To talk about variability of the data in a quantitative manner, we will borrow notions from statistics and probability theory. We can do so because we will assume that our data points were generated by some (unknown) probability distribution ("the nature"). In other words, the data co-ordinates are realisations of some vector random variable. But let us start simple by concentrating on a single coordinate.

2 Mean and Variance

Let us recap ourselves on some fundamental concepts from statistics. Suppose that we have a random variable X and we would like to measure its "variability". We have realisations of the random variable X - the data: $\mathcal{D}_X = \{x^1, x^2, \ldots, x^N\} \subset \mathbb{R}$. We can ask:

1) *What is the 'center of gravity' of X?* and

2) *How much does X fluctuate around this center of gravity?*

Figure 1: Examples of a) low variance and b) high variance.

For question 1), we would like to know what is the average of the data. This is given by the following definition:

Definition 2.1. Let X be a random variable with probability distribution P and event set A. The *expected value* or *mean* of X, denoted as $\mathbb{E}[X]$, is defined by

$$\mathbb{E}[X] = \sum_{x \in A} P(X = x) \cdot x. \tag{1}$$

Equation (1) gives the theoretical quantity of the mean of X, because we assume we know its distribution P. Also, this definition works for discrete random variables. For the continuous case, one would need probability density of X and integral over support of X.

In practice we only can estimate what $\mathbb{E}[X]$ might be using our data. Our estimation of the expected value is given by

$$\widehat{\mathbb{E}[X]} = \frac{1}{N} \sum_{i=1}^{N} x^i. \tag{2}$$

Now we look at question 2), asking ourselves if there is a way to quantify deviations away from the mean. To answer this requires some relatively simple intuition: consider the *square fluctuation* of a value x of X from the mean: $(x - \mathbb{E}[X])^2$. Let $Y = (X - \mathbb{E}[X])^2$ be a random variable representing the square fluctuation of X away from its mean. Our task is then to calculate the expected value of Y.

Definition 2.2. For a random variable X, let A be its event set. Furthermore, let Y be the random variable for the square fluctuations about $\mathbb{E}[X]$. The *variance* of X, denoted as $Var[X]$, is defined by

$$Var[X] = \mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in A} P(X = x) \cdot (x - \mathbb{E}[X])^2. \tag{3}$$

As in the case for the mean, we can estimate the variance using the data set:

$$\widehat{Var[X]} = \frac{1}{N} \sum_{i=1}^{N} (x^i - \widehat{\mathbb{E}[X]})^2 \tag{4}$$

Remark 2.1. Strictly speaking, for our estimator $\widehat{Var[X]}$, we should divide the sum by $N - 1$ instead of N (to obtain an unbiased estimate). However, in practice, as $N >> 1$, we don't need to worry about this.