Module 4

Other perspectives of PCA (optional)

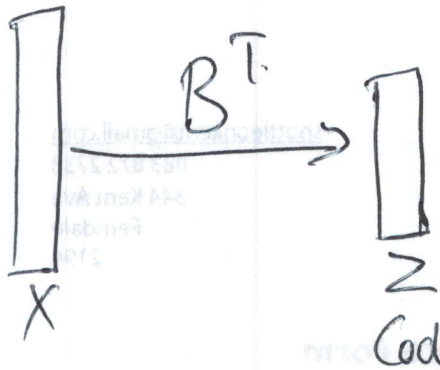
We derived PCA from the perspective of minimizing the average squared reconstruction error.

However PCA can also be interpreted from different perspectives.

In the session we'll have a brief look at some of these interpretations

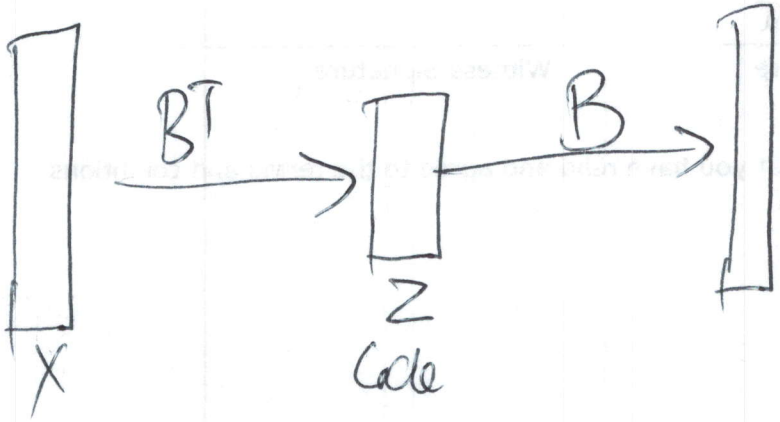
Lets start with a recap what we have done so far:

We took a high dim vector x and we projected it into a low dim representation, z, using the matrix B^T, the columns of this matrix B are the eigenvectors of the data covariance matrix that are associated with the largest eigenvalues

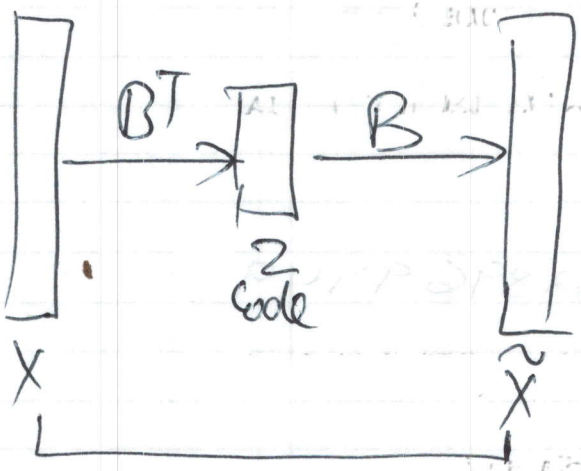


The Z values are the coordinates of our datapoint wrt the basis vectors which span the principal subspace, and that is also called the code of our datapoint.

Once we have that lower Dimensional representation Z, we can get a higher Dim of it, by using the matrix B again, "that is multiplying B onto Z to get a higher Dimensional version of the Z of original data space.

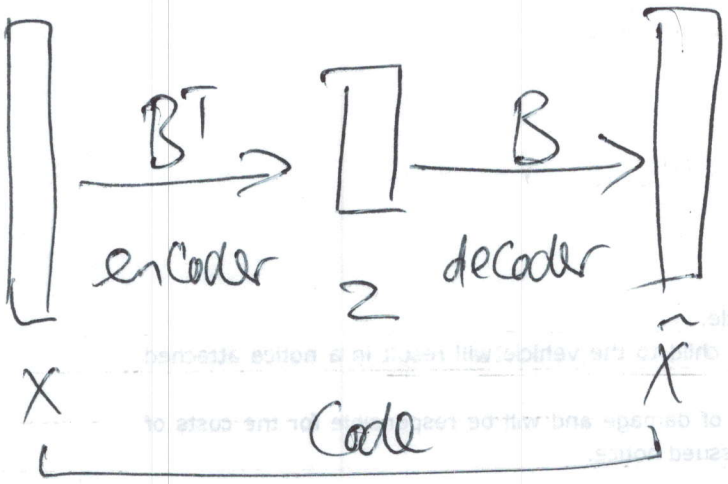


We learned the PCA parameters, such that, "the reconstruction error between X and the reconstruction $\hat{X}$ is minimized"



We can also think of PCA as a linear auto encoder. An auto encoder encodes a data point x, and tries to decode it to something similar to the same data point.

The mapping from the data to the code is called the <u>encoder</u>, and mapping from the code to the original data space is called the <u>decoder</u>



If the encoder and decoder are linear mappings, then we get the PCA solution, when we minimize the squared autoencoding loss.

If we replace the linear mapping of PCA, with a non-linear mapping, we get a non-linear autoencoder.

A prominent example of this is a deep autoencoder with a linear functions of the encoder and decoder are replaced with deep neural networks

Another interpretation of PCA is related to information theory.

We can think of the code as a smaller compressed version of the original data point

When we reconstruct our original data, using the code, we don't get the exact data point back, but a slightly distorted or noisy version of it.

This means that our compression is lossy.

Intuitively, we want to maximise the correlation between the original data and the lower dimensional code

More formally this would be related to the mutual information

We would then get the same solution to PCA we discussed earlier in this course, by maximising the mutual information,

a core concept in information theory.

We we derive PCA, using projections, we reformulated the average reconstruction error loss, as minimizing the variance of the data that is projected onto the orthogonal complement of the principal subspace.

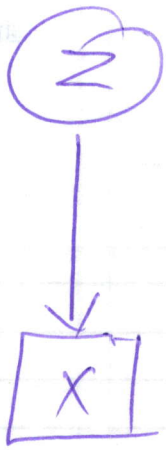
Minimizing that variance is equivalent to maximizing the variance of the data when projected onto the principal subspace.

If we think of variance in the data as information contained in the data, this means that PCA can also be interpreted as a method that retains as much information as possible.

We can also look at PCA from the perspective of a latent variable model.

"We assume that an unknown lower dim code z generates data x, and we assume that we have a linear relationship

between Z and X"



Generally we can then, "write that X B times Z plus $\mu(m)$, and maybe some noise(ε) and assume that the noise is isotropic, with mean 0 and covariance matrix (sigma squared) times $\underline{I}$)"

$$X = BZ + \mu + \varepsilon,$$

$$\varepsilon \sim N(0, \sigma^2 \underline{I})$$

"We also further assume that the distribution of this (Z) is a standardard normal, so PDF Z is gaussian with mean 0, and covariance the I matrix"

$$p(z) = N(0, I)$$

we can now write down the likelyhood of the model, so the likelyhood is p of x given z and that is the Gaussian distribution in x, with mean $B_z + \mu$ and covariance matrix, sigma squared I.

$$p(x|z) = N(x|B_z + \mu, \sigma^2 I)$$

We can also compute the "marginal likelihood as p of x as the integral of p of x given z, times the distribution on z, dz"

$$p(x) = \int p(x|z)\, p(z)\, dz$$

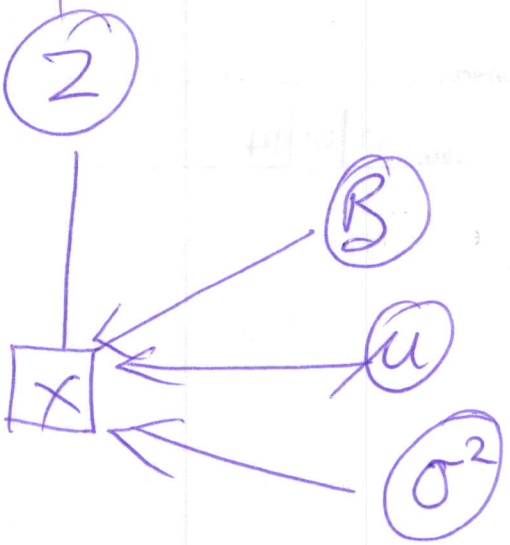
"and that turns out to be" gaussian distribution of X with mean μ and covariance matrix B times B Transpose plus sigma squared $\underline{I}$"

$$= \mathcal{N}(x | \mu, BB^T + \sigma^2 \underline{I})$$

The parameters of this model are:

- μ
- B
- σ^2

and we can write them explicitly down in our model up here



we can now determine the parameters of the model using maximum likelihood estimation.
we will find that μ is the mean of the data and B is a matrix that contains the eigenvectors that correspond to the largest eigenvalues.

To get the lower Dim code of datapoint, we can apply Bayes theorem to invert the linear relationship between z and x

In particular we going to get, "p of z, given x" as p of x given z, that is the likelihood (p of z), times p of z divided by marginal likelihood p of x

$$p(z|x) = \frac{p(x|z)\,p(z)}{p(x)}$$

In this session, we looked at 5 different perspectives of PCA, that lead to different objectives.

1. Minimizing the Squared reconstruction error
2. Minimizing the auto encoder loss
3. Maximizing the mutual information
4. Maximizing the variance of the projected data.
5. Maximizing the likelihood in latent variable model

all the different perspectives give us the SAME solution to the PCA problem

The strengths and weaknesses of individual perspective became more clear and when we consider properties of real data