Principal Component Analysis

The idea of PCA is to rotate the data (choose a new coordinate system) so that we end up with dimensions with low variance, which we can throw away without losing much information.

- PCA is looking for projections with maximum variance
- Or looking for projections which minimise the distance from the original points to their projections
- These are equivalent

A projection from D dimensions to 1 is defined by a D dimensional vector u1 so the projection of x is:

$$u_1^T x = \text{ projection of x} u_1^T u_1 = 1 \text{ unit vector}$$

Eigenvector Projections:

$$\text{Covariance Matrix: } S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T \qquad \bar{x} = \text{mean} \text{The variance of the projected data} = u_1^T S u_1 \text{We}$$

The Eigenvector with the largest eigenvalue is the first principal component. The second principal component is the direction that maximises the projected variance subject to being orthognal to the first principal component. Each subsequent principal component is chosen to maximise variance subject to being orthognal to all previous components. This prinipal components will always be the eigenvectors of the covariance matrix ordered by eigen values.

$$x_n = \sum_{i=1}^{D}(x_n^T u_i)u_i = \sum_{i=1}^{D}\alpha_{ni}u_i$$

So each datapoint is a linear combination of the principal components. We typically want to keep M < D dimensions. When approximating a D-Dimensional datapoint the best approximation accounts for the mean by adding a constant vector:

$$\bar{x} - \sum_{i=1}^{M}(\bar{x}^T u_i)u_i \text{Projected Data: } \tilde{x}_n = \sum_{i=1}^{M}(x_n^T u_i)u_i + \bar{x} - \sum_{i=1}^{M}(\bar{x}^T u_i)u_i$$

Probabailistic PCA:

Reformulate PCA as a maximum likelihood solution to a latent variable model. The latent variable is this case is continuous.

The latent variable z has zero-mean unit-covariance gaussian distribution:

$p(z) = \mathcal{N}(z|0, 1)$The distribution of the observed data conditional on this latent variable is: $p(x|z) = \mathcal{N}(x|Wz$

W is a DxM matrix where D is the dimension of the data and M is the dimension of the PCA space where M <= D. We can then use MLE to maximise the likelihood of X given the parameters above. These will be eigenvectors and eigenvalues of the sample covariance matrix. We may still resort to **Expectation Maximisation** if the sample covariance matrix is huge or if we have missing data.