1 Principal Component Analysis

Given a collection of measured data points in a high dimensional space, it is often desirable to reduce the dimensionality of such data to a lower dimensional latent space, this could be to aid to processing the data or enable visualisation. In such instances, it is desirable for the mapping to the latent space to maintain as much information from the original data as possible.

Principal Component Analysis (PCA) aims to reduce the dimensionality of a collection of data points while maximising the variance in the latent space. Alternatively, PCA aims to find a mapping from the original data to a new latent space which allow for the original data to be reconstructed with minimal error. These two aims are in fact equivalent.

1.1 Title?

Let $\boldsymbol{x_i} \in \mathbb{R}^f$ represent a data point in f dimensional space and $\boldsymbol{y_i} \in \mathbb{R}^d$ represent the point which $\boldsymbol{x_i}$ is transformed to by a linear transformation $\boldsymbol{W}$ equation (1), where $d << f$.

$$\boldsymbol{y_i} = \boldsymbol{W}^\top \boldsymbol{x_i} \tag{1}$$

where,

$$\boldsymbol{y_i} = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{id} \end{bmatrix}, \quad \boldsymbol{x_i} = \begin{bmatrix} \boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{x}_{if} \end{bmatrix}, \quad \boldsymbol{W} = [\boldsymbol{w_1}, \ldots, \boldsymbol{w_d}], \quad \boldsymbol{w_i} = \begin{bmatrix} w_{i1} \\ \vdots \\ w_{if} \end{bmatrix}$$

The optimal transformation $\boldsymbol{W}$ will maximise the variance in the data, where variance is expressed by equation (2). This then follows that the optimal solution is found by maximising equation (4). To prevent the trivial solution where $\boldsymbol{w}_k = \infty$, constrain a fixed magnitude $\boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}$.

$$\sigma_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_{ik} - \mu_k)^2 \tag{2}$$

where,

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} y_{ik}$$

$$\begin{aligned} \boldsymbol{W} &= \arg\max_{\boldsymbol{W}} \frac{1}{N} \sum_{k=1}^{d} \sum_{i=1}^{N} (y_{ik} - \mu_k)^2 \\ &= \arg\max_{\boldsymbol{W}} \frac{1}{N} \sum_{k=1}^{d} \sum_{i=1}^{N} \boldsymbol{w}_k^\top (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{w}_k \\ &= \arg\max_{\boldsymbol{W}} \sum_{k=1}^{d} \boldsymbol{w}_k^\top \boldsymbol{S}_t \boldsymbol{w}_k \end{aligned}$$

where $\boldsymbol{S}_t$ is the covariance matrix,

$$\boldsymbol{S}_t = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \tag{3}$$

and,

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

$$\boldsymbol{W} = \arg\max_{\boldsymbol{W}} \operatorname{tr}[\boldsymbol{W}^\top \boldsymbol{S}_t \boldsymbol{W}], \quad \text{subject to} \quad \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I} \tag{4}$$

By constructing the Lagangian from equation (4) and solving the solution in equation (5) can be obtained. From eigendecomposition, $\boldsymbol{W}$ has columns of columns of eigenvectors which correspond to the d largest non-zero eigenvalues of the covariance matrix, $\boldsymbol{S}_t$.

$$L(\boldsymbol{W}, \boldsymbol{\Lambda}) = \operatorname{tr}[\boldsymbol{W}^\top \boldsymbol{S}_t \boldsymbol{W}] - \operatorname{tr}[\Lambda(\boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{I})]$$

$$\frac{\partial \operatorname{tr}[\boldsymbol{W}^\top \boldsymbol{S}_t \boldsymbol{W}]}{\partial \boldsymbol{W}} = 2\boldsymbol{S}_t \boldsymbol{W}, \quad \frac{\partial \operatorname{tr}[\Lambda(\boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{I})]}{\partial \boldsymbol{W}} = 2\boldsymbol{W}\boldsymbol{\Lambda}$$

$$L(\boldsymbol{W}, \boldsymbol{\Lambda}) = 0$$

$$\boldsymbol{S}_t \boldsymbol{W} = \boldsymbol{W}\boldsymbol{\Lambda} \tag{5}$$

References