PCA : Method for Dimensionality Reduction

Task:

Given $\mathbf{x}_i \in R^d$, $i = 1, \ldots, N$, find one-dimensional represenation of the data. Specifically,

1) Find a line in R^d that "best represents" the data.
2) Assign each data point to a point along that line.

To determine $\mathbf{b}$, $\mathbf{v}$ and $\{\alpha_i\}_{i=1}^N$, solve the optimization problem:

$$\begin{cases} \min\limits_{\{\mathbf{b},\mathbf{v},\alpha_i\}} \dfrac{1}{N}\sum\limits_{i=1}^{N} \|\mathbf{x}_i - \underbrace{(\mathbf{b}+\alpha_i\mathbf{v})}_{\hat{x}_i}\|^2 = R \quad (\text{error}) \\ \text{s.t.} \quad \|\mathbf{v}\|^2 = 1 \end{cases}$$

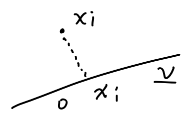
Step 1 : fix $\underline{v}$, with length 1 ($\|v\| = \underline{v}\cdot\underline{v} = 1$). find $\underline{b}$ and $\{a_i\}_{i=1}^N$

$$\min R = \min_{\underline{b},\{a_i\}} \frac{1}{N}\sum_{i=1}^{N} \|\underline{x}_i - \hat{\underline{x}}_i\|_2^2 = \min_{\underline{b},\{a_i\}} \frac{1}{N}\sum_{i=1}^{N} \|\underline{x}_i - (b + a_i\underline{v})\|_2^2$$

(note : $\|\underline{x}\|_2^2 = (\underline{x}, \underline{x}) = \underline{x}^T\underline{x} = \underline{x}\cdot\underline{x}$)

$$\begin{cases} 0 = \dfrac{\partial R}{\partial a_i} = -\dfrac{2}{N}(\underline{x}_i - (\underline{b} + a_i\underline{v}))\cdot\underline{v} \quad \Rightarrow \quad a_i = (\underline{x}_i - \underline{b})\cdot\underline{v} \\ 0 = \dfrac{\partial R}{\partial \underline{b}} = \nabla_b R = -\dfrac{2}{N}\sum\limits_{i=1}^{N}(\underline{x}_i - (\underline{b} + a_i\underline{v})) = -\dfrac{2}{N}\left(\sum\limits_{i=1}^{N}\underline{x}_i - N\underline{b} + \underline{v}\sum\limits_{i=1}^{N} a_i\right) \end{cases}$$

(since $\sum_{i=1}^{N} a_i = (\Sigma x_i - N\cdot b)\cdot v = (0 - Nb)\cdot\underline{v} = -Nb\,\underline{v}$)



$$= -\frac{2}{N}(-N\underline{b} - N(\underline{b}\cdot\underline{v})\,\underline{v}) = 0$$

if $b = 0$, satisfied. $\Rightarrow$ go through the origin. $a_i = \underline{x}_i\cdot\underline{v} = \underline{v}^T\cdot\underline{x}_i$

So $\hat{\underline{x}}_i = \underline{b} + a_i\underline{v} = (\underline{v}^T\underline{x}_i)\,\underline{v}$ $\leftarrow$ orthogonal projection of $\underline{x}_i$ onto $\underline{v}$.

Step 2 : Find $\underline{v}$

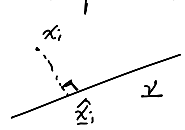
$$R = \frac{1}{N}\sum_{i=1}^{N} \|x_i - (b + a_i\underline{v})\|_2^2 = \frac{1}{N}\sum_{i=1}^{N} \|x_i - (\underline{x}_i\cdot\underline{v})\,\underline{v}\|_2^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i^T\underline{x}_i - 2\,\underline{v}^T\underline{x}_i\underline{x}_i^T\underline{v} + \underline{v}^T\underline{x}_i\underline{x}_i^T\underline{v}) = \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i^T\underline{x}_i - \underline{v}^T\underline{x}_i\underline{x}_i^T\underline{v})$$

$$\Leftrightarrow \begin{cases} \max \ \dfrac{1}{N}\sum\limits_{i=1}^{N}\underline{v}^T\underline{x}_i\underline{x}_i^T\underline{v} = v^T\left(\dfrac{1}{N}\sum\limits_{i=1}^{N}x_ix_i^T\right)\underline{v} = v^T S v \\ \text{s.t.} \ \|\underline{v}\|_2^2 = 1 \end{cases}$$

where S : sample covariance matrix of $\{x_i\}$. symmetric. semi positive definite.

$x_i \in R^d$, $x_i^Tx_i \in R$, $x_ix_i^T$: $d\times d$ matrix



$\Leftrightarrow$ objective function : v^TSv , $\underline{x}_i \rightarrow \hat{\underline{x}}_i = a_i\underline{v} = \underline{v}^T\underline{x}_i\,\underline{v}$, $\{\underline{x}_i\} \rightarrow \{a_i\}$

Sample variance of a_i is

$$\mathrm{var}\{a_i\} = \frac{1}{N}\sum_{i=1}^{N}\left(\{a_i\} - a_0\right)^2 = \frac{1}{N}\sum_{i=1}^{N} a_i^2 = \frac{1}{N}\sum_{i=1}^{N}(\underline{v}^T\underline{x_i})(\underline{x_i}^T\underline{v}) = v^T\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T\right)\underline{v} = \underline{v}^T S\, \underline{v}$$

where a_0 is sample mean of $\{a_i\}$, and $a_0 = \frac{1}{N}\sum a_i = \frac{1}{N}\sum \underline{v}^T \underline{x_i} = \frac{1}{N} v^T\left(\underbrace{\sum_{i=1}^{N} x_i}_{0}\right) = 0$

and $\underline{v}$ = the direction of maximum variation.

$$\Leftrightarrow \begin{cases} \max\limits_{\underline{v}} \ \underline{v}^T S\, \underline{v} \\ \text{s.t.} \ \|v\|^2 = \underline{v}^T\underline{v} = 1 \end{cases}$$

$\Rightarrow \ L(\underline{v}, \lambda) = \underline{v}^T S\, \underline{v} + \lambda(1 - \underline{v}^T\underline{v})$, λ Lagrange multiplier, $\lambda \in R$

$$\begin{cases} 0 = \frac{\partial L}{\partial \lambda} = 1 - \underline{v}^T\underline{v} \\ 0 = \frac{\partial L}{\partial \underline{v}} = \nabla_v L = 2S\underline{v} + \lambda(-2\underline{v}) \end{cases} \Rightarrow S\underline{v} = \lambda\underline{v} \quad \text{(Eigenvalue Problem for } S \text{)}$$

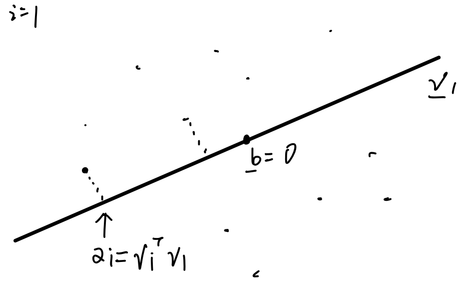
where λ: eigenvalue of S , $\underline{v}$: eigenvector.

$$\begin{matrix} \lambda_1 \geq & \lambda_2 \geq & \cdots \geq & \lambda_d \\ v_1 & v_2 & & v_d \end{matrix}$$

$\Rightarrow \ \max \ \underline{v}^T S\underline{v} = \underline{v}^T(\lambda\underline{v}) = \lambda\, \underline{v}^T\underline{v} = \lambda$, optimal direction is given $\underline{v}_1$

$$\begin{cases} \underline{v} : \text{first principal component axis of } \{x_i\} \\ a_i = \underline{v}_1^T\underline{x_i} : \text{first principal component score of } x_i \end{cases}$$

e.g. $\{\underline{x}_i\}_{i=1}^{N}$



$\mathrm{var}\{a_i\} = \lambda_1$

Reduction to higher dimension

$a^1 v_1 + a^2 v_2 + \cdots + a^p v_p$, where v_p : eigenvector of S corresponding to the p-th largest eigenvalue, p-th PC.

Projection of x_i: $\hat{x}_i = (v_1^T x_i)v_1 + (v_2^T x_i)v_2 + \cdots + (v_p^T x_i)v_p$

To summarize, using PCA, we can

- map the data to p-dimensional subspace, where $p < d$.
  The reduced representation is given by

$$Z_p = U_p X$$

- reconstruct the data from the reduced representation:

$$\hat{X} = U_p Z_p$$

$$\hat{x}_i = (v_1^T x_i) v_1 + (v_2^T x_i) v_2 + \cdots + (v_p^T x_i) v_p = (\underline{v_1}, \underline{v_2}, \cdots, v_p)_{d\times p} \begin{pmatrix} \underline{v_1^T x_i} \\ \vdots \\ \underline{v_p^T x_i} \end{pmatrix}_{p\times 1}, \quad \underline{v_i} \in R^d$$

$$X = (\underline{x_1}, \underline{x_2}, \cdots, \underline{x_n}) \quad d\times N$$

$$S = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T = \frac{1}{N} XX^T \quad d\times d$$

$$U_p = (\underline{v_1}, \underline{v_2}, \cdots, \underline{v_p})_{d\times p}$$

$$Z_p = \begin{bmatrix} \underline{v_1^T x_1} & \cdots & \underline{v_1^T x_N} \\ \vdots & & \vdots \\ \underline{v_p^T x_1} & & \underline{v_p^T x_N} \end{bmatrix}_{p\times N} = U_p^T \cdot X$$

$$\hat{X} = (\hat{x}_1, \hat{x}_2, \cdots \hat{x}_N)_{d\times N} = U_p Z_p$$

SVD of $\underline{X}$ $d\times N$

$(U^T U = UU^T = I, \; V^T V = VV^T = I)$

$X = U \Sigma V^T$, where $U: R^{d\times d}$, $V: R^{N\times N}$, U, V are orthogonal matrix

$$\Sigma = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & 0 \\ 0 & & \sigma_d & \end{pmatrix}_{d\times N}, \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d \geq 0, \text{ signular value of } X$$

$$XX^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma V^T V \Sigma^T U^T = U \Sigma\Sigma^T U^T$$

$$S = \frac{1}{N} XX^T = U\left(\frac{1}{N}\Sigma\Sigma^T\right)U^T = U \Lambda U^T$$

$$\Lambda = \frac{1}{N}\Sigma\Sigma^T = \begin{pmatrix} \sigma_1^2/N & & 0 \\ & \sigma_2^2/N & \\ 0 & & \ddots & \\ & & & \sigma_d^2/N \end{pmatrix}_{d\times d}$$

$$SU = U\Lambda U^T U = U\Lambda \Rightarrow S\underline{u_1} = \frac{\sigma_1^2}{N}\underline{u_1}, \; S\underline{u_2} = \frac{\sigma_2^2}{N}\underline{u_2} \cdots$$

variation of the data along j-th pc

$$\frac{1}{N}\sum_{i=1}^{N}(\underbrace{\underline{v_j}^T \underline{x_i}}_{\text{principle score along } x_i})^2 = \frac{1}{N}\sum_{i=1}^{N} \underline{v_j}^T (\underline{x_i}\,\underline{x_i})^T \underline{v_j} = \underline{v_j}^T \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i^T\right)\underline{v_j} = \underline{v_j}^T S \underline{v_j} = \underline{v_j}^T (\lambda_j \underline{v_j}) = \lambda_j$$

In practice, we use the singular value decomposition (SVD) $X = U\Sigma V^T$ to compute the principal components.

Using the SVD of X, we obtain the eigenvalue decomposition of S:

$$S = \frac{1}{N}XX^T = U\Lambda U^T$$

where $\Lambda = \frac{1}{N}\Sigma\Sigma^T$.

- The j-th largest eigenvlaue of S, λ_j, tells how much variation in the data is captured by the j-th principal component.
- The proportion of the variance captured by the first p principal components is

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_p}{\lambda_1 + \cdots + \lambda_p + \lambda_{p+1} + \cdots + \lambda_d}$$

- Using p principal components, the projection error is given by

$$R = \sum_{j=p+1}^{d} \lambda_j$$