⑤ [PCA]

Data: $\{x_n\}_{n=1}^N$, $x_i \in \mathbb{R}^D$

$u_i u_j = \delta_{ij}$, $u_i \in \mathbb{R}^D$

$$x_n = \sum_{i=1}^{D} (x_n^T u_i) u_i \xrightarrow{\text{approx}} \tilde{x}_n = \sum_{i=1}^{M} z_{ni} u_i + \sum_{i=M+1}^{D} b_i u_i$$

(b_i: const)

Goal is to find such $u_1, \dots, u_D$, b and other parameters, that distortion is minimized.

$$\frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|^2 \longrightarrow \min$$

1) [Let's assume u_is are fixed]

$$F = \frac{1}{N} \sum_{n=1}^{N} \| \sum_{i=1}^{M} (x_n^T u_i - z_{ni}) u_i + \sum_{i=M+1}^{D} (x_n^T u_i - b_i) u_i \|^2 =$$

$$= \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{i=1}^{M} (x_n^T u_i - z_{ni})^2 + \sum_{i=M+1}^{D} (x_n^T u_i - b_i)^2 \right)$$

$$\frac{\partial F}{\partial z_{ni}} = \frac{1}{N} 2 (x_n^T u_i - z_{ni}) = 0 \Rightarrow \boxed{z_{ni} = x_n^T u_i}$$

$$\frac{\partial F}{\partial b_i} = \frac{1}{N} \sum_{n=1}^{N} 2 (x_n^T u_i - b_i) = 0 \Rightarrow \boxed{b_i = \bar{x}^T u_i}$$

2) $$F = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} \left((x_n - \bar{x})^T u_i \right)^2 =$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} u_i^T (x_n - \bar{x})(x_n - \bar{x})^T u_i =$$

$$= \sum_{i=M+1}^{D} u_i^T S u_i , \quad S = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$$

Optimization problem is

$$\begin{cases} \min \sum_{i=M+1}^{D} u_i^T S u_i \\ \text{s.t.} \quad u_i u_j = \delta_{ij} \quad , \; i,j \in \{M+1, \dots, D\} \end{cases}$$

Lagrangian

$$\downarrow \mathcal{L} = \sum_{i=M+1}^{D} \left[u_i^T S u_i + \lambda_i (1 - u_i^T u_i) \right] - \sum_{i=M+1}^{D} \sum_{\substack{j=M+1 \\ j \neq i}}^{D} \lambda_{ij} u_i^T u_j$$

$$\frac{\partial \mathcal{L}}{\partial u_i} = 2 (S u_i - \lambda_i u_i) - \sum_{\substack{j=M+1 \\ j \neq i}}^{D} \lambda_{ij} u_j = 0 \qquad \Big| \; u_i^T \leftarrow$$

$$\underline{u_i^T S u_i = \lambda_i} \qquad \forall i$$

$$\Big| \; u_j^T \leftarrow$$

$$\underline{2 \, u_j^T S u_i = \lambda_{ij}} \qquad \forall i \neq j$$

$$\forall i \; \hookrightarrow \; S u_i = \sum_{j=M+1}^{D} (u_j^T S u_i) \, u_j \quad \in \underbrace{\text{Lin}\{u_{M+1}, \dots, u_D\}}_{\text{invariant subspace of } S}$$

What means that eigenvectors of S can be divided into two sets $v_1, \dots, v_M$ and $v_{M+1}, \dots, v_D$, such that $\mathcal{L}in\{v_{M+1}, \dots, v_D\} = \mathcal{L}in\{u_{M+1}, \dots, u_D\}$

$$\sum_{i=M+1}^{D} u_i^T S u_i \stackrel{?}{=} \sum_{i=M+1}^{D} v_i^T S v_i$$

$$\sum_{i=M+1}^{D} u_i^T S u_i = \sum_{i=M+1}^{D} \left(\sum_j \alpha_{ij} v_j^T\right) S \left(\sum_j \alpha_{ij} v_j\right) =$$

$$\stackrel{v_i v_j = \delta_{ij}}{=\!=} \sum_i \sum_j \alpha_{ij}^2 \lambda_j = \sum_j \left(\sum_i \alpha_{ij}^2\right) \lambda_j \quad (=)$$

// $\alpha \in \mathbb{R}^{(D-M)\times(D-M)} = \{\alpha_{ij}\}_{i,j=M+1}^{D}$; since $u_i u_j = \delta_{ij}$, we get

$\alpha^T \alpha = I$; but since α is a square matrix, then $\alpha \alpha^T = I$, what implies

$$\sum_i \alpha_{ij}^2 = 1 \qquad //$$

$$(=) \sum_j \lambda_j \Rightarrow$$

$$\min \sum_{i=M+1}^{D} u_i S u_i = \sum_{i=M+1}^{D} \lambda_j, \text{ if}$$

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_D.$$

It means that if we choose $u_1, \dots, u_M$ to be eigenvectors corresponding to M largest eigenvalues, then distortion is minimized. Though, it is sufficient to choose $u_1, \dots, u_M$ to be any basis in $\text{Lin}\{v_1, \dots, v_M\}$.

(a) Again, we've got data points $\{x_n\}_{n=1}^{N}$, where $x_n \in \mathbb{R}^D$, and we wish to project them onto a space of lower dimension, such that a variance of data is maximized.

Let $[u_1, \dots, u_M]$ be a basis in this space. Then $U \in \mathbb{R}^{D \times M}$ $\quad [U^T U = I]$

$$y_i = U^T x_i = \sum_{j=1}^{M} (u_j^T x_i) u_j$$

$$\bar{y} = U^T \bar{x}$$

$$\Downarrow$$

$$\text{Variance} = \frac{1}{N}\sum_{n=1}^{N} \|y_n - \bar{y}\|^2 = \qquad (*)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{M} \left(u_j^T (x_n - \bar{x})\right)^2 \to \min$$

$$\text{s.t.} \quad u_i^T u_j = \delta_{ij}$$

Lagrangian:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{M} \left(u_j^T (x_n - \bar{x})\right)^2 + \sum_{j=1}^{M} \lambda_j (1 - u_j^T u_j)$$

$$- \sum_{k \neq e} \lambda_{ke} u_k^T u_e$$

$$\frac{\partial \mathcal{L}}{\partial u_j} = \frac{1}{N}\sum_{n=1}^{N} 2(x_n - \bar{x})(x_n - \bar{x})^T u_j -$$

$$- 2\lambda_j u_j - \sum_{k \neq j} \lambda_{kj} u_k =$$

$$2 S u_j - 2\lambda_j u_j - \sum_{k \neq j} \lambda_{kj} u_k = 0.$$

$$\Downarrow$$

$$u_j^T S u_j = \lambda_j \qquad \forall j,k$$
$$2 u_j^T S u_k = \lambda_{kj}$$

$$\Downarrow$$

$$S u_j = \sum_{\ell=1}^{M} (u_j^T S u_\ell)\, u_\ell \in \underbrace{\text{Lin}\{u_1, \ldots, u_M\}}_{\text{invariant subspace of } X}$$

Again, there exists division of eigenvectors of matrix S into two sets, such that

$$\text{Lin}\{u_1, \ldots, u_M\} = \text{Lin}\{v_1, \ldots, v_M\}.$$

[For now, $v_1, \ldots, v_M$ are not required to correspond to M largest eigenvalues]

But since $\text{Lin}\{u_1, \ldots, u_M\} = \text{Lin}\{v_1, \ldots, v_M\}$

$$\forall n \quad \sum_{j=1}^{M} \left((x_n - \bar{x})^T u_j\right)^2 = \sum_{j=1}^{M} \left((x_n - \bar{x})^T v_j\right)^2$$

(follows from the meaning: both are squared norm of $x_n - \bar{x}$ projection onto the same linear subspace)

$\Rightarrow$

$$\text{Variance} = \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{M}\left(u_j^T(x_n-\bar{x})\right)^2 =$$

$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{M}\left(v_j^T(x_n-\bar{x})\right)^2 =$$

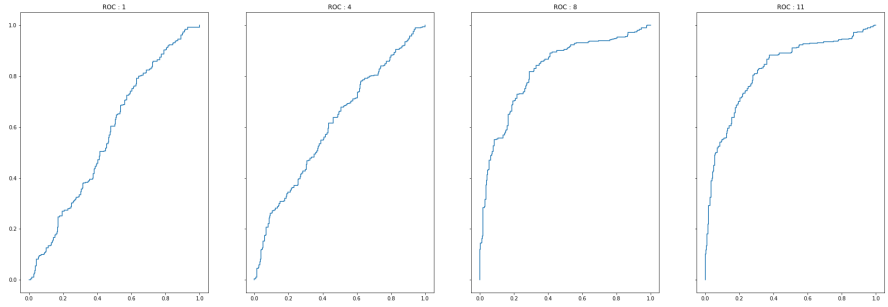
$$= \frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{M} v_j^T(x_n-\bar{x})(x_n-\bar{x})^T v_j =$$

$$= \sum_{j=1}^{M} v_j^T\left(\frac{1}{N}\sum_{n=1}^{N}(x_n-\bar{x})(x_n-\bar{x})^T\right)v_j =$$

$$= \sum_{j=1}^{M} v_j^T S v_j = \sum_{j=1}^{M}\lambda_j \longrightarrow \max$$

Variance is maximized if $v_1,\dots,v_m$, indeed, correspond to m largest eigenvalues.

(c) PCA was launched on wine dataset. Below is ROC curve for 4 different numbers of principal components. [1, 4, 8, 11]



AUC-ROC was computed for all eleven possible values (from 1 till 11)

K	1	2	3	4	5	6	7	8	9	10	11
AUC-ROC	0.5728	0.5987	0.5938	0.6117	0.7815	0.8034	0.8227	0.8223	0.8221	0.8235	0.8234

On this dataset SVM with linear kernel fitted projected train set shows best performance when number of PC is 10.

(2) [SVD]

a) (I) Let us suppose that feature matrix X is already centered, i.e. sum of all rows is zero.

1) PCs of X then are eigenvectors of matrix

$$\frac{1}{N}\sum_{n=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T \overset{\bar{x}=0}{=} \frac{1}{N}\sum_{n=1}^{N} x_i x_i^T =$$

$$= \frac{1}{N} X^T X$$

2) If $X = U\Sigma V^T$ is a singular value decomposition of X, then it is well-known, that V consists of eigenvectors of $X^T X$. Indeed,

$X^TX = V(\Sigma^T\Sigma)V^T$, which is a spectral decomposition of symmetric matrix X^TX. What means columns of V are principal components of X.

(II) 1) PCA projects each data point to a subspace, so that „non-similarity" of data (variance) is maximized.

2) SVD finds important concepts in data, i-th column of V is a „description" of the concept, i-th singular value shows importance of the concept in data, i-th column of U shows „relevance" of concept to each data point.

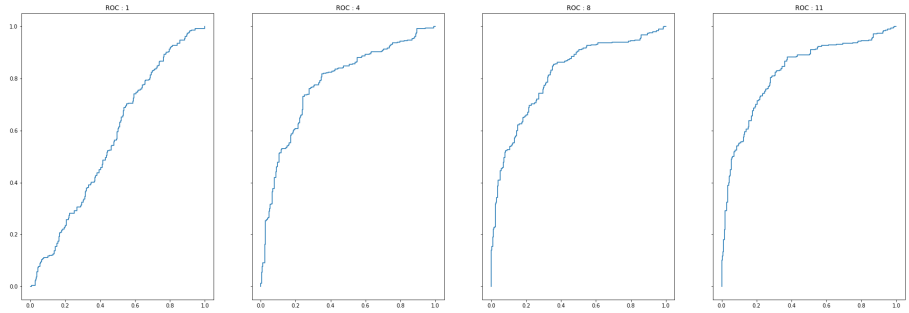
$$X = U\Sigma V^T = \sum_i \sigma_i u_i v_i^T$$

(b) The same experiment was performed using SVD approximation of data.

number of singular values	1	2	3	4	5	6	7	8	9	10	11
AUC_ROC	0.5678	0.6359	0.7324	0.7809	0.7819	0.801	0.8153	0.8143	0.8219	0.8235	0.8234

Below are ROC Curves for 4 different

numbers of singular values. [1, 4, 8, 11]



(SVM was launched on linear kernel with default values of parameters).

As can be seen from the table above, approximating data with 10-ranked matrix shows better performance than using initial feature space.

NB: any basis mentioned in hw is supposed implicitly to be orthonormal;