Lecture 17. Principal Component Analysis

Lecturer: Jie Wang Date: Dec 8, 2021

1 Preliminary

1.1 Singular Value Decomposition

Definition 1. A set of vectors $\{\mathbf{v}_i\}_{i=1}^n$ in $\mathbf{R}^d$ are called orthonormal if

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \begin{cases} 1, \text{ if } i = j, \\ 0, \text{ otherwise.} \end{cases}$$

A matrix $M \in \mathbb{R}^{d\times d}$ is orthogonal if

$$M^\top M = I,$$

where $I \in \mathbb{R}^{d\times d}$ is the identity matrix.

Theorem 1. *Given a matrix* $A \in \mathbb{R}^{m\times n}$*. Suppose that* $\operatorname{rank}(A) = r$*. Then, there exists* n *right singular vectors* $\mathbf{v}_1, \ldots, \mathbf{v}_n$ *that are orthonormal in* $\mathbb{R}^n$*, and* m *left singular vectors* $\mathbf{u}_1, \ldots, \mathbf{u}_m$ *that are orthonormal in* $\mathbb{R}^m$*, such that*

$$A\mathbf{v}_i = \sigma_i \mathbf{u}_i,\ i = 1, \ldots, r, \tag{1}$$

$$A\mathbf{v}_i = 0,\ i = r+1, \ldots, n, \tag{2}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ *are the* r *positive singular values.*

Remark *1.*

1. The last $n - r$ right singular vectors $\mathbf{v}_i$, $i = r+1, \ldots, n$, span the null space of A. The last $m - r$ left singular vectors $\mathbf{u}_i$, $i = r+1, \ldots, m$, span the null space of $A^\top$.

2. Let $V = (\mathbf{v}_1, \ldots, \mathbf{v}_r, \ldots, \mathbf{v}_n)$, $U = (\mathbf{u}_1, \ldots, \mathbf{u}_r, \ldots, \mathbf{u}_m)$, and

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \cdots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma_r & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0. \end{pmatrix}.$$

   We can write Eq. (1) as

$$AV = U\Sigma.$$

3. The singular value decomposition of A is

$$A = U\Sigma V^\top.$$

4. The singular value decomposition of A can be written as a sum of r rank 1 matrix:

$$A = U\Sigma V^\top = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^\top + \ldots + \sigma_r \mathbf{u}_r \mathbf{v}_r^\top.$$

5. Let $V_r = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r)$, $U_r = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)$, and

$$\Sigma_r = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{pmatrix}.$$

The reduced form of the SVD of A is

$$A = U_r \Sigma_r V_r^\top.$$

1.2 Random Vectors

A random vector X takes the form of

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}.$$

The mean of X is

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_d \end{pmatrix} = \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_d) \end{pmatrix}. \tag{3}$$

The covariance matrix Σ, also written as $\mathbb{V}(X)$, is

$$\Sigma = \begin{pmatrix} \mathbb{V}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_d) \\ \mathrm{Cov}(X_2, X_1) & \mathbb{V}(X_2) & \cdots & \mathrm{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_d, X_1) & \mathrm{Cov}(X_d, X_2) & \cdots & \mathbb{V}(X_d) \end{pmatrix}.$$

Suppose that we randomly sample n data instances:

$$\mathbf{x}_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,d} \end{pmatrix}, i = 1, \ldots, n. \tag{4}$$

The sample mean is

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_d \end{pmatrix} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i,$$

which implies that

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{i,j}, \; j = 1, \ldots, d.$$

The sample variance matrix $S \in \mathbb{R}^{d\times d}$ is

$$S = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,d} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d,1} & s_{d,2} & \cdots & s_{d,d} \end{pmatrix},$$

where

$$s_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k).$$

By simple algebraic manipulation, we can see that

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top = \frac{1}{n-1}\widetilde{X}\widetilde{X}^\top, \tag{5}$$

where $\widetilde{X} \in \mathbb{R}^{d\times n}$ and its i^{th} column is $\mathbf{x}_i - \bar{\mathbf{x}}$.

2 Principal Component Analysis

The core idea of PCA is that, we would like to project the data instances into a subspace such that the set of projected data instances preserves as much information as possible.

2.1 The formulation

Suppose that we have a set of data instances $\mathbf{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. Let $\mathbf{g}_k \in \mathbb{R}^d$, $k = 1, \ldots, K$, with $K \leq d$, be a set of vectors such that

$$\langle \mathbf{g}_i, \mathbf{g}_j \rangle = \begin{cases} 1, & i \neq j; \\ 0, & \text{otherwise}, \end{cases}$$

and

$$G = (\mathbf{g}_1, \ldots, \mathbf{g}_K).$$

Then, the projection of the $\mathbf{x}_i$ into the subspace spanned by $\{\mathbf{g}_1, \ldots, \mathbf{g}_K\}$, that is, the column space of G, is

$$\mathbf{z}_i = P_G(\mathbf{x}_i) = GG^\top \mathbf{x}_i. \tag{6}$$

We use the **sample variance** to measure the information carried by the data instances. Thus, the information preserved by the projected data instances is

$$\frac{1}{n-1}\sum_{i=1}^{n}\|\mathbf{z}_i - \bar{\mathbf{z}}\|^2,$$

where

$$\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i. \tag{7}$$

By plugging Eq. (6) into Eq. (7), we have

$$\bar{\mathbf{z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i = \frac{1}{n}\sum_{i=1}^{n} GG^\top \mathbf{x}_i = GG^\top \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\right) = GG^\top \bar{\mathbf{x}},$$

where

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i.$$

Thus, the problem becomes

$$\begin{aligned} \max_{G\in\mathbb{R}^{d\times K}} \ & \frac{1}{n-1}\sum_{i=1}^{n}\|GG^\top \mathbf{x}_i - GG^\top \bar{\mathbf{x}}\|^2, \\ \text{s.t.}\ & G^\top G = I. \end{aligned} \tag{8}$$

Notice that

$$\begin{aligned} \frac{1}{n-1}\sum_{i=1}^{n}\|GG^\top \mathbf{x}_i - GG^\top \bar{\mathbf{x}}\|^2 =& \frac{1}{n-1}\sum_{i=1}^{n}\langle GG^\top(\mathbf{x}_i - \bar{\mathbf{x}}), GG^\top(\mathbf{x}_i - \bar{\mathbf{x}})\rangle \\ =& \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^\top GG^\top GG^\top(\mathbf{x}_i - \bar{\mathbf{x}}) \\ =& \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})^\top GG^\top(\mathbf{x}_i - \bar{\mathbf{x}}) \\ =& \frac{1}{n-1}\sum_{i=1}^{n}\operatorname{tr}\left((\mathbf{x}_i - \bar{\mathbf{x}})^\top GG^\top(\mathbf{x}_i - \bar{\mathbf{x}})\right) \\ =& \frac{1}{n-1}\sum_{i=1}^{n}\operatorname{tr}\left(G^\top(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top G\right) \\ =& \operatorname{tr}\left(G^\top\left(\frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top\right)G\right) \\ =& \operatorname{tr}\left(G^\top S G\right). \end{aligned}$$

Thus, the problem in (8) becomes

$$\begin{aligned} \max_{G\in\mathbb{R}^{d\times K}} \ & \operatorname{tr}(G^\top S G), \\ \text{s.t.}\ & G^\top G = I. \end{aligned} \tag{9}$$

Question 1. Consider the problem in (9).

1. Does the problem always admit a solution?
2. If the problem admit a solution, is it unique?

2.2 Solution to problem (9)

Recall from Eq. (5) that

$$S = \frac{1}{n-1}\widetilde{X}\widetilde{X}^\top.$$

Denote the SVD of $\widetilde{X}$ by

$$\widetilde{X} = U\Sigma V^\top.$$

Thus,

$$S = \frac{1}{n-1}U\Sigma^2 U^\top. \tag{10}$$

Plugging Eq. (10) into the problem in (9) leads to

$$\begin{aligned}\max_{G\in\mathbb{R}^{d\times K}} \ & \mathrm{tr}(G^\top U\Sigma^2 U^\top G), \\ \text{s.t.}\ & G^\top G = I.\end{aligned} \tag{11}$$

Denote

$$Q = U^\top G. \tag{12}$$

We can see that

$$Q^\top Q = I.$$

Thus, the problem in (11) reduces to

$$\begin{aligned}\max_{Q\in\mathbb{R}^{d\times K}} \ & \mathrm{tr}(Q^\top \Sigma^2 Q), \\ \text{s.t.}\ & Q^\top Q = I.\end{aligned} \tag{13}$$

We can see that

$$\mathrm{tr}(Q^\top\Sigma^2 Q) = \sum_{k=1}^{K}\sum_{j=1}^{d}\sigma_j^2 q_{j,k}^2 = \sum_{j=1}^{d}\sigma_j^2\left(\sum_{k=1}^{K} q_{j,k}^2\right).$$

Denote

$$\alpha_j = \sum_{k=1}^{K} q_{j,k}^2. \tag{14}$$

We can see that

$$\begin{aligned}&\alpha_j \in [0,1],\ j = 1,\ldots,d, \\ &\sum_{j=1}^{d}\alpha_j = K.\end{aligned}$$

Thus, we can further transform the problem (13) to

$$
\begin{aligned}
&\max_{\alpha\in\mathbb{R}^d} \sum_{j=1}^{d} \alpha_j\sigma_j^2, \\
&\text{s.t.}\, \alpha_j \in [0,1],\, j=1,\ldots,d, \\
&\qquad \sum_{j=1}^{d} \alpha_j = K.
\end{aligned} \tag{15}
$$

We can solve the above problem by the Lagrange multiplier method. However, we provide an alternative approach. Let

$$
f(\alpha)=\sum_{j=1}^{d} \alpha_j\sigma_j^2.
$$

Recall that we arrange the singular values in decending order, that is,

$$
\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d \geq 0.
$$

As $\sum_{j=1}^{d}\alpha_j = K$, we have

$$
\sum_{j=K+1}^{d} \alpha_j = K - \sum_{j=1}^{K} \alpha_j.
$$

Thus, for any α that is feasible with respect to problem (15)

$$
\begin{aligned}
f(\alpha) =& \sum_{j=1}^{K} \alpha_j\sigma_j^2 + \sum_{j=K+1}^{d} \alpha_j\sigma_j^2 \\
\leq& \sum_{j=1}^{K} \alpha_j\sigma_j^2 + \left(\sum_{j=K+1}^{d} \alpha_j\right)\sigma_{K+1}^2 \\
=& \sum_{j=1}^{K} \alpha_j\sigma_j^2 + \left(K - \sum_{j=1}^{K} \alpha_j\right)\sigma_{K+1}^2 \\
=& \sum_{j=1}^{K} \alpha_j\sigma_j^2 + \left(\sum_{j=1}^{K}(1-\alpha_j)\right)\sigma_{K+1}^2 \\
\leq& \sum_{j=1}^{K} \alpha_j\sigma_j^2 + \sum_{j=1}^{K}(1-\alpha_j)\sigma_j^2 \\
=& \sum_{j=1}^{K} \sigma_j^2 \\
=& f(\alpha^*),
\end{aligned}
$$

where $\alpha^* = (\alpha_1^*,\ldots,\alpha_d^*)$ with

$$
\alpha_j^* = \begin{cases} 1,\, j=1,\ldots,K, \\ 0,\, j=K+1,\ldots,d. \end{cases} \tag{16}
$$

Moreover, it is easy to see that α^* is feasible. Thus, the vector α^* is the optimal solution to problem (15).

We denote the optimal solution to problem (13) by

$$Q^* = (\mathbf{q}_1^*, \dots, \mathbf{q}_K^*).$$

In view of Eq. (14) and Eq. (16), we can see that the last $d-K$ entries of $\mathbf{q}_j$ are 0 for all $j = 1, \dots, K$, that is

$$Q^* = \begin{pmatrix} \widetilde{Q}^* \\ \mathbf{0} \end{pmatrix}_{d \times K},$$

where

$$\widetilde{Q}^* \in \mathbb{R}^{K \times K} \text{ and } (\widetilde{Q}^*)^\top \widetilde{Q}^* = I.$$

Thus, by Eq. (12), we have

$$G^* = UQ^* = (\mathbf{u}_1, \dots, \mathbf{u}_K)\widetilde{Q}^*. \tag{17}$$

That is, the optimal solution G^* to problem (9) is the matrix which shares the same column subspace spanned by the K left singular vectors of $\widetilde{X}$ corresponding to its first K largest singular values.

2.3 Principal components

Notice that, $\widetilde{Q}^*$ in Eq. (17) is an arbitrary $K \times K$ orthogonal matrix. Although G^* is a solution to problem (9) for any orthogonal matrix $\widetilde{Q}^*$, the column vectors are not necessarily the so-called *principal component vectors* of the sampled data $\{\mathbf{x}_i\}_{i=1}^n$.

The column vectors of G^* are the *principal component vectors* of the data $\{\mathbf{x}_i\}_{i=1}^n$ only if $\widetilde{Q}^* = I$, that is

$$G^* = (\mathbf{u}_1, \dots, \mathbf{u}_K),$$

and $\{\mathbf{u}_j\}_{j=1}^K$ are the first K Principal component vectors.

***Remark* 2.** Commonly seen approach to derive the principal component vectors is to first set $K = 1$ and solve the problem in (9). By the same approach in the last section, we can get the first principal component vector as $\mathbf{u}_1$. Then, we fix $\mathbf{u}_1$ and solve the problem in (9) by setting $K = 2$. We can get the second Principal component vector $\mathbf{u}_2$. Repeating this procedure, we can get the first K principal component vectors.

References