PCA under An Optimization Lens

Archer Yang

McGill University

September 24, 2024

PCA IN THE MATRIX FORMAT

Maximization of variance

- In PCA, we want to find the M leading principal components by solving

$$\max_{\boldsymbol{b}_j} \boldsymbol{b}_j^\top \boldsymbol{S} \boldsymbol{b}_j \qquad \text{for } j = 1, \cdots, M$$
$$\text{subject to } \boldsymbol{b}_j^\top \boldsymbol{b}_k = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k. \end{cases} \tag{1}$$

- Show that $\text{tr}(\boldsymbol{B}^\top \boldsymbol{S} \boldsymbol{B}) = \sum_{j=1}^{M} \boldsymbol{b}_j^\top \boldsymbol{S} \boldsymbol{b}_j$ if $\boldsymbol{B} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M) \in \mathbb{R}^{D \times M}$.

- We can rewrite Eq. (1) as

$$\max_{\boldsymbol{B}} \mathrm{tr}(\boldsymbol{B}^\top \boldsymbol{S}\boldsymbol{B})$$

$$\text{subject to } \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_{M\times M}$$

Minimization of reconstruction error

- Alternatively, in PCA, we want to find projection of the data onto an M dimensional subspace spanned by $\boldsymbol{B} = (\boldsymbol{b}_1, \cdots, \boldsymbol{b}_M)$ such that the reconstruction error is minimized

$$\min_{\boldsymbol{B}} \sum_{n=1}^{N} \|\boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_n\|^2 \quad \text{subject to } \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_{M\times M} \tag{2}$$

Minimization of reconstruction error

- We can show that for $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)^\top \in \mathbb{R}^{N \times D}$

$$\sum_{n=1}^{N} ||\boldsymbol{x}_n - \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{x}_n||^2 = \mathrm{tr}((\boldsymbol{X}^\top - \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X}^\top)^\top (\boldsymbol{X}^\top - \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X}^\top))$$

$$= ||\boldsymbol{X}^\top - \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X}^\top||_F^2$$

- Here $|| \cdot ||_F$ is the Frobenius matrix norm. The Frobenius norm of an $m \times n$ matrix $\boldsymbol{A} = (a_{ij})$, denoted $|| \cdot ||_F$, is defined by

$$||\boldsymbol{A}||_F \equiv \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2} = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$$

Minimization of reconstruction error

- Thus we can rewrite Eq. (2) as

$$\min_{\boldsymbol{B}} \frac{1}{N} \|\boldsymbol{X}^\top - \boldsymbol{B}\boldsymbol{B}^\top \boldsymbol{X}^\top\|_F^2 \equiv \frac{1}{N} \left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\right\|_F^2$$

$$\text{subject to } \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_{M\times M}$$

Minimization of reconstruction error

- We can also show that

$$\underbrace{\frac{1}{N}\|\boldsymbol{X}^\top - \boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}^\top\|_F^2}_{\text{reconstruction err.}} = \underbrace{\frac{1}{N}\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)}_{\text{total variance}} - \underbrace{\operatorname{tr}\left(\boldsymbol{B}^\top\boldsymbol{S}\boldsymbol{B}\right)}_{\text{projected variance}}$$

Proof

$$
\begin{aligned}
\frac{1}{N}\left\|\boldsymbol{X}^\top - \boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}^\top\right\|_F^2 &= \frac{1}{N}\left\|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\right\|_F^2 \\
&= \frac{1}{N}\operatorname{tr}\left(\left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\right)^\top\left(\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\right)\right) \\
&= \frac{1}{N}\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) - \frac{2}{N}\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\right) \\
&\quad + \frac{1}{N}\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\right) \\
&= \frac{1}{N}\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) - \frac{2}{N}\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\right) \\
&\quad + \frac{1}{N}\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\right) \\
&= \frac{1}{N}\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) - \frac{1}{N}\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}^\top\boldsymbol{X}^\top\boldsymbol{X}\right) \\
&= \frac{1}{N}\operatorname{tr}\left(\boldsymbol{X}^\top\boldsymbol{X}\right) - \operatorname{tr}\left(\boldsymbol{B}^\top\boldsymbol{S}\boldsymbol{B}\right)
\end{aligned}
$$