

PCA
A tutorial on Principle Component Analysis

Khanh Nguyen

August 2022, March 2023 Revisited

1 Problem Setup

In this section, we introduce the PCA problem

Given n data points, we define *mean* and *variance* as follows

Definition 1 (mean and variance of n data points). *Let $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$ where $x_i \in \mathbb{R}^d$, $x_i \in \mathbb{N} \cap [1, n]$. Define*

$$\mu(\mathcal{X}) = \frac{1}{n} \sum_{i=1}^n x_i$$
$$\sigma(\mathcal{X}) = \frac{1}{n-1} \sum_{i=1}^n \|x_i - \mu(\mathcal{X})\|_2^2$$

If data is centered, i.e $\mu(\mathcal{X}) = 0$, the variance can be rewritten as the sum of squared L2 norm of all data points, i.e $\sigma(\mathcal{X}) = \frac{1}{n-1} \sum_{i=1}^n \|x_i\|_2^2$. Throughout this tutorial, we assume data is centered.

We also denote $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$ as the data matrix of \mathcal{X} where each column of X corresponds to a data point in \mathcal{X} , we can rewrite *mean* and *variance* as follows

Definition 2 (mean and variance of a data matrix). *Let $X = [x_1, x_2, \dots, x_n] \in \mathbb{R}^{d \times n}$ be the data matrix of n data points in \mathbb{R}^d*

$$\mu(X) = \frac{1}{n} X \mathbf{1}_n = \mathbf{0}_d$$
$$\sigma(X) = \frac{1}{n-1} \|X\|_F^2 = \frac{1}{n-1} \text{tr } X^T X$$

The PCA problem attempts to find a k -dimensional subspace of \mathbb{R}^d denoted as \mathcal{U}_k such as the orthogonal projection of \mathcal{X} into \mathcal{U}_k preserves as much variance of \mathcal{X} as possible.

Let $U_k = [u_1, u_2, \dots, u_k] \in \mathbb{R}^{d \times k}$ be the matrix of a orthogonal basis of \mathcal{U}_k , i.e $U_k^T U_k = I_k$. The PCA projection of \mathcal{X} into \mathcal{U}_k can be written as $X \mapsto U_k U_k^T X$.

Definition 3 (Principle Component Analysis). Let $X \in \mathbb{R}^{d \times n}$

$$\text{pca}_k X = \max_{U_k \in \mathbb{R}^{d \times k} \wedge U_k^T U_k = I_k} \sigma(U_k U_k^T X) \quad (1)$$

Some preliminary observations

- If data is centered, the projected data is also centered. $(U_k U_k^T X) \mathbf{1}_n = U_k U_k^T (X \mathbf{1}_n) = \mathbf{0}_d$
- the maximum variance of data after the project is achievable if and only if all data points lie in the subspace \mathcal{U}_k

To elaborate on the second observation, let V_k be the complement subspace of U_k in \mathbb{R}^d , i.e every vector $x \in \mathbb{R}^d$ can be expressed as $x = u + v$ where $u \in U_k$ and $v \in V_k$. Furthermore, the L2 norm of x can be expressed as $\|x\|_2^2 = \|u\|_2^2 + \|v\|_2^2$ (this is well-know Pythagorean theorem). Sum up all data points, we have $\sum_{i=1}^n \|x_i\|_2^2 = \sum_{i=1}^n \|u_i\|_2^2 + \sum_{i=1}^n \|v_i\|_2^2$. Hence, $\sum_{i=1}^n \|x_i\|_2^2 \geq \sum_{i=1}^n \|u_i\|_2^2$. Since, the projected data is also centered, the RHS is the variance of projected data.

In machine learning, we often use the inner product of data points into the k *principle* directions as a dimensionality reduction method for downstream tasks. In this tutorial, we call it *PCA embedding*

Definition 4 (PCA Embedding).

$$X \mapsto \hat{U}_k^T X \in \mathbb{R}^{k \times n} \quad (2)$$

where \hat{U}_k is the optimal value of U_k .

2 Reduction to Trace Optimization Problem

In this section, we find PCA solution by reducing it to Trace Optimization Problem

Definition 5 (Trace Optimization Problem). Given $M \in \mathbb{R}^{d \times d}$ symmetric positive semidefinite with d eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$. Find a matrix $U \in \mathbb{R}^{d \times k}$ with $U^T U = I_k$ such as the trace $\text{tr } U^T M U$ is maximized or minimized

The optimal value of *Trace Optimization* is the sum of k largest / smallest eigenvalues.

$$\begin{aligned} \max_{U \in \mathbb{R}^{d \times k} \wedge U^T U = I_k} \text{tr } U^T M U &= \sum_{i=1}^k \lambda_i \\ \min_{U \in \mathbb{R}^{d \times k} \wedge U^T U = I_k} \text{tr } U^T M U &= \sum_{i=d-k+1}^d \lambda_i \end{aligned}$$

In the case of PCA, the objective can be rewritten as

$$\begin{aligned}
\sigma(U_k U_k^T X) &= \text{tr}(U_k U_k^T X)^T (U_k U_k^T X) \\
&= \text{tr} X^T U_k (U_k^T U_k) U_k^T X \quad (\text{decompose}) \\
&= \text{tr} X^T U_k U_k^T X \quad (\text{orthogonal of } U_k) \\
&= \text{tr} U_k^T (X X^T) U_k \quad (\text{cyclic property of trace})
\end{aligned} \tag{3}$$

Therefore, the solution of PCA can be obtained by solving *Trace Optimization* where $M = X X^T$. Let $X = U \Sigma V^T$ be the *Singular Value Decomposition* of X . We rewrite $X X^T = U \Sigma^2 U^T$. Hence, the solution of PCA is the subspace with basis consists of k left singular vectors corresponding to the k largest singular values.

3 Equivalent to Low-Rank Approximation on Frobenius Norm

3.1 Low-Rank Approximation on Frobenius Norm (LRA-FN)

Given matrix $A \in \mathbb{R}^{m \times n}$, the problem of Low-Rank Approximation on Frobenius Norm (LRA-FN) seeks to find a rank- k approximation of A ($k \leq \min(m, n)$). Formally,

$$\min_{A_k \in \mathbb{R}^{m \times n} \wedge \text{rank } A_k = k} \|A_k - A\|_F^2 \tag{4}$$

The *Eckart-Young-Mirsky theorem* states that the optimality is achievable when the rank- k matrix A is the rank- k SVD of A_k , i.e. $\hat{A}_k = U_k \Sigma_k V_k^T$. The optimal objective value is

$$\min_{A_k \in \mathbb{R}^{m \times n} \wedge \text{rank } A_k = k} \|A_k - A\|_F^2 = \|\hat{A}_k - A\|_F^2 = \sum_{i=k+1}^n \sigma_i^2 \tag{5}$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(m, n)} \geq 0$ are the singular values of A . We can rewrite the rank- k SVD decomposition of A as

$$U_k \Sigma_k V_k^T = U_k U_k^T A \tag{6}$$

Intuitively speaking, the best rank- k approximation of matrix A is achievable by orthogonally projecting its columns into the subspace constructed from k left singular vectors corresponding to k largest singular values. This operation is identical to PCA. In fact, LRA-FN and PCA are equivalent.

Theorem 1. *PCA and LRA-FN are equivalent*

3.2 PCA \rightarrow LRA-FN

In this section, we find the solution of PCA from the solution of LRA-FN

In PCA, we want to find U_k such as $\sigma(U_k U_k^T X)$ is maximized. We can rewrite the objective as

$$\sigma(U_k U_k^T X) = \|U_k U_k^T X\|_F^2 \quad (7)$$

Since $U_k U_k^T$ is a orthogonal projection, $U_k U_k^T x$ and $U_k U_k^T x - x$ are orthogonal for all $x \in \mathbb{R}^d$: $(U_k U_k^T x)^T (U_k U_k^T x - x) = 0$

Apply *Pythagorean theorem* for all columns of $U_k U_k^T X$

$$\begin{aligned} \|U_k U_k^T X\|_F^2 + \|U_k U_k^T X - X\|_F^2 &= \|X\|_F^2 \\ \|U_k U_k^T X\|_F^2 &= \|X\|_F^2 - \|U_k U_k^T X - X\|_F^2 \end{aligned} \quad (8)$$

By *LRA-FN*,

$$\|U_k U_k^T X - X\|_F^2 \geq \|\hat{U}_k \hat{U}_k^T X - X\|_F^2 \quad (9)$$

Where \hat{U}_k is the matrix of k left singular values of X corresponding to the k largest singular values.

Hence,

$$\|U_k U_k^T X\|_F^2 \leq \|X\|_F^2 - \|\hat{U}_k \hat{U}_k^T X - X\|_F^2 = \sum_{i=1}^k \sigma_i^2 \quad (10)$$

where $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(d,n)} \geq 0$ are the singular values of X .

$U_k = \hat{U}_k$ admits equality.

3.3 LRA-FN \rightarrow PCA

In this section, we find the solution of LRA-FN from the solution of PCA

Given any rank- k approximation $A_k \in \mathbb{R}^{m \times n}$ of A , let $U_k \in \mathbb{R}^{m \times k}$ be a orthogonal basis of column space of A . We will prove that $U_k U_k^T A$ gives a better or as good approximation of A as A_k .

Theorem 2 (minimal distance). *For any vector u in a k -dimensional subspace \mathcal{U}_k , the minimal distance to another vector $a \in \mathbb{R}^m$ is achievable when u is the projection of a onto \mathcal{U}_k .*

$$\min_{u \in \mathcal{U}_k} \|u - a\|_2 = \|U_k U_k^T a - a\|_2 \quad (11)$$

where columns of U_k is a orthogonal basis of \mathcal{U}_k

Apply theorem 2 to columns of A_k and A , we have

$$\|A_k - A\|_F \geq \|U_k U_k^T A - A\|_F \quad (12)$$

By *Pythagorean theorem*,

$$\|U_k U_k^T A - A\|_F = \|A\|_F - \|U_k U_k^T A\|_F \quad (13)$$

By PCA, $\|U_k U_k^T A\|_F$ is maximal when U_k is the k left singular vectors corresponding to the k largest singular values.

$$\|\hat{U}_k \hat{U}_k^T A\|_F \geq \|U_k U_k^T A\|_F \quad (14)$$

Hence,

$$\begin{aligned} \|A_k - A\|_F &\geq \|U_k U_k^T A - A\|_F \\ &= \|A\|_F - \|U_k U_k^T A\|_F \\ &\geq \|A\|_F - \|\hat{U}_k \hat{U}_k^T A\|_F \end{aligned} \quad (15)$$

The equality is admitted in both conditions (1) A_k is the orthogonal projection of A in some subspace of dimension k and (2) the subspace is from PCA.

4 Sequential PCA

In machine learning, sometimes, number of data points is very large and they come sequentially. Sequential PCA attempts to approximate the PCA in $O(1)$ time. This section is a discussion on concept drifting in sequential PCA

Suppose there exists an algorithm producing PCA embedding $y_1^{(t)}, y_2^{(t)}, \dots, y_t^{(t)} \in \mathbb{R}^k$ of input data point $x_1, x_2, \dots, x_t \in \mathbb{R}^d$ after receiving data point x_t at time t . Let $U_k^{(t)} \in \mathbb{R}^{k \times d}$ be the approximated k -dimensional projection subspace of PCA at time t . When a new data point come, the algorithm yields a new approximation of the projection subspace $U_k^{(t+1)}$. Generally, the new approximation will be different. The authors in [2] introduced an update to all previous embedding vectors as

$$y_t^{(t_2)} \mapsto U_k^{(t_2)T} U_k^{(t_1)} y_t^{(t_1)} \quad (16)$$

The update can be decomposed into two steps: (1) map the embedding of x_t at time t_1 : $y_t^{(t_1)} \in \mathbb{R}^k$ back to \mathbb{R}^d (2) project the resulting vector / tensor into the new basis $U_k^{(t_2)}$ that yields the embedding of x_t at time t_2 : $y_t^{(t_2)}$

5 Appendix

5.1 A proof of Trace Optimization Problem

We have

$$\begin{aligned}
\text{tr } U^T M U &= \text{tr } M(UU^T) \quad (\text{cyclic property}) \\
&\leq \sum_{i=1}^d \sigma_i(M) \sigma_i(UU^T) \quad (\text{Von Neumann's Trace Inequality}) \\
&= \sum_{i=1}^k \sigma_i(M) \quad (U \text{ is orthogonal rank-}k)
\end{aligned} \tag{17}$$

where $\sigma_i(A)$ is the i -th singular value of A sorted descending.

5.2 A proof of Von Neumann's Trace Inequality

This proof is by user1551 from Mathematics Stack Exchange [1]

The Von Neumann's Trace Inequality is stated as follow:

Theorem 3. *Given two complex matrices $A, B \in \mathbb{R}^{n \times n}$ with singular values $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n \geq 0$ and $\beta_1 \geq \beta_2 \geq \dots \geq \beta_n \geq 0$*

$$|\text{tr } AB| \leq \sum_{i=1}^n \alpha_i \beta_i \tag{18}$$

Lemma 1. *The Von Neumann's Trace Inequality can be reduced to*

$$|\text{tr } D U S V^*| \leq \text{tr } D S \tag{19}$$

such that U and V are unitary and $D = \text{diag}(d_1, d_2, \dots, d_n)$, $S = \text{diag}(s_1, s_2, \dots, s_n)$

Let P_k denotes the orthogonal projection matrix $I_k \oplus 0_{n-k} = \text{diag}(1, 1, \dots, 1, 0, 0, \dots, 0)$ (k times of 1)

We write D and S as the non-negatively weighted sum of P_k s

$$D = (d_1 - d_2)P_1 + (d_2 - d_3)P_2 + \dots + (d_{n-1} - d_n)P_{n-1} + d_n P_n \tag{20}$$

and similarly for S . Conveniently, we write $D = \sum_k \alpha_k P_k$, $S = \sum_l \beta_l P_l$. Inequality 19 becomes

$$\left| \sum_{k,l} \alpha_k \beta_l \text{tr } P_k U P_l V^* \right| \leq \sum_{k,l} \alpha_k \beta_l \text{tr } P_k P_l \tag{21}$$

If we have $|\text{tr } P_k U P_l V^*| \leq \text{tr } P_k P_l$, *Triangle Inequality* implies the inequality 21. ($|a + b| \leq |a| + |b|$)

Indeed, denote $U = [u_1, u_2, \dots, u_n]$, $V = [v_1, v_2, \dots, v_n]$, so that $P_k U P_l = [P_k u_1, P_k u_2, \dots, P_k u_l, 0, \dots, 0]$. Assuming $l \leq k$, we have

$$\begin{aligned}
|\operatorname{tr}(P_k U P_l) V^*| &= |\operatorname{tr} V^* (P_k U P_l)| \quad (\text{cyclic property}) \\
&= \left| \sum_{i=1}^n \langle (P_k U P_l)_i, v_i \rangle \right| \quad (\text{unroll}) \\
&= \left| \sum_{i=1}^l \langle P_k u_i, v_i \rangle \right| \quad (\text{unroll}) \\
&\leq \left| \sum_{i=1}^l \|P_k u_i\| \|v_i\| \right| \quad (\text{Cauchy-Schwarz inequality}) \\
&= \sum_{i=1}^l \|P_k u_i\| \quad (\text{unit vector}) \\
&= \sum_{i=1}^l 1 \quad (\text{orthogonal projection matrix}) \\
&= l \\
&= \operatorname{tr} P_k P_l
\end{aligned} \tag{22}$$

For the other case, $l > k$, we write $|\operatorname{tr}(P_k U P_l) V^*| = |\operatorname{tr} U (P_l V^* P_k)|$ then apply *Cauchy-Schwarz inequality* on row space instead.

5.2.1 Proof of lemma 1

SVD: $A = U_A \Sigma_A V_A^*$, $B = U_B \Sigma_B V_B^*$

$$\begin{aligned}
\operatorname{tr} AB &= \operatorname{tr} U_A \Sigma_A V_A^* U_B \Sigma_B V_B^* \\
&= \operatorname{tr} \Sigma_A (V_A^* U_B) \Sigma_B (U_A^* V_B)^* \quad (\text{cyclic property}) \\
&= \operatorname{tr} DUCV^*
\end{aligned} \tag{23}$$

References

- [1] user1551. matrices - how to show von neumann trace inequality $\operatorname{Tr}(ab) \leq \sum_{i=1}^n \sigma_{A,i} \sigma_{B,i}$? - mathematics stack exchange.
- [2] Shinjae Yoo, Hao Huang, and Shiva Prasad Kasiviswanathan. Streaming spectral clustering. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 637–648, 2016.