NOTES ON PRINCIPAL COMPONENT ANALYSIS

IORDAN GANEV

1. Notation

We use Einstein notation (link) for vectors and matrices. Specifically, suppose $\mathbf{v}$ is a vector in $\mathbb{R}^N$. Unless specified otherwise, we consider $\mathbf{v}$ to be a column vector with entries indicated by upper indices:

$$\mathbf{v}_i = \begin{bmatrix} v^1 \\ v^2 \\ \vdots \\ v^N \end{bmatrix}$$

It will be clear from context whether the upper subscripts denote indices or exponents. To save space, we may also write $\mathbf{v}$ as a tuple: $\mathbf{v} = (v^1, v^2, \dots, v^N)$. The mean of the vector $\mathbf{v}$ is defined as:

$$\text{mean}(\mathbf{v}) = \frac{1}{N} \sum_{i=1}^{N} v^i$$

The covariance of two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^N$ is defined as:

$$\text{cov}(\mathbf{v}, \mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left(v^i - \text{mean}(\mathbf{v}) \right) \left(w^i - \text{mean}(\mathbf{w}) \right)$$

2. The data matrix X

Let $X \in \mathbb{R}^{N \times d}$ be the data matrix. There are N samples, recorded as the rows of X, and each sample has d features, corresponding to the columns of X. We assume $N \geq d$. Following Einstein notation we denote matrix entries using upper indices for the rows and lower indices for the columns:

$$X = \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots & x_d^1 \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_d^1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^N & x_2^N & x_3^N & \cdots & x_d^N \end{bmatrix} \in \mathbb{R}^{N \times d}$$

In other words, x_j^i denotes the (scalar) entry appearing in the i-th column and j-th row of X, for $i = 1, \dots, N$ and $j = 1, \dots, d$. This is the measurement of the d-th feature of the i-th sample. The covariance matrix[1] of X is a d by d matrix whose (j, k) entry is given by

[1] The covariance matrix can be computed in numpy via the command `np.cov(X, ddof=0)`. Having delta degrees of freedom (`ddof`) equal to d amounts to dividing by $N - d$ instead of N. The default value is `ddof` $= 1$, which gives an unbiased estimator for the population covariance.

the covariance of the j-th and k-th columns of X:

$$\mathrm{Covmat}(X)_k^j = \mathrm{cov}\left((x_j^1, \ldots, x_j^N), (x_k^1, \ldots, x_k^N)\right)$$

3. The matrix A

Let $\mathrm{Id}_{N\times N}$ be the identity N by N matrix, and let $\mathbb{1}_{N\times N}$ be the N by N matrix of all ones. Define a new N by N matrix as:

$$A = \mathrm{Id}_{N\times N} - \frac{1}{N}\mathbb{1}_{N\times N}$$

We leave the proof of the following lemma as an exercise:

Lemma 3.1. *We have:*

1. *The matrix A is symmetric, and $A^2 = A$.*
2. *Each column of the matrix AX has mean zero.*
3. *The covariance of two vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^N$ is given by:*

$$\mathrm{cov}(\mathbf{v}, \mathbf{w}) = \frac{1}{N}\mathbf{v}^T A\mathbf{w}$$

4. *The covariance matrix of X is given by:*

$$\mathrm{Covmat}(X) = \frac{1}{N}X^T AX$$

4. Principal Component Analysis

Lemma 3.1.2, shows that the matrix AX is 'centered', i.e., its column means are all zero. Let $AX = USV^T$ be the singular value decomposition of AX. Hence, $U \in \mathbb{R}^{N\times N}$ is an orthogonal matrix, $S \in \mathbb{R}^{N\times d}$ is a diagonal matrix with non-negative, non-increasing entries along the diagonal (the singular values σ_i of AX), and $V \in \mathbb{R}^{d\times d}$ is an orthogonal matrix. Terminology:

- The columns of V are the *pincipal components* of X.

Using Lemma 3.1.4, and the fact that $U^T U = \mathrm{Id}_{N\times N}$, one computes that the covariance matrix of AXV is given by:

$$\mathrm{Covmat}(AXV) = \frac{S^T S}{N} \in \mathbb{R}^{d\times d}$$

Since S is a diagonal matrix, we conclude that $\mathrm{Covmat}(AXV)$ is also a diagonal matrix. Its i-th diagonal entry, denoted λ_i, is the square of the i-th singular value of AX, divided by N, that is, $\lambda_i = \frac{\sigma_i^2}{N}$ for $i = 1, \ldots, d$. Terminology:

- The dataset $R = AXV$ is the *centered, diagonalized* version of X.

Assume all singular values of X are positive (this is usually the case in practice). Then let $\Lambda^{-1/2}$ be the diagonal d by d matrix whose i-th entry is $(\lambda_i)^{-1/2} = \frac{\sqrt{N}}{\sigma_i}$. We compute that the covariance matrix of $AXV\Lambda^{-1/2}$ is the identity matrix[2]:

$$\text{Covmat}(AXV\Lambda^{-1/2}) = \text{Id}_{d\times d}$$

Terminology:

- The dataset $Z = AXV\Lambda^{-1/2}$ is the *whitened* version of X, with mean zero and unit variance.

5. Projections

Let $s \leq d$, and consider the matrix:

$$\pi_s = \begin{bmatrix} \text{Id}_{s\times s} & 0 \\ 0 & 0 \end{bmatrix}$$

This is a projection matrix onto the first s components. We now:

- Project the centered, diagonalized data to obtain $AXV\pi_s$.
- Apply V^T to obtain the projected centered, undiagonalized data: $AXV\pi_s V^T$.
- Add the column means to obtain the *s-truncated* version of X:

$$\hat{X} = AXV\pi_s V^T + \frac{1}{N}\mathbb{1}_{N\times N}X$$

Note that, if $s = d$, then $\hat{X} = X$, and if $s = 0$, then $\hat{X}$ is the matrix of column means. Otherwise, $\hat{X}$ is a lower-dimensional representation of X. One can show that $\hat{X}$ simplifies to:

$$\hat{X} = X - US\left(\text{Id}_{d\times d} - \pi_s\right)V^T.$$

In other words, one zeros out the first s singular values of $AX = USV^T$ and subtract the result from X.

Finally, we compute the error of this lower-dimensional representation. Let $\mathbf{x}^{(i)}$ denote the i-th row of X, representing the i sample. Similarly, let $\hat{\mathbf{x}}^{(i)}$ denote the i-th row of $\hat{X}$. Then the error is:

$$\begin{aligned}
\text{Error} &= \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)})^T(\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}) \\
&= \frac{1}{N}\text{Trace}\left(\left(X - \hat{X}\right)^T\left(X - \hat{X}\right)\right) \\
&= \frac{1}{N}\text{Trace}\left(\left(US\left(\text{Id}_{d\times d} - \pi_s\right)V^T\right)^T\left(US\left(\text{Id}_{d\times d} - \pi_s\right)V^T\right)\right) \\
&= \frac{1}{N}\text{Trace}\left(S^TS - S^TS\pi_s\right) = \sum_{i=s+1}^{d}\lambda_i
\end{aligned}$$

[2]The same is true for $AXV\Lambda^{-1/2}Q$ where $Q \in \mathbb{R}^{d\times d}$ is any orthogonal matrix