Worksheet on "Principal component Analysis"

PRML – CS5691 (Jul–Nov 2023)

October 11, 2023

1. Consider a dataset of N points with each datapoint being a D-dimensional vector in $\mathbb{R}^D$. Let's assume that:

   - we are in a high-dimensional setting where $D >> N$ (e.g., D in millions, N in hundreds).
   - the $N \times D$ matrix X corresponding to this dataset is already mean-centered (so that each column's mean is zero, and the covariance matrix seen in class becomes $S = \frac{1}{N} X^T X$).
   - the rows (datapoints) of X are linearly independent.

   Under the above assumptions, please attempt the following questions.

   (a) Whereas X is rectangular in general, XX^T and $X^T X$ are square. Show that these two square matrices have the same set of non-zero eigenvalues. Further, argue briefy why these equal eigenvalues are all positive and N in number, and derive the multiplicity of the zero eigenvalue for both these matrices.
   (Note: The square root of these equal positive eigenvalues $\{\lambda_i := \sigma_i^2\}_{i=1,\ldots,N}$ are called the singular values $\{\sigma_i\}_{i=1,\ldots,N}$ of X.)

   Solutions by Om Shri Prasath Ramanan S and Kasibhatla Srivatsava from the earlier offering of PRML

Solution:

Let λ be a non-zero eigen value of X^TX and the corresponding eigenvector be v. Thus, we have

$$X^TXv = \lambda v$$

Multiplying by X on both sides, we have

$$XX^TXv = \lambda Xv \implies XX^Tv' = \lambda v'$$

where $v' = Xv$ is the eigenvalue of XX^T with the same eigenvalue of λ. From this we can conclude that any eigenvalue of X^TX with eigenvector v is also an eigenvalue of XX^T with eigenvector Xv.

Note that v cannot be from the kernel of X, because if it was from the kernel of X, then $X^TXv = 0 \implies \lambda v = 0 \implies \lambda = 0$, but we have assumed that $\lambda \neq 0$. **Thus, the matrices X^TX and XX^T have the same eigenvalues.**

To show that the non-zero eigenvalues are positive, we need to show that the matrices are positive semi-definite. A real matrix M of shape $n \times n$ is said to be positive semi-definite if $x^TMx \geqslant 0 \; \forall \, x \in \mathbb{R}^n$. Consider the matrix XX^T, which is a $N \times N$ matrix. Consider a vector $v \in \mathbb{R}^N$. We can see that $v^TXX^Tv = \|X^Tv\| \geqslant 0$. Thus XX^T is a positive semi-definite matrix, which means that **all the non-zero eigenvalues are positive**. And since XX^T and X^TX have same non-zero eigenvalues, we can also say that X^TX has positive non-zero eigenvalues too.

From rank-nullity theorem, we know that the sum of the rank and nullity of a matrix is equal to the number of columns of the matrix. Consider the matrix XX^T, we know that the its rank is N since the rank of X is N (the rows of X are linearly independent) and the number of columns is also equal to N, thus the nullity of the matrix is 0. Since nullity is equal to the number of zero eigenvalues of a matrix, we can say that all the eigenvalues of XX^T is non-zero. Since the number of eigenvalues XX^T is N and from the first result, we conclude that the **non-zero eigenvalues are all positive and is N in number**.

To find the multiplicity of zero eigenvalues, we need to find the nullity of the matrices. For XX^T, as we proved above the nullity is 0, thus **the multiplicity of zero eigenvalue for XX^T is 0**. For X^TX, the rank of the matrix is N and the number of columns is K, thus the nullity is $K - N$. Thus **the multiplicity of zero eigenvalue for X^TX is $K - N$**

(b) We can choose the set of eigenvectors $\{u_i\}_{i-=1,\ldots,N}$ of XX^T to be an orthonormal set and similarly we can choose an orthonormal set of eigenvectors $\{v_j\}_{j=1,\ldots,D}$ for X^TX. Briefly argue why this orthonormal choice of eigenvectors is possible. Can you choose $\{v_i\}$ such that each v_i can be computed easily from u_i and X alone (i.e., without having to do an eigenvalue decomposition of the large matrix X^TX; assume $i = 1, \ldots, N$ so that $\lambda_i > 0$ and $\sigma_i > 0$)?
(Note: $\{u_i\}, \{v_i\}$ are respectively called the left,right singular vectors of X, and computing them along with the corresponding singular values is called the Singular Value Decomposition or SVD of X.)

Solution: The ability to choose orthonormal set of eigenvectors for X^TX and XX^T is because they are **symmetric matrices**. We can prove that symmetric matrices have orthogonal eigenvectors. Let u and v be the eigenvectors of a symmetric matrix A with eigenvalues λ and μ ($\lambda \neq \mu$).

$$\lambda \langle u, v \rangle = \langle \lambda u, v \rangle = \langle Au, v \rangle = \langle u, A^T v \rangle = \langle u, Av \rangle = \langle u, \mu v \rangle = \mu \langle u, v \rangle$$

Thus, we have $\lambda \langle u, v \rangle = \mu \langle u, v \rangle \implies (\lambda - \mu) \langle u, v \rangle = 0$. Since $\lambda \neq \mu$, thus we have $\langle u, v \rangle \implies u, v$ are orthogonal. Thus we see that any two given eigenvectors of a symmetric matrix is orthogonal. We can create a orthonormal set from these orthogonal vectors by normalizing them. **Thus it is possible to select an orthonormal set of eigenvectors for XX^T and X^TX.**

To get the corresponding eigenvectors v_i from u_i, we can follow the step which we used to prove that the non-zero eigenvalues of X^TX and XX^T are same. Let u_i be an eigenvector of XX^T with eigenvalue λ_i. This gives $XX^Tu_i = \lambda_i u_i$. Multiplying both sides by X^T, we get $X^TXX^Tu_i = \lambda X^Tu_i \implies X^TXv_i = \lambda v_i$, which means $v_i = X^Tu_i$. There is a chance that v_i is not an unit length vector, thus we can normalize it as $\boxed{v_i = \frac{X^Tu_i}{\|X^Tu_i\|}}$

(c) Applying PCA on the matrix X would be computationally difficult as it would involve finding the eigenvectors of $S = \frac{1}{N}X^TX$, which would take $O(D^3)$ time. Using answer to the last question above, can you reduce this time complexity to $O(N^3)$? Please provide the exact steps involved, including the exact formula for computing the normalized (unit-length) eigenvectors of S.

Solution: We use the notation in the previous question (b) to denote the orthonormal eigenvectors of matrices X^TX and XX^T. Then, the orthonormal eigenvectors of S will be $\{v_j\}_{j=1,\dots,D}$, with eigenvalues $\frac{\lambda_i}{N}$, for i = 1 to N and the rest are zero eigenvalues.

First, we find solve the eigenvector problem for matrix XX^T. This involves finding the eigenvalues and their corresponding orthonormal unit eigenvectors of the matrix. Since this is a NxN matrix, the time complexity of this computation is $O(N^3)$. The algorithm, called as Jacobi method, can be used to solve this problem for real symmetric matrices. It

The above operation takes $O(DN^2)$ time. This is because, X^T is DxN and every u_i is Nx1, so computing each v_i takes $O(DN)$ time and since there are N such operations, it takes $O(DN^2)$ time.

This will give the N orthonormal eigenvectors of S corresponding to its non-zero eigenvalues. Now, to apply PCA, it is sufficient to know these. The other D-N eigenvectors of S correspond to zero eigenvalue and hence, they don't add any extra variance even if they are considered as principal components.

So, the overall time-complexity is $O(N^3) + O(DN^2)$. It is economical than $O(D^3)$.

2. CMU School of Computer Science (Fall 2008)

3. Given 3 data points in 2-D space, $(1, 1)$, $(2, 2)$, and $(3, 3)$,

   (a) What is the first principle component?

   Solution: The first principle component is $\mathbf{pc} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)' = (0.707, 0.707)'$ (the negation is also correct).

   (b) If we want to project the original data points into 1-D space by the principle component you choose, what is the variance of the projected data?

   Solution:

   $$\text{Projected data} = \begin{bmatrix} -\frac{2}{\sqrt{2}} \\ 0 \\ \frac{2}{\sqrt{2}} \end{bmatrix}$$

   The variance of the projected data is $\frac{4}{3} = 1.33$.

   (c) For the projected data in (b), now if we represent them in the original 2-D space, what is the reconstruction error?

   Solution: The reconstruction error is 0.

4. CMU School of Computer Science (Fall 2008)

5. Given 6 data points in 5-D space, $(1, 1, 1, 0, 0)$, $(-3, -3, -3, 0, 0)$, $(2, 2, 2, 0, 0)$, $(0, 0, 0, -1, -1)$, $(0, 0, 0, 2, 2)$, $(0, 0, 0, -1, -1)$. We can represent these data points by a 6×5 matrix X, where each row corresponds to a data point:

   $$X = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ -3 & -3 & -3 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix}$$

   Note: Use numpy code to compute the eigen vectors. [can also be solved using Singular value decomposition]What is the sample mean of the data set?

   (a) **Solution:** The sample mean of the data set is $[0, 0, 0, 0, 0]$.

   (b) What is the first principle component for the original data points?

   Solution: The first principle component is $\mathbf{pc} = \pm[c, c, c, 0, 0] = \pm[0.577, 0.577, 0.577, 0, 0]$. Intuition: First, we want to notice that the first three data points are co-linear, and so do the last three data points. Also, the first three data points are orthogonal to the rest three data points. Then, we want to notice that the norm of the first three are much bigger than the last three, therefore, the first principle component has the same direction as the first three data points.

(c) If we want to project the original data points into 1-D space by the principle component you choose, what is the variance of the projected data?

Solution: The variance of the projected data is $\frac{\sigma_1^2}{6} = 7$.
Intuition: We just keep the first three data points and set the rest three data points as $[0, 0, 0, 0, 0]$ (since they are orthogonal to the principle component), and then compute the variance among them.

(d) For the projected data in (c), now if we represent them in the original 5-D space, what is the reconstruction error?

Solution: The reconstruction error is $\frac{\sigma_2^2}{6} = 21$.
Intuition: Since the first three data points are orthogonal with the rest three, the reconstruction error is just the sum of the norm of the last three data points $(2 + 8 + 2 = 12)$, and then divided by the total number (6) of data points, if we use the average definition.

6. Consider a dataset consisting of n data points with each datapoint being a D-dimensional vector in $\mathbb{R}^D$. What can you say about the covariance matrix and the principal components if there is no correlation between the features?

Solution: The covariance matrix is diagonal, and the principal components (eigenvectors) of the covariance matrix will be the standard basis vectors.

7. Given a dataset X:

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Calculate the covariance matrix and the corresponding eigenvectors. Determine the minimal number of principal components required to retain at least 90% of the variance in the dataset.

Solution:
Minimal number of principal components required to retain at least 90% of the variance in the dataset = 1