Homework：Principal Component Analysis, PCA

111C51502, CY Chingyao Fu, NTUT

Oct 2023

Reference

A Tutorial on Principal Component Analysis
Jonathon Shlens
Google Research, Mountain View, CA 94043
(Dated: April 7, 2014; Version 3.02)

Introduction

主成分分析（Principal Component Analysis，PCA）是一種用於現代數據分析的主要工具。它是一種簡單的、非參數的方法，用於從混亂的數據集中提取相關信息。目的是找到一個更有意義的基礎來重新表示數據集，希望這個新基礎能夠過濾掉噪聲並揭示隱藏的結構。

1 基本概念

- 主成分（Principal Components）：這些是原始數據變量的線性組合，並且是正交的（即互相獨立）。第一主成分解釋了最多的變異，第二主成分（與第一主成分正交）解釋了次多的變異，依此類推。
- 變異（Variance）：在PCA中，目標是最大化每個主成分上的變異，因為更大的變異通常意味著更多的信息。
- PCA通常用於降低數據的維度，同時儘量保留有用的信息。這是通過保留前k個主成分來實現的，其中k遠小於原始數據的維度
- 數據投影（Data Projection）：一旦找到主成分，原始數據就可以投影到這些主成分構成的新空間中，從而實現降維。
- 特徵值和特徵向量（Eigenvalues and Eigenvectors）：在數學上，主成分是協方差矩陣的特徵向量，而這些特徵向量對應的特徵值表示了變異的大小。

2 數學框架

PCA做出一個嚴格但有力的假設：線性性。大大簡化了問題，因為它限制了潛在基礎的集合。

1. 數據標準化：首先，對每個變量（特徵）進行標準化，使其均值為0，標準差為1。這是為了確保每個變量對結果的影響是一致的。

$$Z = \frac{X - \mu}{\sigma}$$

2. 計算協方差矩陣：協方差矩陣是一個對稱矩陣，其中每個元素表示兩個變量之間的協方差。

$$\Sigma = \frac{1}{n} Z^T Z$$

3. 求解特徵值和特徵向量：對協方差矩陣進行特徵分解，得到特徵值和特徵向量。特徵值表示了該方向上的變異量，而特徵向量則定義了這個方向。

$$\Sigma V = VA$$

4. 排序和選擇主成分：將特徵值由大到小排序，並選擇前k個最大的特徵值對應的特徵向量。這k個特徵向量構成了一個新的基，用於將原始數據投影到低維空間。

$$W = [v_1, v_2, ..., v_k]$$

5. 數據轉換：用k個特徵向量將原始數據投影到新的低維空間。

$$Y = ZW$$

3 實際應用

- 數據可視化：通過降低數據的維度，PCA可以幫助我們更容易地可視化高維數據。
- 特徵選擇和降維：在機器學習和數據分析中，高維數據往往會導致計算成本高和模型過擬合。PCA可以有效地降低數據維度。
- 噪聲過濾：PCA也可以用於噪聲過濾，因為它能夠保留數據中最重要的變異，同時去除不重要的變異（噪聲）。
- 領域應用：除了數據科學和機器學習，PCA還廣泛應用於生物信息學、金融、工程等多個領域。