Notes on Principal Components Analysis

DSBA Mathematics Refresher 2024

Abstract

1 Sample vs. Population

In statistics, it is crucial to distinguish between a **sample** and a **population**:

- **Population:** The entire set of individuals or observations that we are interested in studying. For example, all people in a country.
- **Sample:** A subset of the population that is used to represent the entire population. For instance, 1,000 people surveyed from the population.

We often cannot measure the entire population due to time, cost, or logistical constraints. We rely on samples to make inferences about the population.

2 Mean, Standard Deviation, and Estimators

2.1 Population Mean and Standard Deviation

Given a population with N elements, the **population mean** μ and the **population standard deviation** σ are defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

2.2 Sample Mean and Standard Deviation

For a sample of size n, the **sample mean** $\bar{x}$ and the **sample standard deviation** s are calculated as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2.3 Why is the Variance Estimator Biased?

The sample variance estimator is given by:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

This estimator is biased because it tends to underestimate the true population variance σ^2. The bias arises because $\bar{x}$ is itself a random variable that depends on the sample, and it pulls the variance down slightly. This correction by dividing by $(n-1)$ instead of n is known as Bessel's correction [1].

2.4 When to Use $\frac{1}{n}$ vs. $\frac{1}{n-1}$

Population Data When you have data for the entire population, use:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Here, you divide by N, the total number of observations in the population.

Sample Data When you have data for a sample and wish to estimate the population variance, use:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

This adjustment accounts for the extra variability introduced by using the sample mean $\bar{x}$ instead of the population mean μ.

3 Change of Basis

In linear algebra, a **change of basis** refers to expressing a vector in a different coordinate system. Suppose $\mathbf{v}$ is a vector in a vector space with basis $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$. If we have a new basis $\mathbf{B}' = \{\mathbf{b}'_1, \mathbf{b}'_2, \ldots, \mathbf{b}'_n\}$, we can represent $\mathbf{v}$ in the new basis by finding the coordinates relative to $\mathbf{B}'$.

If $\mathbf{v} = a_1\mathbf{b}_1 + a_2\mathbf{b}_2 + \cdots + a_n\mathbf{b}_n$, then under the new basis $\mathbf{B}'$, the same vector can be written as:

$$\mathbf{v} = a'_1\mathbf{b}'_1 + a'_2\mathbf{b}'_2 + \cdots + a'_n\mathbf{b}'_n$$

[1] See `https://gregorygundersen.com/blog/2019/01/11/bessel/` for more in depth explanation.

The coordinates $\mathbf{a}' = (a'_1, a'_2, \ldots, a'_n)$ are related to the original coordinates $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ by a transformation matrix $\mathbf{P}$:

$$\mathbf{a} = \mathbf{P}\mathbf{a}' \qquad or \qquad \mathbf{a}' = \mathbf{P}^{-1}\mathbf{a}$$

4 Principal Component Analysis (PCA) Theory

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of data while preserving as much variance as possible. It does so by finding a new basis in which the first few dimensions capture the most variance in the data. All but the first few components can be discarded without loosing too much information, while making the data (much) easier to process.

Steps to perform PCA

1. **Standardize the Data:** Subtract the mean and divide by the standard deviation for each feature.
2. **Compute the Covariance Matrix:** For a dataset with n features, compute the $n \times n$ covariance matrix.
3. **Eigenvalue Decomposition:** Perform an eigenvalue decomposition of the covariance matrix to find the eigenvalues and eigenvectors.
4. **Select Principal Components:** The eigenvectors corresponding to the largest eigenvalues are chosen as the principal components.
5. **Transform the Data:** Project the original data onto the principal components to obtain the transformed data in the new basis.

Mathematical Formulation Given a data matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ where m is the number of samples and n is the number of features, the covariance matrix $\mathbf{C}$ is:

$$\mathbf{C} = \frac{1}{m-1}\mathbf{X}^\top \mathbf{X}$$

The eigenvalue decomposition of $\mathbf{C}$ gives:

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$$

where $\mathbf{V}$ is the matrix of eigenvectors and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. The principal components are the columns of $\mathbf{V}$ corresponding to the largest eigenvalues.