1 Principle Component Analysis

1.1 Basic Concept and Background

We will only consider Euclidean space.

A set is a collection of vectors (or, points).

A space $S$ is a set equipped with some metric\footnote{A metric measures ‘distance’ between points.} e.g. Euclidean distance.

We say a set $W := \{w_1, \ldots, w_d\}$ spans $S$ if every point in $S$ can be linearly expressed by $W$.\footnote{By ‘linearly express’, we mean every point in $S$ is a linear combination of elements in $W$ (with probably different coefficients).}

We say $W$ is a basis of $S$ if it spans $S$ and its elements are linearly independent.\footnote{We say a set of vectors are linearly independent if none of them can be linearly expressed by others.}

The dimension of $S$ is the cardinality of its basis.

[Exercise] Give examples of above concepts in $\mathbb{R}^2$.

[Exercise] Prove different bases of $S$ have the same cardinality.

Suppose $W$ is a basis of $\mathbb{R}^d$. This means any point $x \in \mathbb{R}^d$ can be expressed as

$$x = \alpha_1 w_1 + \ldots + \alpha_d w_d$$

(1)

for some coefficients $\alpha_1, \ldots, \alpha_d$ (depending on $x$).

Now, we will introduce the second interpretation of PCA – a minimizer of reconstruction error of data from reduced feature space. Unlike the first interpretation, we will now interpret each projection vector as a ‘basis vector’ of the reduced feature space. Also, we will now start from all basis vectors, instead of from one to another in an incremental manner.

1.2 PCA as a Reconstruction Error Minimizer

Let $W = \{w_1, \ldots, w_p\}$ be a basis of $\mathbb{R}^p$ and $x_i \in \mathbb{R}^p$ be an arbitrary instance. By definition

$$x_i = \sum_{j=1}^{p} \alpha_{ij} w_j,$$

(2)

where $\alpha_{ij}$ are coefficients depending on $x_i$ (thus an additional index $i$).

Since $w_j$’s are orthogonal, we have an important claim:

$$\alpha_{ij} = w_j^T x_i.$$  

(3)

[Exercise] Verify (3).
Now, without loss of generality, reducing \( x_i \)'s dimension to \( d \) means finding a subset \( w_1, \ldots, w_d \) to approximate \( x_i \). Specifically, we can write the approximated instance as

\[
\tilde{x}_i = \sum_{j=1}^{d} \alpha_{ij} w_j + \sum_{j=d+1}^{p} \beta_j w_j,
\]

where \( \alpha_{ij} \) still depends on \( x_i \) whereas \( \beta_j \) remains constant for all instances\(^4\).

PCA aims to find \( w_j, \alpha_{ij} \) and \( \beta_j \) so as to minimize the total approximation error

\[
L(W) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - \tilde{x}_i||^2.
\]

We also consider \( ||x_i - \tilde{x}_i|| \) as reconstruction error as \( \tilde{x}_i \) can be interpreted as instance \( x_i \) being reconstructed from a \( d \)-dimensional space (using only \( w_1, \ldots, w_d \)).

Now we solve the above optimization problem. Let's first write out \( L(W) \) i.e.

\[
L(W) = \frac{1}{n} \sum_{i} x_i^T x_i - \frac{2}{n} \sum_{i,j=1}^{j=d} \alpha_{ij} x_i^T w_j - \frac{2}{n} \sum_{i,j=d+1}^{j=p} \beta_j x_i^T w_j + \frac{1}{n} \sum_{i,j=1}^{j=d} \alpha_{ij}^2 + \frac{1}{n} \sum_{i,j=d+1}^{j=p} \beta_j^2.
\]

[Exercise] Verify (6).

Observe \( L(W) \) is a quadratic function of unknown parameters. We can try critical point method. Solving \( \frac{\partial}{\partial \alpha_{ij}} L(W) = 0 \) for \( \alpha_{ij} \), we have

\[
\alpha_{ij} = x_i^T w_j.
\]

Solving \( \frac{\partial}{\partial \beta_j} L(W) = 0 \) for \( \beta_j \), we have

\[
\beta_j = \bar{x}^T w_j.
\]

where \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \).

[Exercise] Verify (7) and (8).

Plugging \( \alpha_{ij} \) and \( \beta_j \) back to \( L(W) \), together with (2), (3) and (4), we have

\[
x_i - \tilde{x}_i = \sum_{j=d+1}^{p} \left[(x_i - \bar{x})^T w_j \right] \cdot w_j.
\]

Thus

\[
||x_i - \tilde{x}_i||^2 = (x_i - \tilde{x}_i)^T (x_i - \tilde{x}_i)
= \sum_{j=d+1}^{p} \left[(x_i - \bar{x})^T w_j \right]^2 \cdot w_j^T w_j
= \sum_{j=d+1}^{p} w_j^T (x_i - \bar{x})(x_i - \bar{x})^T w_j,
\]

\(^4\)This means the rest basis vectors are not useful in expressing instances, but only serve as a bias term.
[Exercise] Verify (10).

The reconstruction error becomes

\[ L(W) = \frac{1}{n} \sum_{i=1}^{n} ||x_i - \tilde{x}_i||^2 = \sum_{j=d+1}^{p} w_j^T \Sigma w_j, \tag{11} \]

where \( \Sigma = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \) is the data variance matrix.

Above analysis suggests we want to find a basis \( w_1, \ldots, w_d, w_{d+1}, \ldots, w_p \) such that the (supposedly discarded) subset \( w_{d+1}, \ldots, w_p \) can minimize \( L(W) \).

The rest analysis is similar to the first interpretation of PCA.

One can easily verify \( w_p \) is an eigenvector of \( \Sigma \) associated with the smallest eigenvalue. And, incrementally, \( w_{p-1}, \ldots, w_{d+1} \) are eigenvectors of \( \Sigma \) associated with the smallest eigenvalues.

Recall all eigenvectors of \( \Sigma \) form a basis of \( \mathcal{R}^p \).

Thus above analysis suggests we can find all eigenvectors of \( \Sigma \) and approximate instances using only those associated with the largest eigenvalues (thus discarding those associated with the smallest eigenvalues).

1.3 Other Issues

[Q1] How to extract eigenvectors in practice?

[A1] There are standard packages/functions to extract eigenvectors.

[Q2] What is the dimension I should reduce to?

[A2] Depends. For data visualization purpose, one can choose 2 to 3 dimensions. For data compression, a typical choice is the number of leading eigenvectors whose total eigenvalue is 90% - 95% of the total eigenvalue of the covariance matrix.

[Reading] PRML, Chapter 12.1 (12.2)