1 Principle Component Analysis

1.1 Basic Concepts and Background

Let $M \in \mathbb{R}^{p \times p}$ be a squared matrix. We say $x \in \mathbb{R}^p$ is an eigenvector of $M$ if

$$Mx = \lambda x$$

for some scalar $\lambda$. The scalar is an eigenvalue of $M$ associated with $x$.

Let $x \in \mathbb{R}^p$ be an instance and $w \in \mathbb{R}^p$ be a vector. A linear projection of $x$ on $w$ is

$$\tilde{x} = w^T x.$$  \hspace{1cm} (2)

[Demo] Draw a projection of instances.

Suppose an instance is described by $p$ features. It is often useful to reduce its feature dimension for e.g. data visualization, extracting more useful information or speeding up computation. Such set of techniques are generally known as dimensionality reduction techniques.

There are two general strategies: feature selection and feature transformation.

Feature selection selects a subset of features to describe instance.

Feature transformation projects raw feature space onto a lower dimensional space. The projection usually needs to be learned. PCA provides a useful learning criterion.

PCA has two interpretations: maximize data variance in reduced feature space, or minimize data reconstruction error from reduced feature space. We will introduce both.

1.2 PCA as a Data Variance Maximizer

1.2.1 First Principal Component

PCA finds a projection vector $w$ so that data variance in the projected space is maximized – this minimizes data compression loss. Let $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$ be a set of instances and $\mu$ be their mean. Sample variance in the original feature space is

$$\Sigma = \frac{1}{n} \sum_i (x_i - \mu)(x_i - \mu)^T.$$  \hspace{1cm} (3)

Sample variance in the projected space of $w$ is

$$\Sigma_w = \frac{1}{n} \sum_i (w^T x_i - w^T \mu)(w^T x_i - w^T \mu)^T = w^T \Sigma w.$$  \hspace{1cm} (4)

PCA aims to find a \( w \) that maximizes \( \Sigma_w \). Why? It preserves the distinctive relation between instances (for analysis).

[Discussion] What happens if data variance becomes zero in the projected space?

In addition, we can control model complexity by restricting \( ||w||^2_F = w^T w = 1. \)

This leads to the following PCA optimization problem:

\[
\max_w \Sigma_w \quad \text{s.t. } w^T w = 1. \tag{5}
\]

We can solve it using the Lagrange multiplier technique. The augmented objective function is

\[
L(w) = \Sigma w + \lambda(w^T w - 1) = w^T \Sigma w + \lambda(w^T w - 1) = w^T (\Sigma + \lambda I) w - \lambda, \tag{6}
\]

where \( \lambda \) is any coefficient.

Note \( L(w) \) is a quadratic function of \( w \), thus can be solved by critical point method. Setting

\[
\frac{\partial}{\partial w} L(w) = (\Sigma - \lambda I) w = 0, \tag{7}
\]

and solving for \( w \), we have

\[
\Sigma w = \lambda w. \tag{8}
\]

By definition, (9) suggests \( w \) is an eigenvector of \( M \), and \( \lambda \) is the associated eigenvalue.\(^2\)

But a matrix has many eigenvectors. Which one should we choose?

From (4) we see

\[
w^T \Sigma w = \lambda = \Sigma_w. \tag{9}
\]

Remember we want to maximize \( \Sigma_w \) – this means we want \( \lambda \) to be as large as possible – this means we should select an eigenvector \( w \) associated with the large eigenvalue \( \lambda \).

In summary, the optimal projection vector \( w \) is an eigenvector of the covariance matrix \( \Sigma \) associated with the largest eigenvalue. This \( w \) is also known as the first principal component.

### 1.2.2 More Principal Components

Suppose we obtained the first principal component \( w_0 \). It allows us to reduce feature dimension to 1 (by projecting all instances to \( w_0 \)).

But what if we want to reduce feature dimension to 2, 3 or more?

Then we need more projection vectors. PCA finds them in a similar way to \( w_0 \), with an additional constraint that newly reduced feature dimensions are statistically uncorrelated.

Let \( w \) denote the second projection vector. PCA finds it by solving the following problem

\[
\max_w \Sigma_w, \quad \text{s.t. } w^T w = 1, \quad \text{cov}(w^T x, w_0^T x) = 0, \tag{10}
\]

\(^1\)Otherwise, we may result in some trivial solution.

\(^2\)There are many standard packages to find matrix eigenvector and eigenvalue.
where \( \text{cov}(w^T x, w_0^T x) \) is the (estimate of) covariance between reduced feature \( w^T x \) and \( w_0^T w \). In practice, we have

\[
\text{cov}(w^T x, w_0^T x) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - w^T \mu)(w_0^T x_i - w_0^T \mu)^T = 0
\]

\[
\Rightarrow \frac{1}{n} \sum_{i=1}^{n} w^T (x_i - \mu)(x_i - \mu)^T w_0 = 0
\]

\[
\Rightarrow w^T \left( \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T \right) w_0 = 0
\]

\[
\Rightarrow w^T \Sigma w_0 = 0
\]

\[
\Rightarrow w^T (\lambda w_0) = 0
\]

\[
\Rightarrow w^T w_0 = 0,
\]

where the last second line holds because \( w_0 \) is an eigenvector of \( \Sigma \).

Plugging above result to (10), we have a new optimization problem for \( w \):

\[
\max_w \Sigma w, \quad \text{s.t.} \quad w^T w = 1, \quad w^T w_0 = 0.
\]  

(12)

Applying Lagrange Multiplier, we have an augmented objective function

\[
L(w) = \Sigma w - \lambda_1 (w^T w - 1) - \lambda_2 (w^T w_0).
\]

(13)

Setting \( L'(w) = 0 \) and \( \tilde{\lambda}_2 = \frac{1}{2} \lambda_2 \), we have

\[
\Sigma w - \lambda_1 w - \tilde{\lambda}_2 w_0 = 0.
\]

(14)

Here is a trick: if we left-multiply \( w_0 \) into above formula, we have

\[
w_0^T \Sigma w - \lambda_1 w_0^T w - \lambda_2 w_0^T w_0 = 0 \Rightarrow 0 - 0 - \lambda_2 = 0
\]

(15)

This implies \( \lambda_2 = 0 \); thus by (14) we have

\[
\Sigma w - \lambda_1 w = 0 \Rightarrow \Sigma w = \lambda_1 w
\]

(16)

This means the second optimal projection vector \( w \) is also an eigenvector of \( \Sigma \) and \( \lambda_1 \) is its associated eigenvalue. Again, we want \( w^T \Sigma w = \lambda_1 = \Sigma w \) to be as large as possible; this means \( w \) should be the eigenvector associated with the second largest eigenvalue \( \lambda_1 \).

In summary, the second optimal projection vector \( w \) is the eigenvector of \( \Sigma \) associated with its second largest eigenvalue. This \( w \) is also known as the second principal component.

Repeating the above process, we can show the rest principal components are all eigenvectors of the data covariance matrix \( \Sigma \) associated with its largest eigenvalues.

[Reading] PRML, Chapter 12.1 (12.1.1)