PCA

Data preprocessing

- Centering —> eliminate the bias
  - results have 0 mean
- Normalization —> unify the scale
  - e.g. features are of different units
  - results have unit variance
- Whitening —> elinimate the correlation
  - results have 0 covariance

PCA Dimension Reduction

Idea: to transform the data into another basis so that hopefully the majority of the coordinates are small and can be treated as small random fluctuations.

That is, to **reduce the dimension, and retain most of the useful information**.

Before doing PCA ...

- Centering
  - covariance matrix is based on 0-mean data
  - otherwise the primary component's direction may be misleading
- Normalization
  - should be done if data are of different scales
- Whitening
  - no use doing PCA after whitening
  - because by whitening you treat all dimensions the same

How to do PCA

One Way from Wiki (deal with covariance matrix)

Suppose we project each $\mathbf{x}$ to the direction in which the feature values have the largest variance.

Projected feature value for some sample data is $z = \mathbf{v}^T\mathbf{x}$, thus feature variance over the entire dataset is:

$$
\begin{aligned}
var[z] &= \frac{1}{N}\sum_{n=1}^{N} z_n^2 \\
&= \frac{1}{N}\sum_{n=1}^{N} \mathbf{v}^T \mathbf{x}\mathbf{x}^T \mathbf{v} \\
&= \mathbf{v}^T \left(\sum_{n=1}^{N} \mathbf{x}\mathbf{x}^T \right) \mathbf{v} \\
&= \mathbf{v}^T \mathbf{\Sigma} \mathbf{v}
\end{aligned}
\tag{1}
$$

So to maximize the feature variance we'd like to find a projection vector with unit norm such that

$$
\mathbf{v} = \text{argmax}_{\mathbf{v}} \left\{ \frac{\mathbf{v}^T \mathbf{X}^T \mathbf{X} \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \right\}
$$

This is the Rayleigh quotient optimally solved by $\mathbf{v}$ being the eigenvector of the covariance matrix $\mathbf{X}^T\mathbf{X}$ corresponding to its largest eigenvalue.

And the k-th components are found by first subtracting the first $k-1$ components

$$
\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{s=1}^{k-1} \mathbf{X}\mathbf{v}_{(s)}\mathbf{v}_{(s)}^T
$$

And then follow the same way as we find the primary component:

$$
\mathbf{v}_{(k)} = \text{argmax}_{\mathbf{v}} \left\{ \frac{\mathbf{v}^T \hat{\mathbf{X}}_k^T \hat{\mathbf{X}}_k \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \right\}
$$

Another Way from the *LFD* Book (deal with data matrix)

The goal is to transform $\mathbf{x}$ to another coordinate system, i.e., transform from the original system to the new system with orthonormal basis $\{\mathbf{v}_i\}$

$$
\mathbf{x} = \sum_{i=1}^{d} z_i \mathbf{v}_i
$$

and throw away some dimensions

$$
\hat{\mathbf{x}} = \sum_{i=1}^{k} z_i \mathbf{v}_i
$$

So we have to make sure the difference brought by this transformation is minimized so that we do not ignore important information from the original data, that is to minimize

$$\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \sum_{i=k+1}^{d} z_i^2$$

and over the entire dataset, to minimize

$$\sum_{n=1}^{N} \|\mathbf{x}_n - \hat{\mathbf{x}}_n\|^2$$

With the background, the **Eckart-Young** Theorem tells us that the optimal solution for this problem are the **right singular vectors** obtained from the SVD of the matrix $\mathbf{X}$.

After SVD $\mathbf{X} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T$, we have

$$\begin{aligned} \mathbf{U}^T\mathbf{X} = \mathbf{\Gamma}\mathbf{V}^T \rightarrow \mathbf{X}^T\mathbf{u}_i = \lambda_i\mathbf{v}_i \quad (\mathbf{u}_i \text{ are } \textbf{left} \text{ singluar vectors}) \\ \mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{\Gamma} \rightarrow \mathbf{X}\mathbf{v}_i = \lambda_i\mathbf{u}_i \quad (\mathbf{v}_i \text{ are } \textbf{right} \text{ singluar vectors}) \end{aligned} \tag{2}$$

Thus the PCA algorithm is

1. compute SVD of input data matrix $\mathbf{X} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^T$
2. pick first k columns of $\mathbf{V}$, $\mathbf{V}_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$
3. compute transformed features $\mathbf{Z} = \mathbf{X}\mathbf{V}_k$ and restored data matrix $\hat{\mathbf{X}} = \mathbf{X}\mathbf{V}_k\mathbf{V}_k^T$