Lecture 7: Principal Components Analysis

Outline

1. Lecture 7: Principal Components Analysis
Lecture 7: Principal Components Analysis
Recall the covariance

- The frame operator $S$ can be written as
  $$S : H \rightarrow H, \quad v \mapsto \sum_{n=1}^{N} \langle v, x_n \rangle x_n = (PP^*)v,$$
  where $PP^*$ is $D \times D$.

- Hence, up to a scaling factor and a translation, $S$ is the linear operator identified with the $D \times D$ symmetric covariance matrix $C = \frac{1}{N} PP^*$ of the data space, i.e.
  $$C = \frac{1}{N} \left( \sum_{j=1}^{N} x_j[m]x_j[n] \right)^D_{m,n=1}, \quad x_j = (x_j[1], \ldots, x_j[D]) \in \mathbb{R}^D.$$
Principal Component Analysis

- The covariance matrix $C$ we have just defined is certainly symmetric and also positive semidefinite, since for every vector $y$, we have

$$\langle y, Cy \rangle = \frac{1}{N} \sum_{j=1}^{N} |\langle y, x_j \rangle|^2 \geq 0.$$ 

- Thus, $C$ can be diagonalized, and its eigenvalues are all nonnegative. If $K$ denotes the orthogonal matrix that diagonalizes $C$, then we have that $K^*CK$ is diagonal and the whole process of analyzing data using the eigenbasis of covariance matrix is known as **Principal Component Analysis (PCA)**. $K$ is also known as principal orthogonal decomposition or Karhunen-Loeve transform.

- The columns of $K$ are the eigenvectors of $C$. The number of positive eigenvalues is the actual number of uncorrelated parameters, or degrees of freedom in the original data set $X$. Each eigenvalue is the variance of its degree of freedom.


We shall now present a different, data-inspired model for PCA.

- Assume we have $D$ observed (measured) variables: 
  $$y = [y_1, \ldots, y_D]^T.$$ 
  This is our data.

- Assume we know that our data is obtained by a linear transformation $W$ from $d$ unknown variables $x = [x_1, \ldots, x_d]^T$:
  $$y = W(x).$$

Typically we assume $d < D$.

- Assume moreover that the $D \times d$ matrix $W$ is a change of a coordinate system, i.e., columns of $W$ (or rows of $W^T$) are orthonormal to each other:
  $$W^T W = I_d.$$

Note that $WW^T$ need not be an identity.
Given the above assumptions the problem of PCA can be stated as follows:

How can we find the transformation $W$ and the dimension $d$ from a finite number of measurements $y$?

We shall need 2 additional assumptions:

- Assume that the unknown variables are Gaussian;
- Assume that both the unknown variables and the observations have mean zero (this is easily guaranteed by subtracting the mean, or the sample mean).
For a noninvertible matrix, we have its pseudoinverse defined as

\[ W^+ = (W^T W)^{-1} W^T \]

In our case, \( W^+ = W^T \), Thus, if \( y = Wx \), we have

\[ WW^T y = W W^T W x = W l d_d x = y, \]

or, equivalently,

\[ y - WW^T y = 0. \]

With the presence of noise, we cannot assume anymore the perfect reconstruction, hence, we shall minimize the reconstruction error defined as

\[ E_y(\|y - WW^T y\|_2^2). \]

It is not difficult to see that

\[ E_y(\|y - WW^T y\|_2^2) = E_y(y^T y) - E_y(y^T WW^T y). \]
As $E_y(y^T y)$ is constant, our minimization of error reconstruction turns into a maximization of $E_y(y^T WW^T y)$. In reality, we known little about $y$, so we have to rely on the measurements $y(k)$, $k = 1, \ldots, N$. Then,

$$E_y(y^T WW^T y) \sim \frac{1}{N} \sum_{n=1}^{N} (y(n))^T WW^T (y(n)) \sim \frac{1}{N} \text{tr}(Y^T WW^T Y),$$

where $Y$ is the matrix whose columns are the measurements $y(n)$ (hence $Y$ is a $D \times N$ matrix).

Using SVD for $Y$: $Y = V \Sigma U^T$, we obtain:

$$E_y(y^T WW^T y) \sim \frac{1}{N} \text{tr}(U \Sigma^T V^T WW^T V \Sigma U^T).$$

Therefore, after some computations we obtain:

$$\text{argmax}_W E_y(y^T WW^T y) \sim V \text{Id}_{D \times d},$$

and so $x \sim \text{Id}_{d \times D} V^T y$. 
Another approach to PCA is by assuming that the unknown variables are uncorrelated (in a statistical sense). This can boil down in practice to the assumption that the covariance matrix $C$ is diagonal. Since the observed measurements are often corrupted, we may write

$$C_y = E(yy^T) = E(Wxx^TW^T) = WE(xx^T)W^T = WC_xW^T.$$ 

Alternatively, because of the orthogonality in $W$, we have

$$C_x = W^TC_yW.$$ 

Now, we use eigendecomposition of $C_y$ (since we can), to write $C_y = V\Lambda V^T$. This leads to

$$C_x = W^TV\Lambda V^TW.$$ 

This equality can hold only when $W = V Id_{D \times d}$. 