Principal Component Analysis

Wojciech Czaja

November 24, 2015

Norbert Wiener Center
for Harmonic Analysis and Applications

From data to operators

- Given is data set X consisting of N vectors $x_n \in \mathbb{R}^D$. Without loss of generality, assume $\sum x_n = 0$ (subtract mean).
- Let P be $D \times N$ matrix whose columns are the data vectors x_n.
- Let $\mathbb{H} = \text{span}\{x_n\}_{n=1}^N \subseteq \mathbb{R}^D$. Define $L : \mathbb{H} \to \mathbb{R}^N$,

  $$v \mapsto P^* v = L(v) = \{\langle v, x_n \rangle\} \in \mathbb{R}^N,$$

  and its adjoint $L^* : \mathbb{R}^N \to \mathbb{H} \subseteq \mathbb{R}^D$,

  $$w \mapsto L^*(w) = \sum_{n=1}^{N} w[n] x_n, \quad w = (w[1], w[2], \ldots, w[N]).$$

- L is called the *Bessel (analysis)* operator, and L^* is called the *synthesis* operator.

Derivation of Covariance

- The frame operator S can be written as

$$S : \mathbb{H} \to \mathbb{H},\ v \mapsto \sum_{n=1}^{N} \langle v, x_n \rangle x_n = (PP^*)v,$$

  where PP^* is $D \times D$.

- Hence, up to a scaling factor (of $1/N$) and a translation (mean subtraction), S is the linear operator identified with the $D \times D$ symmetric **covariance matrix** $C = \frac{1}{N} PP^*$ of the data X, i.e.

$$C = \frac{1}{N} \left(\sum_{j=1}^{N} x_j[m] x_j[n] \right)_{m,n=1}^{D}, \quad x_j = (x_j[1], \dots, x_j[D]) \in \mathbb{R}^D.$$

Principal Component Analysis

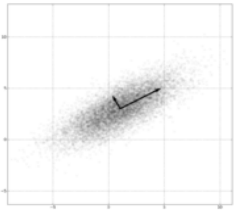
- The covariance matrix C we have just defined is certainly symmetric and also positive semidefinite, since for every vector y, we have

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

- 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, Hoteling transform, or Karhunen-Loève 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 represents the variance of its degree of freedom.

Illustration of Principal Component Analysis

The following plot contains a collection of points obtained from a multivariate Gaussian distribution centered at (1,3) with a standard deviation of 3 in the (0.878, 0.478) direction, and of 1 in the orthogonal direction. The vectors shown are the eigenvectors of the covariance matrix scaled by the square root of the corresponding eigenvalue, and shifted so their tails are at the mean.



Source of imagery: Wikipedia

PCA History

- K. Pearson, *On lines and planes of closest fit to systems of points in space,* Philosophical Magazine, vol. 2 (1901), pp. 559–572
- H. Hoteling, *Analysis of a complex of statistical variables into principal components,* Journal of Education Psychology, vol. 24 (1933), pp. 417–44
- K. Karhunen, *Zur Spektraltheorie stochastischer Prozesse,* Ann. Acad. Sci. Fennicae, vol. 34 (1946)
- M. Loève, *Fonctions aléatoire du second ordre,* in Processus stochastiques et mouvement Brownien, p. 299, Paris (1948)

Data perspective

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

- Assume we have D observed (measured) variables: $y = [y_1, \dots, y_D]^T$. This is our data.
- Assume we know that our data is obtained by a linear transformation W from d unknown variables $x = [x_1, \dots, x_d]^T$:

  $$y = W(x).$$

  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:

  $$W^T W = Id_d.$$

  Note that WW^T need not be an identity matrix.

PCA

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).

PCA minimizing the reconstruction error

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

$$WW^T y = WW^T Wx = WId_d x = y,$$

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

$$E_y(\|y - WW^T y\|_2^2) = E_y(y^T y) - E_y(y^T WW^T y).$$

PCA from minimizing the reconstruction error

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 $y(k)$, $k = 1, \dots, N$. Then,

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

where Y is the matrix whose columns are the measurements $y(n)$ (hence Y is a $D \times N$ matrix).
Using Singular Value Decomposition (SVD) for Y: $Y = V\Sigma U^T$, we obtain:

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

Therefore, after some computations we obtain:

$$argmax_W E_y(y^T WW^T y) \sim V\, Id_{D\times d}.$$

Thus, we have that $W \sim V\, Id_{D\times d}$, and so $x \sim Id_{d\times D} V^T y$.

PCA from maximizing the decorrelation

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

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

Alternatively, because of the orthogonality in W, we have

$$C_x = W^T C_y W.$$

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

$$C_x = W^T V \Lambda V^T W.$$

This equality can hold only when $W = V\, Id_{D\times d}$. Hence, again $x = Id_{d\times D} V^T y$.

Summary

To summarize, we have presented two different approaches which lead to the same result:

If we assume that our data Y consists of N measurements $y(n) \in \mathbb{R}^D, n = 1, \ldots, N$ and is obtained by an orthogonal transformation W from N (a priori) unknown measurements $x(n) \in \mathbb{R}^d, n = 1, \ldots, N$, with $d < D$ (represented by a $d \times N$ matrix X) such that :

$$Y = W(X),$$

and if we make some further auxiliary assumptions, then we can we find the variables X, the transformation W, and the dimension d from N measurements Y by letting

$$X = Id_{d \times D} V^T Y,$$

where V comes from SVD of Y and d is the number of nonzero singular values of Y.