1 Summary of PCA

We summarize the results from Shlens' PCA tutorial.[1]

Consider an experiment on some system whose behavior is not known. Specifically, we don't know how many degrees of freedom the system has, what they are, or how to measure them. We ignorantly take measurements of m different variables. For example, we could measure the positions of a moving particle along m arbitrary and possibly non-orthogonal axes, or m state variables of a thermodynamic system. We take n data points, measuring all m variables each time.

We can group the data by time or by measurement type: let $\mathbf{x_i}$ be the m-dimensional column vector of the m measurements taken at a single point in time, and let $\mathbf{\tilde{x}_i}$ be the n-dimensional row vector of all measurements of a single variable. We arrange our data into an $m \times n$ matrix:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x_1} & \cdots & \mathbf{x_n} \end{pmatrix} = \begin{pmatrix} \mathbf{\tilde{x}_1} \\ \vdots \\ \mathbf{\tilde{x}_m} \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{pmatrix},$$

such that the scalar entry x_{ij} is the jth measurement of the ith variable.

The goal is to analyze $\mathbf{X}$ and recover the *principal components* – the true degrees of freedom of the system. We make two important assumptions:

1. **Linearity.** The principal components are linear combinations of the m variables we measured. This is a strong assumption required for the linear algebra techniques that follow.

2. **Signal spread.** The true degrees of freedom are the directions along which the data has the largest spread. This assumption presumes that the data has a high signal-to-noise ratio: the amplitude of the data is high compared to the amplitude of the noise.

For simplicity, we further assume that the data for each of the m variables has mean zero. This is not a strong assumption: if it is not the case, we can simply subtract off the mean of each measurement type. Let $\langle \mathbf{x} \rangle$ denote the average value of the components of $\mathbf{x}$; then the transformed dataset would be:

$$\mathbf{X}^* = \mathbf{X} - \langle \mathbf{X} \rangle = \begin{pmatrix} \mathbf{\tilde{x}_1} \\ \vdots \\ \mathbf{\tilde{x}_m} \end{pmatrix} - \begin{pmatrix} \langle \mathbf{\tilde{x}_1} \rangle & \cdots & \langle \mathbf{\tilde{x}_1} \rangle \\ \vdots & & \vdots \\ \langle \mathbf{\tilde{x}_m} \rangle & \cdots & \langle \mathbf{\tilde{x}_m} \rangle \end{pmatrix},$$

Let $\mathbf{P}$ be a matrix of row vectors $\mathbf{p_i}$, corresponding to the principal components in a way that will be derived below. We further define a transformed data matrix $\mathbf{Y} \equiv \mathbf{PX}$:

$$\mathbf{P} = \begin{pmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{pmatrix} \quad \text{so} \quad \mathbf{Y} = \mathbf{PX} = \begin{pmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{pmatrix} \begin{pmatrix} \mathbf{x_1} & \cdots & \mathbf{x_n} \end{pmatrix} = \begin{pmatrix} \mathbf{p_1} \cdot \mathbf{x_1} & \cdots & \mathbf{p_1} \cdot \mathbf{x_n} \\ \vdots & & \vdots \\ \mathbf{p_m} \cdot \mathbf{x_1} & \cdots & \mathbf{p_m} \cdot \mathbf{x_n} \end{pmatrix}.$$

Note that each column of $\mathbf{Y}$ is an m-dimensional vector whose components are the projections of $\mathbf{x_i}$ along the vectors $\mathbf{p_j}$. This represents a change of basis: the $\mathbf{p_j}$ are the new basis vectors. This shows us how to tranform the data $\mathbf{X}$ into a dataset $\mathbf{Y}$ with a different basis.

We now wish to choose a basis that identifies the principal components – the independent degrees of freedom of the system. This independence can be measured by the covariance matrix:

$$\mathbf{C_X} \equiv \frac{1}{n} \mathbf{X}\mathbf{X}^\mathbf{T} = \frac{1}{n} \begin{pmatrix} \mathbf{\tilde{x}_1} \\ \vdots \\ \mathbf{\tilde{x}_m} \end{pmatrix} \begin{pmatrix} \mathbf{\tilde{x}_1} & \cdots & \mathbf{\tilde{x}_n} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \mathbf{\tilde{x}_1} \cdot \mathbf{\tilde{x}_1} & \cdots & \mathbf{\tilde{x}_1} \cdot \mathbf{\tilde{x}_m} \\ \vdots & & \vdots \\ \mathbf{\tilde{x}_m} \cdot \mathbf{\tilde{x}_1} & \cdots & \mathbf{\tilde{x}_m} \cdot \mathbf{x_m} \end{pmatrix}.$$

Note that the ijth entry of this matrix is a dot product of the ith variable's measurement vector with the jth variable's, measuring how correlated these two variables are.

[1] J. Shlens, *A Tutorial on Principal Component Analysis*, <http://arxiv.org/pdf/1404.1100.pdf>

If the different variables are uncorrelated, the off-diagonal terms will be zero and the correlation matrix will be diagonal. This is what we desire for the principal components. Thus, the problem of finding principal components has been reduced to finding the transformation matrix $\mathbf{P}$ such that $\mathbf{C_Y}$ is diagonalized.

Since $\mathbf{Y} = \mathbf{PX}$, it is easy to algebraically show that:

$$\mathbf{C_Y} = \mathbf{P C_X P^T}.$$

From several important results about diagonalization of matrices in linear algebra (proved in Shlens' paper and various other sources), we know that if we make the rows of $\mathbf{P}$ the eigenvectors of $\mathbf{C_X}$, then $\mathbf{C_Y}$ will be diagonalized. Thus, our computational steps are:

1. Transform $\mathbf{X}$ as mentioned above so that each row has mean zero.
2. Compute $\mathbf{C_X}$.
3. Find the eigenvectors $\mathbf{p_i}$ of $\mathbf{C_X}$.
4. Construct the transformation matrix out of the eigenvectors:
$$\mathbf{P} = \begin{pmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{pmatrix}.$$
5. Return the matrix $\mathbf{Y} = \mathbf{PX}$, which expresses the data in terms of the principal components.

2 Testing

Two tests of the above algorithm were performed with the functions `test_linear` and `test_multi` in `pca.py`.

For the two-variable test, an uniform range of x_i's was selected, and y_i's were generated by $y_i = \kappa(2+4x_i)$, where κ is a noise factor with mean 1 and standard deviation 0.05. Repeated PCA analysis of various samples yields one principal component with spread ~ 12 and another with spread ~ 0.2.

For the five-variable test, an uniform range of t_i's was selected, and five variables were generated as above – with a linear dependence on t and a noise factor. Repeated PCA analysis yields one principal component with spread ~ 18 and four others with spread below 1.

In each case, the fact that the system had only one non-noise degree of freedom was correctly identified.