An example of Principal Component Analysis

In PCA, one begins with a (large) number k of variables associated with some population, and a set of sample data. The goal is to find a new variable X_1 that is a linear combination of the given variables and that provides the best way (amongst all linear combinations) of separating the members of the population.

STEP 1: Write the data in a table whose columns are labelled by k variables and whose rows give the results of n samples. For an example, consider the data shown below on the frequency of 4 particular letters occur in various languages (this data is taken from `http://en.wikipedia.org/wiki/Letter_frequency`).

Frequency of Letters	a	k	p	t
French	7.6	0.05	3	7.2
German	6.5	1.2	0.8	6.1
Turkish	11.7	4.7	0.8	3.0
Mean	8.6	1.98	1.53	5.43

STEP 2: "Center" the data by subtracting the mean of each column from each entry in that column. The result is a matrix A, with the property that the mean of the entries in each column is 0.

$$A = \begin{pmatrix} -1 & -1.93 & 1.47 & 1.77 \\ -2.1 & -0.78 & -0.73 & 0.67 \\ 3.1 & 2.71 & -0.73 & -2.43 \end{pmatrix}$$

This defines a linear transformation $A : \mathbb{R}^k \to \mathbb{R}^n$. The images $A(\mathbf{e}_i)$ of the basis vectors are a collection of k points in $\mathbb{R}^n$ that can be visualized as a "cloud" centered at the origin.

STEP 3: For the matrix $S = A^T A$ (or alternatively the *covariance matrix* $C = AA^T/(n-1)$ where n is the number of samples).

$$S = \begin{pmatrix} 56.5 & -14.7 & -30.7 & -11.0 \\ -14.7 & 20.6 & 9.3 & -15.2 \\ -30.7 & 9.3 & 22.1 & -0.72 \\ -11.0 & -15.2 & -0.72 & 27.0 \end{pmatrix}$$

STEP 4: Diagonalize S. In fact, since S is symmetric, we know there is an orthogonal basis $\{\mathbf{e}_i\}$ in which S is diagonal with its real eigenvalues $\lambda_1, \cdots, \lambda_k$ on the diagonal. The largest eigenvalue is called the *first principal value* and the span of the corresponding eigenvector is the *first principal direction*. The next larges is the *second principal value*, etc.

For our matrix S above, one finds that the eigenvalues are

$$\lambda_1 = 80.0 \qquad \lambda_2 = 39.9 \qquad \lambda_3 = 6.3 \qquad \lambda_4 = 0$$

and the eigenvectors are

$$\mathbf{e}_1 = \begin{pmatrix} 0.83 \\ -0.26 \\ -0.48 \\ -.09 \end{pmatrix} \qquad \mathbf{e}_2 = \begin{pmatrix} -0.14 \\ -0.57 \\ -0.09 \\ 0.8 \end{pmatrix} \qquad \mathbf{e}_3 = \begin{pmatrix} 0.19 \\ -0.60 \\ 0.71 \\ -.31 \end{pmatrix} \qquad \mathbf{e}_4 = \begin{pmatrix} -0.5 \\ -0.5 \\ -0.5 \\ -.5 \end{pmatrix}$$

STEP 5: The first eigenvector gives the direction of the largest dispersion. The corresponding variable

$$X = 0.83F_a - 0.26F_k - 0.48F_p - 0.9F_t \qquad \begin{cases} F_a = \text{Measured Frequency of a} \\ F_k = \text{Measured Frequency of k} \\ F_p = \text{Measured Frequency of p} \\ F_t = \text{Measured Frequency of t} \end{cases}$$

(that is, $X = \mathbf{e}_1 \cdot F$) is the best distinguish between these three languages using data that samples these 4 letters.

STEP 6: Including more principal directions gives more information. Select a small number m and consider the subspace V_m spanned by the first m eigenvalues of S. The projection onto V_m is given by m variables

$$X_1, \cdots X_m$$

like the one above, and the resulting scatter plot in $\mathbb{R}^m$ is the best m-dimensional way to between these three languages using data that samples these 4 letters.

What should one take for m? This is playoff between amount of data and accuracy. One can get a sense of how much information is captured as follows.

The total standard deviation is $\sigma = \sqrt{\sum \lambda_i}$, and the fraction of the total standard deviation that is accounted for by the variation in the first m principal directions is

$$\sqrt{\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_k}}$$

For the case above, one principal direction already accounts for a large percentage of the standard deviation, and the first two account for even more, namely

$$\sqrt{\frac{80}{80 + 39.9 + 6.3}} \approx 79.9\% \qquad \sqrt{\frac{80 + 39.9}{80 + 39.9 + 6.3}} \approx 97.5\%.$$

More Background. The best-fitting ("maximum likelihood") Gaussian probability distribution for our data then has the form

$$P(\mathbf{x}) = C\, e^{-\frac{\mathbf{x} \cdot S^{-1}\mathbf{x}}{n-1}} = C\, e^{-\frac{|A^{-1}\mathbf{x}|^2}{n-1}}$$

where n is the number of samples ($n = 7$ in the case), S^{-1} is the inverse of the covariance matrix, and C is the constant that makes the integral of this function over all of $\mathbb{R}^3$ equal to 1. If S is diagonal in an orthonormal basis $\{\mathbf{e}_i\}$ with eigenvalues $\{\lambda_i\}$, then T is diagonal with eigenvalues $\{\lambda_i^{-1}\}$, and the above probability distribution has the form

$$P(\mathbf{x}) \;=\; C' \, e^{-\frac{Q(\mathbf{x})}{n-1}}$$

where Q is the quadratic form whose value on $\mathbf{x} = \sum_i x_i \mathbf{e}_i$ is

$$Q(\mathbf{x}) = \sum_i \lambda_i^{-1} x_i^2.$$

Picturing the graph of $P(\mathbf{x})$, one sees that the first eigenvector $\mathbf{e}_1$ gives the direction of the largest spread in the data; this is called the *first principal direction*. Similarly, the first two eigenvectors span the plane with the largest spread in the data, etc.