7.5: Statistics: Mean, Covariance, and Principal Component Analysis

Look at the photos on p. 448. In particular, the first photo in the bottom row has by far the best resolution. How was it constructed? They took the individual pixels, measured the greyscale intensity at each point, and constructed a huge set of vectors in $\mathbb{R}^3$ out of this data. They then found the linear combination of data that accounted for the greatest variation. As the caption indicates, this linear combination accounted for 93.5% of all variation.

In general, suppose we are given a *sample* consisting of a set of N vectors $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N$ in $\mathbb{R}^p$. Think about these vectors as consisting of p different measurements made on the same set of individuals. In the case of the photos, N was the total number of pixels, and for each pixel there was a 3-vector of greyscale values in the different spectra. The data might well be temperature, pressure, and humidity measured at various different points in a region, or height, weight, shoe size, and hat size for a group of males.

The *sample mean* is given by $\mathbf{M} = \frac{1}{N}(\mathbf{X}_1 + \mathbf{X}_2 + ... + \mathbf{X}_N$. Often we transform our data to *mean-deviation form* by subtracting $\mathbf{M}$ from each of the data vectors to obtain vectors $\hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, ...\hat{\mathbf{X}}_N$; of course the sample mean of this new set of vectors is $\mathbf{0}$, by design. See Figure 1 on p. 483 and Figure 2 on p. 484 to see the effect of this transformation. Now form the matrix $B = [\hat{\mathbf{X}}_1\, \hat{\mathbf{X}}_2 \,...\, \hat{\mathbf{X}}_N]$. The *covariance matrix*, also sometimes called the *variance-covariance matrix*, is given by $S = \frac{1}{N-1}BB^T$.

What can we say about S? Well, first of all, S is symmetric, since BB^T is. (Why?) Better yet, S is positive-semidefinite, because for any vector $\mathbf{x} \in \mathbb{R}^p, \mathbf{x}^T(BB^T)\mathbf{x} = (\mathbf{x}^T B)(B^T\mathbf{x}) = (B^T\mathbf{x})^T(B^T\mathbf{x})$, which is the inner product of the vector $B^T\mathbf{x}$ with itself, which is always greater than or equal to 0.

Example. Suppose we have two measurements made on four individuals, giving observation vectors $\mathbf{X}_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \mathbf{X}_3 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \mathbf{X}_4 = \begin{bmatrix} 4 \\ 3 \end{bmatrix}$. Then

$$\mathbf{M} = \frac{1}{4}\left(\begin{bmatrix} 2 \\ 3 \end{bmatrix} + \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ 4 \end{bmatrix} + \begin{bmatrix} 4 \\ 3 \end{bmatrix}\right) = \begin{bmatrix} 2 \\ 3 \end{bmatrix},$$

so $\hat{\mathbf{X}}_1 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \hat{\mathbf{X}}_2 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \hat{\mathbf{X}}_3 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \hat{\mathbf{X}}_4 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, and

$$B = \begin{bmatrix} -1 & 0 & -1 & 2 \\ 0 & -1 & 1 & 0 \end{bmatrix}.$$

Therefore

$$S = \frac{1}{3}BB^T = \frac{1}{3}\begin{bmatrix} -1 & 0 & -1 & 2 \\ 0 & -1 & 1 & 0 \end{bmatrix}\begin{bmatrix} -1 & 0 \\ 0 & -1 \\ -1 & 1 \\ 2 & 0 \end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix} 6 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}.$$

The i^{th} diagonal entry of the S matrix is the *variance* of the variable x_j, that is, the variable of scalars constituting the j^{th} components of the $\mathbf{X}$ vectors. Notice that the variance measures how "spread out" the x_j-values are, since it is calculated by adding the squares of the differences of the x_js from the mean. We can see that in the example: The x_1 values are more spread out than the x_2 values, and that is reflected in a larger entry in the upper left of S than in the lower right. The sum of the diagonal entries is called the *total variance* of the data.

The off-diagonal entries are called the *covariances* between the various component variables. In the example we did, there is only one covariance, namely that between x_1 and x_2. In general, the covariance between x_i and x_j appears in both the i, j and the j, i entries of S. Variables with a covariance of 0 are said to be *uncorrelated*. This means roughly that the way in which one of the variables deviates from its mean has no effect on how the other variable can be expected to deviate from its mean. As an example, let's look aback at Figure 3. The variables $\hat{w}$ and $\hat{h}$ are not uncorrelated here, because for most of the data points the coordinates are either both positive or both negative. This will result in the off-diagonal entries of S being positive.

See also Example 3, pp. 484-5. It's pretty clear from looking at the original data that the x_3 variable is the most spread out, followed by the x_1; that phenomenon is reflected in the size of the diagonal entries. Also, the -8 corresponding to the covariance between x_2 and x_3 comes from the fact that their overall trends are opposite to one another, that is, larger x_2-values tend to come with smaller x_3-values.

From now on, let's assume that our $\mathbf{X}$-variables are already put in mean-deviation form, so that will be our first step anytime we encounter a data set. Our goal is going to be to effect a change of variables so that the resulting variables are uncorrelated; that is, so that the S-matrix gets changed to a diagonal one. But guess what? We already know how to do that! S is symmetric and positive-semidefinite, so we can find an orthogonal change of variables $\mathbf{X} = P\mathbf{Y}$ such that $P^TSP = D$ the diagonal matrix of eigenvalues of S. That is, we can rename each of our data vectors $\mathbf{X}_k$ as $P\mathbf{Y}_k$, which is just the coordinate vector of X_k relative to the columns of P. Now, what is the covariance matrix of $\mathbf{Y}$? Well, we can write $B = [\mathbf{X}_1\, \mathbf{X}_2 \ldots \mathbf{X}_N] = [P\mathbf{Y}_1\, P\mathbf{Y}_2 \ldots P\mathbf{Y}_N] = P[\mathbf{Y}_1\, \mathbf{Y}_2 \ldots \mathbf{Y}_N]$; write this as PC, and note that $C = P^TB$. We want to calculate $\dfrac{1}{N-1}CC^T$, which is equal to $\dfrac{1}{N-1}(P^TB)(P^TB)^T = P^TBB^TP = P^TSP = D$. Now, the unit eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_p$ of S, corresponding to columns of P, are the *principal components* of the data; they are uncorrelated, because they correspond to the pure vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_n$ in the new coordinates, and these vectors are clearly uncorrelated.

All this may sound involved, but here's all you need to bear in mind: Calculate S; find its eigenvalues and eigenvectors; change variables by $\mathbf{X} = P\mathbf{Y}$. Then the new

variables $\mathbf{Y}$, calculated as $P^T\mathbf{X}$, are uncorrelated. What are the component variables $y_1, y_2, ..., y_p$? Well, if $\mathbf{u}_1$ is the first column of P, then $\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix} = P^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$, so y_1 is just $\mathbf{u}_1^T\mathbf{X} = u_1x_1 + u_2x_2 + ... + u_px_p$.

To see how this works, let's look at Example 4 on pp. 486-487; it refers back to the photograph we discussed at the beginning of the hour. They formed S as described above and found eigenvalues and eigenvectors using the appropriate software. Notice that they chose the eigenvalues in descending order. Now, the first eigenvalue is so much bigger than the others that it captures a huge fraction, over 93%, of the total variation in the data.

If time, do #9, p. 489, in groups.