Principal Component Analysis

This expository article is a brief overview of principal component analysis. Recall that the general quadratic form in two variables

$$ax_1 + bx_1x_2 + cx_2^2$$

is equal to

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b/2 \\ b/2 & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

The matrix in the middle is a symmetric matrix; we'll call it S. It has a diagonalization

$$S = Q\Lambda Q^{-1}$$

where Q is orthogonal and Λ is diagonal:

$$Q = [\mathbf{u}_1\ \mathbf{u}_2] \quad \Lambda = \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix}$$

with $\mathbf{u}_1, \mathbf{u}_2$ orthogonal unit vectors. Thus, $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthonormal eigenvectors of A, with eigenvalues λ_1 and λ_2. Furthermore, the λ_i's are real. We arrange the λ_i's so that $\lambda_1 \geq \lambda_2$.

Now we use $\mathbf{u}_1$ and $\mathbf{u}_2$ to create new axes. In terms of the variables u_1 and u_2, the quadratic form

$$z(x_1, x_2) = ax_1^2 + bx_1x_2 + cx_2^2$$

is equal to

$$z(x_1, x_2) = \lambda_1 u_1^2 + \lambda_2 u_2^2.$$

Then λ_1 is the largest value of z on the unit circle and λ_2 is the least value of z on the unit circle.

We will apply these facts to a data set. Suppose there are six people, and two quantities, Quantity 1 and Quantity 2, are measured. We get the following matrix of observations

$$A = \begin{bmatrix} 19 & 22 & 6 & 3 & 2 & 20 \\ 12 & 6 & 9 & 15 & 13 & 5 \end{bmatrix}$$

Given a vector $\langle t_1, ..., t_N \rangle$, we define its **sample mean** to be

$$m = \frac{1}{N}(t_1 + \cdots + t_N)$$

and the **sample variance** to be

$$\frac{1}{N-1}\sum_{i=1}^{N}(t_i - m)^2 = \frac{1}{N}\langle t_1 - m, ..., t_N - m\rangle \cdot \langle t_1 - m, ..., t_N - m\rangle.$$

For example, in row 1, $N = 6$, and

$$m = \frac{1}{6}(19 + 22 + 6 + 3 + 2 + 20) = 12.$$

In row 2, $N = 6$ and $m = 10$.

We form the matrix B by subtracting the row means from each row:

$$B = \begin{bmatrix} 7 & 10 & -6 & -9 & -10 & 8 \\ 2 & -4 & -1 & 5 & 3 & -5 \end{bmatrix}$$

Suppose that $\mathbf{u}$ and $\mathbf{v}$ are vectors in $\mathbf{R}^N$ with mean 0. We define the **sample covariance** of $\mathbf{u}$ and $\mathbf{v}$ to be

$$\frac{1}{N-1}\mathbf{u} \cdot \mathbf{v}.$$

Notice that the sample covariance of $\mathbf{u}$ and $\mathbf{u}$ is just the sample variance of $\mathbf{u}$.

We define the **sample covariance matrix** to be

$$S = [s_{ij}] = [\mathrm{Cov}(\mathbf{u}_i, \mathbf{u}_j)].$$

Then

$$S = \frac{1}{N-1}BB^T.$$

Notice that S is symmetric in general. In the example above,

$$S = \begin{bmatrix} 86 & -27 \\ -27 & 16 \end{bmatrix}$$

By definition, the sample variance of $c\mathbf{u} + d\mathbf{v}$ is $\frac{1}{N-1}(c\mathbf{u} + d\mathbf{v}) \cdot (c\mathbf{u} + d\mathbf{v})$. Now $c\mathbf{u} + d\mathbf{v}$ is precisely

$$\begin{bmatrix} c & d \end{bmatrix} B,$$

so the sample variance of $c\mathbf{u} + d\mathbf{v}$ is

$$\frac{1}{N-1}[c\ d]BB^T \begin{bmatrix} c \\ d \end{bmatrix},$$

a quadratic form! Let λ_1 and λ_2 be the eigenvalues of S, with $\lambda_1 \geq \lambda_2$. Suppose we restrict (c, d) to be a unit vector. By what we have seen:

- The largest sample variance is λ_1, occurring at $\mathbf{u}_1$;
- The least sample variance is λ_2, occurring at $\mathbf{u}_2$.

The **total variance** is just the sample variance of $\mathbf{u}$ plus the sample variance of $\mathbf{v}$. This is the sum of the diagonal entries of the sample covariance matrix. In other words, it's the *trace* of the sample covariance matrix. By the general theory, this is the sum of the eigenvalues. The total variance of the data is $\lambda_1 + \lambda_2$. We can phrase what we've found as follows:

- Out of all possible unit vectors, the direction $\mathbf{u}_1$ captures the largest amount of variance in the data.

- The direction $\mathbf{u}_2$ captures the next largest amount of variance in the data in the possible directions orthogonal to $\mathbf{u}_1$.
- The percentage of the total variance captured by the direction $\mathbf{u}_i$ is $\dfrac{\lambda_i}{\sum_{j=1}^{N}\lambda_j}$.

In this scenario, we get

$$\mathbf{u}_1 = \begin{bmatrix} .95 \\ -.32 \end{bmatrix}, \quad \lambda_1 = 95.2$$

and

$$\mathbf{u}_2 = \begin{bmatrix} .32 \\ .95 \end{bmatrix}, \quad \lambda_2 = 6.8.$$

The total variance in the data is $95.2 + 6.8 = 102$. (Check that this is, in fact, the sum of the diagonal entries of S!)

The direction $\mathbf{u}_1$ captures $\dfrac{95.2}{102} \approx 93.3\%$ of the total variance. The direction $\mathbf{u}_2$ captures approximately 6.7% of the total variance.

We call $\mathbf{u}_1$ and $\mathbf{u}_2$ the **principal components** of the data (in the matrix of observations A). The **first principal component** is the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix S, the **second principal component** is the eigenvector corresponding to the second largest eigenvalue, and so on. *Principal component analysis* is based on the spectral theorem for symmetric matrices.

We have learned about lines of best fit using least squares. Principal component analysis is equivalent to something called orthogonal regression, as we now explain. Recall that if B has row means of 0, we have the sample covariance matrix $S = \dfrac{1}{N-1}BB^T$. Suppose that $\mathbf{u}$ is a vector in $\mathbf{R}^N$. Then

$$\begin{aligned}
\mathbf{u}^T S\mathbf{u} &= \frac{1}{N-1}\left([x_1\ x_2][\mathbf{b}_1\ \mathbf{b}_2\ \ldots\ \mathbf{b}_N]\right)\left([x_1\ x_2][\mathbf{b}_1\ \mathbf{b}_2\ \ldots\ \mathbf{b}_N]\right)^T \\
&= \frac{1}{N-1}[\mathbf{b}_1\cdot\mathbf{u}\ \ldots \mathbf{b}_N\cdot\mathbf{u}]\begin{bmatrix}\mathbf{b}_1\cdot\mathbf{u}\\ \vdots \\ \mathbf{b}_N\cdot\mathbf{u}\end{bmatrix} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{b}_i\cdot\mathbf{u})^2.
\end{aligned}$$

Now suppose that $\mathbf{u}$ is a unit vector. Then $\sum_{i=1}^{N}(\mathbf{b}_i\cdot\mathbf{u})^2$ is the sum of the squares of the lengths of the projections of the $\mathbf{b}_i$ on the line spanned by $\mathbf{u}$.

Notice that

$$||\mathbf{b}_i||^2 = (\mathbf{b}_i\cdot\mathbf{u})^2 + (\mathbf{b}_i - \mathbf{b}_i\cdot\mathbf{u})^2$$

expresses the square of the magnitude of $\mathbf{b}_i$ as the square of the length of the projection of $\mathbf{b}_i$ onto $\mathbf{u}$ and the square of the distance from $\mathbf{b}_i$ to the line spanned by $\mathbf{u}$. Therefore,

$$\sum_{i=1}^{N}||\mathbf{b}_i||^2 = \sum_{i=1}^{N}(\mathbf{b}_i\cdot\mathbf{u})^2 + \sum_{i=1}^{N}\mathrm{dist}(\mathbf{b}_i,\ell_{\mathbf{u}})^2.$$

Since the left side is a fixed constant (because the $\mathbf{b}_i$'s are given, we see that maximizing $\sum_{i=1}^{N}(\mathbf{b}_i\cdot\mathbf{u})^2$ is equivalent to minimizing $\sum_{i=1}^{N}\mathrm{dist}(\mathbf{b}_i,\ell_{\mathbf{u}})^2$. But maximizing $\sum_{i=1}^{N}(\mathbf{b}_i\cdot\mathbf{u})^2$ is

equivalent to maximizing $\mathbf{u}^T S\mathbf{u}$! We have seen that the $\mathbf{u}$ that does this is the first principal component. Therefore, the line $t\mathbf{u}$ is the best approximation to the data, in the sense that the sum of the squares of the *orthogonal* distances to the line is minimized. For this reason, principal component analysis is equivalent to what is termed *orthogonal regression*.

June 29, 2025