CLARK UNIVERSITY

Principal component analysis
Math 218, Mathematical Statistics

D Joyce, Spring 2016

The data. We start with n observations $\mathbf{x}_{i*}$, $i = 1, \ldots, n$. Each observation consists of k components $\mathbf{x}_{i*} = (x_{i1}, \ldots, x_{ik})$, so altogether we have n points in $\mathbf{R}^k$, the data space.

For example, you might ask $n = 100$ people $k = 10$ questions, and each question can be answered on some linear scale (like from 1 to 5). That gives you 100 points in $\mathbf{R}^{10}$. The $n = 100$ people may be stratified (i.e., classified) in one or more ways. If they're college students, they might be stratified by class year: freshman, sophomore, junior, or senior, and you may be interested in how their responses vary by class year.

This information can be collected in one $n \times k$ matrix X whose ij^{th} entry is the j^{th} component of the i^{th} observation. Each of the n rows is one of the observations $\mathbf{x}_{i*}$. (Some authors reverse the orientation of the matrix.)

The j^{th} column in this matrix gives the j^{th} data components for all n observations, $\mathbf{x}_{*j} = (x_{1j}, \ldots, x_{nj})$. If you like, you can standardize the data in each of these data components. You could subtract the sample average,

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

of the n j^{th} components from each x_{ij}. Making the mean 0 simplifies later computations.

The data component $\mathbf{x}_{*j}$ also has a sample variance

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j).$$

If you divide $\mathbf{x}_{*j}$ by the sample standard deviation s_j, then the standardized data for each of the k coordinates has standard deviation 1. By making the standard deviation 1, it makes each of the coordinates equally important, however you may want to treat some data components as more important than others.

Principal component analysis (PCA). This analysis will find an orthogonal coordinate system in $\mathbf{R}^k$ so that the first coordinate, the first principal component, accounts for much of the variation in the data, the second accounts for less variation, and so on. Usually two or three principal components account for most of the variation, the remaining ones a smaller amount. Choosing just the first two principal components, you can display the data on a planar graph, while three gives a display on a spacial graph.

You can then use your stratifications to visually tell if there's any difference among the strata. Color code the dots on your graph according to the strata—one color for freshman, one for sophomores, etc.

The covariance matrix. The mathematics involves some linear algebra using the covariance matrix. It has various other names including variance-covariance matrix.

The sample covariance matrix S is defined as the $k \times k$ matrix whose j^{th} diagonal entry S_{jj} is the sample variance

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$$

of the j^{th} data component, while the off-diagonal entry $s_{j_1 j_2}$ is sample covariance

$$s_{j_1 j_2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij_1} - \overline{x}_{j_1})(x_{ij_2} - \overline{x}_{j_2})$$

between the j_1^{st} and the j_2^{nd} data components.

If the components have been standardized by subtracting their means, there's an easier way to

define S using matrix multiplication. Namely, S can be defined as the product

$$S = \frac{1}{n-1} X^{\mathsf{T}} X$$

where X^{T} denotes the transpose of the matrix X.

However you define it, S is a symmetric matrix. Symmetric matrices are all diagonalizable and they have real eigenvalues. Furthermore, $X^{\mathsf{T}}X$ is a positive semidefinite matrix. That means that the eigenvalues of a matrix $X^{\mathsf{T}}X$ are all nonnegative. We'll denote the eigenvalues in decreasing order

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0.$$

Finally, the eigenspaces corresponding to these eigenvalues are all orthogonal.

The principal components. The λ_1-eigenspace is called the first principal component, the λ_2-eigenspace is called the second principal component, etc. When the data points are given these component coordinates, since the eigenspaces are orthogonal, they'll be independent.

The eigenvalues themselves indicate how much of the variance in the data comes from variance in the corresponding principal component.

Geometric interpretation. The first principal component is a line in $\mathbf{R}^k$. Of all the lines through the origin, it's the one spreads the data out most when you project the data on to it.

If you then subtract from each data point its projection on to this line (the first eigenspace), the result will be a point on the $k-1$-dimensional hyperplane perpendicular to the line. The original point has been projected onto that hyperplane.

Now, if you take that reduced data as a starting point and find its principal component, you're actually finding the second principal component of the original data. You'll find the line in that hyperplane that spreads out the reduced data the most when the reduced data is projected on it.

Again, you can subtract from each reduced data point its projection onto this second eigenspace, you'll get a point in a $k-2$-dimensional subspace. And so forth. Each stage you're squeezing more information out of the original data.

Dimension reduction. The different k principal components account for different amounts of variance in the data, each dimension less than the previous. Ignoring the last few dimensions won't lose a lot of information. The first two or three will contain a lot of it. You can display the data in those two or three dimensions to get an intuition for the data, perhaps enough to answer your questions or spur you into asking different questions.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`