Principal component analysis with linear algebra

Jeff Jauregui

August 31, 2012

Abstract

We discuss the powerful statistical method of principal component analysis (PCA) using linear algebra. The article is essentially self-contained for a reader with some familiarity of linear algebra (dimension, eigenvalues and eigenvectors, orthogonality). Very little previous knowledge of statistics is assumed.

1 Introduction to the problem

Suppose we take n individuals, and on each of them we measure the same m variables. We say that we have n samples of m-dimensional data. For the ith individual, record the m measurements as a vector $\vec{x}_i$ belonging to $\mathbb{R}^m$.

For instance, we might ask 30 people their height (in meters), their weight (in kilograms), and their IQ. In this case, $n = 30$ and $m = 3$. The measurement $\vec{x}_1$ might look like

$$\vec{x}_1 = \begin{bmatrix} 1.8 \\ 70.3 \\ 105 \end{bmatrix}.$$

You could visualize this data as a plot of 30 points in $\mathbb{R}^3$.

Principal component analysis, or PCA, is a powerful statistical tool for analyzing data sets and is formulated in the language of linear algebra. Here are some of the questions we aim to answer by way of this technique:

1. Is there a simpler way of *visualizing* the data (which *a priori* is a collection of points in $\mathbb{R}^m$, where m might be large)? For instance, in the above example, are the points in $\mathbb{R}^3$ essentially clustered around a plane?

2. Which variables are *correlated*? In the example, we would probably expect to see little correlation between height and IQ, but some correlation between height and weight.

3. Which variables are the most *significant* in describing the full data set? Later, we will see more precisely what this means.

2 Linear algebra background

Let A be an $m \times n$ matrix of real numbers and A^T its transpose. The following theorem is one of the most important in linear algebra.

Theorem 1. *If A is symmetric (meaning $A^T = A$), then A is orthogonally diagonalizable and has only real eigenvalues. In other words, there exist real numbers $\lambda_1, \ldots, \lambda_n$ (the eigenvalues) and orthogonal, non-zero real vectors $\vec{v}_1, \ldots, \vec{v}_n$ (the eigenvectors) such that for each $i = 1, 2, \ldots, n$:*

$$A\vec{v}_i = \lambda_i \vec{v}_i.$$

This is a very powerful result (often called the Spectral Theorem), but it is limited by the fact that it applies only to symmetric matrices. Nevertheless, we can still get some use out of the theorem in general with the following observation:

Exercise 1. *If A is any $m \times n$ matrix of real numbers, then the $m \times m$ matrix AA^T and the $n \times n$ matrix A^TA are both symmetric.*

Thus, we can apply the theorem to the matrices AA^T and A^TA. It is natural to ask how the eigenvalues and eigenvectors of these matrices are related.

Proposition 1. *The matrices AA^T and A^TA share the same **nonzero** eigenvalues.*

Proof. Let $\vec{v}$ be a (nonzero) eigenvector of A^TA with eigenvalue $\lambda \neq 0$. This means:

$$(A^TA)\vec{v} = \lambda \vec{v}.$$

Now, multiply both sides on the left by A, and group the parentheses as follows:

$$AA^T(A\vec{v}) = \lambda(A\vec{v}).$$

This is precisely the statement that the vector $A\vec{v}$ is an eigenvector of AA^T, with eigenvalue λ. The only technical point we must check is that $A\vec{v}$ is not the zero vector (since eigenvectors aren't allowed to be zero). But from the first equation, if $A\vec{v}$ were zero, then $\lambda\vec{v}$ would be zero as well. However, we specifically said that $\vec{v} \neq \vec{0}$ and $\lambda \neq 0$, so this can't happen.

We conclude that the nonzero eigenvalue λ of A^TA is also an eigenvalue of AA^T. Moreover, we learned that to get from an eigenvector $\vec{v}$ of A^TA to an eigenvector of AA^T, you just multiply $\vec{v}$ on the left by A. (And it is worth checking that to get from an eigenvector $\vec{w}$ of AA^T to an eigenvector of A^TA, you just multiply $\vec{w}$ on the left by A^T.) □

This proposition is very powerful in the case that m and n are drastically different in size. For instance, if A is 500×2, then there's a quick way to find the eigenvalues of the 500×500 matrix AA^T: first find the eigenvalues of A^TA (which is only 2×2). The other 498 eigenvalues of AA^T are all zero!

Proposition 2. *The eigenvalues of AA^T and A^TA are nonnegative numbers.*

Proof. First, recall that length squared of a vector $\vec{w}$ is given by the dot product $\vec{w} \cdot \vec{w}$, which equals $\vec{w}^T\vec{w}$.

Let $\vec{v}$ be an eigenvector of A^TA with eigenvalue λ. We compute the length squared of $A\vec{v}$:

$$\begin{aligned} \|A\vec{v}\|^2 &= (A\vec{v})^T(A\vec{v}) \\ &= \vec{v}^T(A^TA)\vec{v} \\ &= \lambda\vec{v}^T\vec{v} \\ &= \lambda\|\vec{v}\|^2. \end{aligned}$$

Since lengths are nonnegative, we see that λ is nonnegative. Replacing A with A^T, we get the corresponding statement for AA^T. □

3 Statistics background

Suppose we're measuring a single variable A (such as the height of randomly selected individuals) n times (so $m = 1$). Let the n measurements be denoted $a_1, \ldots, a_n$. The most basic quantity in statistics is the *mean* of a variable A. However, the mean is rarely known in practice, so we estimate the mean using the *sample average*:

$$\mu_A = \frac{1}{n}(a_1 + \ldots + a_n).$$

In this article, we will be a little sloppy and not distinguish between the mean and sample average.

The mean tells us where the measurements are centered. The next question we would like to ask is: how spread out are the measurements? This is commonly quantified with the *variance* of A (again, generally unknown in practice), which is estimated by the sample variance:

$$\mathrm{Var}(A) = \frac{1}{n-1}\left((a_1 - \mu_A)^2 + \ldots + (a_n - \mu_A)^2\right).$$

The square root of the variance, is called the *standard deviation*, but we will only use the variance. We will not distinguish here between the variance and the sample variance.

What if we're measuring two variables A, B in a population? It's natural to ask if there's some relationship between A and B. (For instance, you'd expect to see a significant relationship between height and weight, but not necessarily height and IQ.) One way to capture this is with the *covariance* of A and B, defined as:

$$\mathrm{Cov}(A, B) = \frac{1}{n-1}\left((a_1 - \mu_A)(b_1 - \mu_B) + \ldots + (a_n - \mu_A)(b_n - \mu_B)\right).$$

If the covariance is negative, it indicates that when variable A is larger, variable B tends to be smaller. Also, notice that $\mathrm{Cov}(A, B) = \mathrm{Cov}(B, A)$.

If we're measuring three or more variables, notice that we can talk about the variance of any variable, and covariance of any two different variables.

4 Principal component analysis

Using the notation from the introduction, we can store the mean of all m variables as a single vector in $\mathbb{R}^m$:

$$\vec{\mu} = \frac{1}{n}\left(\vec{x}_1 + \ldots + \vec{x}_n\right). \tag{1}$$

It's common to "re-center" the data so that the mean is zero. (In other words, shift the cluster of data points in $\mathbb{R}^m$ so their center of mass is the origin.) This is accomplished by subtracting the mean $\vec{\mu}$ from each sample vector $\vec{x}_i$. Let B be the $m \times n$ matrix whose ith column is $\vec{x}_i - \vec{\mu}$:

$$B = [\vec{x}_1 - \vec{\mu} \,|\, \ldots \,|\, \vec{x}_n - \vec{\mu}\,]. \tag{2}$$

Define the *covariance matrix* S (which will be $m \times m$) as

$$S = \frac{1}{n-1} BB^T. \tag{3}$$

By exercise 1, we see that S is symmetric. Let's investigate what the entries of S look like in an example. Suppose

$$\vec{x}_1 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix}, \quad \vec{x}_2 = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}, \quad \vec{x}_3 = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}, \quad \vec{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix},$$

so that

$$B = \begin{bmatrix} a_1 - \mu_1 & b_1 - \mu_1 & c_1 - \mu_1 \\ a_2 - \mu_2 & b_2 - \mu_2 & c_2 - \mu_2 \\ a_3 - \mu_3 & b_3 - \mu_3 & c_3 - \mu_3 \\ a_4 - \mu_4 & b_4 - \mu_4 & c_4 - \mu_4 \end{bmatrix}.$$

Then, for instance, the $1,1$ entry of S is

$$S_{11} = \frac{1}{3-1}\left((a_1 - \mu_1)^2 + (b_1 - \mu_1)^2 + (c_1 - \mu_1)^2\right),$$

which is precisely the variance of the first variable. As another example, consider the $2,1$ entry of S:

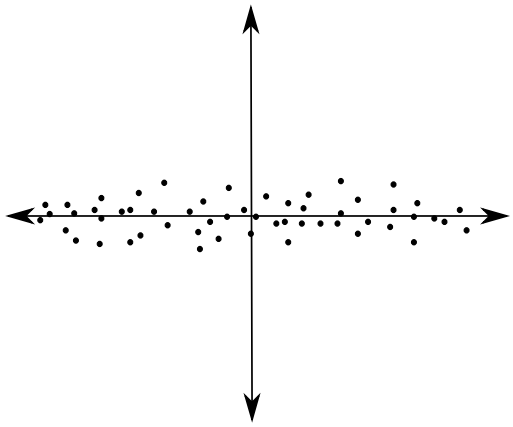
$$S_{21} = \frac{1}{3-1}\left((a_1 - \mu_1)(a_2 - \mu_2) + (b_1 - \mu_1)(b_2 - \mu_2) + (c_1 - \mu_1)(c_2 - \mu_2)\right),$$

which is the covariance of the first and second variables.

We generalize these observations as follows. For $1 \leq i, j \leq m$:

- The ith entry on the diagonal of S, namely S_{ii}, is the variance of the ith variable.
- The ijth entry of S, S_{ij}, with $i \neq j$, is the covariance between the ith and jth variables.

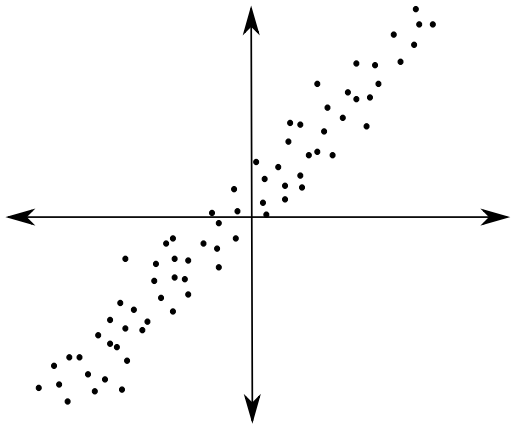
Example 1. *Suppose that $m = 2$, so that each of the n individuals' measurements forms a point in $\mathbb{R}^2$. Suppose the plot of the n data points looks like:*



In the first variable (corresponding to the horizontal axis), there is quite a lot of variation, so we expect S_{11} to be large. However, in the second variable (corresponding to the vertical axis), there is little variation by comparison; we expect S_{22} to be much smaller. How about the covariance? Well, there is very little relationship between the two variables: knowing where you are on the horizontal axis tells you essentially nothing about where you are on the vertical axis. So, perhaps our covariance matrix might look like

$$S = \begin{bmatrix} 95 & 1 \\ 1 & 5 \end{bmatrix}.$$

On the other hand, suppose our data points look like:



Now the horizontal and vertical directions have approximately the same variance, and there is a strong, positive correlation between the two. So the covariance might look something like:

$$S = \begin{bmatrix} 50 & 40 \\ 40 & 50 \end{bmatrix}.$$

The two data sets of the last example in some sense are very similar: they both essentially form a thin rectangular strip, clustered along a line. However, their covariance matrices are completely different. **PCA will provide a mechanism to recognize this geometric similarity through algebraic means.**

Since S is a symmetric matrix, it can be orthogonally diagonalized by Theorem 1. This connection between statistics and linear algebra is the beginning of PCA.

Apply the theorem, and let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$ be the eigenvalues of S (in decreasing order) with corresponding orthonormal eigenvectors $\vec{u}_1, \ldots, \vec{u}_m$. These eigenvectors are called the *principal components* of the data set. (Remark: you can always replace any of the $\vec{u}_i$ with their negatives.)

Observation: on one hand, the trace of S is the sum of the diagonal entries of S, which is the sum of the variances of all m variables. Let's call this the *total variance*, T of the data. On the other hand, the trace of a matrix is equal the sum of its eigenvalues, so $T = \lambda_1 + \ldots + \lambda_m$.

The following interpretation is fundamental to PCA:

- The direction in $\mathbb{R}^m$ given by $\vec{u}_1$ (the first principal direction) "explains" or "accounts for" an amount λ_1 of the total variance, T. What fraction of the total variance? It's $\frac{\lambda_1}{T}$. And similarly, the second principal direction $\vec{u}_2$ accounts for the fraction $\frac{\lambda_2}{T}$ of the total variance, and so on.

- Thus, the vector $\vec{u}_1 \in \mathbb{R}^m$ points in the most "significant" direction of the data set.

- Among directions that are orthogonal to $\vec{u}_1$, $\vec{u}_2$ points in the most "significant" direction of the data set.

- Among directions orthogonal to both $\vec{u}_1$ and $\vec{u}_2$, the vector $\vec{u}_3$ points in the most significant direction, and so on.

Below we describe one of the possible uses of this technique. The example on birds in the next section indicates additional uses.

4.1 Dimension reduction

It is often the case that the largest few eigenvalues of S are much greater than all the others. For instance, suppose $m = 10$, the total variance T is 100, and $\lambda_1 = 90.5, \lambda_2 = 8.9$ and $\lambda_3, \ldots, \lambda_{10}$ are all less than 0.1. This means that the first and second principal directions explain 99.4% of the total variation in the data. Thus, even though our

data points might form some cloud in $\mathbb{R}^{10}$ (which seems impossible to visualize), PCA tells us that these points cluster near a two-dimensional plane (spanned by $\vec{u}_1$ and $\vec{u}_2$). In fact, the data points will look something like a rectangular strip inside that plane, since λ_1 is a lot bigger than λ_2 (similar to the previous example). We have effectively reduced the problem from ten dimensions down to two.

Warning: don't forget that we subtracted μ from the vectors $\vec{x}_1, \ldots \vec{x}_n$. Then to make the last statement completely accurate, the data points would be clustered around the plane passing through μ and spanned by directions parallel to $\vec{u}_1$ and $\vec{u}_2$.

4.2 Summary

In short, here is how to perform PCA on a data set.

1. Gather the n samples of m-dimensional data $\vec{x}_1, \ldots, \vec{x}_n$, vectors in $\mathbb{R}^m$. Compute the mean μ (equation (1)), build the matrix B (equation (2)), and compute S (equation (3)).

2. Find the eigenvalues $\lambda_1, \ldots, \lambda_m$ of S (arranged in decreasing order), as well as an orthogonal set of eigenvectors $\vec{u}_1, \ldots, \vec{u}_m$.

3. Interpret the results: are a small number of the λ_i much bigger than all the others? If so, this indicates a dimension reduction is possible. Which of the n variables are most important in the first, second, etc. principal components? Which factors appear with the same or opposite sign as others?

The last couple of questions will become clearer after the main example on birds in section 5.

First we return to example 1. In the first case, in which

$$S = \begin{bmatrix} 95 & 1 \\ 1 & 5 \end{bmatrix},$$

we find (using a computer) that (approximately) $\lambda_1 = 95.011$, $\lambda_2 = 4.99$, and

$$\vec{u}_1 = \begin{bmatrix} 0.9999 \\ 0.0111 \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} -0.0111 \\ 0.9999 \end{bmatrix}.$$

So in particular, $\vec{u}_1$ essentially points along the x-axis, the "main direction" of the data, and $\vec{u}_2$ essentially points along the y-axis.

Exercise 2. *In the second case of example 1, in which*

$$S = \begin{bmatrix} 50 & 40 \\ 40 & 50 \end{bmatrix},$$

verify (by hand) that $\lambda_1 = 90$, $\lambda_2 = 10$, *and*

$$\vec{u}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}.$$

Draw these vectors in the corresponding figure, and verify that $\vec{u}_1$ *points in the "main direction" of the data.*

5 Bird example

This example and data is courtesy of Adam Kapelner, from Wharton Statistics. Adam used Sibley's Bird Database of North American birds to gather data on a simple random sample of 100 bird species. Three factors were measured: length (inches), wingspan (inches), and weight (ounces). Thus, $m = 3$ and $n = 100$, so B is a 3×100 matrix, and S is 3×3, given below:

$$S = \begin{bmatrix} 91.43 & 171.92 & 297.99 \\ & 373.92 & 545.21 \\ & & 1297.26 \end{bmatrix}.$$

As is customary, the entries below the diagonal were omitted, since the matrix is symmetric. Also, S was computed without dividing by $n-1$ (also a common practice).

We can use MATLAB or octave[1], for instance, to compute the eigenvalues and orthonormal eigenvectors. In this case:

$$\lambda_1 = 1626.52, \qquad \lambda_2 = 128.99, \qquad \lambda_3 = 7.10$$

and

$$\vec{u}_1 = \begin{bmatrix} 0.22 \\ 0.41 \\ 0.88 \end{bmatrix}, \quad \vec{u}_2 = \begin{bmatrix} 0.25 \\ 0.85 \\ -0.46 \end{bmatrix}, \quad \vec{u}_3 = \begin{bmatrix} 0.94 \\ -0.32 \\ -0.08 \end{bmatrix}.$$

The first thing to notice is that λ_1 is much larger than λ_2 and λ_3. In fact, the first principal component $\vec{u}_1$ accounts for $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = 92.28\%$ of the variation in the data, and the second $\vec{u}_2$ accounts for 7.32%. The remaining principal component, explaining only 0.40% of the data, is negligible compared to the first two.

Now, how to interpret all of this? In studying the sizes (length, wingspan, weight) of North American birds, there are apparently only two factors that are important (corresponding to $\vec{u}_1$ and $\vec{u}_2$). We might think of $\vec{u}_1$ as giving a generalized notion of "size" that incorporates length, wingspan, and weight. Indeed, all three entries of $\vec{u}_1$ have the same sign, indicating that birds with larger "size" tend to have larger length, wingspan, and weight.

[1] octave is a free, open source alternative to MATLAB.

We could also ask: which of the factors (length, wingspan, weight) is most significant in determining a bird's "size"? In other words, does the first principal component $\vec{u}_1$ point the most in the direction of the length axis, the wingspan axis, or the weight axis in $\mathbb{R}^3$? Well, the third entry, weight, of $\vec{u}_1$ is the largest, so weight is the most significant. This means a change in one unit of weight tends to affect the size more so than a change in one unit of length or wingspan. The second entry of $\vec{u}_1$ is the next largest, which corresponds to wingspan. Thus, wingspan is the next most important factor in determining a bird's size (followed lastly by length).

Now, what does the second principal component mean? It is mostly influenced by wingspan and weight, as these entries in $\vec{u}_2$ have the greatest absolute values. However, they also have opposite signs. This indicates that $\vec{u}_2$ describes a feature of birds corresponding to relatively small wingspan and large weight, or vice versa. We might call this quality "stoutness."

For each of these birds, is the "size" large or small? Is the degree of "stoutness" large or small?

In other words, to a very good approximation, this sample of North American birds is described by only two parameters: the "size" (most important) and the "stoutness" (less important).

6 Eigenfaces: facial recognition

A very exciting application of PCA that was first considered by Sirovich and Kirby [2] and implemented by Turk and Pentland [3] is to the field of human facial recognition by computers. The basic idea of this "Eigenface" method is to collect a database of (black and white, say) images of faces of a wide variety of people, say n of them. Store each image as a huge vector of length m by sticking the second column of pixels below the first column, then the third column below that, etc. Each entry represents the brightness of a pixel (say where 0.0 is black and 1.0 is white). So we can view every possible face as a vector in some $\mathbb{R}^m$, where m is very large. Running PCA on this data set gives us principal components, which can be converted back into images. These are what are known as *eigenfaces*, each of which is a composite (linear combination) of faces from the data set. Generally, it is found that a relatively small number of eigenfaces (say around 100) are needed to give a basis, up to a good approximation,

of all other faces in the data set. (In terms of PCA, the first 100 variances were much larger compared to the other remaining ones.)

In Spring 2012, I performed an eigenface demonstration using my students' roster photos for the database. Here is what the first eigenface (i.e., the first principal component) looked like:

How could this be used for face recognition? Suppose you are a casino operator and have a database of images of everyone that is banned from your casino. For each of these, you use the Gram–Schmidt method (i.e., orthogonal projection) to approximate the image as a linear combination of eigenfaces. For each individual that walks through the doors, your security system would take a picture of their face, break that down into its eigenface components, and compare it to all images in the database. If it is close enough to one in the database, perhaps that person is banned from the casino.

A number of resources for and examples of the implementation of Eigenfaces are available on the web.

For additional reading on PCA in general, we refer the reader to section 7.5 of Lay's textbook [1]. We also encourage the reader to search for additional applications of PCA (of which there are many).

References

[1] D. Lay, *Linear Algebra and its applications*, 4th ed., Pearson, 2012.

[2] L. Sirovich and M. Kirby, *Low-dimensional procedure for the characterization of human faces*, Journal of the Optical Society of America A, vol. 4, no. 3, 1987.

[3] M. Turk and A. Pentland, *Eigenfaces for recognition*, Journal of Cognitive Neuroscience vol. 3., no. 1.