COMPUTATIONAL BIOLOGY

Introduction to Statistical Methods to Analyze Large Data Sets: Principal Components Analysis

Neil R. Clark and Avi Ma'ayan*

This Teaching Resource provides lecture notes, slides, and a problem set for a series of lectures from a course entitled "Systems Biology: Biomedical Modeling." The materials are a lecture introducing the mathematical concepts behind principal components analysis (PCA). The lecture describes how to handle large data sets with correlation methods and unsupervised clustering with this popular method of analysis, PCA.

Introduction

This Teaching Resource is intended for use by instructors who have some knowledge of statistics and linear algebra. This introductory material is appropriate for those with limited experience in math and statistics. Familiarity with the R programming environment is useful. Principal components analysis (PCA) is an analysis method that is increasingly gaining acceptance for use in high-dimensional data integration and analysis.

Lecture Notes

PCA: An Introduction

PCA is a method of extracting information from data that keeps only what is most important and finds the underlying trends. The data may be high dimensional and of a random nature, which can make the patterns difficult to see. A simplified example of the type of data that this method can be applied to comes from the following hypothetical experiment, which generated the data below:

$$\textit{Genes}\left\{\overbrace{\begin{pmatrix} 0.1 & 0.5 & 0.3 & 0.5 \\ 2.2 & 1.5 & 0.5 & 0.1 \\ 1.2 & 1.0 & 2.0 & 0.3 \\ 0.3 & 2.4 & 2.0 & 7.0 \\ 0.2 & 0.1 & 1.0 & 0.2 \end{pmatrix}}^{\textit{Experiments}}\right.$$

These data show the relative extents of expression of five different genes in four experiments. We placed the data in a table of numbers enclosed in parentheses in anticipation of using matrices to handle our data (Slide 2). This is a simplified example, and real data, for example, from a microarray experiment, would have more rows for the analysis of more genes. In each experimental observation, we may measure m variables, which might be a large number, and we expect that some of the variables might be codependent. With PCA, we can find a small number of new variables that mostly describe the variation within the data. These new variables will be independent of each other, and they will be created from a linear combination of the original variables. We may also be able to interpret the meaning of these new variables and to understand the original data in terms of the new variables.

We will demonstrate how PCA works by performing the analysis on the simple data set. With this example, we will understand the workings of the method. Here, we will perform the analysis by hand to see how it works; however, when you apply the method to experimental data, we recommend the use of MATLAB or other similar software. Before we proceed to perform the PCA, we need to briefly cover two mathematical concepts, one from basic statistics and the other from matrix or linear algebra.

Basic Statistical Concepts

Statistics extract trends from data wherever there is some randomness and uncertainty. In many areas of science, it is not possible to know or measure, given some initial conditions, exactly what will happen to the system's variables over time; however, it is possible to predict the general trend of the behavior with statistics. Statistics are critical in biology because experiments do not produce exact outcomes and they are subject to measurement errors and biological noise. Measurement errors in experiments are a source of randomness because nothing can be measured perfectly accurately and because the error in measurement can be random or biased. In biology, and specifically in regulatory molecular systems biology, it is well established that noise or randomness plays a critical role in regulatory mechanisms.

We will introduce three variables that describe the trends in some random data (Slide 3). These are (i) the mean, defined as the middle of the data; (ii) the variance, which is the spread of the data; and (iii) the covariance, which is defined as the degree of codependence of two variables. We will cover each of these variables in turn.

The Mean

To make things clear, we will look at an example (Slide 4). We have a set of numbers, which we will call S, which is defined as

$S = (1.1 \quad 0.5 \quad 2.6 \quad 0.3 \quad 2.0)$

In this set, we have five numbers. We can refer to an individual element of our set of numbers, S_i, as the i^{th} element of the set. So, for example we can pick out the third number as follows:

$S_3 = 2.6$

These data could be from five experiments to measure some

Department of Pharmacology and Systems Therapeutics and Systems Biology Center New York, Mount Sinai School of Medicine, New York, NY 10029, USA.

*Corresponding author. E-mail, avi.maayan@mssm.edu

quantity; for example, they could be from several measurements of the relative abundance of a protein in a specific cell line after treatment with different drugs. If we wanted to approximate the typical amount of the protein, we can compute the mean. The mean of any set, S, is written as $\bar{S}$, and is given by

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n} S_i$$

where n is the number of elements in the set. We can calculate the mean of the example set above as follows:

$$\begin{aligned}\bar{S} &= \frac{1}{5}(1.1 + 0.5 + 2.6 + 0.3 + 2.0)\\ &= 1.3\end{aligned}$$

Note that we never measured the protein abundance as 1.3 units; however, the value 1.3 represents the average abundance of the protein that we measured (Slide 4). A histogram plot of these data has a peaked distribution (Slide 4). The position of the mean is illustrated with the dotted line and shows that the mean is located roughly in the middle of the data.

The Variance

The variance (Slide 5) quantifies the spread in the data. The variance of the set S is given by

$$Var(S) = \overline{S^2} = \frac{1}{n-1}\sum_{i=1}^{n}(S_i - \bar{S})^2$$

Note that we use $n – 1$ rather than n. This is because, given a sample of data, $n – 1$ rather than n gives a closer approximation to the true variance of the distribution.

The variance of our sample data set can be calculated as

$$\begin{aligned}Var(S) &= \frac{1}{5-1}\begin{bmatrix}(1.1-1.3)^2 + (0.5-1.3)^2 + (2.6-1.3)^2\\ +(0.3-1.3)^2 + (2.0-1.3)^2\end{bmatrix}\\ &= \frac{1}{4}\begin{bmatrix}(-0.2)^2 + (-0.8)^2 + (1.3)^2 + (-1.0)^2\\ +(0.7)^2\end{bmatrix}\\ &= \frac{3.86}{4}\end{aligned}$$

So, in this case, our variance is ~1.0.

An illustration of the meaning of the variance of a set of data (Slide 5) shows two distributions that have the same mean, but the distribution with the larger variance is more spread out.

The Covariance

The covariance (Slide 6) is a measure of the degree of codependency of two variables. If the two variables are unrelated, such that the value of one does not depend on the value of the other, then they are said to be "uncorrelated," and their covariance will be zero. For example, under the same experimental conditions, if the expression of one gene is completely independent of the expression of another, then the covariance of several measurements of the expression of each gene will have a covariance of zero, or very close to zero.

If the two variables do depend on each other to some extent, then they are said to be "correlated," and their covariance will be nonzero. The covariance will increase in size as the degree of correlation between the two variables increases (Slide 6, compare panels a and b). The size of the covariance will also increase as the variance of each of the two variable increases. If an increase in one of the variables corresponds to an increase in the other, then the covariance will be positive (Slide 6, panel c). If an increase in one of the variables leads to a decrease in the other variable, then the covariance will be negative (Slide 6, panel d).

The covariance of two variables, X and Y (Slide 6), is given by

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

You can understand some of the features of the covariance by carefully studying this equation. If the variables are independent of each other, then, for any given $(X_i - \bar{X})$ the other variable $(Y_i - \bar{Y})$ is as likely to be positive as it is to be negative, and so the sum of all the terms will cancel out, resulting in a variance that is zero. If on the other hand, the two variables are positively correlated, then when $(X_i - \bar{X})$ is positive, $(Y_i - \bar{Y})$ is more likely to be positive, and so their product is most likely to be positive. When $(X_i - \bar{X})$ is negative, $(Y_i - \bar{Y})$ is more likely to be negative, and again their product is most likely to be positive. So, the sum of all of the terms will tend to be positive when they are positively correlated, and the covariance will be a positive number.

Note that if you calculate the covariance of a variable with itself, you get the variance, thus:

$$\begin{aligned}Cov(X,X) &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})\\ &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\\ &= Var(X)\end{aligned}$$

These are all of the statistics that we need to understand PCA. Next, we will look at the matrix algebra that we will need for PCA.

Matrix Algebra

We will convert our data into the form of a matrix. A matrix is a number-filled grid that obeys certain simple mathematical rules (Slide 7).

Here are a few examples of matrices:

$$A = \begin{pmatrix}1 & 2 & -1\\ 0 & 3 & -4\end{pmatrix}; \quad B = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix};$$

$$C = \begin{pmatrix}6 & -4 & 1\\ 0 & 7 & 5\\ 8 & -2 & 4.5\\ 7.25 & 0 & 1\end{pmatrix}$$

A given matrix that has m rows and n columns is referred to as an $m \times n$ matrix (Slide 7). We can refer to individual elements of a matrix by their row number, i, and their column number, j. So, if we wanted to pick out the element in the i^{th} row and the j^{th} column from the matrix A, then we would call it, A_{ij}. For example, from the matrix shown above, $A_{23} = -4$

So, a general 3 × 3 matrix looks like this,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

A matrix with a single column is called a column vector, for example:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

This is an n-dimensional column vector. As another example, here is a two-dimensional (2D) column vector:

$$\begin{pmatrix} 4 \\ -1 \end{pmatrix}$$

Two matrices are said to be equal when they have the same number of rows and columns and when each of their corresponding elements are equal. If two matrices have the same numbers of rows and columns, we can add them together by adding each of their corresponding elements (Slide 8), for example:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}; \ \mathrm{B} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix};$$

$$\text{then } A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix}$$

Let us do an example. Add $\begin{pmatrix} 1 & -2 \\ 4 & 0 \end{pmatrix}$ $\begin{pmatrix} -1 & 3 \\ 1 & 2 \end{pmatrix}$

$$\begin{pmatrix} 1 & -2 \\ 4 & 0 \end{pmatrix} + \begin{pmatrix} -1 & 3 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 1-1 & -2+3 \\ 4+1 & 0+2 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 5 & 2 \end{pmatrix}$$

We can also multiply matrices by numbers simply by multiplying every element, as follows:

$$2A = 2\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} 2a_{11} & 2a_{12} \\ 2a_{21} & 2a_{22} \end{pmatrix}$$

We can multiply two matrices together like so, AB, if the number of columns of A is equal to the number of rows of B. The matrix that results from multiplying these matrices has elements that are generated from multiplying the elements from one row of A by the elements of one column of B, and summing. So, for example:

$$AB = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

For an example, let us also multiply a 2 × 2 matrix by a column vector, as follows:

$$\begin{pmatrix} 1 & 2 \\ -1 & 4 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 1 \times (-1) + 2 \times 2 \\ (-1) \times (-1) + 4 \times 2 \end{pmatrix} = \begin{pmatrix} -1+4 \\ 1+8 \end{pmatrix} = \begin{pmatrix} 3 \\ 9 \end{pmatrix}$$

There is a special matrix called the identity matrix, I, which, when it multiplies any matrix simply gives that same matrix back again: $IA = A$

The 2 × 2, and 3 × 3 versions of the identity matrix are shown below:

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad I_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The transpose of a matrix is the matrix that is made by swapping the row and column of each element. So the element A_{ij} when transposed becomes A_{ji}. The transpose of matrix, say A, is usually written as A^{T}. Here is an example:

$$\text{If } A = \begin{pmatrix} 1 & -3 \\ 0 & 5 \end{pmatrix}, \text{ then } A^{\mathrm{T}} = \begin{pmatrix} 1 & 0 \\ -3 & 5 \end{pmatrix}$$

We are now going to take a little detour into the world of coordinate transformations. We will use what we have learned about matrices to do some simple rotations. This will be helpful for us to understand exactly what goes on when we do PCA.

Coordinate Transformations: An Example of Matrix Algebra

We are going to use the matrix algebra we just learned to do something that may seem a little abstract but that will be useful for understanding what occurs when we use PCA. We will consider a space and locate points on this space with the coordinates x_1 and x_2. This space may represent our data later, but for now just consider this as an abstract space (Slide 9). We will represent points in our space, defined by the coordinates x_1 and x_2, with column vectors, thus:

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

For example, the point with coordinates $x_1 = 1$ and $x_2 = 2$ is represented by the column vector: $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$

The point that this vector represents can be shown on a graph (Slide 9). This way of labeling the points is called a coordinate system. There are many ways to label the points—there are many different coordinate systems. We will consider one alternative. Suppose we have another set of axes that are rotated compared with our original axes (Slide 10). This new set of axes gives us a new way of labeling the points in our space. Points that were labeled $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ have a different column vector in the new coordinate system, which is given by $\begin{pmatrix} x_1' \\ x_2' \end{pmatrix}$.

How can we relate these two coordinates? If we know that, for example

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, then what is the value of $\begin{pmatrix} x_1' \\ x_2' \end{pmatrix}$?

To calculate this, we just need a matrix. The points in one coordinate system can be expressed in another coordinate system by multiplying the column vector by a matrix. If we call the matrix that transforms these coordinates, T, then the coordinates are related by $Tx = x'$

Here is an example of such a transformation. Suppose that our new coordinate system has axes that are rotated at 45° to the original axes (Slide 11). The matrix that we need to transform the coordinates from the original system to the new rotated system is given by

$$T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

So, let us pick a point, which is identified in the original coordinates as $\begin{pmatrix} 1 \\ 2 \end{pmatrix}$. Then in our new rotated coordinates, this point has the column vector:

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 3 \\ 1 \end{pmatrix}$$

So when $x_1 = 1$ and $x_2 = 2$, then in the rotated coordinates,

$x_1' = \frac{3}{\sqrt{2}}$ and $x_2' = \frac{1}{\sqrt{2}}$.

The type of matrix that performs rotated-axis coordinate transformations such as this one is called an orthogonal matrix.

In our example, we could identify points with the coordinates $x_1 = 1$ and $x_2 = 2$, or we could use a different set of coordinates that are related to the original ones by

$$x_1' = \frac{1}{\sqrt{2}}(x_1 + x_2)$$
$$x_2' = \frac{1}{\sqrt{2}}(-x_1 + x_2)$$

The two sets of coordinates are related by an orthogonal matrix, which represents a rotation of the coordinate axes, and the two sets of coordinates are just two different ways of describing the same space. We now turn to the final ingredient of PCA, which is a special kind of vector called an eigenvector, which has an associated number called an eigenvalue.

Eigenvectors and Eigenvalues

We have just seen an example of a matrix that transforms a column vector. There is a special equation whereby the matrix maps a vector to a multiple of itself, thus:

$Tx = \lambda x$ where λ is a number and T is a square matrix (Slide 12). If the vector x has n dimensions, then the square matrix must be $n \times n$. There are then n column vectors that satisfy this equation; these are the eigenvectors of the matrix T. Each eigenvector satisfies the equation with a particular value of λ, which is the eigenvalue associated with the eigenvector.

We could demonstrate how to find the eigenvalues and eigenvectors of a matrix, but that will have to wait for another day. For now, you can use MATLAB (or free software such as OCTAVE or R), which will calculate these values for you.

If you calculate the eigenvectors of a symmetrical matrix (a matrix that is equal to its own transpose) and place each of these column vectors side by side to make another square matrix, then this resultant matrix will be orthogonal and so will transform coordinates by a rotation of the axes. Armed with this knowledge, you are now ready to understand the workings of PCA. This will be our next topic.

PCA

We will work through the method of PCA by applying it to a simple example (Slides 13 to 18). We will consider a very simple data set, which consists of a 2D set of points. We will have several measurements of the variables, which we call x_1 and x_2. We take the following ten points as our data:

x_1 = 2.5; 0.5; 2.2; 1.9; 3.1; 2.3; 2.0; 1.0; 1.5; 1.1
x_2 = 2.4; 0.7; 2.9; 2.2; 3.0; 2.7; 1.6; 1.1; 1.6; 0.9

Earlier, we represented a single point in a 2D space with the column vector $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

We will represent our ten points by positioning all of the column vectors that represent each data point beside each other in a matrix, thus:

$$D = \begin{pmatrix} 2.5 & 0.5 & 2.2 & 1.9 & 3.1 & 2.3 & 2.0 & 1.0 & 1.5 & 1.1 \\ 2.4 & 0.7 & 2.9 & 2.2 & 3.0 & 2.7 & 1.6 & 1.1 & 1.6 & 0.9 \end{pmatrix}$$

In this way, our data are represented as the matrix D. This data matrix is written in the standard configuration in which the different variables run down the rows and the different observations of these variables run across the columns. We then plotted these data (Slide 14).

Our data contain a certain degree of random scatter, so we will start to use some statistics. First, we calculate the mean of each variable and subtract it from each observation value. This involves taking the average of the numbers in each row and subtracting the result from the elements in each row.

For our data, the mean of the first variable is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} D_{1i} &= \frac{1}{10}(2.5+0.5+2.2+1.9+3.1\\ &\quad +2.3+2.0+1.0+1.5+1.1)\\ &=1.81\end{aligned}$$

The mean of the second row is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} D_{2i} &= \frac{1}{10}(2.4+0.7+2.9+2.2+3.0\\ &\quad +2.7+1.6+1.1+1.6+0.9)\\ &=1.91\end{aligned}$$

We then subtract these means from each row, to generate the matrix D':

$$D' = \begin{pmatrix} 0.69 & -1.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.81 & -0.31 & -0.71 \\ 0.49 & -1.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -1.01 \end{pmatrix}$$

This process simply moves the data so that they are centered on the origin of our coordinate system (Slide 15). The next step is to calculate the covariance matrix, C, which is calculated from the matrix D' as follows:

$$C = \frac{1}{n} D' D'^{\mathrm{T}}$$

This is a square matrix in which the element in the i^{th} row and j^{th} column is the covariance of the i^{th} and j^{th} variable. We have the two variables x_1 and x_2, so the covariance matrix is given by

$$\begin{pmatrix} Cov(x_1,x_1) & Cov(x_1,x_2) \\ Cov(x_2,x_1) & Cov(x_2,x_2) \end{pmatrix}$$

If we put in the numbers, we generate

$$\begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

The next step is to calculate the eigenvalues and eigenvectors of this covariance matrix. For our data, the eigenvalues are 0.0490833989 and 1.28402771

The corresponding eigenvectors are

$$\begin{pmatrix} -0.735178656 \\ 0.677873399 \end{pmatrix} \text{ and } \begin{pmatrix} 0.677873399 \\ 0.735178656 \end{pmatrix}$$

Note that these eigenvectors are unit eigenvectors. This means that their length is 1.0 (the sum of the squares of their elements is unity). This is important. If you obtain eigenvectors that do not have length of 1.0, then you need to rescale them.

To do this rescaling, you need to compute the length, also called the norm, of the eigen vector as follows:

$$\text{length} = \sqrt{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}$$

Then, use the length to rescale the vector as follows:

$$e_{\text{scaled}} = \frac{1}{\text{length}} e$$

We then place the eigenvectors side by side to make a square matrix, thus:

$$\begin{pmatrix} -0.735178656 & 0.677873399 \\ 0.677873399 & 0.735178656 \end{pmatrix}$$

We then swap the columns of this matrix around so that they are in order of the size of their corresponding eigenvalues, with the largest eigenvalue to the left. Because our columns are ordered such that the eigenvector with the largest eigenvalue is on the right, we need to swap the columns in our matrix of eigenvectors to generate the matrix, which we will call W:

$$W = \begin{pmatrix} 0.677873399 & -0.735178656 \\ 0.735178656 & 0.677873399 \end{pmatrix}$$

We will use the transpose of this matrix, W^{T}, to perform a coordinate transformation:

$$W^{\mathrm{T}} = \begin{pmatrix} 0.677873399 & 0.735178656 \\ -0.735178656 & 0.677873399 \end{pmatrix}$$

We will now pause during our PCA to consider what is happening. Remember that we said that the eigenvectors of a symmetrical matrix, when placed side by side, make a new matrix, which is orthogonal. We also saw how orthogonal matrices make coordinate transformations from one set of coordinates to a new set with axes that are rotated.

We have our data in a coordinate system in which our data points are labeled with the coordinates x_1 and x_2. The matrix of eigenvectors that we have constructed above, W^{T}, gives us a coordinate transformation. Any data point in our original coordinates can be transformed into their new coordinates by multiplying the column vector by our matrix above. As we noted earlier, this matrix is orthogonal, and so it corresponds to a rotation of the axes.

We can then plot the data represented in the matrix D' again (Slide 18, top graph), but this time we include the rotated axes of the new coordinate system. We choose to call these new coordinates x'_1 and x'_2. We transform our data from the original coordinates to these new rotated coordinates by multiplying by the transformation matrix, W^{T}, thus: $W^{\mathrm{T}}x = x'$

For our matrix W, we can write this as

$$W^{\mathrm{T}}x = \begin{pmatrix} 0.677873399 & 0.735178656 \\ -0.735178656 & 0.677873399 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$
$$= \begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix}$$

Writing this out, we can express the new coordinates in terms of the original ones, like so:

$$x'_1 = 0.677873399x_1 + 0.735178656x_2$$
$$x'_2 = -0.735178656x_2 + 0.677873399x_2$$

We can express our data in the new rotated coordinates by multiplying our transformation matrix W^{T} by our shifted data matrix D', like so:

$D_{\mathrm{PCA}} = W^{\mathrm{T}}D'$where D_{PCA} is the matrix of data expressed in the new rotated coordinates. So, for our data,

$$D_{\mathrm{PCA}} = \begin{pmatrix} 0.677873399 & 0.735178656 \\ -0.735178656 & 0.677873399 \end{pmatrix} \begin{pmatrix} 0.69 & -1.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.81 & -0.31 & -0.71 \\ 0.49 & -1.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -1.01 \end{pmatrix}$$
$$= \begin{pmatrix} 0.83 & -1.8 & 0.99 & 0.27 & 1.8 & 0.91 & -0.099 & -1.1 & -0.44 & -1.2 \\ -0.18 & 0.14 & 0.38 & 0.13 & -0.21 & 0.18 & -0.35 & 0.046 & 0.018 & -0.16 \end{pmatrix}$$

These are our data in the new coordinates (Slide 18, bottom graph).

The coordinate transformation that we generated with this method has rotated the axes so that the first coordinate axis lines up with the data in such a way that most of the variation is in that coordinate. The data were correlated in the original coordinates, but they are not correlated in the new rotated coordinates. The aim of PCA is to derive a new set of coordinates for the data that are uncorrelated and that are in the order of the degree of variation in that coordinate.

In our example, we may decide that the new coordinate x'_1 captures most of the information in the data and that the coordinate x'_2 can be discarded. In this case, we would reduce our data to a single dimension. How many of the new coordinates (also called components) are kept is an arbitrary choice; however, the intention is to keep only enough components to capture the essence of the data. So, now you understand what the C in PCA stands for.

A typical method of selecting the components to keep is to sum all of the eigenvalues and then keep only those components with the largest eigenvalues, which sum up to no less than 90% of the total. This is done because the larger the eigenvalue, the greater amount of variation of the data in the direction of the corresponding eigenvector. In PCA, we choose to keep only those components that carry most of the variation of the data. The discarded components are removed by removing the corresponding columns from the W matrix. Then, the transformed data,

$D_{\mathrm{PCA}} = W^{\mathrm{T}}D'$will only have a number of rows equal to the number of retained coordinates. Hence, the transformed data in the new coordinates will have a reduced dimension. If we decide to keep only the first component in our example, then we must keep only the first column in the W vector:

$$W \rightarrow \begin{pmatrix} 0.677873399 \\ 0.735178656 \end{pmatrix}$$

Then, if we calculate the PCA-transformed data, we obtain

$$D_{\mathrm{PCA}} = \begin{pmatrix} 0.677873399 & 0.735178656 \end{pmatrix} \begin{pmatrix} 0.69 & -1.31 & 0.39 & 0.09 & 1.29 & 0.49 & 0.19 & -0.81 & -0.31 & -0.71 \\ 0.49 & -1.21 & 0.99 & 0.29 & 1.09 & 0.79 & -0.31 & -0.81 & -0.31 & -1.01 \end{pmatrix}$$
$$= \begin{pmatrix} 0.83 & -1.8 & 0.99 & 0.27 & 1.7 & 0.91 & -0.099 & -1.1 & -0.44 & -1.2 \end{pmatrix}$$

You can see that we have reduced the dimension of our data to one, which contains most of the variance of the data.

Summary

The main points about PCA can be summarized as follows: (i) The data are placed in a matrix, D, with the variables running down the rows and the observations running across the columns. (ii) The means of each variable are found and subtracted from each variable, which generates the matrix D'. (iii) The covariance matrix is constructed by $C = (1/n)D'D'^{\mathrm{T}}$, where n is the number of variables. This matrix has the covariance of the i^{th} j^{th} variables in the element C_{ij}. (iv) The eigenvalues and eigenvectors of the covariance matrix are found and placed in order of the size of the eigenvalues. (v) The eigenvectors that correspond to the eigenvalues whose sum is no less than 90% of the total are arbitrarily retained. (vi) The resulting eigenvectors are placed side by side into a matrix, W, which describes a new coordinate system with the axes rotated so that they align with the greatest variation of the data. The first components carry the most variation because they have larger eigenvalues. And (vii), the data are expressed in the new coordinate system by multiplying the D' by the transpose of W, thus: $D_{\mathrm{PCA}} = W^{\mathrm{T}}D'$

In this way, we can move the data into a new coordinate system of variables that are independent and have a lower dimension because we have kept only those variables that carry most of the variation of the data. The concept behind PCA is that the system of variables with reduced dimension carries the main trends of the data and is easier to interpret and visualize than the original data. We may begin with a large number of variables; however, through PCA, we are able to represent most of the features of the data in a just a few variables.

The example we have used here is simple and has only two variables; however, the same method may be applied to much higher dimensional data, with a much larger number of data points. This method could be used for comparing gene expression microarrays or RNA-seq, proteomics, phosphoproteomics, or any other type of high-dimensional data collected in systems biology. The observations made under different conditions, cell types, or time points may be treated as the variables, or the genes, proteins, or other molecular species that are measured may be treated as variables. Both are valid ways of exploring the data (Slide 19). For example, we have used the PCA approach to visualize the similarity between Nestin$^+$ hematopoietic niche cells isolated from the bone marrow and other relevant cell types that were profiled by different groups (*1*). Our analysis placed the gene expression profile in Nestin$^+$ cells in the context of those of other similar cell types previously isolated from the bone marrow, and we showed that this cell population was distinct from the others. Careful analysis of the genes that distinguished between Nestin$^+$ cells and the other cell types revealed less expression of cell cycle–related genes and increased expression of genes involved in metabolic pathways, supporting the proposed role of the Nestin$^+$ cells as niche cells (Slide 20).

Problem Set

Suppose you are given the results of a microarray experiment. The experiment measures the expression of five genes (in practice, this number will be much larger than five), which are labeled G1, G2, G3, G4, and G5. The expression of these genes is measured in nine samples, labeled A to I. The data that you are given are shown below. The experimentalist hypothesizes that these samples should fall into three separate categories. The following questions will lead to PCA of these data. You can use this to reveal whether the hypothesis is correct, and if so, identify which genes in the sample belong in which category.

	A	B	C	D	E	F	G	H	I
G1	1.4553	0.01416	1.50532	1.36762	0.446724	1.31581	0.154451	1.35969	1.35211
G2	0.944012	0.861993	0.78199	0.177722	1.08446	1.01952	1.09119	0.102935	0.0632143
G3	1.35651	0.204572	1.33507	1.10833	0.495636	1.33565	0.136843	1.04292	1.05102
G4	0.291371	0.20958	0.439246	1.215	0.263377	0.320713	0.0174625	1.50438	1.31253
G5	1.74347	0.282323	1.69477	0.616727	0.157345	2.06409	0.392771	0.514497	0.651402

You can create a text file containing these data as a matrix. It is recommended that you use the programming language "R" to solve the following problems because the answers are given in that format; however, you can use any other software tool that can perform the analysis. Answer each of the questions that follow to perform the PCA:

Question 1. What is the mean expression of each gene? Generate a new data matrix that is shifted to have a mean of zero.

Question 2. Compute the covariance matrix of the shifted data matrix from Q. 1.

Question 3. Compute the eigenvalues of the covariance matrix from Q. 2. How many principal components do you need to capture at least 90% of the variation in the data?

Question 4. Compute the eigenvectors of the covariance matrix and construct a matrix composed of these vectors.

Question 5. Project the data onto the new coordinate axes by multiplying the transposed matrix of eigenvectors (that is, the transpose of the matrix calculated in Q. 4).

Question 6. Generate a plot of the first two principal components. Identify the samples that belong to any of the clusters that you might find. Was the experimentalist correct in hypothesizing that there are three clusters?

Educational Details

Learning Resource Type: Lecture, assignment, PowerPoint
Context: Graduate
Intended Users: Teacher, learner
Intended Educational Use: Learn, plan, teach
Discipline: Biochemistry; biocomplexity; bioinformatics,

genomics and proteomics; biostatistics; biotechnology; cell biology; molecular biology; pharmacology; proteomics; systems biology; theoretical biology

Keywords: Cell signaling, computational biology, principal components analysis, dimensionality reduction, clustering Analysis

Technical Details

Software: R
Requirements: Platform-independent open-source
Download: http://www.r-project.org/index.html

Supplementary Materials

(http://stke.sciencemag.org/cgi/content/full/sigtrans;4/190/tr3/DC1)

Slides: Introduction to Statistical Methods for Analyzing Large Data Sets: Principal Components Analysis
Problem set key is available upon request.

References and Notes

1. S. Méndez-Ferrer, T. V. Michurina, F. Ferraro, A. R. Mazloom, B. D. Macarthur, S. A. Lira, D. T. Scadden, A. Ma'ayan, G. N. Enikolopov, P. S. Frenette, Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* **466**, 829–834 (2010).

2. **Acknowledgments:** We thank S. L. Jenkins for comments and suggestions. **Funding:** This work was supported by NIH grants 5P50GM071558-03, 1R01DK088541-01A1, KL2RR029885-0109, and RC2OD006536-01.

10.1126/scisignal.2001967

Citation: N. R. Clark, A. Ma'ayan, Introduction to statistical methods to analyze large data sets: Principal components analysis. *Sci. Signal.* **4**, tr3 (2011).

Introduction to Statistical Methods to Analyze Large Data Sets: Principal Components Analysis
Neil R. Clark and Avi Ma'ayan (September 6, 2011)
Science Signaling **4** (190), tr3. [doi: 10.1126/scisignal.2001967]
originally published online September 6, 2011