Timothy Arnold

Professor Little

Mont 105N-01

20 April 2013

Principal Component Analysis and its Applications

 Principal component analysis is a variable reduction method built upon a covariance or dispersion matrix. In summary, it uses the original data to form new variables through the creation of linear combinations, or principal components, of the original variables. This process involves finding the eigenvalues and corresponding eigenvectors of the covariance matrix of the original data set. “The orthonormal eigenvectors are used for constructing the principal components and the eigenvalues are the variances of the corresponding principal components” (Narayanaswamy 309). The covariance matrix, and its corresponding eigenvalues and eigenvectors are the backbone of principal component analysis.

Before getting into the details of Principal component analysis, one must have a background in certain concepts of mathematics and statistics. Among these many concepts is standard deviation and variance. The standard deviation, or SD, of a data set is a measure of how spread out the data is, or the average distance of the mean of the data set to a point. Although the standard deviation is not directly used in PCA, variance, or the square of the standard deviation is. The formula for variance is the average of the squared distances from the mean. It is from this formula that one of the backbones of PCA, covariance, is derived.

 Covariance is very similar to variance, but instead of one dimension of variance, covariance incorporates a second dimension or variable into the formula, so one can see how spread the data is in terms of two different variables, and see how they change together. Instead of the point minus the average squared as in variance, the covariance formula uses the x point minus the x average times the y point minus the y average, thus creating a measure of the spread according to two variables. With these concepts of statistics as a background, one can begin the process of PCA.

 Upon finding the covariance with each combination of variables or dimensions, the covariance matrix is ready to be made. The covariance matrix must always be a square matrix with n dimensions. The rows of this matrix would consist of each covariance with the same first variable, and the columns would consist of the same second variable. This would produce a diagonal of covariance values with the same variable such as cov(x,x), or in other words, just the variance of x. This diagonal would run from the top left of the square matrix to the bottom right, and every value not on this diagonal, would have a reflection of an identical value across from it, such as cov(x,y) and cov(y,x). Since multiplication is cumulative, or can be written as A times B or B times A, these two covariance values will be equal.

 This matrix of covariance values, once found, is used first to find eigenvalues and then eigenvectors of the matrix. If a matrix is multiplied by a vector of the same amount of dimensions, then the vector is an eigenvector if the answer is the same or a multiple of the original vector. If the answer is a multiple of the original vector, the number factored out to make the original vector is known as an eigenvalue. All eigenvectors of a given matrix are perpendicular, or orthogonal to one another. This becomes an important factor later on in PCA when the original data is re-written in terms of new axes. However, finding these eigenvalues and eigenvectors can be one of the most difficult parts of PCA if the amount of dimensions is very high and computational resources are not available as an aid. The computations for eigenvectors of two by two or three by three matrices are much simpler and easier to compute, but higher dimensionalities can become nearly impossible to find eigenvectors without a computer to help.

 Upon finding the covariance matrix, one must use Linear Algebra techniques to compute the eigenvalues and eigenvectors. First, one must compute the eigenvalues by using the concept of Ax=λx. This says the matrix times a vector equals a multiple of that vector. Changing this formula to (A-λI)x=0 helps us to find our eigenvalues. A-λI is the covariance matrix subtracted by the identity matrix times a certain number lambda. The identity matrix is a matrix containing all zeros except the diagonal from the top left to the bottom right which contain ones. The result of this subtraction of matrices leaves the variances of the diagonal minus the number lambda. At this point, we find the determinant of the matrix (Strang 283).

 The determinant of a matrix is a special value that, in a way, sums up the values of a square matrix into a single number or equation. The determinant is found in an (n x n) matrix by multiplying the first number of the first row by the determinant of the numbers not found in that numbers row or column. The same thing is to be done for every other number in the first row, subtracting every other value. For example, in a three by three matrix,

the determinant would be a(ei-fh)-b(di-fg)+c(dh-eg). Using the determinant, we will be able to find the eigenvalues of the matrix, or the lambda values.

 The amount of eigenvalues of a given matrix depends on the amount of dimensions. When taking the determinant of the covariance matrix minus the lambda identity matrix, lambdas will be multiplied together to yield an equation containing a lambda to the power of how many dimensions the matrix had. If the matrix had three dimensions, the equation would be cubic, giving three eigenvalues. The eigenvalues can be found once the equation is set equal to zero. Everything up to this point can be done by hand for high dimension matrices, but solving this last equation is when help is needed. If fifty dimensions are used, finding all the values of an equation containing a power of fifty will be almost impossible. Technological resources are needed at this point if the dimensions are high enough (Strang 283-4).

 Once the eigenvalues are found, one can use them to find the eigenvectors of the matrix. The eigenvectors of this covariance matrix are called the Principal components. Reusing the equation used earlier, (A-λI)x=0, we can plug in the eigenvalues previously found using the determinant to find the eigenvectors. Plugging an eigenvalue in for lambda will result in a matrix containing only constants with different values along the diagonal. When this new matrix is multiplied by a vector x, the result is zero. Upon multiplying, a system of equations is developed. Solving for all three variables will yield the eigenvector for the eigenvalue used. The eigenvalue with the largest value will yield the Principal component Vector, or the vector that accounts for the most variability of the data set (Strang 283-4).

 The principal component vector is like the line of best fit in the data set. This shows how the two data sets can be compared long that vector. The second principal component, as previously stated, is perpendicular or orthogonal to the principal component. This line gives us information as to how deviated from the principal component the data is. As the eigenvalues get smaller, the eigenvectors become less significant. This is why this process is good for comparing data sets with a high number of variables. It can be reduced to account for only one or two components. This whole process up to this point has given us lines that describe how the data is related. The remainder of PCA is used to rewrite the original data in terms of these new principal components.

 At this point, one must decide how many principal components to be used. The greater the amount of components left out, the greater the amount of information that is lost. Although more information is lost, some of the information is so miniscule and insignificant that it can be left out, especially with higher dimensional data set. If we started with n dimensions, and chose the first p dimensions, we would finish with a data set of only p dimensions. Upon selecting the components, one must put the vectors into a matrix, with the rows being the components of the vectors, and the columns being the same component of the vector, but different vectors. This matrix of vectors is known as the feature vector. This feature vector is to be multiplied by the transformed data set found in the second step, where the mean is subtracted. This step gives the original data in terms of only the eigenvectors we chose to keep.

 This step is best explained in Lindsey Smith’s Tutorial,

Our original data set had two axes, x and y, so our data was in terms of them. It is possible to express data in terms of any two axes that you like. If these axes are perpendicular, then the expression is the most efficient. This was why it was important that eigenvectors are always perpendicular to each other. We have changed our data from being in terms of the axes x and y, and now they are in terms of our 2 eigenvectors. In the case of when the new data set has reduced dimensionality, i.e. we have left some of the eigenvectors out; the new data is only in terms of the vectors that we decided to keep (Smith 17).

In this paragraph, Smith is trying to give an image of this transformation. Her example used two dimensions, so they can be plotted along the x and y axes. This is why it was important to have eigenvectors orthogonal to one another. Upon reaching this step, the data was transformed so that the data followed the principal component, and deviated from it by values given by the second principal component. If only the principal component was used, the new values would all be along the principal component.

 The final step of this long process of Principal component analysis is getting the original data back. In order to accomplish this final step, the eigenvectors chosen must be multiplied by the data, and added to the original mean of the data. The formula for this last step can be given as

RowOriginalData=(RowDataAdjust x RowFeatureVector^(-1))+OriginalMean. Before this step we have data in terms of the principal components as the axes. This last step will rewrite the data in terms of the original axes, but with the data following along our eigenvectors with the amount of variability we chose in the previous step. If we were to use Smith’s example of two dimensions, getting the original data back with one dimension would yield a graph of the data that does not deviate from the principal component. The second components variability is gone, and all of the data is written on the line of best fit (Smith 19-20).

 After having a basic understanding of how Principal component analysis works, an example can be worked out using basic technology like excel and eigenvector calculators. For this example, I will use a three-dimensional data set from online comparing life expectancy, people per television, and people per physician. First, find the mean of each of the columns using the AVERAGE() function in excel. After that, transform the data by subtracting the respective mean off each of the values in the columns. Next, I either squared those values, or multiplied it by another column to give me six new columns. Finding the sum and dividing by one less than the amount of values gives the covariance and variance values needed to find the covariance matrix.

 Using an eigenvalue calculator online, I calculated the eigenvalues and eigenvectors of the matrix, because the determinant equation of three dimensions became too long and complicated. Upon finding these values I transformed the data to create the data in terms of the principal components as the axes. Finally, I reduced dimensions and brought back the original data using one variable and two variables to compare the results on the graphs in Excel. When graphed the original data had almost no correlation, but after transformed using PCA, it can be seen in a linear format using dimension reduction.

 The column people per television had the highest degree of randomness in the data set, so when the eigenvalues and eigenvectors were calculated, the television vector had the smallest eigenvalue associated with it, making it insignificant. Life expectancy had the highest eigenvalue, making it the principal component. All of the values are plotted along this component, either on it as in one variable, or along it as if a line of best fit.

 Although the invention and development of computers have opened the widespread use of principal component analysis, there are instances where it is hindered by computational limitations. For example, there are several models that try to explain the pricing of securities. One model suggests that the return of securities is linearly correlated to specific economic factors. In order to perform an empirical test upon this statement, one needs the first several dominant eigenvalues of the securities dispersion matrix. The number of securities within an empirical test of the sort would be astonishingly large, making the matrix very large. In this case, as well as others, the matrix is too large for even a computer to calculate the corresponding orthonormal eigenvalues and eigenvectors (Narayanaswamy 309).

 Professor Ian Joliffe goes into detail in a series of articles describing different methods of examining and deciding which variables are meant to be discarded in a large data set of ten or more variables. Variables are often left unchanged if a subset of fewer variables is used. He goes on to say that very little has been done in the area of Principal component analysis to determine which variables may be discarded; however, sometimes the variables are reduced prior to using component analysis. Extra variables merely complicate data, and often do not give any extra information; so if possible, one should eliminate them wherever he or she can. Time and money are often also spared in variable reduction, because of less computing time and in the future, fewer variables would need to be measured (Joliffe *Artificial* 160).

 Joliffe generated artificial data to experiment with various variable reduction techniques. The techniques he used abandoned all of the redundant variables within the large data set. In a later paper, Joliffe used these techniques on real data, making these reduction techniques more applicable to the real world. In this real data variable reduction, Joliffe uses five techniques, two of which involve principal component methods of reductions. The principal component methods used each associate a variable with a principal component. This part differs among the two techniques. After relating the variables to the principal components, the variables associated with the last components are then rejected. The number of variables rejected is equal to the number of eigenvalues less than .7 (Joliffe *Real* 21-22).

 Jorge Cadima and Ian Joliffe performed an experiment using sixty-three crayfish and thirteen variables. They wanted to see whether the principal components of the size of an organism could determine its shape. These two men took measurements of thirteen different factors of the fish, and determined that the first component accounted for 64% of the total variation. This component can be used to determine the fish’s overall size. They concluded that all of the other principal components could all be used to shape the fish. The second and third components were claw shape and claw to body ratios respectively. Using these variables and principal components, they were able to see which variables were negligent and which determined size or shape (Cadima 710-11).

 In conclusion, principal component analysis can be used to make a random data set into one that shows linear patterns which can be much easier to read. As the amount of variables increase, principal component analysis becomes more useful, eliminating those variables which have almost no purpose, and emphasizing those that do. Using PCA can make hundreds of variables into a mere few, especially with the right computational technology. Even technology has its limitations however, but principal component analysis is useful most of the time. As in the example with the crayfish, Cadima and Joliffe were able to determine which variables of the length and width of the organism’s body parts were most important in playing a role in determining the size and shape of the organism. Principal component analysis has many uses and applications, and can only become a more useful tool to organize data as time goes on.

Works Cited

Cadima, Jorge F. C. L., and Ian T. Jolliffe. "Size- and Shape-Related Principal Component Analysis." *Biometrics* 52.2 (1996): 710-6. Web.

Jolliffe, I. T. "Discarding Variables in a Principal Component Analysis. I: Artificial Data." *Journal of the Royal Statistical Society.Series C (Applied Statistics)* 21.2 (1972): 160-73. Web.

---. "Discarding Variables in a Principal Component Analysis. II: Real Data." *Journal of the Royal Statistical Society.Series C (Applied Statistics)* 22.1 (1973): 21-31. Web.

Narayanaswamy, C. R., and D. Raghavarao. "Principal Component Analysis of Large Dispersion Matrices." *Journal of the Royal Statistical Society.Series C (Applied Statistics)* 40.2 (1991): 309-16. Web.

Smith, Lindsay I. "A Tutorial on Principal Component Analysis." *Http://www.cs.otago.ac.nz/*. University of Otago, 26 Feb. 2002. Web. 18 Apr. 2013.

Strang, Gilbert. "Chapter 6: Eigenvalues and Eigenvectors." *Introduction to Linear Algebra*. Wellesley, MA: Wellesley-Cambridge, 2009. 283-97. Web.