Biol 206/306 – Advanced Biostatistics
Lab 7 – Principal Component Analysis
Fall 2016

By Philip J. Bergmann

0. Laboratory Objectives

1. Learn what sort of data can be analyzed using Principal Component Analysis (PCA)
2. Learn how PCA works: procedure, assumptions, results
3. Learn how PCA differs from DFA
4. Learn how to do PCA in R

1. Principal Component Analysis

We continue with multivariate statistics by studying principal component analysis (PCA). In many ways, PCA is very similar to MANOVA/DFA in that it repartitions the variance present in multiple response variables to produce a set of eigenvectors that describe how each original variable contributes to each new component. It also allows for the calculation of factor scores for each individual sample, which then allows you to visualize how individuals are related in a simplified multivariate space.

How PCA differs from DFA is in how it repartitions the variance. DFA requires that you assign each individual to a group *a priori*, and then calculates eigenvectors so as to maximize the discriminatory power of the first discriminant function. In contrast, PCA treats each individual equally, essentially making the *a priori* assumption that all individuals are sampled from the same population (this is not a formal assumption, and is readily violated in many applications of PCA). In this framework, the first principal component (PC-1) is calculated so that it explains the greatest possible amount of variation in the dataset. Interestingly and usefully, the PC-1 axis corresponds to the multivariate version of the reduced major axis or major axis (depending on approach) of the data, relating the technique to model II regression. Subsequent PCs are then orthogonal (independent) to all other PCs. If there are *p* response variables, then there are *p* PCs that explain ever-decreasing amounts of variance, and so only the first few are interpreted.

PCA has a number of important applications. First, it is viewed as an exploratory technique that can help the investigator determine how response variables are related to one another. What this means is that there is no specific null hypothesis that is tested by PCA – the technique simply repartitions variance. Second, PCA is used as a data reduction technique. Since variance is repartitioned so that most of the variance present in *p* response variables is explained by only a few PCs, you can go from trying to make sense of a dozen or even more variables to trying to interpret a few (often one to four) PCs. This makes data handling and visualization much easier. Third, the variables produced by PCA are orthogonal, having a correlation of ~0 with one another. The implication of this is that PCA can be used to deal with multicollinearity. PCA is a great technique, but one limitation is that each PC is an amalgam of all of the original variables, and so interpretation of just what PC-1 or PC-2 means can be difficult. One may also wonder what the point of PCA is if it doesn't test any hypothesis. This may be a strength of the technique because once you have the PCs defined and factor scores calculated, you can do tests on the factor scores (e.g., test for sexual dimorphism in factor scores using a t-test).

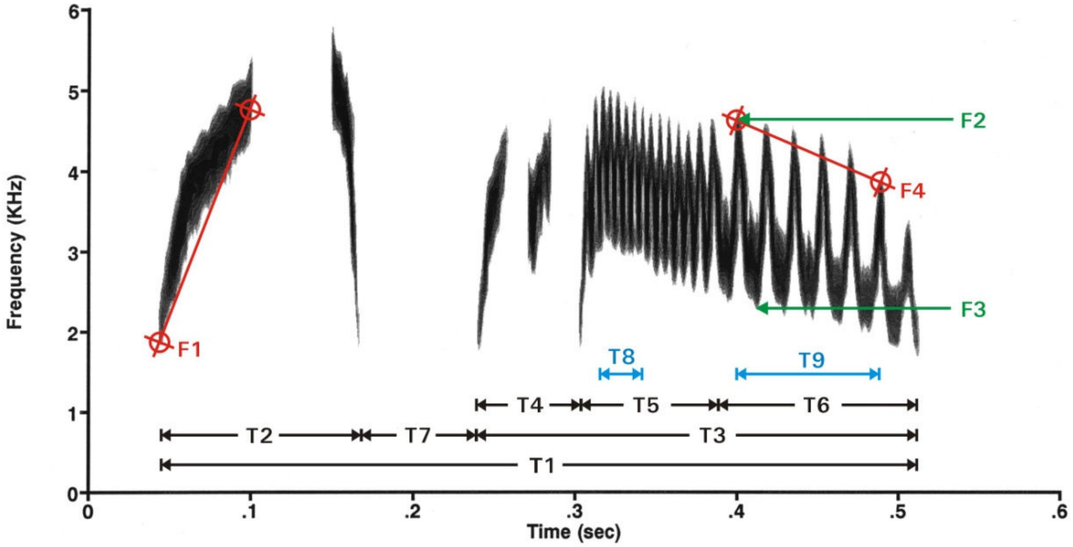
PCA comes in several flavors. While a MANOVA/DFA is done on the matrix of **sums of squares and cross products (SSCP)** of the original data, PCA can either be done on the **correlation matrix** or the **covariance matrix** for the original variables. Using the covariance matrix yields a PC-1 that is equivalent to the multivariate version of a major axis regression. The covariance matrix does not standardize for the magnitude of the original variables, and so is typically used when all of your variables have similar amounts of error associated with them, and are measured in the same units (e.g., skull measurements all taken in millimeters). The correlation matrix yields a PC-1 that is equivalent to the reduced major axis of the data, standardizes the data because correlations are being studied, and so is appropriate when the original variables have different amounts of error associated with them, often when taken in different units (e.g., some variables are in seconds, others in Hz).

Finally, PCA makes several assumptions, just like all statistical techniques. First, the data should be randomly sampled. As always, this is best ensured by designing your study properly. Second, the data are assumed to follow a multivariate normal distribution (mnd). As with DFA, the technique is robust to violations of this assumption, which cannot be easily tested. Whether individual variables are normally distributed has little bearing on whether the data together follow mnd, and this approach can be downright misleading. Third, PCA assumes that the variables are linearly related. This can be tested simply by plotting your variables in a pairwise manner. A log-transformation is often applied (especially with morphometric datasets) to the data to improve linearity. Often some of the variables are not correlated and so it can be hard to evaluate linearity. In all honesty, this assumption is rarely tested, and the most important issue is that variables are not related in a highly non-linear manner (e.g., a tight but curved relationship). Related to the assumption of linearity, PCA deals poorly with situations where there are zeros in the dataset. When this is the case, it is best to use a transformation of some sort that eliminates the zeros, which distort the factor space. Finally, it is important to note that PCA does not make the assumption that variables are uncorrelated, something that DFA does. In fact, PCA is often used with highly multicollinear data.

List as many ways in which PCA and DFA differ as you can.

2. Doing PCA in R

The dataset that we will be analyzing using PCA is one that has taken measurements of male bird songs. The dataset was generated in the lab of Dr. M.R. Lein, at the University of Calgary, and consists of about ten individual "Fitz-Bew" calls from each of thirteen male Willow Flycatchers (*Empidonax traillii*) from a population in the Canadian Rockies. The Willow Flycatcher is morphologically very similar to other sympatric species in the genus, and so the Fitz-Bew call is useful for identifying the species in the field. Because there is variation in the call, it is possible that Willow Flycatchers can identify individuals by their call. This is something we can evaluate using PCA.



Above is a sonogram of a Fitz-Bew call with the variables in the dataset mapped out on it. Some of the variables (T1 to T9) are temporal, representing the length of different parts of the song in seconds. For example, T2 represents the length of the first syllable (Fitz), while T3 represents the length of the second syllable (Bew). Variables F1 to F4 are frequency variables, measured in Hertz, with F1 and F4 being changes in frequency, and F2 and F3 being maximum and minimum frequency, respectively. Many other variables would be possible – those chosen represent measurements that are obtainable from most songs and that are objectively repeatable. If you are interested in what the call sounds like, a recording is available here: http://www.appliedbioacoustics.com/Repertoires/Passeriformes/Tyrannidae/Empidonaxtraillii/bird.html.

Given the description of the dataset, would you do a PCA using a covariance matrix or a correlation matrix? Justify your answer.

Start by downloading and viewing the dataset in Excel, saving it to a text file, and importing it into R as an object named "fb_data". Examine the structure of the imported data frame. Note that the variable "male" is an integer instead of a factor. You do not need this variable to do the PCA, but will need it later, represented as a factor. Create an object called "maleID" that contains the "male" variable as a factor.

Visualize the data by making pairwise plots of all variables. You can do this at once with the following useful multivariate plot:

```
> plot(fb_data[,3:15])
```

Do you notice any variables that obviously are related in a curved, non-linear manner? Note that some may be correlated, and this is okay – you are only looking for blatant violations of linearity.

Doing a principal component analysis is straightforward, but a lot can be done with the output. Do a PCA on the 13 song variables using the following function, saving the resulting model in an object "fb_pca1":

```
> princomp(x, cor=FALSE, scores=TRUE, covmat=NULL)
```

In this function, x is the dataset, consisting of all the variables to be analyzed; cor=F is a logical argument, where the default is a PCA done on a covariance matrix and cor=T indicates the use of a correlation matrix; scores=T tells the function whether or not to calculate factor scores; and covmat=NULL can be used to specify a user-supplied covariance matrix – this is if you would like to do the PCA done using a covariance matrix as input, as opposed to your raw variables.

Do your PCA, choosing which method is appropriate for this dataset. You can omit the "scores" and "covmat" arguments, leaving them as default. You can get the PCA output using:

```
> summary(pca_object, loadings=FALSE, cutoff=0.1)
```

Here, "pca_object" is the output from the "princomp" function. Loadings=F specifies whether the component loadings should be provided. Although default is FALSE, you almost always want these, so should specify TRUE. Finally, cutoff=0.1 specifies the value below which a component loading is not displayed in the table of loadings. ***Try the summary function on your PCA object, specifying loadings=T and leaving cutoff=0.1. Then repeat, also specifying cutoff=0.0001. How does the output compare?***

What other output are you given by the summary statement?

This gives you most of the output that you need to interpret the PCA, but note that the eigenvectors (loadings) are not standardized. You can standardize the component loadings by dividing each value by the square root of the eigenvalue. You are given the square root of the eigenvalue as the "standard deviation" of each PC in the output. We will not be standardizing loadings today, but we will be using eigenvalues (square of the standard deviation). The eigenvalues are also important because they can be used to calculate the proportion of variance explained by each component (as the quotient of the eigenvalue and the sum of all eigenvalues), and determine how many PCs to interpret. ***Square the standard deviations in your output to get eigenvalues. What are the first ten eigenvalues using a correlation matrix and the Fitz-Bew song data? What can you type at the R prompt to get all of these at once?***

	PC-1	PC-2	PC-3	PC-4	PC-5	PC-6	PC-7	PC-8	PC-9	PC-10
λ =										

There are a number of ways of deciding how many PCs to interpret, and this is a controversial topic. One approach, if interpreting a PCA done on a correlation matrix is to consider any PC with an eigenvalue >1. This does not work when a covariance matrix is used because its eigenvalues all tend to be <1. Another approach is to consider all the PCs that together explain 90% of total variance. 90% is a completely arbitrary number, with some people using 95% and others 80%, making this approach difficult to justify. A third approach, probably the most used, is to view a screeplot and select the PCs that explain considerable variation. A screeplot is a plot of eigenvalues against PC number. You inspect the plot to see a sudden drop in eigenvalue from one PC to the next, and then interpret all PCs before the drop. You can make a screeplot using the following function, which labels the y-axis as "variance", which is equivalent to eigenvalue (you can confirm this for yourself):

```
> screeplot(pca_object, type=c("barplot", "lines")
```

How many PCs should you interpret with the Fitz-Bew dataset using each approach?

Method:	Eigenvalue > 1	90% cumulative variance	Screeplot
# PCs:			

Each component explains an independent component of variance in the original data and all original variables contribute in different proportions to each PC. The component loadings tell you how the original variables contribute. The higher the magnitude of a loading, the more that variable contributes. Positive and negative signs show you how the variables are related. If a loading is negative, then as the PC value increases, the value for the original variable in question decreases, and vice versa.

Try to interpret the first two PCs. In the assignment section at the end of the lab, fill out the table with the loadings for the first three PCs. Add the eigenvalues and the percentage of variance explained by each. The write the interpretation for the first two PCs to the right of the table. To help you, consider only loadings >0.3 and identify these in bold.

The next step in our PCA is to examine the factor scores. The factor scores give us the coordinates of each sampling unit in PC space, and we can visualize this easily by plotting factor scores for pairs of PCs in a scatter plot. ***What are the sampling units in the current dataset?***

You can obtain the scores quite easily from your estimated PCA object because they are already calculated. ***Look at the structure of the fb_pca1 object. What would you type to get the factor scores output to the screen?***

Create a data frame that contains your factor object "maleID" and all of the factor scores. Change the row names to match the song identifier. Create a scatter plot of PC-1 against PC-2 with the maleID color-coded. Save the figure and insert it at the end of this lab, in the assignment section. Are the males recognizable by their song characteristics? How many males seem readily identifiable?

3. Post hoc Analysis

In the previous section, you completed a PCA of the Willow Flycatcher call data. Part of the versatility of PCA is that it produces a set of new, independent variables, called principal components (PCs) that can then be further analyzed using your favorite statistical techniques. Analysis options are really just limited by creativity and molded by the biological questions you are interested in. To demonstrate some of what can be done, we will do some *post hoc* analyses of the factor scores that you plotted in the last section.

Two questions of biological importance come to mind. First, are males distinguishable from one another based on call characteristics? This question could be addressed just as easily with DFA, but the appeal of PCA is that it does not presuppose that males are different. This question could be biologically quite interesting because if males are distinguishable by their call, then other males may recognize stronger males before engaging in an altercation, and females may be able to distinguish their mate from others. Second, we may be interested in whether the calls that some males produce are more variable than songs of other males. One could imagine a situation where having a consistent call is important, especially when one needs to be recognized by other individuals.

Assignment: (10 points)

How would you test the two biological questions described above using your PCA and other statistical tests? Be sure to mention what tests you would use (Hint: It may be a good idea to refer to past labs for ideas).

Complete the table below to report your PCA results. Identify loadings >0.3 with bold font, and include a written interpretation of PC-1 and PC-2 to the right of the table. Insert your plot of factor scores for PC-1 and PC-2 below the table.

Variable	PC-1	PC-2	PC-3
T1			
T2			
T3			
T4			
T5			
T6			
T7			
T8			
T9			
F1			
F2			
F3			
F4			
Eigenvalue			
% explained			

For all of the analyses in this section, consider the first two PCs. Start by testing the first biological question using your chosen analysis. Insert a table to report the results and write a sentence or two interpreting the results. Make sure that you do a complete analysis and present the statistics necessary for a reader to fully interpret it. When considering which males are different, a diagram or graphic may help. Are there any males that are completely unique on the PC? Give a biological conclusion.

Now do the analysis necessary to answer the second biological question. Report the results and interpretation in the space provided.

You can take a further look at how variances compare among males for the first two PCs by calculating the variances and plotting them. Do the following steps, referring to past labs as needed:

1. ***Use the tapply function to calculate variance for each male for PC-1 and PC-2, saving the results for each PC in a separate object.***
2. ***Note that you should get a vector with variances, with the maleID number as the "name" of each cell.***
3. ***Plot the variance against maleID and include the plots here. Write a sentence or two describing your findings.***