Running PCA

Programs that implement PCA: Eigenstrat (https://github.com/DReichLab/EIG/tree/master/EIGENSTRAT), KING (http://people.virginia.edu/~wc9c/KING/), R, and many, many, many others.

Additional reading:

* Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904.
* Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... & Stephens, M. (2008). Genes mirror geography within Europe. *Nature*, *456*(7218), 98.
* McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS genetics*, *5*(10), e1000686.

Running it in R (**bold-underlined** statements are R commands you can use)

Steps

1. Open R
2. Type **?prcomp** to learn about this function and how to use it
3. Let’s read in some data
   1. Use **data=read.table(“”)** with the name of the file in the quotes
   2. One example is **data=read.table(“SimDat.A0.0.rdat”)**
   3. The first column are population labels so let’s pull them out and give them their own data structure:
      1. **labels = data[,1]**
      2. **data=data[,2:length(data[1,])]**
   4. Let’s get ready some colors. I use which, “which” probably isn’t the best way but it works
      1. **col=rep("black",length(labels))**
      2. **col[which(labels=="P2")] = "red"**
      3. **col[which(labels=="P4")] = "blue"**
      4. **col[which(labels=="A3")] = "purple"**
      5. **col[which(labels=="A5")] = "orange"**
4. Run prcomp
   1. You will need the “x” output so specify retx (“return x”), though it is in the default as well
   2. Centering and scaling are also suggested.
   3. So that leaves us with:
      1. **pca = prcomp(data,center=TRUE,retx=TRUE,scale.=TRUE)**
5. So what do you have?
   1. What is x?
   2. If you plot x what do you see?
      1. **plot(x=pca$x[,1],y=pca$x[,2],col=col,pch=19)**
   3. How much variation does each PC explain?
      1. sdev is the square root of the eigenvalues for each PC. When we divide the absolute value of each eigenvalue by the sum of all the values we get the proportion of variance that the PC explains.
      2. **pca$sdev[1]\*\*2/sum(pca$sdev\*\*2)**
      3. **barplot(pca$sdev[1:10]\*\*2/sum(pca$sdev\*\*2))** for the first 10 PCs (beyond that it gets difficult to look at)
      4. This is a scree plot
      5. Which PCs explain most of the variation?
      6. Which don’t seem to add much?
6. What do you think the “A” in the file names stands for?
   1. Run the other data sets and see how they differ.