When trying to find the number of clusters, as is known, I get different results when I retain different numbers of PCs.

As background, I have samples from 180 individuals over 11 different sites, and am trying to find the best structure.

In the tutorial, it says that when you run find.clusters there is no reason for keeping small numbers of principle components here. When I run with n.pca.max = 60 (so, n/3), using xval I get pretty consistently that the good number of PCs to retain is 50.



When I run find.cluster using 50 PCs I get anywhere between 7 and 9 clusters, mostly telling the same story for the data. However, when I run find.cluster with over 100 PCs I consistently get k = 4 or 5, and the plot is much cleaner. In addition, however, when I look at my variance explained plots, they don’t really asymptote, either for find.cluster or for dapc.

Both of the variance explained plots look like



Using the scaled dataset

mat <- scaleGen(Stickle8c10NoOdds, NA.method="mean")

I use 120 PCs, and get



Running with 90 PCs



However, if I run find.cluster and choose 50 PCs, I get



> head(NumClust$Kstat, 11)

 K=1 K=2 K=3 K=4 K=5 K=6 K=7 K=8 K=9 K=10 K=11

1492.620 1472.790 1455.980 1448.216 1443.735 1442.909 1440.166 1440.344 1440.867 1441.979 1443.101

Are the xval procedure results (i.e., 50 PCs in my case) meant to be used only at the dapc1 <- dapc(mat, NumClust$grp) stage? And, do my variance explained plots concern you at all given that they don’t asymptote?