Progress report: *Emys marmorata* sub-species identification and classification

April 3, 2013

1 Methods

No-free lunch theorem. Try lots of things because we don't understand everything.

1.1 Unsupervised

Underlying structure in data?

1.1.1 Gap-based clustering

Comparison of gap statistic results for partitioning around medoids (PAM) divisive clustering. Confidence intervals are determined via bootstrap. The higher the gap statistic, the better the clustering result. Standard errors of the gap statistic were estimated from 100 resamples.

1.1.2 Evidence Accumulation Clustering

Choosing an optimal number of partitions is hard, which is why gap-based cluster selection was used above. An alternative method is to look at the co-occurrence frequency, that is how frequently any two samples occur in the same partition. Repeating this process over and over again creates the frequency, or "vote" for how the data set should be partitioned and which specimens should be in the same cluster.

EAC was originally devised using k-means clustering, but I've extended it to use PAM clustering instead. The hope is to determine underlying structure in the data given a wide enough partition range and a high enough number of iterations. Dissimilarity based EAC was performed using a range of 1 though 200 possible partitions and based on 10,000 iterations.

1.2 Supervised

How well does data conform to predetermined structure?

- multinomial logistic regression
- feed-forward neural networks
- random forests

2 Preliminary results

2.1 Unsupervised

Comparison of gap statistic over a very wide range of plausible partitions indicates that as the number of partitions increase, there is a marginal increasing gap statistic until approximately 31 after which there is a marginal decrease in gap statistic (Fig. 1). It is notable that the standard errors around the gap statistic values are very large, and the marginal increases in gap statistic with an increased number of partitions may not be important. Additionally, all gap statiic values are within 0.0065 of each other meaning that there is little over all consensus for how many clusters are present when comparing gap statistics.

Dissimilarity based EAC estimated approximately 4 optimal partitions. This corresponds to the number of partitions in both the Spinks et al. and shXXX classification schemes. However, the number of individuals assigned to each class is very different.

	tmorph.dac
1	33
2	3
3	730
4	10

Table 1: Number of specimens assigned to optimal number of partitions as determined by dissimilarity based EAC. Each column corresponds to a different partition, with the number assigned directly below it.

2.2 Supervised

In the interest of space/time, I'm only going to display results from one of the selected models from each method for each classification scheme.

I'm currently holding back on showing the tables because there are a lot of them.

But I can show the comparison of the predictive accuracies.

3 Miscellaneous affairs

3.1 Evolution 2013

Machine learning approaches for recognizing cyptic diversity: a case study using *Emys marmorata*.

3.2 Grants

3.2.1 Hinds Fund

Resubmit previous application with the results from this study in the prelim-results section?

Get network for one turtle? This should take long, just requires time.

3.2.2 Paleontological Society

Why not? Need to wait till next year. Should probably join PaleoSoc anyway.

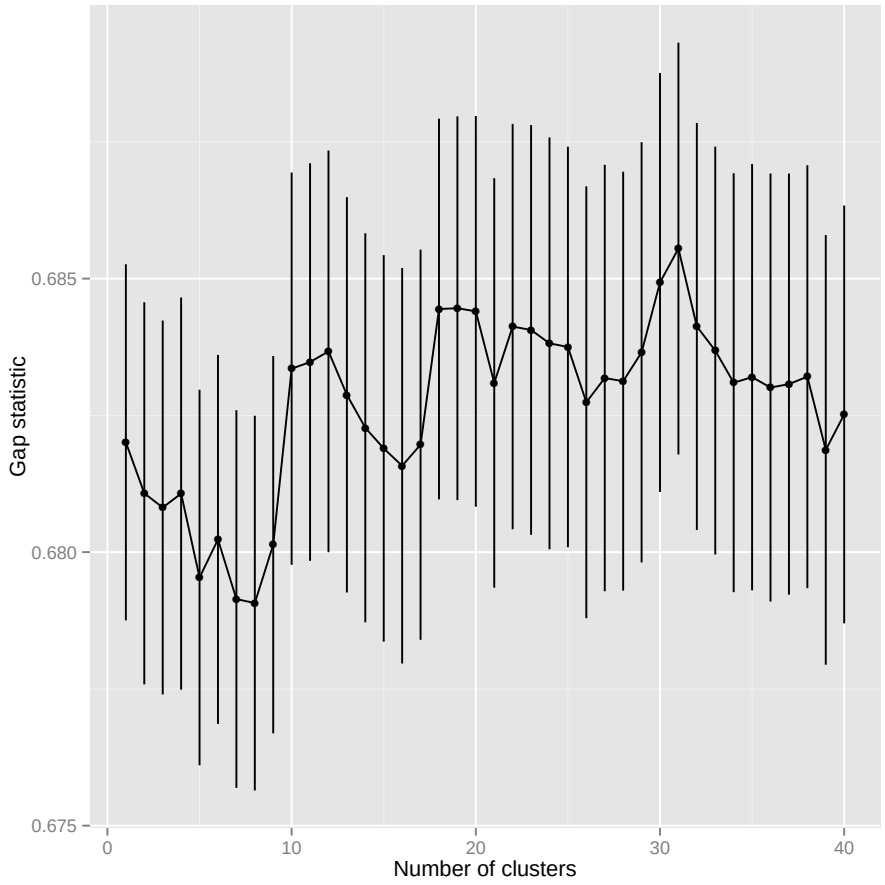

Figure 1: Gap statistic values for multiple PAM-based clustering configurations of the Riemmanian shape distances of the *Emys marmorata* plastra. Higher values indicate greater clustering. Standard errors are estimated from 100 bootstrap resamples.