Informal explination of study "Ecotype detection in (geometric) morphometric data"

Peter D Smits Benjamin Frable

January 12, 2013

1 Introduction

Stuff

2 Materials and Methods

2.1 Fish

Four families of Characiformes: Anostomidae, Chilodontidae, Curimatidae, and Prochilodontidae.

Total 78 specimens for four families of Characiformes: 18 Anostomidae, 4 Chilodontidae, 48 Curimatidae, and 8 Prochilodontidae.

2.2 Morphometric analysis

35 two-dimensional landmarks and the head and neck. Landmarks were selected for some reason and were collected using TPSDig2. Generalized Procrustes analysis was used to remove the effects of orientation, rotation, and scaling CITATION. Following this, points were projected onto tangent space using principal components analysis CITATION. This was done in the R statistical programming language CITATION using the "shapes" CITATION and "geomorph" packages.

2.3 Machine learning

The predictive accuracy of shape for family as assessed using multinomial logistic regression and a training set of 75% of all samples and a test set of 25% of

all samples. Multinomial logist regression is a generalization of the standard logistic regression where instead of a binary response variable, there are multiple categorical responses. In this case, the multinomial response variable was family while predictors were the pricipal components of shape.

Pricipal components analysis produced 70 eigenscores for every taxa and there are only 78 taxa sampled, it is necessary to select only from a subset of the possible predictors. Models were compared using the second-order Akaike's information criterion (AICc) where lower values indicate a better bias-variance trade off between the number of predictors and the likelihood of the model CITATIONS while taking into account sample size.

Model selection is not without uncertainty, as some models may have very close AICc values. Model selection was done using δAICc and Akaike weight values CITATION. δAICc values are calculated as the difference between the AICc of a model and the AICc of the AICc-best model, or the model with the lowest AICc. Akaike weights sum to 1 and represent the propotional amount of information explained by a particular model and are approximately the model selection probability. Models with a δAICc of less than 2 were considered approximately identical in information to the AICc best model, while models with a δAICc of less than 6 were considered as plausible but suboptimal. Models with a δAICc of greater than 10 were considered extremely poor.

I will eventually be making predictions from a model using averaged parameters of the first 95% confidence set of models, as measured by Akaike weights CITATION.

The models compared varied between 1 and 7 pricipal component predictors, producing models with 6, 9, 12, 15, 18, 21, 24 parameters respectively.

I did some model training using "caret" and "e1071" packages. Need to explain.

Following model training and model comparison, the AICc-best model was used to predict the family of the testing data set described above.

3 Results

3.1 Morphometric analysis

3.2 Machine learning

4 Discussion

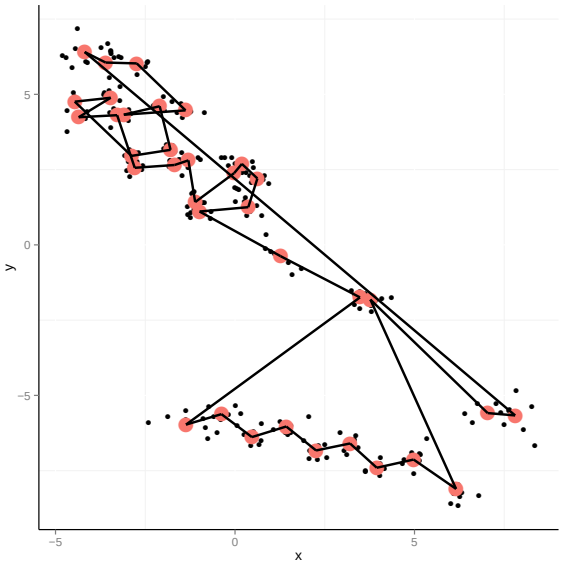

Figure 1: Procrustes fit landmark dispersion for all taxa in comparison to mean landmark position in red.

	(Intercept)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	family	df	logLik	AICc	delta	weight
2	+	+	+							9.00	-14.34	50.35	0.00	0.97
3	+	+	+	+						12.00	-13.18	57.15	6.80	0.03
4	+	+	+	+	+					15.00	-11.46	64.09	13.74	0.00
5	+	+	+	+	+	+				18.00	-8.82	70.73	20.39	0.00
1	+	+								6.00	-33.71	81.03	30.68	0.00
6	+	+	+	+	+	+	+			21.00	-7.10	81.17	30.82	0.00
7	+	+	+	+	+	+	+	+		24.00	-0.04	83.38	33.04	0.00

Table 1: Model selection table.

	ano	chi	cur	pro
ano	4	0	0	0
chi	0	1	0	0
cur	0	0	12	0
pro	0	0	0	2

Table 2: Confusion matrix.

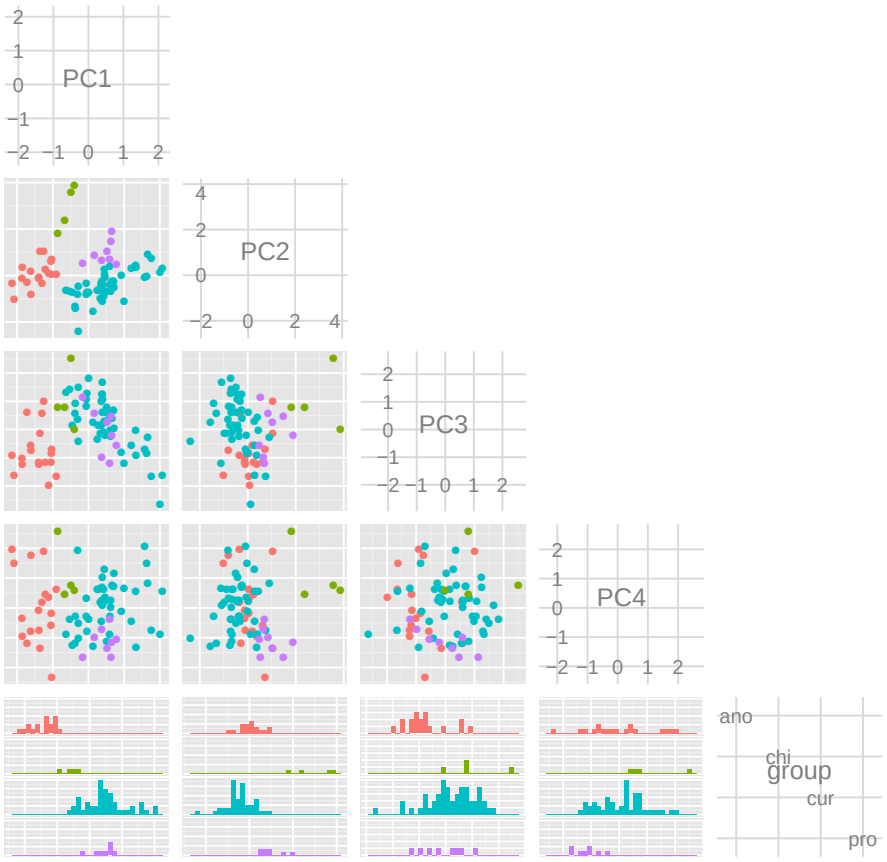

Figure 2: Comparison of first four pricipal components of Procrustes fit landmarks. Familes are highlighted.

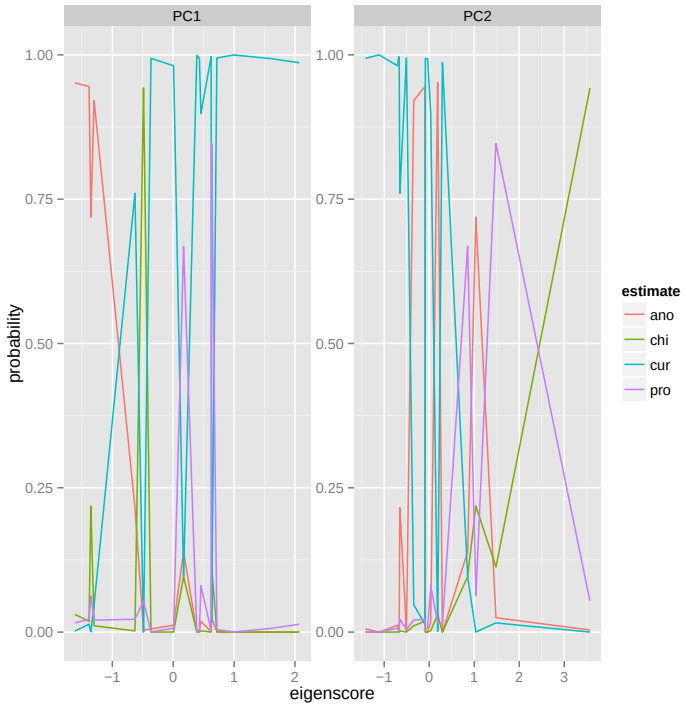

Figure 3: Relative probabilities of Characiform family identification from the first two principal components of shape. The first two pricipal components are plotted here because they are the predictors of family from the AICc-best model.