9. Principal Component Regression

FISH 560: Applied Multivariate Statistics for Ecologists

Topics

- Using the scores from a PCA in multiple regression

R Packages: vegan
R Source: biostats

BACKGROUND

Principal component analysis has played an important role in ecology by providing an indispensable analytical tool for elucidating structure in inherently complex datasets. Principal component scores (i.e., the set of orthogonal or independent coordinates derived from the principal component analysis) or scores from PCoA are often visually inspected in reduced-ordination space, or used to summarize the major patterns of variation in the data prior to conducting post hoc statistical analyses (e.g., regression analysis, ANOVA designs: see Jackson 1991; Jolliffe 2002).

Increasingly, ecologists are using the principal components rather than the original explanatory variables in regression analyses; an approach called principal component regression (PCR). In community ecology, for example, fish-habitat relationships have been examined by regressing a biological variable (e.g., species abundance, species richness) against the principal components derived from an environmental dataset (e.g., Kodric-Brown & Brown 1993; Zampella & Bunnell 1998; Marsh-Matthews & Matthews 2000; Romanuk & Kolasa 2002 are just a few examples). For example, Marsh-Matthews & Matthews (2000) used PCR to explore the association between fish species richness and landscape factors (over 30 variables summarized by a few principal components) for streams in midwestern United States. The use of composite scores from a subset of principal components (see Jolliffe 1982; 2002 for a review of principal component selection in regression) has the prospective advantage of alleviating a number of statistical problems associated with regression analysis, including the deleterious effects of multicollinearity (e.g., biased and imprecise estimates of the regression coefficients), biased model selection (e.g., erroneous selection of random variables in the regression model when examining large numbers of variables) and the interaction between the two (e.g., failure to identify significant variables). Therefore, it has been suggested that PCR may provide a simple solution to the negative influence of multicollinearity in regression analysis, and result in models with more stable estimates of variable importance and greater predictive power compared to using the original explanatory variables (Jolliffe 2002).

Here, let's conduct a simple Principal Component Regression that quantifies the relationship between stream habitat characteristics and fish species richness in the class dataset.

SET-UP

In this exercise you will be working with the MAHA species abundance dataset. But first remember to set-up your R work session by defining the current work directory to your folder of choice and loading the vegan library. Also, make sure to source the BIOSTATS file from the *File* pull-down menu. You can also do this using the functions **`setwd`**, **`library`** and **`source`**. Import the dataset by typing:

```
speocc <- read.csv('MAHA_speciesocc.csv',header=TRUE, row.names=1)
envdata <- read.csv('MAHA_environment.csv', header=TRUE, row.names=1)
```

Let's transform the data before diving into the analysis. We will do this because the species abundance dataset is highly skewed and contains some rather large values which are valid but highly influential. Remember, the log of zero is undefined so we'll add 1 to each value in our data set.

PRINCIPAL COMPONENT REGRESSION

First, let's conduct a PCA on the environment dataset following the instructions provided in Chapter 6.

```
env.pca<-prcomp(envdata, scale=TRUE)
summary(env.pca)
```

This returns the following information:

```
Importance of components:
                          PC1   PC2   PC3   PC4    PC5    PC6    PC7    PC8
Standard deviation     1.655 1.377 1.203 1.168 0.8916 0.7472 0.7225 0.5870
Proportion of Variance 0.274 0.190 0.145 0.136 0.0795 0.0558 0.0522 0.0345
Cumulative Proportion  0.274 0.463 0.608 0.745 0.8241 0.8799 0.9321 0.9666
```

Note that the first 4 PCs explain almost 75% of the variation in the original environmental dataset, therefore we will use the principal component scores from these axes to represent the dominant gradients of environmental variation. Recall from Chapter 6 that we can extract these scores by simply typing:

```
env.scores<-env.pca$x
```

Next, let's conduct a regression analysis of fish species richness against the composite PCs representing the majority of the variation in the original environmental dataset.

Calculate species richness for each site:

```
sperich<- rowSums(speocc)
```

Merge the species richness vector with the PC scores:

```
data<-data.frame(cbind(sperich,env.scores))
```

Conduct a simple linear regression relating the first four PCs to species richness:

```
reg<-lm(sperich~PC1+PC2+PC3+PC4,data)
summary.lm(reg)
```

This will result in the following regression output

```
Call:
lm(formula = sperich ~ PC1 + PC2 + PC3 + PC4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3941 -2.3419 -0.1691  1.9075  8.5136

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.7333     0.5232  16.693  < 2e-16 ***
PC1          -0.3869     0.3197  -1.210 0.233248
PC2          -0.6622     0.3843  -1.723 0.092577 .
PC3           0.9513     0.4398   2.163 0.036559 *
```

```
PC4           1.6824     0.4530   3.714 0.000623 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.51 on 40 degrees of freedom
Multiple R-squared: 0.3641,     Adjusted R-squared: 0.3005
F-statistic: 5.726 on 4 and 40 DF,  p-value: 0.0009677
```

OPTIONAL READINGS

Jolliffe, I.T. 1982. A note on the use of principal components in regression. Applied Statistics 31: 300-303.

Jolliffe, I.T. 2002. Principal Component Analysis, 2nd ed. Springer-Verlag, New York.

EXERCISE

Purpose

Upon completion of this chapter, you should be able to complete a Principal Component Regression.

Tasks

- Conduct a Principal Component Regression to analyze the relationship between fish species richness and derived PCs based on stream environmental characteristics.
    - What does the output from the regression tell you regarding the fish richness-habitat relationship?
    - Can you determine which environmental variables are the most important for predicting species richness?
    - Can you better predict species richness using a Principal Component Regression approach or a traditional regression using the raw environmental variables?