Multivariate Regression, 2B-PLS and Bootstrapping

Last Lecture
Project presentations
Project write-ups
Review
Regression
Multivariate Regression
2-Block partial least squares
Bootstrapping
Project Presentations

**Length:** 12 minutes (10 minutes for presentation, 2 for questions)

**Slides:** e-mail to me (pdpolly@indiana.edu) by 12 pm on Friday

**Content:** succinct background to project, methods and data, results, interpretation

Project Reports

**Length:** 10 to 15 pages (guideline only)

**Deadline:** Due Wednesday, May 2 (“exam date”)

**Content:** Written like journal article, with intro, methods and data, results, interpretations, conclusions, bibliography, figures (remember recommendations about good graphic style)
Review

1. Collect landmark coordinates
   1. 2D landmarks or outlines -> photographs
   2. 3D landmarks or outlines -> Microscribe or other device
   3. Convert to TPS format or CSV, import with tpsImport[] or Import[]

2. Standardization of Shape Coordinates
   1. Landmarks -> Procrustes
   2. Outlines -> Phi function (Eigenshape), Procrustes (semilandmarks)
   3. EDMA -> Interlandmark Distance Matrix

3. Analysis
   1. Principal Components -> Shape scores (shape coordinates) for further work
   2. Regression -> comparison with continuous variables
   3. ANOVA -> comparison with meristic variables
   4. Tree building or clustering -> examination of heirarchical patterns
   5. Projection of tree to shape space -> evolution of shape
   6. Two-block Partial Least Squares -> comparison with many continous variables
Mathematica Functions for Shape Analysis

EDMA[landmarks1, landmarks2, landmarklabels]
Euclidean Distance Matrix Analysis: useful for studying which parts of a shape are more variable than others

PrincipalComponentsOfShape[data, {PC1, PC2}, labels]
Finds major axes of shape variation in a single data set and constructs models of the distribution of shape in the associated morphospace for 2D landmarks

PrincipalComponentsOfShape3D[data, {PC1, PC2}, labels]
Finds major axes of shape variation in a single data set and constructs models of the distribution of shape in the associated morphospace for 3D landmarks

TreeToMorphospace[proc, labels, PCs, tree]
Projects a phylogenetic tree into morphospace by reconstructing the ancestral nodes based on Brownian motion null model of evolutionary change

ShapeRegress[proc, variable (, PCs)]
Does multivariate regression of entire shape onto a continuous variable

TwoBlockPartialLeastSquares[data1, data2, mode (, PLS)]
Analysis of correlation between two multivariate data sets
Review

Eigenvalues
variance on each PC axis
(In Mathematica: \textit{Eigenvalues[CM]})

Eigenvectors
loading of each original variable on each PC axis
(In Mathematica: \textit{Eigenvectors[CM]})

Scores (=shape variables)
location of each data point on each PC axis
(In Mathematica: \textit{PrincipalComponents[resids]})

resids are the residuals of the Procrustes coordinates
CM is the covariance matrix of the residuals
Statistical analysis: partitioning variance

In GMM, variance / covariance = variation in shape

Purpose of statistical analysis is to ascertain to what extent part of that variance is associated with a factor of interest, aka partitioning variance.

1. **P-value**: indicates whether the association is greater than expected by random chance

2. **Regression parameters (slopes, intercepts)**: indicate the axis in shape space associated with the factor, useful for modeling the aspect of shape associated with the factor

3. **Correlation coefficient \((R)\)**: indicates the strength of the association between the variance and the factor

4. **Coefficients of determination \((R^2)\)**: indicates the proportion of the variance that is associated with the factor
Univariate linear regression

\[ Y = a \, X + b + E \]

Linear regression of one Y variable onto one X variable.
We have regressed one principal component onto one explanatory variable.

\[ a = 2.0 \]
\[ b = 0.5 \]
\[ R^2 = 0.85 \]

\( R^2 \) also ranges from 1.0 (100% explained) to 0.0 (0% explained).
Regression - a multivariate look

Univariate Linear Regression

Multiple Linear Regression

Multivariate Linear Regression
ShapeRegress[$proc, variable (, PCs)]

Function for multivariate regression of entire shape onto a single independent variable

- $proc$ is a matrix of Procrustes superimposed landmark coordinates with objects in rows and coordinates in columns.
- $variable$ is the variable onto which $proc$ is to be regressed. It is a vector containing observations for each object from a continuous variable.
- $PCs$ is an optional parameter specifying which PC should be shown in the graph. By default the regression on PC1 is shown.
Example of \texttt{ShapeRegress[]}

\texttt{ShapeRegress[proc, x, 2]}

\begin{verbatim}
R-square (all PCs) = 0.21
P[R-square is random] = 0.34
\end{verbatim}

\begin{tabular}{l|cccc}
 & PC1 & PC2 & PC3 & PC4 \\
\hline
Intercept & -0.405353 & -0.63799 & -0.022762 & -0.07787 \\
Slope & 0.550962 & 0.867165 & 0.0309385 & 0.105842 & -0.0315015 \\
Univariate R-square & 0.09 & 0.83 & 0.00 & 0.07 & 0.01 \\
\end{tabular}

Graph shows one dimension of the regression

R-square shows total amount of shape explained by \texttt{x} variable

Table gives regression coefficients and univariate r-squared
Two-Block Partial Least Squares

Shape is inherently multivariate

Independent variables, such as diet, vegetation, precipitation, and temperature, are likely to be correlated with one another

The two correlated independent variables will have overlapping correlation with shape (if vegetation and precipitation are correlated with each other, then both will be correlated with shape if one of them is)

One needs to control for the correlations between independent variables in order to properly test many of them
Two-block partial least squares (2B-PLS)
Multivariate regression lines (major axes of correlation)

PLS axis one is best regression of data set one on data set two (where best means that it explains the most of both data sets)
**Example from Rohlf and Corti paper**

**Regression axes**

**Table 2. Results of partial least-squares analysis of the correlation matrix in Table 1.**

<table>
<thead>
<tr>
<th>Matrix</th>
<th>Variable</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>$F_1$</td>
<td>Skull length</td>
<td>0.737</td>
</tr>
<tr>
<td></td>
<td>Skull breadth</td>
<td>0.676</td>
</tr>
<tr>
<td>$F_2$</td>
<td>Fibula</td>
<td>0.520</td>
</tr>
<tr>
<td></td>
<td>Tibia</td>
<td>0.495</td>
</tr>
<tr>
<td></td>
<td>Humerus</td>
<td>0.479</td>
</tr>
<tr>
<td></td>
<td>Ulna</td>
<td>0.505</td>
</tr>
<tr>
<td></td>
<td>Singular value</td>
<td>1.619</td>
</tr>
<tr>
<td></td>
<td>Correlation</td>
<td><strong>0.670</strong></td>
</tr>
</tbody>
</table>

Correlation between blocks on each axis.

Block 1

Block 2
Example from Rohlf and Corti paper

### Table 3

Results of partial least-squares analysis of correlation matrix in Table 1. Correlations within and among columns of F₁ and F₂ as identified in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>F₁</th>
<th>F₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>F₁</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.000</td>
<td>-0.062</td>
</tr>
<tr>
<td>2</td>
<td>-0.062</td>
<td>1.000</td>
</tr>
<tr>
<td>F₂</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.670</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>0.106</td>
</tr>
</tbody>
</table>

**Block 1**

**Block 2**

**Regression axes**
Example from Rohlf and Corti paper

TABLE 6. Results of a partial least-squares analysis of the covariance between the square roots of the numbers of different kinds of prey items found in the stomach and the partial warps for the *Plethodon* data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acarina</td>
<td>0.028</td>
<td>0.039</td>
<td>0.037</td>
</tr>
<tr>
<td>Eggs</td>
<td>-0.004</td>
<td>0.139</td>
<td>-0.171</td>
</tr>
<tr>
<td>Isoptera</td>
<td>0.020</td>
<td>-0.190</td>
<td>0.071</td>
</tr>
<tr>
<td>Collembola</td>
<td>0.015</td>
<td>-0.118</td>
<td>0.141</td>
</tr>
<tr>
<td>Cheloneithida</td>
<td>0.026</td>
<td>0.004</td>
<td>0.042</td>
</tr>
<tr>
<td>Hymenoptera</td>
<td>-0.578</td>
<td>0.005</td>
<td>0.656</td>
</tr>
<tr>
<td>Gastropoda</td>
<td>0.106</td>
<td>0.134</td>
<td>0.278</td>
</tr>
<tr>
<td>Larvae</td>
<td>-0.073</td>
<td>0.409</td>
<td>0.226</td>
</tr>
<tr>
<td>Coleoptera</td>
<td>0.679</td>
<td>-0.342</td>
<td>0.472</td>
</tr>
<tr>
<td>Diptera</td>
<td>-0.291</td>
<td>-0.185</td>
<td>0.169</td>
</tr>
<tr>
<td>Araneida</td>
<td>0.078</td>
<td>0.130</td>
<td>0.078</td>
</tr>
<tr>
<td>Isopoda</td>
<td>0.260</td>
<td>0.754</td>
<td>0.152</td>
</tr>
<tr>
<td>Orthoptera</td>
<td>-0.051</td>
<td>0.061</td>
<td>-0.221</td>
</tr>
<tr>
<td>Diplopoda</td>
<td>0.042</td>
<td>0.009</td>
<td>-0.054</td>
</tr>
<tr>
<td>Oligochaeta</td>
<td>0.153</td>
<td>-0.001</td>
<td>0.224</td>
</tr>
<tr>
<td>Chilopoda</td>
<td>-0.017</td>
<td>0.057</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Singular values: 0.0547, 0.0096, 0.0076
Correlations: 0.725, 0.557, 0.442
TwoBlockPartialLeastSquares[proc1, proc2, {"Shape", "Shape"}]

This function performs a two-block partial least squares analysis following the methodology of Rohlf and Corti (2000). Two blocks of data are given to the function, along with a list of two strings indicating the type of data. Allowable types are "Shape" (Procrustes superimposed coordinates), "Standardized" (independent variables with different units of measurement that need to be standardized), and "Unstandardized" (independent variables with the same unit of measurement that do not need to be standardized). An optional argument is the number of the PLS axis to plot in the output graph. By default PLS 1 is plotted.

Arguments:

- **data1 and data2** are two blocks of variables, one or both of which can be a matrix of Procrustes superimposed landmark coordinates with objects in rows and coordinates in columns.

- **{type1, type2}** is a list of two data types, in quotation marks. Allowable types are "Shape" (Procrustes superimposed coordinates), "Standardized" (independent variables with different units of measurement that need to be standardized), and "Unstandardized" (independent variables with the same unit of measurement that do not need to be standardized).

- **PLS** is an integer indicating which PLS axis to plot.
TwoBlockPartialLeastSquares[\textit{data1, data2, \{type1, type2\}, (, PLS)]

Proportion of Actual Squared Covariance Explained by PLS 1: 0.48
Proportion of Total Possible Squared Covariance Explained by All PLS Axes: 0.08

Scatter plot of scores of each data block on their shared PLS axis

Shape models for the positive end of the PLS axis, showing what aspects of the two shapes are correlated
Statistics, Tests, and Bootstrapping

Statistic – a measure that summarizes some feature of a set of data (e.g., mean, standard deviation, skew, coefficient of variation, regression slope, correlation, covariance, principal component, eigenvalue, f-value).

Statistical parameter – the value of a particular statistic for the entire population.

Statistical estimate – the value of a particular statistic for a sample of a population.

Statistical test – an assessment of how likely a null hypothesis is to be true given the data at hand, or how probable it is that the data fit the null hypothesis given a random sample of the population.

Null hypothesis – usually the hypothesis that the estimated statistics from two or more samples result from two or more random samplings of the same population. In other words, the hypotheses that the sample do not come from different populations.

Confidence interval or standard error – a metastatistic that expresses how closely a statistical estimate is likely to match the population parameter.
Bootstrapping and randomization tests

Used when assumptions of ordinary (parametric) statistical tests are not met, or when they are not known

These ‘Monte Carlo’ tests randomize the data with respect to the statistic being measured

Randomization is repeated a large number of times (e.g., 10000) and a distribution of the randomized statistic is generated.

Observed value from the real data is compared to see whether it falls within the range of randomized values

These tests take biases, non-normality, etc. into account automatically
Types of randomization tests

Bootstrap. Random resampling of original data, recalculation of test statistics to determine standard errors.

Jackknife. Same as bootstrap, but where each individual data point is left out in turn and the test statistic recalculated each time to determine standard error.

Randomization. Randomizing original data, test observed compared to randomized samples.

Monte Carlo. Data are simulated based on a particular hypothesis or model, real data are tested against the simulated data to see if the model holds.
Example: Test for difference in mean

1. Choose statistic that describes difference in mean:

\[ D = \sqrt{\text{Mean}[\text{sample1}] - \text{Mean}[\text{sample2}]}^2 \]

2. Pool samples and randomly draw new sample 1 and 2 with replacement.

3. Calculate D for randomly drawn samples

4. Repeat 10,000 times

5. Compare real D with randomized D distribution

6. P-value is the proportion of randomized D smaller than real D.
Bootstrap replicates vs. theoretical normal density

**Fig. 2.** Histogram of $B = 1000$ bootstrap replications of $\hat{\theta}^*$ for the law school data. The normal theory density curve has a similar shape, but falls off more quickly at the upper tail.

Efron and Tibshirani, 1986