USING SAS(r) SOFTWARE TO QUANTIFY DISSIMILARITY AMONG ENVIRONMENTAL SAMPLES

John M. Rinehart, RMC Environmental Services

Abstract

This paper provides computational support for a data analysis technique proposed by Ramsey and Marsh in Biometrics 40(3) (1984). Their paper, "Diet Dissimilarity," outlined a method of quantifying the relative importance of prey taxa in determining overall dissimilarity of diets between groups of predators, using a Bayes classification procedure. The method can be extended for use as a general dissimilarity index between any two groups of multivariate environmental samples. SAS code is provided to implement the dissimilarity calculations under a variety of sampling models. It is hoped that this will stimulate further research into the theory and applicability of this technique.

Introduction

Environmental research often generates data which are essentially multivariate in character, with each "sample" consisting of a vector of recorded values, one for each of a set of descriptive variables. Many environmental samples are meant to describe some collection of organisms, such as a natural plant or animal community, a parasitic infestation in a particular host, or a predator's diet. In such cases, the descriptive variables usually represent some taxonomic classification, with values indicating the number, proportion, or presence of organisms in the sample.

Comparisons and contrasts among such samples are made with a variety of multivariate analysis techniques, the most common being the calculation of some sort of similarity index. Huhti (1979) and Green (1980) provide good reviews of the many similarity indices which have been proposed. All of them, when used in appropriate circumstances, provide some measure of "likeliness" among environmental samples. However, it is frequently more interesting to ask, even of very similar samples, just what in fact makes them different.

Ramsey and Marsh (1984) have recently proposed a new index of dissimilarity to help quantify differences between groups of multivariate samples and identify which descriptive variables are most important in determining those differences. This paper provides SAS Software code to implement their proposal.

The Dissimilarity Technique

In their 1984 paper, Ramsey and Marsh give a detailed description of their "coefficient of dissimilarity", including theory and model parameter estimation. I attempt only a brief outline here. The interested reader is encouraged to consult the original text.

It is first observed that the dissimilarity between two groups of environmental samples should be directly related to our ability to correctly classify an individual sample into one of the groups, based on the observed values of its descriptor variables. The problem is then posed as a Bayes classification procedure. The probability of belonging to class s is proportional to

\[ \pi_s f_s(x) \]

where x is the sample vector of descriptor values, \( f_s(x) \) is the joint probability density function of the descriptor variables for class s, and \( \pi_s \) is the prior probability of belonging to class s. Then samples might be classified according to the sign of the logit score

\[ \lambda(1|x) = \log \left( \frac{\pi_1 f_1(x)}{\pi_2 f_2(x)} \right) \]

Ramsey and Marsh suggest that certain functions of the exponential family may reasonably be used as models for the joint probability density functions of environmental samples. In particular, they suggest the independent Poisson model for descriptor variables representing counts, the multinomial model for proportions, and the independent binomial model for presence/absence data. They further discuss and provide parameter estimate formulae for each of the models, which allow calculation of the following statistics from sample data.

Since the magnitude of the logit score for each sample is directly related to the probability of correct allocation, a reasonable "coefficient of dissimilarity" (DIS) might be the average logit score for correctly classifying all samples,

\[ \text{DIS} = \frac{1}{2} \left\{ \frac{1}{n_1} \sum_{i=1}^{c_1} \lambda(1|x_{i1}) - \frac{1}{n_2} \sum_{i=1}^{c_2} \lambda(1|x_{i2}) \right\} \]

That is, a higher average logit score indicates a greater probability of correct classification, thus a greater dissimilarity between the groups of samples. This statistic is presented on an antilogit scale, ranging from 0.0 (identical groups) to 1.0 (disjoint groups).
Ramsey and Marsh then note that the logit score for each sample may be partitioned to reflect a composite of factors associated with each descriptor variable,
\[ \lambda(I|x) = \sum_{d=1}^{D} \lambda(I|x_d) \]
where \( \lambda(I|x_d) \) represents a logit score based on a model using only one descriptor variable. Then a "model factor" for each descriptor may be constructed by contrasting logits based on extreme values,
\[ \lambda_d = \lambda(I|x_d=1) - \lambda(I|x_d=0) \]
Model factors represent a measure of the difference in descriptor values between sample groups. Descriptor variables with large model factors thus may contribute to the overall dissimilarity between groups. Again, the sign of each model factor will indicate the likely group association of samples having relatively large values for the corresponding descriptor variables.

However, descriptors with large model factors but low, though contrasting, values in both groups may be unimportant in determining group differences. Ramsey and Marsh further suggest calculation of a "data factor" for each descriptor variable, which is dependent on the magnitude as well as the group differences of the descriptor values,
\[ \lambda_d = (X_{d1} - X_{d2}) \lambda_d / 2 \]

It turns out that summing the data factors over all descriptors is also equal to DIS. Thus data factors may be expressed as a percentage of DIS. Those descriptors with large data factors are then considered to be the most important in determining group dissimilarity, and each descriptor will be associated with one group or the other. 

Implementation

The SAS Software code provided at the end of this paper implements the DIS calculations according to the Ramsey and Marsh model parameter estimation formulae. The present version uses that procedures *proceude not print*, PROC MATRIX. Hopefully it will translate into the new matrix manipulation procedures with little difficulty. It is written as a MACRO, to be called with several parameters. Several DATA and PROC steps are involved, so any titles for printed output should be specified before the MACRO call.

Printed output includes class sample sizes, descriptor variable mean values for each class and combined, the DIS coefficient, and model and data factors for each descriptor. Input data and logit score listings are optional, as are plots and histograms of the various results. The MACRO also generates several SAS datasets containing sample logit scores, model factors, and data factors.

**Syntax:**

```
TITLE ' ... ';
%DIS (DATA = <sasdata>)
  CLASS = <var>
  ID = <var ...>
  MODEL = MULTINOM / POISSON / BINOMIAL
  TRANS = PROPORTIONS / PRESENCE / NONE
  PLOT = YES / NO
  LIST = NO / YES
  SCORES = NO / YES
  MU = 1 <0.1 to 10> );
```

**Operands:**

- **DATA=** Any valid SAS dataset name. Default is _LAST_. The dataset should contain one observation per "sample," a variable to identify group or class membership, and one or more numeric sample descriptor variables.
- **CLASS=** Required, the name of the variable used to identify group or class membership of each observation.
- **ID=** Optional, the name of a variable used to identify individual samples within each group.
- **VARS=** Optional, a list of variables to be used as descriptors. Default is all numeric variables.
- **MODEL=** Functional model used for descriptor joint pdf. Three models are implemented in this code. Default is MULTINOM.
- **TRANS=** Input data transformation function. Default is _TRANS_ descriptor values to PROPORTIONS within each sample. Other options include a PRESENCE/absence (0/1) transformation, and NONE (input data unchanged).
- **PLOT=** Requests data plots as part of printed output. Includes a histogram of individual sample logit scores, a scatterplot of descriptor model and data factors, and a histogram of descriptor data factors as a percentage of DIS. Default is YES.
- **LIST=** Requests listing of input and transformed data. Default is NO.
Requests listing of logit scores for individual samples. Default is NO.

Sets prior precision constant for logit calculations. Default is 1. Ramsey suggests values between 0.1 and 10; larger values give more weight to prior group membership probabilities, and result in "more conservative" dissimilarity estimates.

The Ramsey and Marsh paper included an example of the dissimilarity analysis on a dataset of shorebird diet data. The stomach contents of eighteen and nineteen individuals respectively from two bird species were examined, and the samples were classified as two groups of "predators." The descriptor variables were counts of twelve invertebrate "prey" taxa found in the predator's diets. Several of their results are reproduced here, generated by this implementation.

Figure 1 represents a frequency distribution of individual logit scores within each predator class, calculated using the multinomial model on proportionally transformed data. While only 34 out of 37 samples were correctly classified by the sign of their logit score, there is clearly a separation of the two predators. The average logit score, or DIS, was 0.81.

The mean values of each prey taxon within each predator class are plotted in Figure 2. This reduced dataset would be required for most traditional similarity analyses. Two commonly used indices, percent similarity = 46.7 and Spearman's rank correlation = 0.08 indicate a low similarity between predator diets. On this figure, the distance of each point from the diagonal (X=Y) would be proportional to that descriptor's model factor, the difference in average value between the two groups of samples. On the other hand, descriptor data factors would also take into account the magnitude of the values, giving more weight to those descriptors farther from the origin.

Descriptor (prey) model and data factors are plotted in Figure 3, with the logit signs indicating the association of each prey with a particular predator. While most prey taxa have large model factors, indicating a differing presence in the two predator diets, only two taxa were consumed at rates high enough to generate significant data factors. All data factors, as a percentage of DIS, are given in Figure 4, clearly showing that a large proportion of the mussel Mytilus in a predator diet is a strong indicator the surfbird, while the presence of Callinella suggests that the predator is a black turnstone.

A different example also uses macroinvertebrate taxa (families) as descriptor variables, but this time representing proportions of organisms colonizing an artificial substrate sampling device in a small stream. Samples are classified by area, those potentially impacted by a power plant effluent and those in an unaffected reference area. Six replicate samples are available from each area.

Mean proportions of each taxon in each area are plotted in Figure 5. Clearly, invertebrate communities are quite similar between the two areas. Percent similarity = 85.4, Spearman's rank correlation = 0.87 and DIS is very low at 0.54. Nevertheless, there are some interesting differences between the areas, and the data factors in Figure 6 indicate that the caddisfly family Hydropsychidae is associated with impacted areas, while the Leptophlebiidae mayflies are more common in reference samples.

Discussion

Ramsey and Marsh have given us an interesting new tool for environmental data analysis. In addition to quantifying dissimilarity among samples, a complement to the many widely used similarity measures, it also helps to identify which factors contribute to differences, even between very similar samples. As a new procedure, it will need to be applied and tested in many diverse situations to determine its usefulness in increasing our understanding of environmental data. It is hoped that the availability of the procedure in this implementation will encourage such use and testing.

Although the procedure was primarily developed for use with taxonomic descriptor variables, it would seem to be useful with other kinds of multivariate data as well, assuming the availability of an appropriate pdf model with estimable parameters. Further research into alternate models, model applicability to various kinds of data sets, and the procedure's sensitivity to model specification would be most useful.

Another interesting feature of this dissimilarity procedure is that while most similarity measures simply compare two samples (often, each a composite of many replicates), DIS is intended to compare two groups of samples. Intuitively, this would seem to make much better use of information contained in replicate sampling, usually performed with an awareness of environmental variability. Much more work needs to be done on the effects of sample size and unequal sizes between groups, as well as optimum allocation of sampling resources. This also suggests that some measure of uncertainty for the DIS coefficient might be reasonably developed.
I invite any comments and questions on this work, directed to me at

RHC Environmental Services
Muddy Run Ecological Laboratory
1921 Elver Road, P.O. Box 10
Drumore, PA 17518
(717) 588-2121

Acknowledgements

This work was developed in part under consulting contracts with Philadelphia Electric Company and Pennsylvania Power and Light Company. Data for the macroinvertebrate examples were provided by PPL.

SAS is the registered trademark of SAS Institute, Inc., Cary, NC, USA.

References


Figure 3. SHOREBIRD DIETS

Figure 4. SHOREBIRD DIETS

Figure 5. MACROINVERTEBRATES - JUNE

Figure 6. MACROINVERTEBRATES - JUNE

Figure 3. SHOREBIRD DIETS

Figure 4. SHOREBIRD DIETS

Figure 5. MACROINVERTEBRATES - JUNE

Figure 6. MACROINVERTEBRATES - JUNE

Figure 3. SHOREBIRD DIETS

Figure 4. SHOREBIRD DIETS

Figure 5. MACROINVERTEBRATES - JUNE

Figure 6. MACROINVERTEBRATES - JUNE

Figure 3. SHOREBIRD DIETS

Figure 4. SHOREBIRD DIETS

Figure 5. MACROINVERTEBRATES - JUNE

Figure 6. MACROINVERTEBRATES - JUNE