Paper TT06-2015

Using SAS@ Programs to Conduct Discriminate Analysis

Mengyu Liu, University of Southern California, Los Angeles, CA

ABSTRACT

This paper describes the use of DISCRIM procedure in SAS® to conduct a discriminate analysis. A Seal Vocalization data set containing the recordings of calls, which is comprised of eight features (one "response" variable and seven quantitative variables), of harp seals in three herds is used. The goal of this analysis is to determine whether the vocalization data can be used to construct a rule which discriminates between the three herds of seals. The STEPDISC procedure is used to select a subset of the quantitative variables for use in discriminating among the groups. Multivariate normality is checked in each of the three herds. Option POOL=TEST of PROC DISCRIM is used to test whether the same variance-covariance matrix of response across different groups. A quadratic discriminant function is derived based on the result of equal variance test. Moreover, since the multivariate normal assumptions are not satisfied, a nonparametric method based on kernel density estimates is also applied. Error rates (misclassification rates) are compared across different methods, and rules with error rates that smaller than the rate if randomly assigned are considered. This example illustrates discriminate analysis in SAS® using a research design for users don't familiar with basic procedures of such analysis in SAS®.

INTRODUCTION

This paper presents an example of PROC DISCRIM to perform a discriminate analysis, suitable for users familiar with the basic concepts of discriminate analysis but unfamiliar with procedures in SAS®. Discriminant analysis is designed to classify data into known groups. Discriminate analysis is a multivariate statistical technique used to build a predictive/descriptive model of group discrimination based on observed predictor variables and to classify each observation into one of the groups [(1)]. In the analysis, each observation in a training data is assigned a probability of belonging to a given group or class based on the distance of its discriminant function from that of each class mean. Stepwise, canonical and discriminant function analyses are commonly used discriminate analysis techniques available in the SAS® systems STAT module [(2)].

The research study is concerned with harp seals, and in particular the herds from Jan Mayen Island, Gulf of St. Lawrence and Front. The data used is from Dr. Jack M. Terhune, Department of Biology, University of New Brunswick, St. John[(3)]. Tagging and morphometric studies suggest little exchange between the three harp seal herds. Vocalization differences among these three herds would provide evidence that the groups are reproductively isolated [(3)]. Data are obtained from underwater recordings of harp seals in three herds. One thousand calls from each of the three herds are recorded, and eight features of each recording are noted. The goal of the study is to determine whether the vocalization data can be used to construct a rule which discriminates between the three herds of seals.

VARIABLES

The following variables are used in this analysis:

RESPONSE VARIABLE

Herd (v8): this is the herd from which the recordings are obtained (1-Jan Mayen Island,2-Gulf of St. Lawrence, 3-Front).

QUANTITIVE VARIABLES

ELEMDUR (v1): this is the duration of a single element of a harp seal underwater vocalization.
INTERDUR(v2): this is the time between elements in multiple element calls.
NO_ELEM (v3): this is the number of elements of the call.
TARTFREQ (v4): this is the pitch at the start of the call or the highest pitch if the call has an extremely short duration (call shape 0 below). Measure by Hertz (Hz) and converted to octaves using the formula: octave=log2(Hz)
ENDFRE(V5): this is the pitch at the end of the call or the lowest pitch if the call has an extremely short duration (call shape 0).
WAVEFORM (v6): this codes a series of waveform shapes (a plot of amplitude vs time) which lie more or less along a continuum.
CALLSHAP(v7): this codes a series of call shapes as they would appear in a sonogram spectral analysis (a plot of frequency vs time). The shapes lie along a continuum

There are totally 3000 observations in the data and the numbers of observation in each herd group is 1000. Near half of observations in v2 are missing (1340 out of 3000), so that v2 is dropped from dataset. Variables v1-v5 are continuous variables and v6- v7 are categorical variables. Since v6 and v7 code a series of waveform or call shapes that along a continuum, they are treated as ordinal variables. Log2 transformations are applied to v4 and v5 to

change the units from hertz to octave, which is the normal way mammals hear. Table 1 shows the data format before analysis.

Obs	v1	v3	log2v4	log2v5	v6	v7	Herd
1	131	4	7.7004	7.6439	7	5.0	1
2	118	4	7.8580	5.3219	2	2.5	1
3	37	4	8.4512	6.0000	2	0.0	1
…							

Table 1. Example of Data Format Before Analysis

STATISTICAL ANALYSES

The first analytical step is to determine which variables are significantly related to the "response" variable, herd (v8), which is coded to reflect group membership. PROC STEPDISC is applied to the dataset. It is found that all six variables are significant, so that all variables are included.

The next step is to check the multivariate normality in each of the three herd group since the estimated minimum total probability of misclassification (TPM) rule used is under the assumption of multivariate normality. The %MULTNORM macro, downloaded from SAS® Knowledge Base [(4)] is used to test multivariate.

For seals in Jan Mayen Island (Herd =1), none of the six variables satisfy the univariate normal assumption based on the Shapiro-Wilk test ($p<0.0001$, Output 1). A chi-square quantile-quantile (Q-Q) plot also showed a violation of multivariate normality (Output 1). Similar results are found for seals in the other two herds. Univariate normality is also checked for each variable since the test given by %MULTNORM macro may be too stringent. It turns out that the distribution of the log2-transformed v4 and v5 are approximately normally distributed based on Q-Q plots. However, four other variables, v1, v3, v6 and v7 need proper transformations.

Since multivariate normal assumptions are not satisfied due to skewed distributions for variables, transformations of data are considered. A common used method is the power transformation, and Box-Cox transformations are used in this data. The formula is showed as follows:

$$x^{\lambda} = \begin{cases} \dfrac{x^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ lnx & \lambda = 0 \end{cases}$$

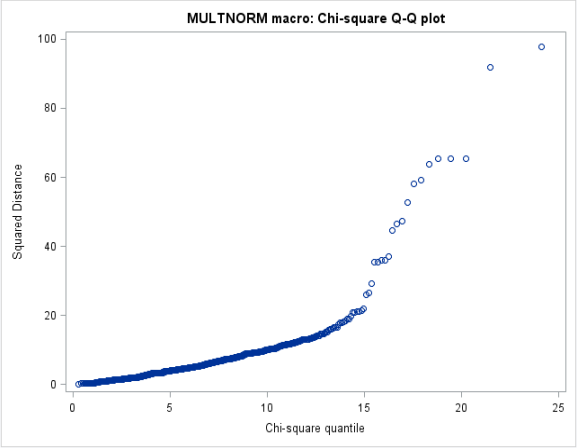
Where $x_1, x_2, \ldots x_n$ are observations, λ is the power.

The %BCTRANS2 macro written by Steven M. LaLonde et al [(5)] is used to perform the transformation. Since Box-Cox transformations require positive value of a variable, v7n=v7+1 is used to in the transformation. Results of powers and transformed variables are showed in table 2 below.

Variables	Optimal Power λ	Transformed Variables
v1	-0.1; approximately equal to 0	v1t=ln(v1)
v3	-0.7	v3t=[v3^(-0.7)-1]/(-0.7)
v6	1.5	v6t=[(v6)^1.5-1]/1.5
v7n	1.1; approximately equal to 1	v7t=[(v7n)^1-1]/1=v7n-1=v7+1-1=v7

Table 2. Powers and Transformations of Variables

Normality Test			
Equation	Test Statistic	Value	Prob
v1	Shapiro-Wilk W	0.77	<.0001
v3	Shapiro-Wilk W	0.71	<.0001
log2v4	Shapiro-Wilk W	0.98	<.0001
log2v5	Shapiro-Wilk W	0.95	<.0001
v6	Shapiro-Wilk W	0.87	<.0001
v7	Shapiro-Wilk W	0.84	<.0001
System	Mardia Skewness	5421	<.0001
	Mardia Kurtosis	82.32	<.0001
	Henze-Zirkler T	97.41	<.0001



Output 1. Selected Output for Test of Multivariate Normality for Seals in in Jan Mayen Island (Herd=1)

Multivariate normality is future checked for the data set including transformed variables. The Q-Q plots show that normality assumption is better satisfied, even though results of Shapiro-Wilk tests are still statistically significant (p<0.0001). Histograms overlaid by normal curves indicated that normality assumption is improved for transformed variables of v1, v3, and v6 (v1 shows in Figure 1) and multivariate normality is assumed in this data.

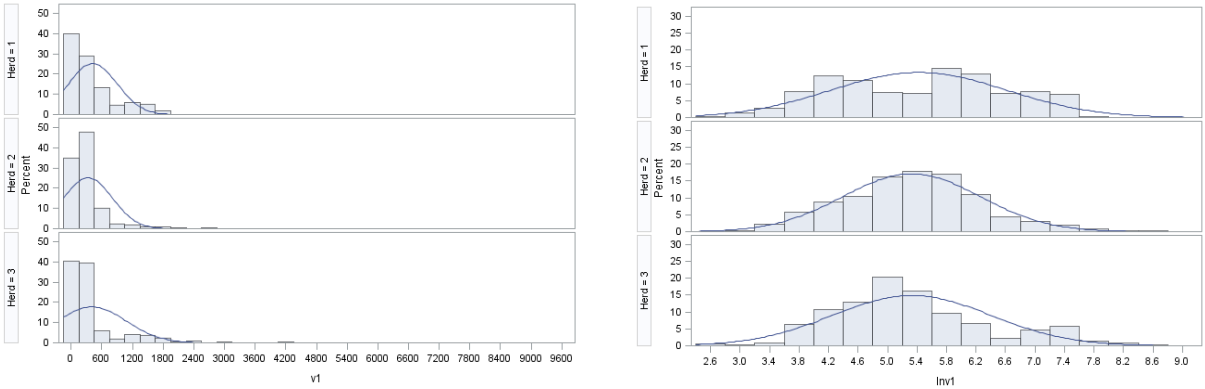


Figure 1. Histograms of Elemdur (v1) by Herd for Seals (Left: Before Transformation; Right: After Natural log Transformation)

The next step is to conduct a discriminate analysis using PROC DISCRIM. Since the multivariate normal distribution within each herd group is assumed, a parametric method would be used and a linear discriminant analysis (LDA) or a quadratic discriminant analysis (QDA) would be conducted. LDA assumes same variance-covariance matrix of the responses across all herd groups while QDA assumes each group has a unique variance structure. Test of equal variance is done by using option POOL=TEST. Equal prior probabilities for each herd group are assumed. The code for this step is as follows:

```
proc discrim data=Sealn pool=test;
class Herd;
var lnv1 v3t log2v4 log2v5 v6t v7;
priors equal;
run;
```

It is showed that Chi-Square value of equal variance test is statistically significant (p<0.0001, Output 2), so that a quadratic discriminant function would be used, which is specified by POOL=NO option in PROC DISCRIM.

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
1076.204414	42	<.0001

Output 2. Selected Output from PROC DISCRIM that Test Equal Variance Assumption

One important way of evaluating the performance of a classification procedure is to calculate the error rates, or misclassification probabilities. The error rates obtained from the proper designed rule should be smaller than that if observations are assign into groups at random (66.7% in Seal Vocalization data). Moreover, in order to reduce biases, each observation in the dataset is classified using a discriminate function computed from other observations, excluding the observation being classified. Therefore, the CROSSVALIDATE option in PROC DISCRIM is applied and cross-validated classification errors are calculated. This error rate, for moderate samples, is a nearly unbiased estimate of the expected actual error rate. The DISCRIM procedure is as follows:

```
proc discrim data=Sealn pool=no crossvalidate;
class Herd;
var lnv1 v3t log2v4 log2v5 v6t v7;
priors equal;
run;
```

Number of observations classified in each herd and classification error rate and cross-validated result are shown in Output 2. It is found that total error rate without cross-validation classification is slightly lower than that computed from cross-validation method (50.37% vs. 51.00%, Output 3). Furthermore, the total error rates and the error rate in each herd group are all smaller than the rate if assigned randomly (66.7%), which indicates that quadratic discriminate function can be properly used to discriminate seals in the three herds.

Classification Summary for Calibration Data: WORK.SEALN
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Herd

From Herd	1	2	3	Total
1	470 47.00	256 25.60	274 27.40	1000 100.00
2	115 11.50	636 63.60	249 24.90	1000 100.00
3	172 17.20	445 44.50	383 38.30	1000 100.00
Total	757 25.23	1337 44.57	906 30.20	3000 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Herd

	1	2	3	Total
Rate	0.5300	0.3640	0.6170	0.5037
Priors	0.3333	0.3333	0.3333	

Classification Summary for Calibration Data: WORK.SEALN
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Herd

From Herd	1	2	3	Total
1	467 46.70	256 25.60	277 27.70	1000 100.00
2	115 11.50	629 62.90	256 25.60	1000 100.00
3	176 17.60	450 45.00	374 37.40	1000 100.00
Total	758 25.27	1335 44.50	907 30.23	3000 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Herd

	1	2	3	Total
Rate	0.5330	0.3710	0.6260	0.5100
Priors	0.3333	0.3333	0.3333	

Output 3. Selected Output from PROC DISCRIM using Quadratic Discriminate Function (Left: Resubstitution summary; Right: Cross-validation Summary)

The coefficients of quadratic discriminate function (constant, linear and quadratic terms) can also be obtained from OUTSTAT option in PROC DISCRIM procedure. However, there are six quantitative variables in this study and many terms are derived. It's hard to interpret the function so the outputs of coefficients are omitted. However, for analysis that use linear discriminate function and contain a small number of quantities variables, the function with coefficients can be presented.

The third step is to conduct a non-parametric approach. Nonparametric discriminant methods are based on nonparametric estimates of group-specific probability densities, either a kernel method or the k-nearest-neighbor (KNN) method can be used. Multivariate normality is not fully satisfied, so that a non-parametric method is used and specified by METHOD=NPAR in PROC DISCRIM. A kernel method is used since in this data, some observations won't be classified in the three herd groups (an "other" group appears) if using KNN. Normal kernel is assumed (KERNEL=NOR) and a bandwidth of 0.6 is assumed (R=0.6). Cox-Box transformations are not applied to variables in this analysis. Code is showed as follows:

```
proc discrim data=Sealn method=npar kernel=nor R=0.6 crossvalidate;
class Herd;
var v1 v3 log2v4 log2v5 v6 v7;
priors equal;
run;
```

Number of observations classified in each herd and classification error rate and cross-validated results are shown in Output 4. The total error rate without cross-validation classification is a little lower than that obtained from cross-validation method (36.37% vs. 40.00%, Output 4). The error rates in groups 1, 2, and 3 are 23.50%, 39.80%, and 56.7.0% respectively in corss-validated summary (Output 4). The total error rate computed from cross-validation using normal kernel density (40.00%, Output 4) is lower than that obtained from the parametric method using quadratic discriminant function (51.00%, Output 2), and similar patterns were found for error rates in each three herd group. However, in the nonparametric method, the chosen of parameter may be not optimized and the results can be misleading.

Classification Summary for Calibration Data: WORK.SEALN
Resubstitution Summary using Normal Kernel Density

Number of Observations and Percent Classified into Herd				
From Herd	1	2	3	Total
1	801 80.10	141 14.10	58 5.80	1000 100.00
2	208 20.80	636 63.60	156 15.60	1000 100.00
3	202 20.20	326 32.60	472 47.20	1000 100.00
Total	1211 40.37	1103 36.77	686 22.87	3000 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Herd				
	1	2	3	Total
Rate	0.1990	0.3640	0.5280	0.3637
Priors	0.3333	0.3333	0.3333	

Classification Summary for Calibration Data: WORK.SEALN
Cross-validation Summary using Normal Kernel Density

Number of Observations and Percent Classified into Herd				
From Herd	1	2	3	Total
1	765 76.50	156 15.60	79 7.90	1000 100.00
2	221 22.10	602 60.20	177 17.70	1000 100.00
3	216 21.60	351 35.10	433 43.30	1000 100.00
Total	1202 40.07	1109 36.97	689 22.97	3000 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for Herd				
	1	2	3	Total
Rate	0.2350	0.3980	0.5670	0.4000
Priors	0.3333	0.3333	0.3333	

Output 4.Selected Output from PROC DISCRIM Using Kernel Density Estimates (Left: Resubstitution summary; Right: Cross-validation Summary)

CONCLUSION

A parametric quadratic discriminant analysis (QDA) and a nonparametric kernel method (bandwidth=0.6) are applied to discriminate between three herds of seals based on the available vocalization data. The misclassification rates (both total error rate and rates in each herd group) obtained from normal kernel method are smaller than that obtained from parametric quadratic method. Since classification error rate obtained for the validation data in kernel method is relatively larger than that for the data without validation, this method performed poorly in validating the independent validation dataset. Using other types of density options might do a better job in classifying the validation dataset. As a result, misclassification rates obtained from QDA would be reported. The error rates in groups 1, 2, and 3 are 53.30%, 37.10%, and 62.6.0% respectively in corss-validated summary and the total weighted error rate is 51.00%.

The STEPDISC procedure is used to perform a stepwise discriminant analysis to select a subset of the quantitative variables for use in discriminating among the herd groups. The %MULTNORM macro is used to test multivariate morality. Box-Cox transformations are applied to selected variables to improve normality. The DISCRIM procedure is used to conduct discriminate analysis. Both parametric and nonparametric methods are used to compare discriminate rates and finally a parametric method used QDA is selected. As we see in this example, the DISCRIM procedure is very powerful tool to classify data into known groups.

There are some future improvements of this analysis. First, v2 is dropped in the analysis for the whole dataset since about half of the value is missing. Other better methods may be applied to deal with large missing values in v2 and keep v2 in the analysis. Second, v6 and v7 are categorical variables and they are treated as ordinal variables. Box-Cox transformation on those ordinal variables is used, which changes the original scale between two ordinals in the variable. The quadratic discriminant functions constructed in this study assumed normality, which is actually contradictory to categorical nature of the two variables. Third, the nonparametric method (normal kernel) used in this

study didn't yield a very good result. The parameter chose for kernel may be not optimal and other methods such as sparse discriminant analysis and logistic regression could be future considered.

REFERENCES

1. G.C.J. Fernandez. (2002). Discriminant analysis: a powerful classification technique in data mining. Proceedings of the SAS® Users International Conference, paper 27.

2. SAS Institute Inc. (1999). SAS/STAT Users Guide, Version 8, Cary NC .SAS Institute Inc.

3. Terhune, J.M. (1994). Geographical variation of harp seal underwater vocalizations, Can. J. Zoology 72(5) 892-897.

4. SAS Institute Inc. (2007). Sample 24983: Macro to test multivariate normality, Version 1.4, Cary NC .SAS Institute Inc. Available at http://support.sas.com/kb/24/983.html

5. Steven M. LaLonde, (2012).Transforming variables for normality and linearity. Proceedings of the SAS® Global 2012 Forum, paper 430

ACKNOWLEDGMENTS

Mengyu would thank Prof. Mark Krailo of Department of Preventive Medicine, University of Southern California for providing guidance and supporting for this project and this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mengyu Liu
Biostatistics Division, Department of Preventive Medicine
University of Southern California
2001 N Soto St.
Los Angeles, CA 90032
213-806-0181
liucathy09@gmail.com