SUBSETTING AND RANKING DATA USING THE DISCRIM PROCEDURE
Kevin M. Smith, The Johns Hopkins University Applied Physics Laboratory

Introduction
The Johns Hopkins University Applied Physics Laboratory provides post-mission evaluation and analysis for the Navy to maintain the effectiveness of the operational systems onboard naval vessels. After a naval ship's return from a deployment, it provides post-mission evaluation and analysis for the Navy to develop statistical tests in order to assess the performance of maintaining the effectiveness of the operational systems onboard. The SAS* language is used extensively to develop statistical tests in order to assess the performance of these systems. SAS procedures, such as GLM, TEST, CLUSTER, etc., have all been very useful for performance analysis. In particular, the NLIN procedure has been useful in reconstructing or estimating a target contact's base course, speed of advance and range from a naval ship, given a set of bearings to the contact and a set of maneuvers by the ship, over the time that the contact was detected, held and lost by the onboard systems. These track reconstructions are one important element in analyzing a particular system's overall performance.

It must be pointed out, however, that many contacts can be detected over a period of time. Reconstructing a large set of contacts can be a very arduous process and not all contacts can be successfully reconstructed due to either a lack of bearing change generated by the contact or a lack of sufficient maneuvers generated by the ship. Therefore, a method was developed, which is based on the SAS procedure DISCRIM, to select and rank, from the contacts detected on a given deployment, a small subset of contacts (e.g., 40 out of 1000) which can be expected to be the most likely contacts to yield a desired number of successful reconstructions. This method permits the use of reconstructions to obtain a satisfactory sample of range statistics to analyze a system's performance. This paper will present the method of subsetting and ranking using PROC DISCRIM, followed by a discussion on the development of the linear discriminant used in the method.

The Method
When a ship returns from sea, sensor data is collected for each detected contact. Suppose 1000 contacts are detected for an analyzed deployment. Further suppose that 40 contacts need track reconstructions in order to obtain a satisfactory sample of contacts to compute range statistics. The method of subsetting and ranking the best 40 contacts out of 1000 detected contacts follows:

A. A SAS data set, DEPLOY, is formed from the submitted contact data. Two SAS variables, DELTAB and STDCRS, are calculated for each contact or observation. The variables are defined as follows:

DELTAB = the difference in true bearing at initial detection and at final loss of each contact, i.e., a measure of the contact's total bearing change.

STDCRS = the standard deviation of the ship's course during the time contact was held, i.e., a measure of the ship's deviation from constant course.

B. A calibration data set, CALIBRAT, is formed from historical data. This cumulative data set contains contacts previously detected for which contact track reconstructions were attempted. Four of the variables stored for each contact in this data set are DELTAB, STDCRS, SUCCESS, and PRSUCCES. The SUCCESS and PRSUCCES variables are defined as follows:

SUCCESS = 1, if the attempted track reconstruction was successful.
0, if the attempted track reconstruction was not successful.

PRSUCCES = the probability, previously assigned by PROC DISCRIM, of a contact (observation) having a successful reconstruction.

C. Based on the CALIBRAT data set, PROC DISCRIM computes a linear discriminant function to model SUCCESS. The discriminant is of the form

\[ \text{SUCCESS} = w_1 \times \text{DELTAB} + w_2 \times \text{STDCRS} \]

where \( w_1 \) and \( w_2 \) are weights associated with DELTAB and STDCRS, to be computed by PROC DISCRIM.

Note: In order to obtain \( w_1 \) and \( w_2 \) or the linear discriminant function, both the reconstructible and non-reconstructible class of contacts are considered to be normally distributed with common dispersion matrices.

D. Based on the linear discriminant function formed from the CALIBRAT data, the DISCRIM procedure classifies each detected contact in the DEPLOY data set and computes the posterior probability (POSTPROB) of the contact belonging to either the "reconstructible" or "nonreconstructible" class. The following SAS code is used:

\[
\text{PROC DISCRIM DATA } = \text{ CALIBRAT}
\text{ TESTDATA } = \text{ DEPLOY TESTLIST;}
\]

The TESTLIST option produces a list of contacts with their posterior probability (POSTPROB) of being assigned to either the "reconstructible" or "nonreconstructible" class.

E. A subset of contacts can be chosen from those contacts with the highest posterior probability of being assigned to the "reconstructible" class. However, it is important to note that not all of the high probability contacts will be successfully reconstructed. In order to determine the number of contacts needed to obtain a given number of desired successful reconstructions, an "a priori selection rate" is first calculated from the CALIBRAT data set. The selection rate, RTSELE, is defined as:
RTSELECT = Σ PRSUCCES / Σ SUCCESS

where Σ PRSUCCES = the sum of the probabilities of each contact in CALIBRAT being associated to the "reconstructible" class, i.e., the "expected" number of successes (SAS variable SUMPROBS).

and Σ SUCCESS = the total number of actual successful reconstructions (SAS variable SUMSUCCS).

The following SAS code is used to calculate the "selection rate":

```sas
PROC SUMMARY DATA = CALIBRAT;
VAR PRSUCCES SUCCESS;
OUTPUT OUT = SUMSTATS
SUM = SUMPROBS SUMSUCCS;
DATA SELECT (KEEP = RTSELECT);
SET SUMSTATS;
RTSELECT = SUMPROBS/SUMSUCCS;
```

F. Contacts in the DEPLOY dataset are ranked (high to low) based upon their posterior probability (POSTPROB) of being assigned to the "reconstructible" class. An estimated number of contacts to subset, N_est, can be chosen such that

\[ N_{est} = RTSELECT \times DESIRED \]

where DESIRED = the desired number of successfully reconstructed contacts needed to obtain a satisfactory sample of range statistics.

However, a more realistic number of contacts to subset, \( N_{real} \), can be chosen such that

\[ \sum_{i=1}^{N_{real}} POSTPROB = RTSELECT \times DESIRED \]

where \( \sum_{i=1}^{N_{real}} POSTPROB \) = the sum of the probabilities associated with those contacts, in the DEPLOY dataset, having the highest probability of being reconstructed (SAS variable DEPSUMPR).

\( N_{real} \) takes into account any deviations that the DEPLOY dataset may have from the CALIBRAT dataset, based on the POSTPROB. The following SAS data step produces a realistic number of subsetted candidates that are the most likely to be successfully reconstructed:

```sas
DATA NREAL;
IF _N_ = 1 THEN SET SELECT;
SET DEPLOY;
IF DEPSUMPR > (RTSELECT * DESIRED) THEN STOP;
DEPSUMPR = DEPSUMPR - PRSUCCES;
```

For example: Suppose, based on the contacts in the CALIBRAT dataset, that RTSELECT = 1.2 and DESIRED = 25, a minimum number of contacts (Nreal) in the DEPLOY dataset that could be chosen for reconstruction would be Nreal = 1.2 * 25 = 30. However, a more realistic number of contacts could be chosen in the following manner:

```
Suppose that Table 1 is a partial list of 1000 contacts detected on a given deployment. The list of contacts are ranked (high to low) based on POSTPROB (DEPSUMPR is also shown). If one were to choose contacts until DEPSUMPR = 30, then 40 contacts would be chosen (Nreal = 40). Based on the contacts' posterior probabilities of being assigned to the "reconstructible" class, and additional 10 contacts would be subsetted to ensure that 25 successful reconstructions are achieved. These 40 contacts and the results of the success or failure of the track reconstruction for each contact are then appended to the CALIBRAT data set to ensure that the linear discriminant model is current based on data from recent deployments.
```

### Table 1: Example DEPLOY Dataset

<table>
<thead>
<tr>
<th>Obs</th>
<th>PRSUCCES</th>
<th>SUCCESS</th>
<th>DEPSUMPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.90</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.88</td>
<td>1.97</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.65</td>
<td>2.02</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.62</td>
<td>3.64</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.50</td>
<td>3.07</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.49</td>
<td>4.74</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.30</td>
<td>6.63</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.18</td>
<td>6.51</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.18</td>
<td>7.37</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.15</td>
<td>7.22</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.12</td>
<td>8.05</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.02</td>
<td>6.97</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.01</td>
<td>10.66</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.01</td>
<td>11.40</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.00</td>
<td>12.29</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.00</td>
<td>13.06</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.78</td>
<td>13.66</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.77</td>
<td>14.64</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.76</td>
<td>15.41</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>0.75</td>
<td>16.11</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>0.74</td>
<td>17.08</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.73</td>
<td>18.39</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>0.72</td>
<td>19.11</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.72</td>
<td>19.83</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>0.71</td>
<td>20.54</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>0.70</td>
<td>21.24</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>0.69</td>
<td>21.53</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>0.68</td>
<td>22.52</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>0.68</td>
<td>23.37</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0.67</td>
<td>23.97</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>0.66</td>
<td>24.63</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.65</td>
<td>25.33</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.65</td>
<td>26.07</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.64</td>
<td>27.21</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>0.64</td>
<td>28.24</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.63</td>
<td>29.46</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>0.62</td>
<td>30.28</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.61</td>
<td>31.46</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>0.61</td>
<td>32.04</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>0.60</td>
<td>32.62</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0.59</td>
<td>33.19</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0.59</td>
<td>33.77</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>0.58</td>
<td>34.35</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.57</td>
<td>34.93</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0.57</td>
<td>35.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Development of the Linear Discriminant

The method of subsetting and ranking a small set of contacts from a large number of contacts detected on a ship's deployment heavily depends on the linear discriminant function. The discriminant function was first developed based on a data base of 465 contacts for which track reconstructions were attempted. Of these 465 contacts, 223 were successfully reconstructed and 242 were not reconstructible. The linear discriminant function is currently updated with contacts detected on recent deployments. In developing a discriminant function, the stepwise discriminant analysis by forward selection, backward...
elimination or stepwise selection of variables useful for discriminating among several classes. The following paragraphs present an alternative method for developing the linear discriminant function, highlighting various SAS procedures used to choose the discriminant variables, DELTAB and STDCRS, and to derive the linear discriminant function.

Step 1 - Ten variables, including DELTAB and STDCRS, were initially considered in formulating a model or discriminant for classifying contacts for track reconstructions. Simple descriptive statistics were calculated (using PROC UNIVARIATE) for the classes of successful and unsuccessful reconstruction. The means were compared by application of PROC TTEST to determine whether a significant difference existed between the two classes for each variable. If a significant difference did not exist between the two classes, the variable was not considered to be a good variable for discrimination. Four of the initial ten variables did not exhibit significant differences between the two classes of contacts and were eliminated from consideration for the discriminant model.

Step 2 - It was thought that some of the remaining six variables contained the same information. Correlation coefficients produced by PROC CORR were calculated for the six variables. Two additional variables were eliminated from consideration for the discriminant function.

Step 3 - When there are only two groups under consideration (i.e., whether a contact was reconstructible or nonreconstructible), a discriminant model can be written in the form of an equation of the hyperplane

\[ Y = w_1x_1 + w_2x_2 + \ldots + w_nx_n \]

where \( x_1, x_2, \ldots, x_n \) are the \( n \) variables, \( w_1, w_2, \ldots, w_n \) are weights assigned to the variables, and \( Y \) is a binary variable indicating group membership (equal to 0 or 1).

The discriminant weights \( (w_1, w_2, \ldots, w_n) \) are proportional to the weights for a multiple regression equation of a binary-group membership variable \( Y \) on the predictors \( x_1, x_2, \ldots, x_n \). Thus the discriminant analysis reduces to a multiple regression analysis. Therefore, a stepwise regression (using PROC STEPWISE) was performed on the remaining four variables. This analysis indicated that the variables, DELTAB and STDCRS, were sufficient for selecting candidates for reconstruction as the addition of the two remaining variables did not produce significant improvements in the regression.

Step 4 - With the two variables DELTAB and STDCRS established as the discriminatory variables for the model, a discrimination based on the line

\[ \text{SUCCESS} = w_1 \times \text{DELTAB} + w_2 \times \text{STDCRS} \]

could be used for the linear discriminant function. However, it was thought that the relationship between \( \text{SUCCESS} \) and each of the two variables, DELTAB and STDCRS, may not have been linear and that the performance of the model could have been improved by exploiting these relationships. In order to investigate the possibility of a better predictive model or discriminant, the data were sorted into subsets, based on DELTAB and STDCRS values, and the average values of the variable \( \text{SUCCESS} \) (which is equivalent to the percentage of \( \text{SUCCESS} \) for each subset) were plotted by the DELTAB and STDCRS variables (see Figures 1 and 2).

Analysis of the data in Figures 1 and 2 indicates that a nonlinear relationship exists between each of the two variables, DELTAB and STDCRS, and the \( \text{SUCCESS} \) variable. By applying a nonlinear regression to the data (using PROC NLR), relationships were found to support a logarithmic dependency for DELTAB and a square root dependency for STDCRS.

Step 5 - Figure 3 presents a linear discrimination based on the straight line model

\[ \text{SUCCESS} = w_1 \times \text{DELTAB} + w_2 \times \text{STDCRS} \]

Figure 4 presents a discrimination based on the curved line model

\[ \text{SUCCESS} = w_1 \times \log(\text{DELTAB}) + w_2 \times \text{STDCRS}^k \]

The discriminant line presented in Figure 4 shows a more robust fit to the data. Based on the results calculated by PROC DISCRIM, more contacts were correctly classified using the curved line discriminant than by using the straight line discriminant. Effectively, the curved line correctly classified more nonreconstructible contacts near the end boundaries of the discriminant, and more reconstructible contacts in the
middle of the model. The discriminant model presented in
Figure 4 is the preferred model for classifying contacts into
the "reconstructible" and "nonreconstructible" classes.

Figure 3: Straight Line Discrimination

* Reconstructible Contact 0 Nonreconstructible Contact

Figure 4: Curved Line Discrimination

* Reconstructible Contact 0 Nonreconstructible Contact

Conclusion

A method of subsetting and ranking data was presented,
based on the posterior probability of an observation being
assigned to a specific class, using the DISCRIM procedure. As
an alternative to the STEPDISC procedure, other SAS
procedures can be used to select discriminant variables in
order to formulate a discriminant equation. It was also shown
that modeling the relationship between the dependent
variable (or classification variable) and the independent
variables of the discriminant equation can improve classifi­
cation results of the discriminant analysis.

Acknowledgments

The author would like to recognize Kenneth T. Ryals of The
Johns Hopkins University Applied Physics Laboratory for his
invaluable contributions to the development of this method.

References

Bard, Yonathob (1974), Nonlinear Parameter Estimation,
Academic Press.

Kendall, Sir Maurice, Stuart, Alan and Ord, J. Keith (1983),
The Advanced Theory of Statistics, Volume 3, Fourth

Tatsuoka, Maruice M. (1971), Multivariate Analysis:
Techniques for Educational and Psychological Research,
J. Wiley and Sons, Inc.


For further information, please contact:

Kevin M. Smith
The Johns Hopkins University Applied Physics Laboratory
Johns Hopkins Road
Laurel, Maryland 20723-6099

SAS® and SAS/GRAPH® are registered trademarks of SAS
Institute Inc., Cary, North Carolina, USA.