

# Discriminant Analysis and Probability Forecasts

The goal of discriminant analysis is to classify a new object, with a given set of properties, as belonging to one of (at least) two mutually exclusive populations in a way that maximizes the rate of correct classifications (e.g., Kendall et al. 1983). One of the advantages to discriminant analysis is that it can simultaneously consider multiple variables (measurements). The first step in calculating the discriminant function is to estimate the probability density function,  $f$ , defined as

$$P(\mathbf{x}_a < \mathbf{x} < \mathbf{x}_b) = \int_{\mathbf{x}_a}^{\mathbf{x}_b} f(\mathbf{x}) d\mathbf{x}, \quad (1)$$

where  $P(\mathbf{x}_a < \mathbf{x} < \mathbf{x}_b)$  is the probability that a measurement falls between  $\mathbf{x}_a$  and  $\mathbf{x}_b$ . A new object is classified as belonging to the population with the higher probability density:

$$q_1 f_1(\mathbf{x}) \geq q_2 f_2(\mathbf{x}) \Rightarrow \text{predict population 1.}$$

$$q_1 f_1(\mathbf{x}) < q_2 f_2(\mathbf{x}) \Rightarrow \text{predict population 2.}$$

where  $q_j$  is the prior probability of belonging to population  $j$ . That is, given some measurements of an object, the discriminant function predicts to which population the object is most like to belong.

The discriminant boundary occurs where there is equal probability of belonging to each population. This can be thought of as a 50% probability forecast. Instead of making a binary prediction that a new object will belong to a particular population, one can instead estimate the probability that the object will belong to a given population. Using Bayes's theorem, the probability that an object belongs to population 1 when it is observed to have properties  $\mathbf{x}$  is

$$P_1(\mathbf{x}) = \frac{q_1 f_1(\mathbf{x})}{q_1 f_1(\mathbf{x}) + q_2 f_2(\mathbf{x})}. \quad (2)$$

In practice, one estimates the properties of the populations from finite samples, and uses the properties of the samples to estimate the probability density function and construct the discriminant function. This can be done by either making an assumption about the functional form of the probability density function (parametric discriminant analysis) or by estimating the probability density function directly (nonparametric discriminant analysis).

For the cases we will be considering, typically the two populations will be active regions which produce an event (flare, CME, etc.) within a given time, and those which do not. The measurements of the active regions will characterize their photospheric and/or sub-photospheric properties.

## 1. Linear Two-Population Discriminant Analysis

The simplest form of discriminant analysis assumes that each variable has a normal (Gaussian) distribution, and that the populations have the same covariance matrices (essentially that both populations have the same standard deviation for a given variable). In this case, the discriminant function is linear in all the variables, so the phase space is divided into two by a plane. In more sophisticated implementations of discriminant analysis, more general forms for the distribution are assumed, or a nonparametric estimate is made for the probability density function (Silverman 1986). In practice, the results of discriminant analysis tend not to be very sensitive to assumptions about the functional form of the probability density (see Leka & Barnes 2007, for some examples in flare forecasting).

Given  $i = 1, \dots, n_l$  measurements  $\{x_{ik}^{(l)}\}$ , of variables  $k = 1, \dots, p$  in groups  $l = 1, 2$ , the probability density function is given by a normalized multivariate Gaussian:

$$\hat{f}_j(\mathbf{x}) = \frac{|\mathbf{C}|^{-1/2}}{(2\pi)^{p/2}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}}^{(j)})' \mathbf{C}^{-1}(\mathbf{x} - \bar{\mathbf{x}}^{(j)}) \right] \quad (3)$$

and the linear discriminant function is given by

$$f(\mathbf{x}) = \mathbf{x} \mathbf{C}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \mathbf{C}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) - \log \left( \frac{q_2}{q_1} \right) \quad (4)$$

where  $q_l$  is the prior probability of belonging to population  $l$  and is usually estimated as  $q_l = n_l$ ,  $\bar{\mathbf{x}}^{(l)}$  is the mean of the measurements in group  $l$ , and  $\mathbf{C}^{-1}$  is the inverse of the covariance matrix,

$$\mathbf{C} = \frac{n_1 \mathbf{C}^{(1)} + n_2 \mathbf{C}^{(2)}}{n_1 + n_2 - 2} \quad (5)$$

with  $\mathbf{C}^{(l)}$  the covariance matrix for population  $l$ , which is estimated from the samples as

$$C_{ij}^{(l)} = \sum_{k=1}^p (x_{ik}^{(l)} - \bar{x}_k^{(l)})(x_{jk}^{(l)} - \bar{x}_k^{(l)}) / (n_l - 1). \quad (6)$$

The phase space is divided into two regions bounded by  $f(\mathbf{x}) = 0$ . A new case at  $\mathbf{x}$  is then classified based on whether the discriminant function at that point is positive or negative. Examples of linear discriminant analysis are shown in Figure 1.

A related quantity is the Mahalanobis distance, which is a measure of the distance between the means of the two samples, and is given by

$$D^2 = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) \mathbf{C}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}). \quad (7)$$

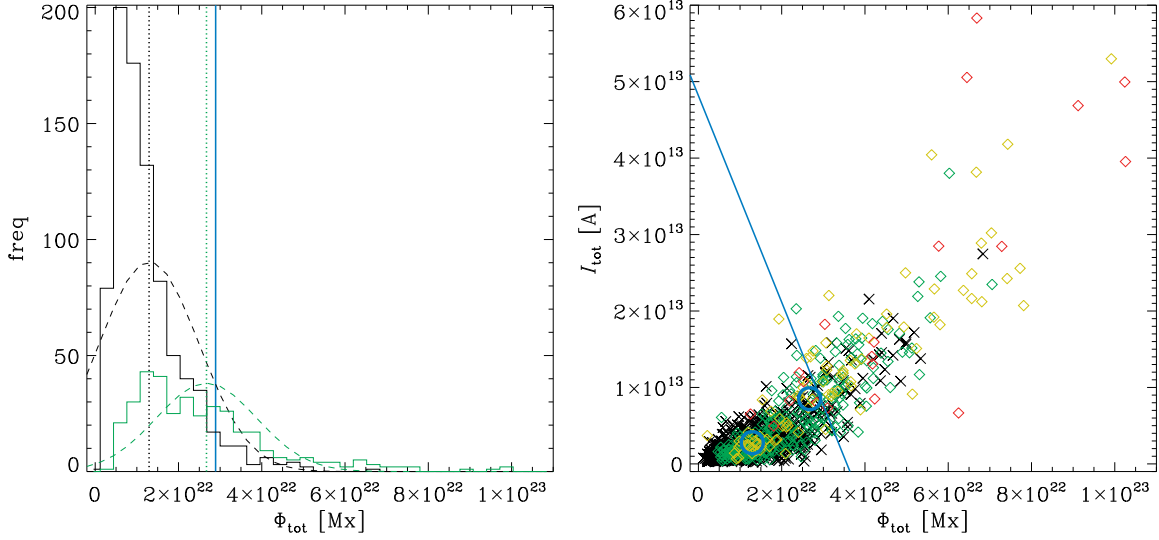


Fig. 1.— Example linear discriminant functions. *Left:* Single variable discriminant analysis. The histograms show the flaring (green) and non-flaring (black) probability densities. The normal distribution fits to the data (dashed lines), and the means (vertical dotted lines) are also shown. The point at which the two density estimates are equal is shown with a blue line that corresponds to the location of the discriminant boundary. An active region with a measured value of  $\Phi_{\text{tot}}$  falling to the right of the discriminant boundary would be predicted to flare; for a measurement falling to the left, the active region would be predicted to be flare-quiet. *Right:* Two variable discriminant analysis. Non-flaring regions ( $\times$ ) and flaring regions ( $\diamond$ ) are shown, with the largest flare in any 24 hr period (C, M, X) indicated by color (green, yellow, red respectively). The mean of each sample is shown as a blue circle, and the discriminant boundary is the blue line. An active region with measured values of  $\Phi_{\text{tot}}$  and  $I_{\text{tot}}$  falling above and to the right of the discriminant boundary would be predicted to flare; for a measurement falling below and to the left, the active region would be predicted to be flare-quiet. [Adapted from Leka & Barnes (2007).]

From the Mahalanobis distance, construct the quantity

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2, \quad (8)$$

which has Hotelling’s  $T^2$ -distribution with  $n_1 + n_2 - 2$  degrees of freedom (Kendall et al. 1983). Under the same assumptions as the linear discriminant function is calculated, this quantity can be used to test the hypothesis that the two samples come from the same population.

## 2. Judging the Results: Verification Statistics

Given samples from known populations, one would like to determine how successful a discriminant function is in classifying new objects. The most straightforward way to do this is to use the discriminant function to classify each point in the samples, to produce a classification table. The  $i, j$  element of the classification table is the number of objects predicted to be in population  $i$  and known to be in population  $j$ . Thus the number of correct predictions is given by the sum of the diagonal elements of the table,  $n_{11} + n_{22}$ , while the number of incorrect predictions is given by the sum of the off-diagonal elements,  $n_{12} + n_{21}$ , and the success rate is given by  $(n_{11} + n_{22}) / (n_{11} + n_{22} + n_{12} + n_{21})$ .

In our typical example, the elements will be as follows:

$n_{11}$ : the number of active regions predicted to flare which did produce a flare.

$n_{12}$ : the number of active regions predicted to flare which did not produce a flare.

$n_{21}$ : the number of active regions predicted to not flare which did produce a flare.

$n_{22}$ : the number of active regions predicted to not flare which did not produce a flare.

Because each data point is used in classifying itself, this approach is biased (will tend to give a higher success rate than is really attainable). An *unbiased* way to construct the classification table is to remove one object from the samples, use the remaining objects to construct a discriminant function, and classify the excluded point using this discriminant function. By repeating this procedure, excluding each object in turn, one arrives at an unbiased classification table (Hills 1966).

The success rate is frequently quoted as a measure of how well a forecasting method works. In the case that events are rare (i.e.,  $n_1 \ll n_2$ ), this can be misleading. It is easy to get a high success rate by simply forecasting that no event will ever occur. There are many alternative ways of evaluating the success of a forecasting method (see Murphy 1996, for a summary of some ways, and why they are needed). One possibility is the Heidke skill score, which indicates the *improvement* of the forecasts over always predicting that no event will occur. In terms of the elements of the classification table (and assuming  $n_1 < n_2$ ), the skill

score is given by

$$\text{SS} = \frac{n_{11} + n_{22} - n_2}{n_1}. \quad (9)$$

It is normalized so that perfect forecasts give a skill score of 1, and uniform forecasts (i.e., always forecasting no event) give a skill score of 0. Negative skill scores are possible, and indicate worse performance than a uniform forecast.

A similar skill score can be constructed for probability forecasts:

$$\begin{aligned} \text{SS}(f, x) &= 1 - \text{MSE}(f, x) / \text{MSE}(\langle x \rangle, x) \\ &= 1 - \text{MSE}(f, x) / \sigma_x^2, \end{aligned} \quad (10)$$

where  $\text{MSE}(f, x)$  is the mean square error,

$$\text{MSE}(f, x) = \langle (f - x)^2 \rangle. \quad (11)$$

This skill score has the same properties as the Heidke skill score for binary classifications except that it measures the improvements of the forecasting method over climatology, which always predicts the same (non-zero) probability of an event occurring. The climatological probability of an event occurring is  $n_1 / (n_1 + n_2)$ .

A graphical way to represent the performance of the probability forecasts is a reliability plot, which shows the observed probability as a function of the forecast probability. A reliability plot is constructed by first dividing the forecasts into probability bins. For a bin containing  $S$  total forecasts, of which  $R$  were observed to have at least one event, then the observed probability is  $p = (R + 1) / (S + 2)$ , with an associated uncertainty  $\delta p = [p(1 - p) / (S + 3)]^{1/2}$ . With this definition, a perfect forecasting scheme would result in the diagonal line of observed probability equal to forecast probability. The reliability plot shows where the forecasting method is likely to underpredict (points lying above the diagonal) or overpredict (points lying below the diagonal). Note, however, that the climatological (uniform probability) forecast will result in one point lying on the diagonal. Qualitatively, a good forecasting method will have most forecast probabilities close to either one or zero *and* will have points in the reliability plot close to the diagonal. An example reliability plot is shown in Figure 2.

### 3. Notes on Code

This documentation should be accompanied by two IDL procedures: `da.pro` and `df.pro`, plus an example IDL saved structure, `da.sav`. The procedure `da.pro` calls the function

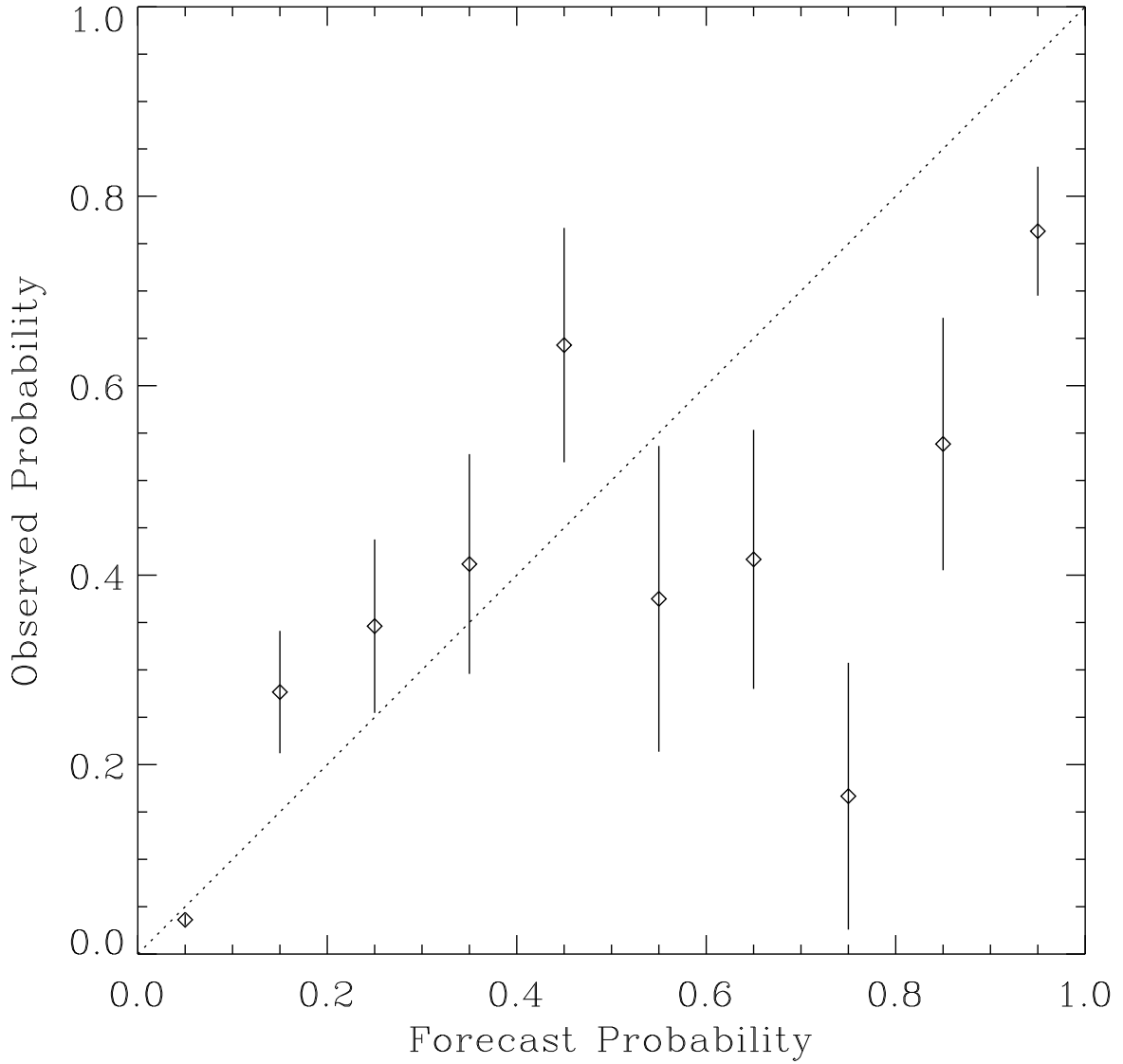


Fig. 2.— Reliability plot for vector magnetic field forecasts of M1.0 and greater flares. For a perfect forecast, all points lie along the line; points lying above the line (most of the small forecast probability bins) indicate an underprediction, points lying below the line (large forecast probability bins) indicate an overprediction. Error bars reflect the number of datapoints in each bin. The large number of forecasts in the smallest probability bin is a sign of a good forecast, but the relatively small number in the largest bin suggests that there is not much improvement over climatology. [Adapted from Barnes et al. (2007).]

df.pro to evaluate the discriminant function, and calculates a range of verification statistics and produces plots. An IDL session using the example data might look something like the following:

```
IDL> restore,/verb,'da.sav'
% RESTORE: Portable (XDR) SAVE/RESTORE file.
% RESTORE: Save file written by graham@pueo, Tue Jun  3 11:09:40 2008.
% RESTORE: IDL version 7.0 (linux, x86).
% RESTORE: Restored variable: STRUC.
IDL> da,struct,tags=["FLUX_TOT"],/color
           pop 0 mean    pop 1 mean
FLUX_TOT      129.436      267.245

discriminant function coefficients
FLUX_TOT      0.00952898
constant      2.75542

Mahalanobis distance squared      1.31319
probability that samples are from different populations:      1.00000

rate of correct classification:      0.772277
classification table
      134      225
      51      802
Heidke skill score (climatology):      0.231198
Heidke skill score (random):      0.364647
rate of correct classification (n-1):      0.772277
classification table (n-1)
      134      225
      51      802
Heidke skill score (climatology, n-1):      0.231198
Heidke skill score (random, n-1):      0.364647

sample 0 size      853
sample 1 size      359
<f>      0.270881
<x>      0.296205
Median f      0.179486
```

<code>sigma_f</code>	0.220306
<code>sigma_x</code>	0.456772
<code>&lt;f x=1&gt;</code>	0.429713
<code>&lt;f x=0&gt;</code>	0.204034
<code>SD f x=1</code>	0.268208
<code>SD f x=0</code>	0.153802
<code>MAE(f,x)=&lt; f-x &gt;</code>	0.312520
<code>MSE(f,x)=&lt; f-x ^2&gt;</code>	0.163510
<code>SS(f,x)</code>	0.216309

The results of this should match §4.2 of Leka & Barnes (2007) (for a C1.0 threshold), although some of the scaling of the plots will differ. For explanation of the verification statistics printed at the end, see Wheatland (2005); Barnes et al. (2007).

#### 4. Acknowledgements

If you publish results using this code, please acknowledge G. Barnes and K.D. Leka, and that the code was developed with funding from AFOSR under contracts F49620-00-C-0004, F49620-03-C-0019, and from NASA under contract NNH07CD25C.

#### REFERENCES

- Barnes, G., Leka, K. D., Schumer, E. A., & Della-Rose, D. J. 2007, *Space Weather*, 5, 9002
- Hills, M. 1966, *J. R. Statist. Soc. B*, 28, 1
- Kendall, M., Stuart, A., & Ord, J. K. 1983, *The Advanced Theory of Statistics*, 4th edn., Vol. 3 (New York: Macmillan Publishing Co., Inc)
- Leka, K. D. & Barnes, G. 2007, *ApJ*, 656, 1173
- Murphy, A. H. 1996, *Wea. Forecasting*, 11, 3
- Silverman, B. W. 1986, *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall)
- Wheatland, M. S. 2005, *Space Weather J.*, 3, S07003