

Photospheric Magnetic Field Properties of Flaring vs. Flare-Quiet Active Regions III: Discriminant Analysis of a Statistically Significant Database

K. D. Leka, Graham Barnes

**Colorado Research Associates Division,
NorthWest Research Associates, Inc.**

Solar active regions are often evaluated for their potential to produce energetic events based on their magnetic morphology. Quantitatively this information is available using vector magnetic field data, (presently) routinely gathered from photospheric observations. Recently we demonstrated a method of parameterizing vector field information such that descriptions of the magnetic morphology and complexity were contained in the statistical description of (as examples) the vertical current or shear angles; it was also demonstrated that even with high-cadence data, no single parameter consistently and uniquely displayed pre-event variations (Leka & Barnes 2003a). We also showed, however, that with Discriminant Analysis (Leka & Barnes 2003b), it is possible to distinguish between an event-imminent photospheric magnetic state and an event-quiet state – but only by considering multiple variables simultaneously. The limitations of that demonstration were primarily due to small-number statistics given the dataset used.

In the present work, Discriminant Analysis is applied to a very different dataset: the daily "survey" magnetograms obtained by the U. Hawai'i/Mees Solar Observatory Imaging Vector Magnetograph. In this manner, the problem of small-number statistics is mitigated and advantages available by DA are explored. However, given the daily temporal cadence the focus shifts toward detecting parametric thresholds rather than pre-event specific evolution. Nonetheless, the central question remains how to distinguish a region which is primed for an energetic event, applicable to modeling efforts by providing empirical discriminating information as to the pre-eruption state of the boundary magnetic field.

This effort is funded by contract F49620-03-C-0019 through the Air Force Office of Scientific Research.

PROJECT SUMMARY:

⊙ Goals are two-fold:

- Determine the magnetic state of the flare-productive photosphere.
- Develop a quantitative tool for flare forecasting using data available presently and in near-future.
- Build on previous work and answer a slightly different question.

● Previous: Time-Series data searching for pre-flare signatures.

Leka & Barnes 2003 a, b

- Demonstrate Discriminant Function Analysis and Hotelling's T^2 test
- Focused sample on flaring active regions to discriminate the “flare-imminent” magnetic state.
- Vector field time-series data, divided into hour-long epochs.
- Explicitly considered the null hypothesis by including “flare-quiet” epochs.
- Primarily a demonstration due to small sample-size.
- help yourself to an offprint...
- Here: “Survey” data → search to parameterize the flaring photosphere.
 - Data comprised of “daily” magnetograms
 - Focus on flare productivity in 24-hours post-observation.
 - Much larger database, approaching statistical significance

STATISTICAL ANALYSIS

⊙ Hotelling's T^2 Test:

- Quantifies probability that the samples come from distinct populations
 - Measures the distance between the sample means, relative to the sample variance.
 - A high probability of different parent populations can come from samples with large overlap.

Kendall, Stuard & Ord
1983

Anderson 1984

Leka & Barnes 2003b

⊙ Discriminant Function Analysis

- Given measurements *known* to come from two populations (e.g., flaring and flare-quiet), a Discriminant Function divides parameter space:
 - For 2-variable functions, the Discriminant is a line (see Figure 1).
 - The magnitudes of the coefficients of the variables in standardized form in the Discriminant Function give the relative predictive power of those variables.
 - The constructed Discriminant Function then predicts from which parent population a (new) data point originates, according to its location relative to the DF.
 - Maximizes the correct predictions given equal probabilities for errors of both types (i.e., off-axis elements in the classification table; see below).

DATA & ANALYSIS

⊙ Imaging Vector Magnetograph, Mees Solar Observatory, Hawai'i

Mickey *et al.* 1996

● Instrument and Data Reduction:

- Imaging Fabry-Perot tunes to 30 samples of FeI 6302.5Å $g_L = 2.5$ line, full Stokes sampling at each.
- Large, AR-sized Field of View ($280''^2$), binned to $1.1''$, ≤ 5 min. cadence capable.
- Default mode: A “survey” of each NOAA numbered active region, followed by high-cadence observations the rest of the day.
- “Survey” data reduction optimized for real-time data availability.
- Inversion based on a wavelet spectrum analysis.
- Ambiguity Resolution: simulated annealing approach to minimize J_z , $\nabla \cdot B$, with an initial comparison to a “best-fit” linear force-free solution.

● ANALYSIS

- Relevant Quantities derived from $B(x, y)$ with the purpose of quantifying the state of the active region's magnetic field (its morphology and complexity).
- Variables discussed below are parameterizations of spatial distributions, using moment analysis (mean, standard deviation, skew and kurtosis) and summations where appropriate (see Table).
- Variables of interest gleaned from literature; always interested in new ideas....

● TIME-SERIES Data

- 7 Active Regions,
- 14 Flare-Quiet epochs, 10 Flare epochs, where a flare epoch ended at the GOES *start* time.

● SURVEY Data

- “Daily” single magnetograms of each NOAA-numbered active region visible on the disk.
- Only $\mu < 0.5$ and $\max(|\mathbf{B}|) < 500\text{G}$ (Limb grams and spotless regions) not considered.
- Classified as “Flaring”/“Non-Flaring” if it did/did not produce *any* flare (C-class or above) in the 24 hr following the magnetogram time.
- Initial data from 2002 January, June and July, 2003 January, June.
- 75 “Flare” magnetograms, 270 “Flare-Quiet” magnetograms.

ONE-DAY SAMPLE of SURVEY DATA

Magnetogram		NOAA AR	Flare Start		Peak
Day	Time	Number	Day	Time	X-Ray Flux
20020107	1734	09773	20020107	2049	C2.9
//	//	//	20020107	2243	C2.8
//	//	//	20020108	1255	C2.5
20020107	1740	09761
20020107	1745	09772
20020107	1750	09767	20020108	1713	C7.2
20020107	1754	09771

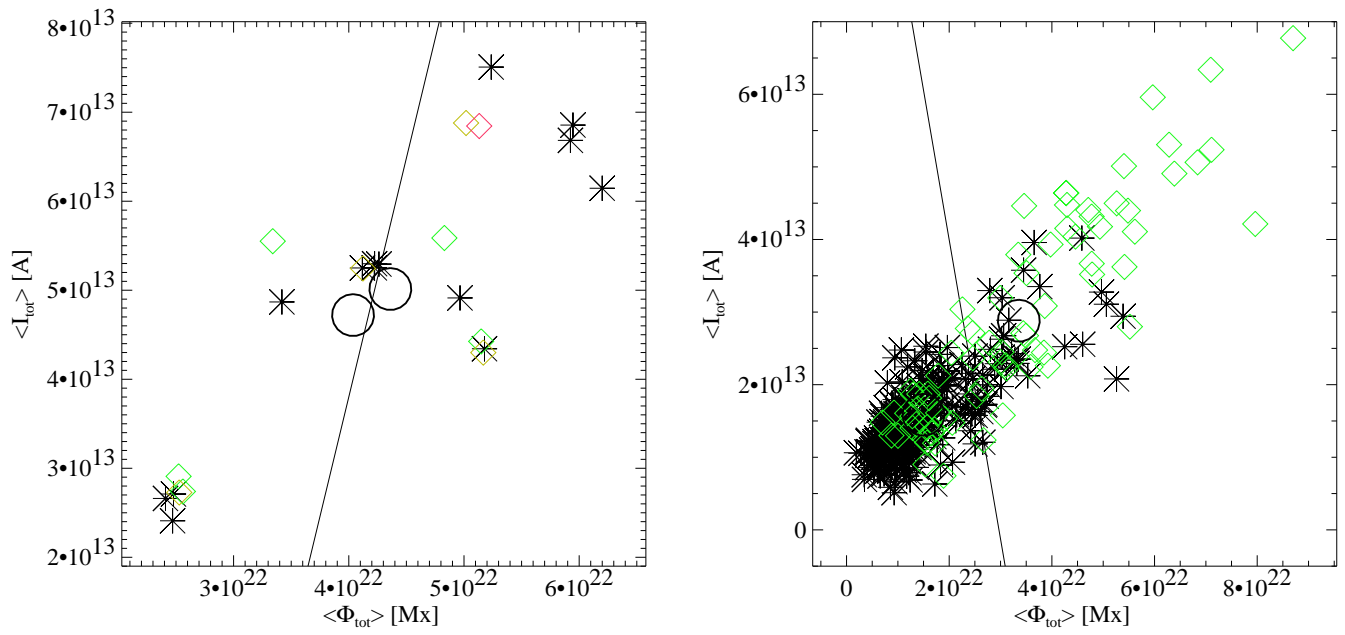


Figure 1: Two-Variable Discriminant Functions [Total Unsigned Magnetic flux vs. Total Unsigned Vertical Current] for (left) original “time-series” data described in Leka & Barnes 2003, and (right) “survey data”. Note that in the former the flares (\diamond) which occurred are color-coded for peak SXR strength while in the latter the flares were not differentiated by size; flare-quiet data are indicated by (*). Also note that for both datasets, these two variables are correlated as seen by the DF’s non-perpendicular orientation with respect to the sample means.

COMPARISON: The Probability Sort

- ⊙ **Q: Which variables are the “best”?**
 - **Small sample size precluded an “all-variable” DF** (see below, however).
 - **Proxy was developed: the N-Variable Sort.**
 - Relied on all permutations of N-var DF where N was appropriate for the sample size.
 - With larger sample size, proxy can be evaluated against “the real thing”.
- ⊙ **Single-Variable Sort: Ranks discriminatory power of each considered variable separately.**
 - **Time-series data:**
 - Consider only means here, not temporal variations
 - **Most discriminating: abs. value of net magnetic flux $|\Phi_{net}|$**
 - Probability that samples are from different populations, T^2 test: **0.6732**
 - Success rate (from classification table): **0.6667**
 - Success rate (from “n - 1”): **0.6667**
 - **Survey data:**
 - **Most discriminating: total current helicity $H_c^{tot} = \sum |h_c| dA$**
 - T^2 probability: **1.0000**
 - Success rate (from classification table): **0.8260**
 - Success rate (from “n - 1”): **0.8260**
 - **Summary:**
 - Results differ between data sets (no big surprise).
 - No single variable can perfectly predict subsequent flare activity (no big surprise).
 - The best single-variable DF from the Survey data is significantly better than from the time-series data (moderate surprise).

⊙ **4-Variable Sort:**

● **DFs calculated for *all* 4-variable permutations**

- Results ranked according to T^2 test.
- In 500 best DFs, which variables appear most frequently?
- The frequency ranking serves as a proxy for the standardized coefficients in a single all-variable DF.

● **Best 4-Var DF:**

○ **Time-Series Data:**

$$f = 0.42 - 7.48 \sigma(B_h) + 9.40 \varsigma(|\nabla_h B|) - 2.28 \overline{J_z} - 12.12 A(\Psi > 80^\circ)$$

- T^2 probability: **0.999195**

Success rate (from classification table): **0.958333**

Success rate (n-1): **0.833333**

○ **Survey data:**

$$f = 0.903 + 1.03 \Phi_{tot} + 0.529 \overline{|\nabla_h B_z|} + 0.990 H_{c,tot} - 0.305 \varsigma(\psi_{NL})$$

- T^2 probability: **1.00000**

Success rate (from classification table): **0.852174**

Success rate (n-1): **0.852174**

Top 10 Variables from 4-Variable Probability-Sort Results

TIME SERIES		SURVEY	
Variable	Frequency	Variable	Frequency
$\sigma(B_h)$	354	Φ_{tot}	458
$\varsigma(\nabla_h B_h)$	168	$H_{c,tot}$	331
$\sigma(\psi)$	162	$\overline{ \nabla_h B }$	178
$L(\Psi_{NL} > 80^\circ)$	132	$\sigma(\Psi_{NL})$	117
$L(\psi_{NL} > 80^\circ)$	126	$\overline{ \nabla_h B_z }$	97
$\kappa(\nabla_h B_h)$	124	$\sigma(\nabla_h B_z)$	80
$L(\psi_{NL} > 45^\circ)$	81	$A(\psi > 45^\circ)$	53
$A(\Psi > 45^\circ)$	56	$\sigma(\nabla_h B)$	47
$L(\Psi_{NL} > 45^\circ)$	47	$\overline{B_h}$	38
$\sigma(B_z)$	39	$\sigma(\nabla_h B_h)$	33

● **Summary:**

- Little overlap between results (moderate surprise).
- Clear that correlated variables are playing (too) large a role and need to be eliminated (“to do” list).

COMPARISON: The “Probability Sort” vs. “‘n – 1’ Sort”

- **Comparison of 4-variable sort results:**
 - Sort on “T²-test probability of different populations”
vs.
 - Sort on “n – 1” success rate.
- **SURVEY data only: “n–1” requires large-N sample.**
- **Top 4-var DF:**
 - **Probability Sort:**
 - (see “Best 4-Var DF” for Survey data, above)
 - **“n–1” Sort:**

$$f = 0.47 + 0.23 s + 0.06 \overline{B}_z + 1.44 I_{tot}^h - 0.06 \kappa(\psi_{NL})$$
 - T² probability: **1.00000**
 - Success rate (from n-1 and classification table): **0.869565**

Top 10 Variables from 4-Variable Sort Results

Sort on PROBABILITY		Sort on “n – 1 ERROR	
Variable	Frequency	Variable	Frequency
Φ_{tot}	458	I_{tot}^h	456
$H_{c,tot}$	331	I_{tot}	325
$ \nabla_h B $	178	s	122
$\sigma(\Psi_{NL})$	117	$\sigma(J_z^h)$	77
$ \nabla_h B_z $	97	$\kappa(B_z)$	57
$\sigma(\nabla_h B_z)$	80	I_{net}^B	46
$A(\psi > 45^\circ)$	53	$\kappa(\nabla_h B)$	46
$\sigma(\nabla_h B)$	47	$\sigma(J_z)$	45
\overline{B}_h	38	E_f	41
$\sigma(\nabla_h B_h)$	33	$\varsigma(J_z)$	40

- **Summary:**
 - Once again, no overlap
 - **Embarrassing: seeing appears as # 3!**
 - Jury is still out.

ANOTHER PROXY

⊙ “Top-Down” Approach:

- A method to sort and *remove*, poorly performing or correlated variables:
 - Given N variables, construct and rank all possible $N - 1$ DFs; that variable missing from the best DF is then dropped from consideration.
 - Procedure is repeated until run out of variables.
- Example: Weeding the Worst
 - Start with 20 different Variables, 10 best & 10 worst according to the 4-variable “n-1” sort (plus a few due to overlap).
 - at $N = 10$, the following were left:

4-Variable Sort vs. Recovered from “Top-Down”

Input Best 10	Input Worst 10	Status at N=10
I_{tot}^h	$\varsigma(J_z^h)$	I_{tot}^h
I_{tot}	$\kappa(J_z)$	I_{tot}
s	$\kappa(h_c)$	I_{net}^B
$\sigma(J_z^h)$	$L(\Psi_{NL} > 45^\circ)$	E_f
$\kappa(B_z)$	I_{tot}	$\kappa(J_z)$
I_{net}^B	$L(\psi_{NL} > 45^\circ)$	$\kappa(h_c)$
$\kappa(\nabla_h B)$	$\sigma(h_c)$	$L(\Psi_{NL} > 45^\circ)$
$\sigma(J_z)$	$\sigma(J_z)$	$L(\psi_{NL} > 45^\circ)$
E_f	$\kappa(\Psi)$	Φ_{tot}
$\varsigma(J_z)$	Φ_{tot}	$\overline{ \nabla_h B_z }$

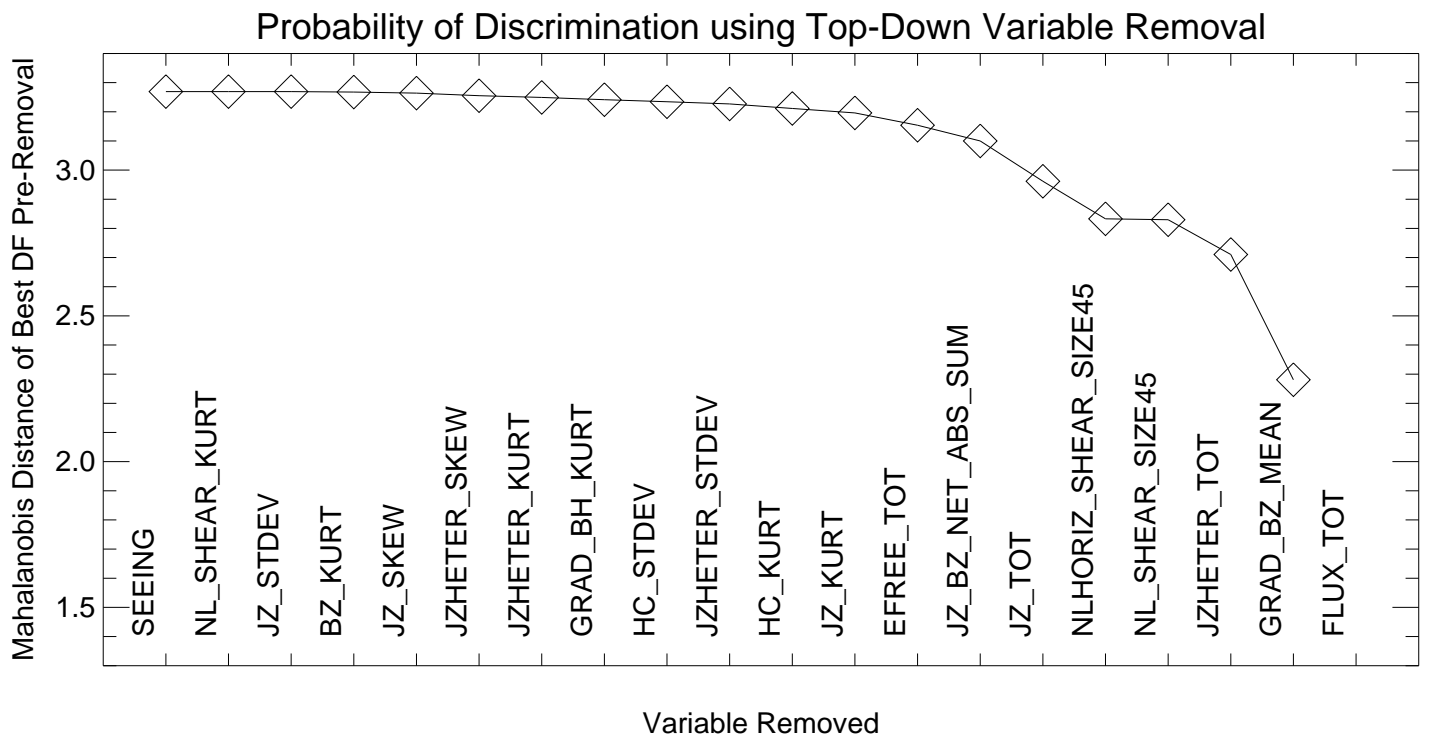


Figure 2: Weeding the Variables: Y-axis shows the Mahalanobis distance from the resulting DF, the best of all combinations using the variables available (a large Mahalanobis indicates a higher probability that the two populations can be distinguished). The X-axis indicates the variable which was removed because it did not appear in that best DF.

- **Summary:**

- Seeing is thrown out first! (yeah!)
- after 10 variables are discarded, only 4/10 of the “best” input variables are recovered.
- At this point, however, the Mahalanobis distance starts to decrease rapidly.

- **Clear demonstration that:**

- Multiple variables perform better than one/few.
- Beyond a certain number of variables, the results do not improve significantly.

THE REAL THING

- ⊙ **The Goal: an All-Variable Discriminant Function:**
 - Only possible with large enough samples size; approaching that now (?).
 - Best if variables are uncorrelated
 - (definitely not the case now)
 - Recall the sorting approach was constructed to be a proxy for the all-variable DF:
 - Time-series data: coefficients of 10-variable DF generally agreed with frequency in best/worst 4-variable sorts.
 - Now can evaluate the proxy directly against an all-variable Discriminant Function.

Best and Worst Comparison from Sort vs. All-Variable DF

4-Var Probability Sort		4-Var “n – 1” Sort		All-Variable DF	
Variable	Frequency	Variable	Frequency	Variable	Coefficient
Φ_{tot}	458	I_{tot}^h	456	$\bar{\psi}$	-10.49
$H_{c,tot}$	331	I_{tot}	325	$L(\Psi_{NL} > 45^\circ)$	-9.645
$ \nabla_h B $	178	s	122	$L(\psi_{NL} > 45^\circ)$	9.112
$\sigma(\Psi_{NL})$	117	$\sigma(J_z^h)$	77	$\bar{\Psi}$	7.100
$ \nabla_h B_z $	97	$\kappa(B_z)$	57	$A(\psi > 45^\circ)$	6.966
Worst 5		Worst 5		Worst 5	
Φ_{tot}	167	I_{tot}	156	J_z^h	0.0766
I_{tot}	187	$L(\psi_{NL} > 45^\circ)$	162	$\kappa(J_z)$	-0.0569
$\kappa(h_c)$	196	$\kappa(h_c)$	202	I_{net}	-0.0384
$\varsigma(J_z^h)$	197	$\kappa(J_z)$	204	$\varsigma(J_z^h)$	0.0270
$\kappa(J_z)$	213	$\varsigma(J_z^h)$	215	I_{net}^h	0.0223

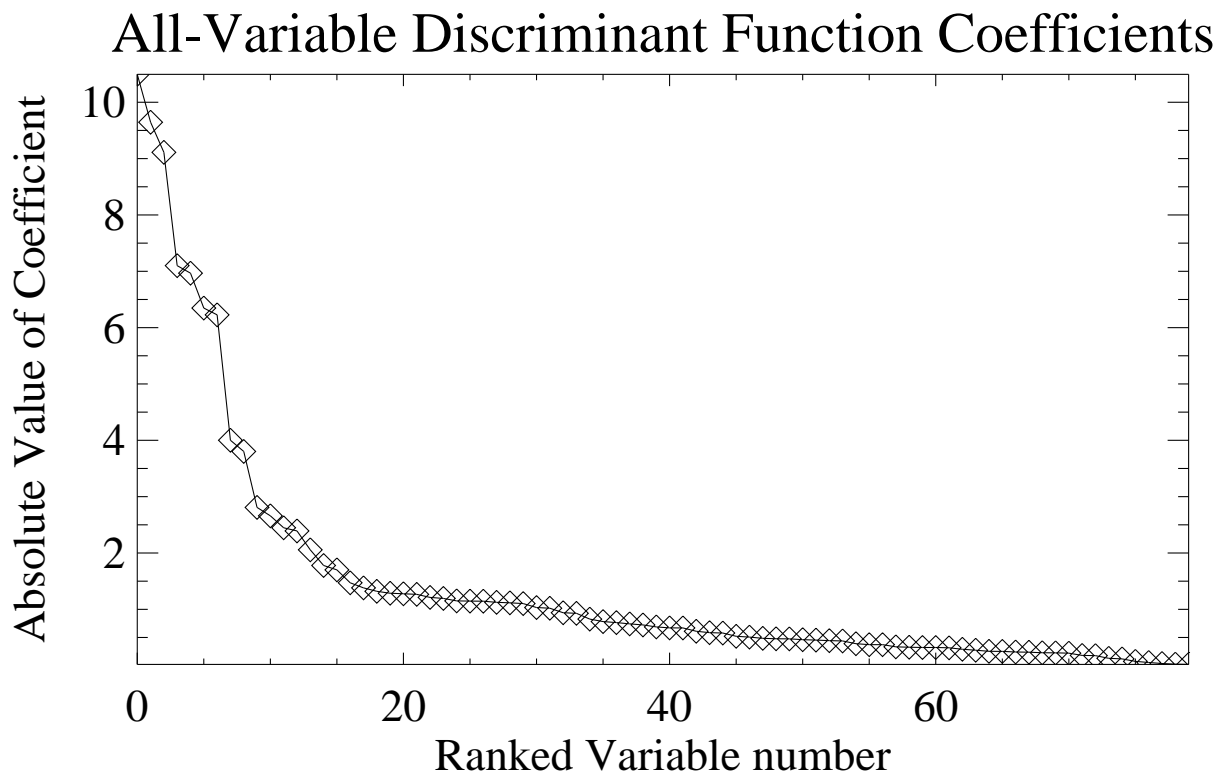


Figure 3: Plot of Standardized Coefficients from all-variable discriminant function from Survey data and 80 variables. Remarkable drop-off in coefficient magnitude may (or may not) reflect how many variables are truly needed for robust discriminant function.

- **Summary:**

- Correlated variables swamping the direct comparison between the all-variable DF and 4-var sort.
- Standardized Coefficient results consistent with hope that at a determinable level, the importance or predictive power of many variables is sharply lower than the best 20 or 30.
- Large enough data-base → all-variable DFs will be *a good thing* once correlated variables are removed..

SUCCESS/ERROR RATE ESTIMATION

- ⊙ **Goal:** Quantitatively estimate how well a discriminant function performs.
- ⊙ **Standard Approaches:**
 - “Truth” (“Classification”) Table is constructed from the Discriminant Function and the input “learning” data.
 - Fraction of incorrect classifications gives an estimate of the error rate.
 - **N.B., the error rate from “Truth” Tables is always an underestimate!**
 - The “ $n - 1$ ” approach: Unbiased estimate of error rate derived by constructing Discriminant Functions using $n - 1$ of the original n “learning” data points.
 - By then classifying the missing data point and iterating for all n data points, an error rate is derived.
 - In large-sample limit, provides a good estimator for the error rate for all n points: this is evident with the new larger sample.

Hills 1966

○ Effect of Unequal Population Sizes

- **A Discriminant Function is constructed to:**

- Obtain mis-classification errors in proportion to the population sizes (or other “cost” requirements)
 - then –
- Minimize the total mis-classification error

- **But: May not provide the lowest error.**

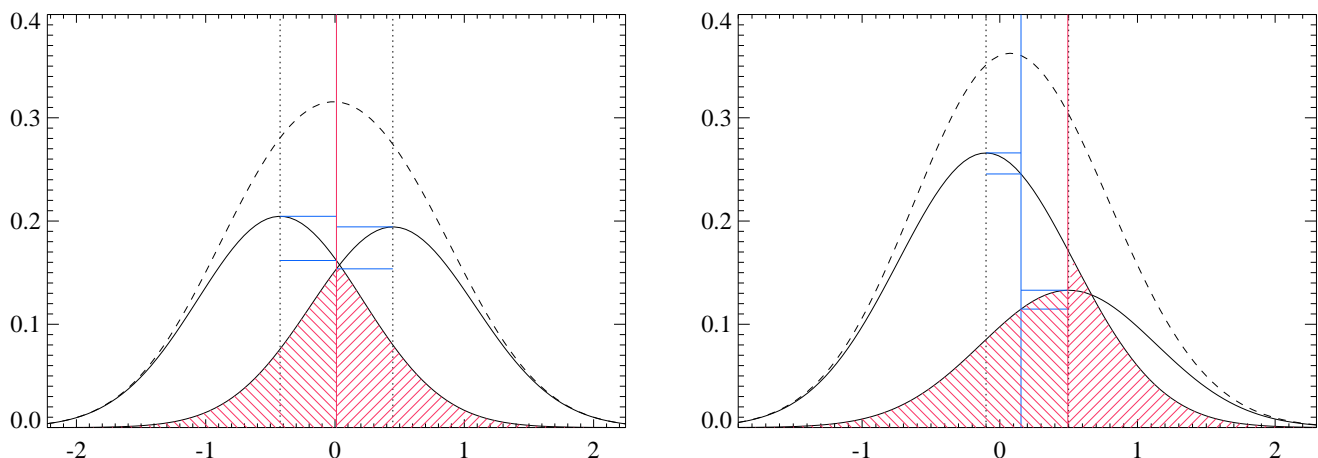


Figure 4: Two populations are shown with the same variance but displaced means (\cdots), with nearly equal (left) and very unequal (right) population sizes (the latter size ratio is comparable to this Survey-data study). Two options of where to place the DF are indicated, equal probability of making errors of either type (red line) or to produce errors in proportion to the relative population sizes (blue line). For the error-minimization, the mis-classification regions are indicated by the hatch patterns.

- **Why are the population sizes unequal?**

- There is no selection effect in the data acquisition: the IVM obtains survey data in order of NOAA AR number, and only spotless and limb-inclusive data are excluded from this study.
- The Sun is providing the unequal population.
- Q: How does this population difference vary with solar cycle?

SUMMARY

⊙ Goal: Two-fold:

- Empirically determine the magnetic state of a flare-ready photosphere.
- Develop a quantitative tool for event forecasting using presently-available data.

⊙ Applying DF analysis to Statistically Significant Database:

● Initial results:

- Work in progress.....
- With larger sample size, tools available from the statistical analysis become available (all-variable DF, *etc.*).
- Systematics in data may play a role in apparent “ceiling” of classification success rates ($\lesssim 0.90$).
- On the other hand, it might be how the Sun simply *is*...
- Redundant and correlated variables obviously influencing results.
- Subtle statistical aspects of the unequal population sizes are becoming apparent.

● To Do list:

- Remove redundant and worst correlated variables for best all-variable DF.
- Fully understand the effects of unequal population on interpretation.
- Explore data from different times of the solar cycle: populations may be unequally unequal (*groan....*).
- (other non-public bug fixes....)

● See you at the SPD!