PROBABILITY PLOTTING OF MULTIPLE DISCRIMINANT RATIOS

Kenneth W. Wachter, Harvard University

This report is about some spinoff from the plotting methods for principal components described at the Ninth Interface Symposium (Wachter (1976a)). The results go back to joint work of Colin Mallows and the present author in 1969 at Bell Telephone Laboratories, Mallows & Wachter (1970). What is new is a rigorous proof of the theorem, stated below, on which the probability plotting methods for multiple discriminant ratios and related quantities from large multivariate data sets can be based. The proof is given in a Harvard Research Memo, Wachter (1976c), which is now being expanded for publication. The present report is a brief sketch of the content, uses, and limitations of the results.

A typical problem which our methods address arises when we have groups of observations on, say, a dozen variables and we want to judge whether discriminant analysis is worth pursuing. We offer a quick graphical method for comparing the data against a standard null hypothesis under which the variables contain no information relevant for discriminating among groups.

In all cases our methods rely on calculations of the asymptotic empirical measure of a set of eigenvalues or singular values or other roots. Figure 1 displays the distribution function of an empirical measure, a step function with a step of size 1/p at each of p values. We deal with asymptotic situations in which, in one way or another, the number of steps, p, is going to infinity, yielding an increasingly gentle staircase like Figure 1 in contrast to the few big jumps of Figure 2.

Figure 1.

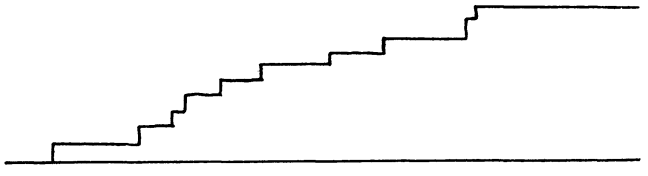
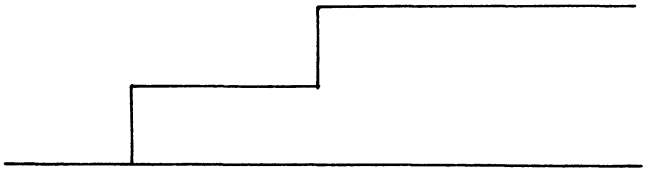


Figure 2.



* * *

The particular p values of interest to us are random variables $L_1 \ldots L_p$ whose joint density is proportional to

$$\Pi_{i=1}^{p} L_i^{(m-p-1)/2} (1-L_i)^{(n-p-1)/2} \Pi_{j<i} |L_i - L_j|$$

We call them the Fisher-Hsu-Roy-Girshick-Mood roots. They are the solutions to a familiar determinental equation in random matrices; their joint distribution was discovered simultaneously in 1939 by these five men. They play a role in multivariate analysis of variance and in canonical correlations as well as in discriminant analysis and go under many names. The equivalence of null hypotheses in these areas to the set of assumptions in Theorem 1 follows easily from the expositions in standard textbooks like Anderson (1958) or Dempster (1969).

Theorem 1 brings us the happy news that the empirical measure of the Fisher-Hsu-Roy-Girshick-Mood roots converges in distribution to a fixed limit with a density in simple closed form. The convergence takes place as the dimension and degree-of-freedom parameters p,m and n in the expression for the joint density go to infinity while their ratios approach fixed parameters.

Theorem 1: Suppose

1. Z or Z(n) is a p(n) by n dimensional matrix whose columns are mean-centered independent exchangeable multivariate normal random vectors and whose transpose is Z*.
2. J is an n by n projection matrix of rank m.
3. K_n is the empirical measure of the p(n) positive solutions x to $\det |ZJZ^* - xZZ^*| = 0$.
4. Prob $\mathbb{R}$ is the space of probability measures on the real line with the topology of weak convergence.
5. As $n \to \infty$, $p(n)/n \to \beta$ and $m(n)/n \to \mu$ with $\beta \leq \mu < 1$. Then as $n \to \infty$, the random element K_n in Prob $\mathbb{R}$ converges in distribution to the fixed element of Prob $\mathbb{R}$ concentrated on $[A^2, B^2]$ with density

$$\sqrt{(y-A^2)(B^2-y)} \; / \; (2\pi\beta(1-y)y)$$

where

$$A = \sqrt{\mu-\mu\beta} - \sqrt{\beta-\mu\beta} \quad \text{and} \quad B = \sqrt{\mu-\mu\beta} + \sqrt{\beta-\mu\beta}.$$

The difficulty of this theorem, responsible for the seven year gap between the discovery by Dr. Mallows and the author of the limit formula and the present proof, is stochastic degeneracy of the limit. For each triplet of finite values of p,m, and n, the empirical distribution function is a random step function. It is easy to imagine a limit

which is still random rather than fixed. That the limit be fixed, that is, stochastically degenerate, is equivalent by Wachter (1974), 3.2, to the assertion that the roots be asymptotically independent. In effect we need this independence to simplify expressions for moments. The proof in Wachter (1974c) insures against random limits by a conditioning argument and throws the onus onto keeping convergence uniform over the conditioning events. It thus depends upon the new uniform convergence theorems for random matrix spectra proved in Wachter (1976b).

The assumptions in Theorem 1 are the standard null assumptions of the areas of application, but they are highly restrictive. In particular, they require multivariate normality. It is possible that uninteresting departures from normality, instead of interesting departures from the exchangeability of groups, could be responsible for bad fit of data to null hypothesis. Unfortunately, normality, via rotational invariance, is crucial to the proof of Theorem 1. On the other hand, the parallel theorems for principal components in Wachter (1976a and b) hold without any distributional assumptions beyond weak moment bounds, tempting us to conjecture the appropriateness of the limit formula even in the absence of normality.

No simulations have yet been done to test the speed of convergence in Theorem 1 itself, but simulations have been completed in the principal component case which suggest convergence rapid enough to make plots for p=12 or more informative.

Figure 3 shows densities for the limiting empirical measures of Theorem 1 for a variety of values of μ and β. Each graph is positioned in the figure according to its value of μ (for the x-axis) and β (for the y-axis). The operation of replacing μ by $1-\mu$ and reflecting each graph about the center would produce further cases.

The densities in Figure 3 give a more intuitive sense of how the Fisher-Hsu-Roy-Girschick-Mood roots tend to spread out when there are many of them than does, for instance, the expression for their joint density. Some surprises lurk in the graphs. Notice, for instance, that for small enough β and large enough μ most of the roots (which have an interpretation as squared correlations) clump above .99. Without theoretical guidance one might wrongly sieze on correlations this large as grounds for rejecting the null hypothesis. Notice also that when μ and β are near 1/2 the roots tend to separate into two bunches. Procedures which include linear discriminators in an analysis until a large gap between roots is reached would operate merrily in this situation, even though in fact the null hypothesis is true.

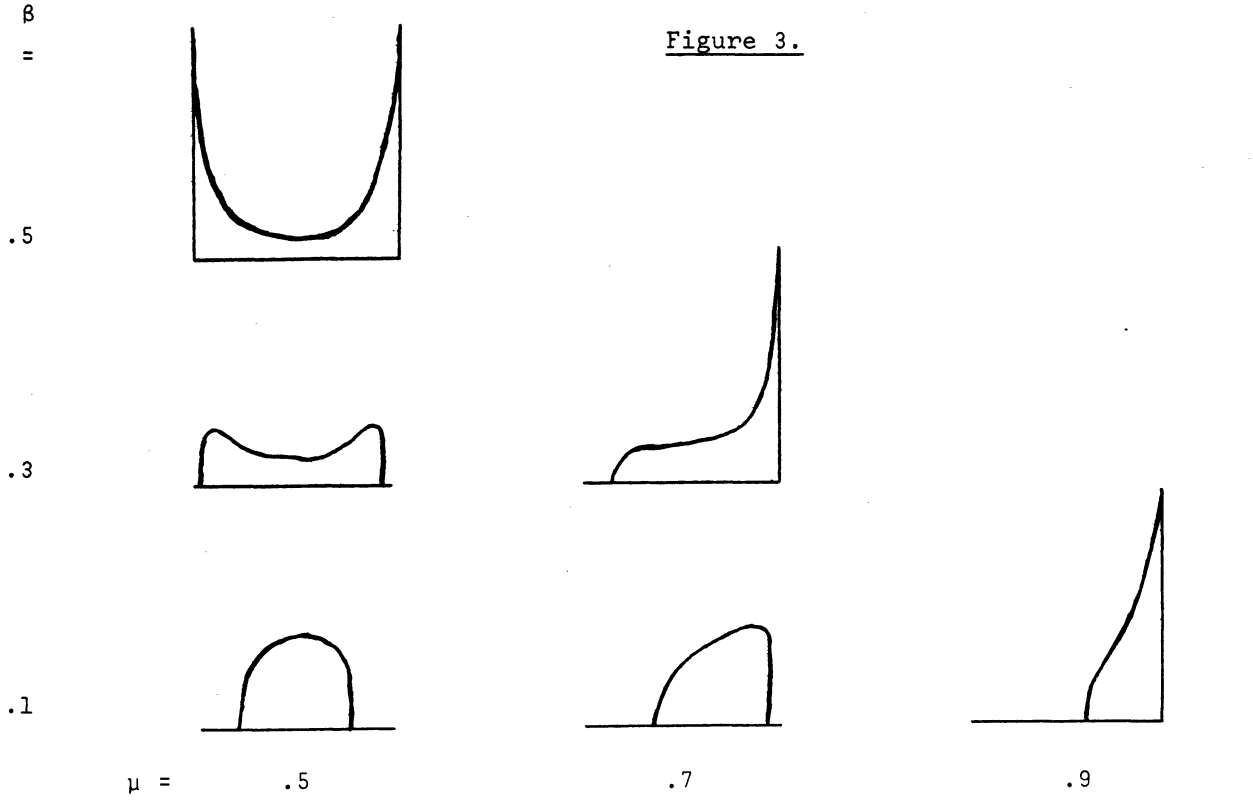


Figure 3.

Besides instructing our intuition, Theorem 1 provides plotting points for quantile-quantile plots of sets of roots from data against theoretical values under the null hypothesis. We illustrate this method with an example of a discriminant analysis, performed during research into talker identification at Bell Telephone Laboratories described in Bricker et al. (1971). We begin with measurements of p variables replicated k times on each of m groups. In this case the p=19 variables are spectral measurements on utterances. There are m=172 groups of utterances consisting of k=4 utterances each. Each group corresponds to a single talker repeating the same digit four times. The aim of the analysis is to discriminate between groups, with the hope of finding variables which will efficiently classify a new utterance into one of the groups, that is, assign an utterance to the person who spoke it.

Like much multivariate research, the data set in this problem is large. Many alternative analyses need to be tried and the cost of computing new variables in any one analysis can mount up. Thus there is a premium on methods which help us recognize hopeless from promising analyses at an early stage. It is precisely under such circumstances that our methods have something to offer.

Let W be the sample within-group covariance matrix pooled across groups and let B be a "between-group" covariance matrix, the sample covariance matrix of the group centroids or means. Our roots $L_1 \ldots L_p$, sometimes called discriminant ratios, are the p eigenvalues of $(W+B)^{-1}B$, though we can calculate them via singular value decompositions without actually forming W and B. The ordered roots for this data set appear in the second column of the table in Figure 4. Without a standard of comparison these eigenvalues may not seem especially illuminating. A standard of comparison is just what Theorem 1 provides.

Suppose R is the distribution function for the density in Theorem 1. In this problem (trading one degree of freedom for the grand mean) n=172*4-1 and we set β=19/n=.028 and μ=172/n=.249. For each i define X_i to be the value such that $R(X_i)=i/(p+1)$. The computer routines listed in Wachter (1976a) can be used to find X_i if the subroutines DENSIT and RANGE are altered in an obvious fashion, but for pencil-and-paper accuracy a hand-calculator suffices. The values $X_1 \ldots X_{19}$ for this μ and β appear in Column C. We plot L_i on the y-axis against X_i on the x-axis for i=1,2...p=19. Agreement between the data and the null hypothesis would manifest itself by points lying along a straight line through the origin at 45°.

Figure 4

i	L_i	X_i
1	.949	.369
2	.907	.349
3	.901	.331
4	.867	.317
5	.843	.301
6	.825	.289
7	.814	.278
8	.803	.265
9	.782	.252
10	.754	.241
11	.722	.229
12	.710	.219
13	.676	.207
14	.659	.198
15	.620	.185
16	.550	.175
17	.473	.162
18	.428	.151
19	.403	.138

Table of Eigenvalues L_i from Talker Identification Data and Quantiles X_i from Null Hypothesis with β=.028 ; μ=.249.

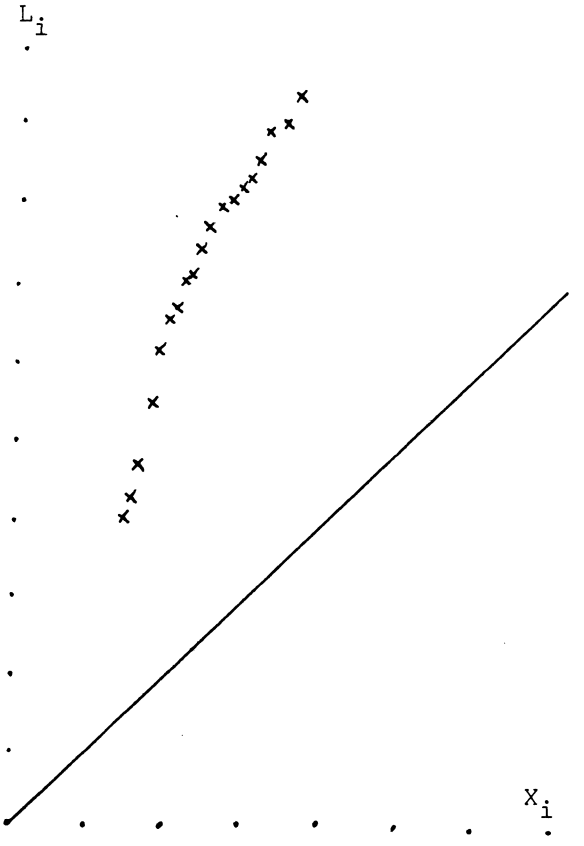
* * * * *

It is evident that the resulting plot, in Figure 5, shows little agreement with the null hypothesis. The points depart radically from the 45° line and their locus curves distinctly. In this problem this is welcome information, for it suggests that talkers are by no means exchangeable, so that additional labour invested in the calculation of discriminant coordinates ought to prove justified. In fact, as described in Bricker et al. (1971), identification strategies using five linear discriminators turned out to be highly successful at identifying new utterances by talkers in this population.

It would be desirable to go further and explore the departure from the null hypothesis indicated by Figure 5 by constructing plots of the L_i against quantiles derived for various alternative hypotheses of structure in the data useful for discrimination. For the parallel methods for principal components described in Wachter (1976a) quantiles under alternative hypotheses are available, but for discriminant ratios Theorem 1 does not furnish any information except under the null hypothesis. It offers no guidance as to how many discriminant coordinates we should retain in an analysis or how we should employ them. In this regard the present methods are severely limited in scope.

On the other hand, the quick graphical check of the data against the null hypothesis based on Theorem 1 becomes feasible for just those high-dimensional

Figure 5.



Q-Q Plot: Data versus Null Hypothesis

* * *

cases where tables of standard test statistics become sparse and intuition requires instruction. A by-product of more extensive results for principal components and general random matrix spectra, these methods fill a small but nagging gap in our collection of statistical tools for approaching multiple discriminant analysis and related fields.

REFERENCES

T.W. ANDERSON (1958). An Introduction to Multivariate Statistical Analysis, Wiley, N.Y.

BRICKER, GNANADESIKAN, MATHEWS, PRUZANSKY, TUKEY, WACHTER & WARNER (1971). "Statistical Techniques for Talker Identification," Bell System Technical Journal, 50, 1427-1454.

A.P. DEMPSTER (1969). Elements of Continuous Multivariate Analysis, Addison-Wesley, Reading, Mass.

C. MALLOWS & WACHTER (1970). "The Asymptotic Configuration of Wishart Eigenvalues," An Abstract," Annals Math. Statist. 41, 1384.

K.W. WACHTER (1974). "Exchangeability and Random Matrix Spectra," in Progress in Statistics, edited by Gani, Sarkadi, & Vince, North Holland Press.

K.W. WACHTER (1976a). "Probability Plotting Points for Principal Components," in Computing and Statistics: Proceedings of the Ninth Symposium on the Interface, edited by Hoaglin & Welsch.

K.W. WACHTER (1976b). "The Strong Limits of Random Matrix Spectra for Sample Matrices of Independent Elements, Parts I and II," typescript awaiting publication.

K.W. WACHTER (1976c). "The Limiting Empirical Measure of Multiple Discriminant Ratios," Research Memo W-76-2, NS-335, Harvard University Department of Statistics.

ACKNOWLEDGEMENT

This work was facilitated by Grant SOC-75-15702 from the National Science Foundation.