

Review of the Williams - Bartlett - Kshirsagar Work
on Direction and Collinearity Factors
in Discriminant Analysis

by

A. M. Kshirsagar

Technical Report No. 107
Department of Statistics ONR Contract

May 1971

Research sponsored by the Office of Naval Research
Contract N00014-68-A-0515
Project NR 042-260

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

This document has been approved for public release
and sale; its distribution is unlimited.

DEPARTMENT OF STATISTICS
Southern Methodist University

A. M. Kshirsagar
 Southern Methodist University

Consider k random samples of sizes $n_1, n_2, \dots, n_k$ from $k = q + 1$ independent p -variate normal populations Π^α ($\alpha = 1, 2, \dots, k$) with means $\bar{\mu}^\alpha$ and the same variance-covariance matrix Σ_1 . The multivariate analysis of variance table will be

Source	d.f.	S.S. & S.P. matrix
Between groups	q	$B = C_{XY} C_{YY}^{-1} C_{YX}$
Within groups	$n - q$	$A = C_{XX} \cdot Y$
Total	n	$A + B = C_{XX}$

In the above table $n = \sum_{\alpha=1}^k n_\alpha - 1 = N - 1$, C_{XX} is the matrix of corrected

S.S. (sum of squares) and S.P. (sum of products) of the N observations on the p variables $\bar{X}$. One can define q dummy variables $Y_1, \dots, Y_q$ forming the vector (column) Y , as $Y_\alpha = 1$ if an observation $\bar{X}$ comes from Π^α ($\alpha = 1, 2, \dots, q$) and $= 0$ otherwise. Then C_{XX} is the matrix of the corrected S.P. of $\bar{X}$ and C_{YY} is the corrected S.S. and

S.P. matrix for Y .

The dimensionality of the k means $\bar{\mu}^\alpha$ is the rank of the non-centrality matrix $(\bar{\mu}^\alpha) = \frac{1}{N} \sum_{\alpha=1}^k n_\alpha \bar{\mu}^\alpha$

$$U = \sum_{\alpha=1}^k n_\alpha (\bar{\mu}^\alpha - \bar{\mu})(\bar{\mu}^\alpha - \bar{\mu})', \quad (1)$$

and is estimated by s , the number of significant roots, out of the

roots $r_1^2 > r_2^2 > \dots > r_s^2$ ($r = \min(p, q)$) of the equation

$$(2) \quad | -r^2_2 (A + B) + B | = 0.$$

Then $\ell_{i,x}^1$ ($i = 1, 2, \dots, s$) are an adequate set of discriminant functions

(sample), where the vectors $\ell_{i,x}^1$ satisfy

$$(3) \quad [-r^2_1 (A + B) + B] \ell_{i,x}^1 = 0 \quad (i = 1, 2, \dots, s).$$

$\ell_{1,x}^1, \ell_{2,x}^1, \dots, \ell_{s,x}^1$ are the canonical variables corresponding to the

canonical correlations $r_1, r_2, \dots, r_s$ between $\bar{x}$ and $\bar{y}$. If $s = 1$,

the means are collinear and only are discriminant function $\ell_{1,x}^1$, corresponding to r_1 will be good enough. The remaining canonical variables

$\ell_{s+1,x}^1, \dots, \ell_{p,x}^1$ corresponding to the insignificant roots are called

"null functions". Canonical variables of the y -space, corresponding to

$r_1^2, \dots, r_s^2$ are called discriminant functions from the space of the dummy

variables and they determine contrasts among $\mu_1, \dots, \mu_k$, that are

significant.

In the case of $k = 2$ groups, Fisher derived an exact test of

goodness of fit of a hypothetical discriminant function. Williams [26]

attempted generalization of this test to the case of several groups.

If the hypothetical or assigned discriminant function is $\bar{a}'\bar{x}$, and

if it alone is adequate to discriminant, it means that the relation

of $\bar{x}$ with $\bar{y}$ is due to $\bar{a}'\bar{x}$ alone and when it is removed, no

residual relationship remains. Williams [26], therefore, defined

'residual' canonical correlations θ_i between $\bar{x}$ and $\bar{y}$, eliminating

$\bar{a}'\bar{x}$. Williams derived θ_i 's by considering the intersection of an

ellipsoid and a hyperplane and then derived their distribution and used

it to obtain an exact test of goodness of fit of $\bar{a}'\bar{x}$. Kshirsagar [4]

replaced the geometrical argument in Williams' derivation by an analytical

one. Williams had expressed $\bar{a}'\bar{x}$ as $w_1 x_1^* + \dots + w_p x_p^*$, where

* $x_1^*, \dots, x_p^*$ are true canonical variables. Williams incorrectly

assumed w_1 to be $NI(0,1)$. Kshirsagar gave their correct distribution.

Fortunately Williams' final results hold inspite of this. However,

Williams' test was obtained for $p = 2$ only. Bartlett [2] generalized

this to the case of p variables, used

$$(4) \quad V' = V \frac{\bar{a}'(A+B)\bar{a}}{\bar{a}'A\bar{a}},$$

where $V = |A|/|A+B|$ as the criterion. This V is Wilks' V to test

the independence of $\bar{x}$ and $\bar{y}$. V' is the residual V , when $\bar{a}'\bar{x}$ is

eliminated. Bartlett [2] gave an "exact" factorization $V' = V_1 V_2$ and

used V_1 to test the "Direction" part of the hypothesis that $\bar{a}'\bar{x}$, the

assigned function agrees in direction with the true discriminant function

$\bar{a}'\bar{x}$ and V_2 to test the "collinearity" (partial with respect to V_1)

part that only one discriminant function is necessary. Bartlett gave an

alternative factorization $V' = V_3 V_4$ also, where V_3 is the collinearity

factor and V_4 is the "partial" direction factor. Bartlett expressed

these factors in terms of canonical correlations and canonical variables,

while Williams [29] then expressed these in terms of a , A and B , and also

in terms of his residual roots θ_1 . Kshirsagar [12] gave an analysis of

variance type break up and expressed these factors in much simpler forms

and gave a completely analytical derivation of their distributions, by

employing matrix transformations. Recently Kshirsagar [21] has provided

an alternative method of derivation of these distributions, which enables

one to express each of these V_1, V_2, V_3, V_4 as $|P|/|P+Q|$, where P, Q are

independent Wishart matrices, showing thereby directly, their independence and their Wilks' Λ type distributions.

Williams [29] tried to extend this factorization to the goodness fit of a proposed "null function" also but Bartlett showed how it is incorrect.

Williams has now [32] generalized Bartlett's factorization to the case of goodness of fit of s ($s > 1$) hypothetical discriminant functions also and Kshirsagar [17] [21] has derived their distributions analytically. Williams [31] considered the goodness of fit test of a discriminant function from the dummy variables space also and used it to analyze Barnard's data on Egyptian skulls, in order to find out whether "time" alone can be a good discriminator. Kshirsagar [9] gave the direction and collinearity factors in such a situation in a more general case and he [20] has solved the distributional problem also, even though the dummy variables Y do not have a normal distribution.

The matrix L defined by $A = CLC'$, $A + B = CC'$ generalizes the concept of correlation coefficient between two variables x and y to two vectors $\bar{x}$ and $\bar{y}$. When the population means $\bar{\mu}_x$ are collinear, the matrix B has a non-central (linear) Wishart distribution.

Kshirsagar [3] [10], studied the effect of non-centrality on the Bartlett decomposition of a Wishart matrix and derived the distribution of this L also, which is a non-central matrix Beta distribution [6]. Its usefulness in discriminant analysis is also considered by him. He [18] has generalized Ruben's [25] result about a sample correlation coefficient to the correlation between two vectors also and discussed its relationship to Wilks' Λ , Hotelling's generalized T_0^2 and Pillai's criterion.

A method of obtaining confidence intervals for discriminant function coefficients has been given by Kshirsagar [11]. It is based on the direction factor λ_1 . Kshirsagar [14] has given an expository article on Wilks' λ , while Williams [33] has written an excellent review article about the analysis of association among many variates. Radcliffe [24] gives modifications to the Bartlett-correction factors to be attached to the direction and collinearity factors, in order to improve the χ^2 tests based on them.

In all this discussion, $\bar{x}$ is normal and y is dummy. In certain cases, both $\bar{x}$ and y are vectors of dummy variables. Such a situation arises in the analysis of a contingency table and then the canonical variables $\bar{y}'\bar{x}$ and $\bar{m}'\bar{y}$, corresponding to the significant canonical correlations provide optimum scores to be assigned to the categories of the two attributes of the contingency table. This quantifies the qualitative data. Bartlett [2] has illustrated this with Taylor's Blood serological data, showing how the reactions — , ? , + , (+) , w can be represented on a linear scale. Williams [28] has discussed analysis of a contingency table thoroughly and Kshirsagar [19] has provided a goodness of fit test for an assigned set of scores for the categories. Both Williams [27] and Kshirsagar [15] have investigated the applicability of direction and collinearity factors, in investigating factorial experiments where the interactions are multiplicative and not additive. Williams [30] has used this analysis in tests associated with concurrency and collinearity of several regression lines also.

The goodness of fit test of a single discriminant function was carried into the area of principal components analysis by Kshirsagar [7].

If λ has only one eigen value $\theta > 1$ and all other eigenvalues are $= 1$, the principal component corresponding to θ is called the single non-isotropic principal component. He derived an exact test for the goodness of fit of such a hypothetical principal component and then partitioned the χ^2 statistic into a direction χ^2 and a collinearity χ^2 . He expressed these in terms of "residual" roots similar to Williams' residual canonical correlations. Kshirsagar [13] has derived the exact distribution (null) of these residual roots and he [16] has also derived the non-null distribution of the direction χ^2 also, in such problems. Kshirsagar and Gupta [22] have extended this to more than one non-isotropic principal components too.

The usefulness of "null functions" defined earlier, in econometrics has been known since Bartlett [1] used it to derive structural equations. Kshirsagar [8] has shown their usefulness in prediction from Wold's implicit causal chain models.

Recently Kshirsagar and Young [23] have extended the factorization of Wilks' Λ to the Wilks' Λ associated with k complex p -variate normal distributions.

REFERENCES

- [1] Bartlett, M. S. (1948). A note on the Statistical estimation of demand and supply relations from time series, Econometrica, 16, 323-329.
- [2] Bartlett, M. S. (1951). The goodness of fit of a single hypothetical discriminant function in the case of several groups. Ann. Eugen., 16, 199-214.
- [3] Kshirsagar, A. M. (1959). Bartlett decomposition and Wishart distribution, Ann. Math. Stat., 30, 239-241.
- [4] Kshirsagar, A. M. (1960). A note on the derivation of some exact multivariate tests, Biometrika, 47, 480-482.
- [5] Kshirsagar, A. M. (1961). Some extensions of the multivariate t-distribution and the generalization of the distribution of regression coefficient, Proceedings of the Cambridge Philosophical Society, 57, 80-85.
- [6] Kshirsagar, A. M. (1961). The non-central multivariate Beta distribution, Ann. Math. Stat., 32, 104-111.
- [7] Kshirsagar, A. M. (1961). The goodness of fit of a single (non-isotropic) hypothetical principal component, Biometrika, 48, 397-407.
- [8] Kshirsagar, A. M. (1962). Prediction from the simultaneous equations system and Wold's implicit causal chain model, Econometrica, 30, 801-811.
- [9] Kshirsagar, A. M. (1962). A note on direction and collinearity factors in canonical analysis, Biometrika, 49, 255-259.

- [10] Kshirsagar, A. M. (1963). Effect of non-centrality on the Bartlett decomposition of a Wishart matrix, Annals of the Institute of Statistical Mathematics, Japan, 14, 217-228.
- [11] Kshirsagar, A. M. (1963). Confidence intervals for discriminant function coefficients, Journal of the Indian Statistical Association, 1, 1-8.
- [12] Kshirsagar, A. M. (1964). Distribution of the direction and collinearity factors in discriminant analysis, Proceedings of the Cambridge Philosophical Society, 60, 217-225.
- [13] Kshirsagar, A. M. (1964). Distribution of the residual roots in principal components analysis, Defence Science Journal, 14, 288-294.
- [14] Kshirsagar, A. M. (1964). Wilks' Λ , Journal of the Indian Statistical Association, 2, 1-20.
- [15] Kshirsagar, A. M. (1965). A note on the use of canonical analysis in factorial experiments, Journal of the Indian Statistical Association, 3, 165-169.
- [16] Kshirsagar, A. M. (1966). The non-null distribution of a direction statistic in principal components analysis, Biometrika, 53, 590.
- [17] Kshirsagar, A. M. (1969). Distributions associated with Wilks' Λ , Journal of the Australian Mathematical Society, 10, 269-278.
- [18] Kshirsagar, A. M. (1969). Correlation between two vector variables, Journal of Royal Statistical Society B, 31, 477-485.
- [19] Kshirsagar, A. M. (1970). Goodness of fit of an assigned set of scores for the analysis of association in a contingency table, Annals of the Inst. Stat. Maths., 22, 295-307.

- [20] Kshirsagar, A. M. (1971). Goodness of fit of a discriminant function from the vector space of dummy variables. To appear in Journal of Royal Statistical Society B
- [21] Kshirsagar, A. M. (1970). An alternative derivation of the direction and collinearity statistics in discriminant analysis, to appear in Cal. Stat. Assoc. Bull., 19, 123-134.
- [22] Kshirsagar, A. M. and Gupta, R. P. (1965). Goodness of fit of two or more hypothetical principal components Annals of the Institute of Statistical Mathematics, Japan, 17, 347-356.
- [23] Kshirsagar, A. M. and Young, John C. Discrimination among several complex multivariate normal populations. Sent for publication.
- [24] Radcliffe, J. (1966) Factorizations of the residual likelihood criterion in discriminant analysis Proceedings of the Cambridge Philosophical Society, 62, 743-752.
- [25] Ruben, H. (1966). Some new results on the distribution of the sample correlation coefficients Journal of Royal Statistical Society B, 28, 513-525.
- [26] Williams, E. J. (1952a). Some exact tests in multivariate analysis. Biometrika, 39, 17-31.
- [27] Williams, E. J. (1952b). The interpretation of interactions in factorial experiments. Biometrika, 39, 65-81.
- [28] Williams, E. J. (1952c). Use of scores for the analysis of association in contingency tables. Biometrika, 39, 274-289.
- [29] Williams, E. J. (1955). Significance tests for discriminant functions and linear functional relationships. Biometrika, 42, 360-381.

- [30] Williams, E. J. (1958). Simultaneous regression equations in experimen-
tation. Biometrika, 45, 96-110.
- [31] Williams, E. J. (1959). Regression Analysis. New York: Wiley.
- [32] Williams, E. J. (1961). Tests for discriminant functions. Journal of
Australian Mathematical Society, 2, 243-252.
- [33] Williams, E. J. (1967). The analysis of association among many variates.
Journal of Royal Statistical Society B, 29, 199-228.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)

2a. REPORT SECURITY CLASSIFICATION

UNCLASSIFIED

2b. GROUP

UNCLASSIFIED

SOUTHERN METHODIST UNIVERSITY

3. REPORT TITLE

"Review of the Williams - Bartlett - Kshirsagar Work on Direction and Collinearity Factors in Discriminant Analysis."

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

A. M. Kshirsagar

6. REPORT DATE

May 1971

8a. CONTRACT OR GRANT NO.

N00014-68-A-0515

b. PROJECT NO.

NR 042-260

c.

d.

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

OFFICE OF NAVAL RESEARCH

13. ABSTRACT

In discriminant analysis, the problem of examining the dimensionality of the configuration of the means of various graphs and of testing the adequacy of the directions and number of assigned discriminant functions are of some importance. A great deal of work is done in this and related areas (principal components and assignment of scores to categories) as done by Williams, Bartlett, Kshirsagar, and others. The whole area is reviewed in this paper.