Canonical Correlation and Its Relationship to Discriminant Analysis and Multiple Regression

HARRY R. GLAHN[1]

Weather Bureau, ESSA, Silver Spring, Md.

(Manuscript received 26 June 1967)

ABSTRACT

Canonical correlation analysis is concerned with the determination of a linear combination of each of two sets of variables such that the correlation between the two functions is a maximum. Under certain conditions this analysis is equivalent to discriminant analysis and under other conditions it is equivalent to multiple regression. In this paper the relationships among these techniques are discussed, equations relating to prediction by canonical variates are derived, a generalized correlation coefficient is proposed, and an example of canonical correlation analysis is presented.

1. Introduction

Multiple regression has been used to relate a dependent variable (predictand) to a set of independent (predictor) variables for many years. Twenty years ago Hotelling (1936) introduced the concepts of canonical correlation and canonical variates for the analysis of relationships between two sets of variables. At about the same time Barnard (1935) and Fisher (1936) proposed discriminant analysis as a means of using one set of variables to discriminate between two categories of another variable; later, this analysis was extended to more than two predictand groups by Brown (1947), Bryan (1950) and others. Miller (1964) generalized to several predictand categories a specialized application of regression which had been used for two predictand categories by Mook[2] and Lund (1955).

It is not generally recognized by meteorologists that all of the above statistical techniques are embodied in Hotelling's canonical correlation analysis. Proofs of this exist but not in the meteorological literature. In this paper the relationships among these techniques are discussed, equations relating to prediction by canonical variates are derived, a generalized correlation coefficient is proposed, and an example of canonical correlation analysis is presented.

2. Canonical variates relationships

Suppose that there exist n observations of each of p variables X_i $(i=1, 2, \cdots, p)$ and of q variables Y_i $(i=1, 2, \cdots, q)$. These observations represent points in a $(p+q)$ dimensional space and can be arranged in the matrices ${}_n\mathbf{X}_p$ and ${}_n\mathbf{Y}_q$. The variables have means $\bar{X}_i$ and $\bar{Y}_i$, respectively, and deviations from the mean are given by $x_i=X_i-\bar{X}_i$ and $y_i=Y_i-\bar{Y}_i$. New variables ${}_n\mathbf{x}_p\mathbf{A}_i$ and ${}_n\mathbf{y}_q\mathbf{B}_i$ $(i=1, 2, \cdots, r)$, where r is $\leq$ the smaller of p and q, can be formed such that their means are zero and

$$ {}_r\mathbf{A}_p'\mathbf{x}_n'\mathbf{x}_p\mathbf{A}_r=n_r\mathbf{I}_r, \tag{1} $$

$$ {}_r\mathbf{B}_q'\mathbf{y}_n'\mathbf{y}_q\mathbf{B}_r=n_r\mathbf{I}_r, \tag{2} $$

$$ {}_r\mathbf{A}_p'\mathbf{x}_n'\mathbf{y}_q\mathbf{B}_r=n_r\mathbf{A}_r, \tag{3} $$

where $\mathbf{I}$ is the unit matrix,

$$ {}_r\mathbf{A}_r=\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & 0 & \\ & 0 & \ddots & \\ & & & \lambda_r\end{bmatrix}, \tag{4} $$

and $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_r$.

Eqs. (1) and (2) state that the variance of each of the new variables is unity and each is uncorrelated with all others in its respective set. Eqs. (3) and (4), together with (1) and (2), state that each ${}_n\mathbf{x}_p\mathbf{A}_i$ is uncorrelated with each ${}_n\mathbf{y}_q\mathbf{B}_j$ except when $i=j$ and then the correlation is λ_i.

It can be shown [for instance, see Anderson (1958)] that the ${}_p\mathbf{A}_i$ $(i=1, 2, \cdots, r)$ can be found from

$$ ({}_p\mathbf{S}_{11p}^{-1}\mathbf{S}_{12q}\mathbf{S}_{22q}^{-1}\mathbf{S}_{21p}-\lambda_i^2{}_p\mathbf{I}_p){}_p\mathbf{A}_i=\mathbf{0}, \tag{5} $$

(providing ${}_p\mathbf{S}_{11p}$ and ${}_q\mathbf{S}_{22q}$ are not singular), where the λ_i satisfy the determinantal equation

$$ |{}_p\mathbf{S}_{11p}^{-1}\mathbf{S}_{12q}\mathbf{S}_{22q}^{-1}\mathbf{S}_{21p}-\lambda^2{}_p\mathbf{I}_p|=0, \tag{6} $$

[1] A major portion of this work was accomplished while the author, employed by the Techniques Development Laboratory of the Weather Bureau, was on Reserve training with the Computer Application Division, U. S. Air Force.

[2] Mook, C. P., 1948: An objective method of forecasting thunderstorms for Washington, D. C., in May. Unpublished manuscript. (Copy in Atmospheric Sciences Library, ESSA, Washington, D. C.)

and where

$$ {}_p\mathbf{S}_{11p}=\frac{1}{n}{}_p\mathbf{x}_n'\mathbf{x}_p, \tag{7} $$

$$ {}_p\mathbf{S}_{12q}={}_p\mathbf{S}_{21q}'=\frac{1}{n}{}_p\mathbf{x}_n'\mathbf{y}_q, \tag{8} $$

$$ {}_q\mathbf{S}_{22q}=\frac{1}{n}{}_q\mathbf{y}_n'\mathbf{y}_q, \tag{9} $$

are the variance-covariance matrices. Then the ${}_q\mathbf{B}_i$ can be found from

$$ {}_q\mathbf{B}_r={}_q\mathbf{S}_{22q}^{-1}\mathbf{S}_{21p}\mathbf{A}_r\mathbf{\Lambda}_r^{-1}. \tag{10} $$

Alternatively, use can be made of

$$ ({}_q\mathbf{S}_{22q}^{-1}\mathbf{S}_{21p}\mathbf{S}_{11p}^{-1}\mathbf{S}_{12q}-\lambda_i^2{}_q\mathbf{I}_q){}_q\mathbf{B}_i=\mathbf{0}, \tag{11} $$

$$ |{}_q\mathbf{S}_{22q}^{-1}\mathbf{S}_{21p}\mathbf{S}_{11p}^{-1}\mathbf{S}_{12q}-\lambda^2{}_q\mathbf{I}_q|=0, \tag{12} $$

$$ {}_p\mathbf{A}_r={}_p\mathbf{S}_{11p}^{-1}\mathbf{S}_{12q}\mathbf{B}_r\mathbf{\Lambda}_r^{-1}. \tag{13} $$

The latter equations are to be preferred if $q<p$ because the matrix which must be diagonalized is then of a lesser dimension.

The "first" pair of functions, defined by the first column of each of ${}_p\mathbf{A}_r$ and ${}_q\mathbf{B}_r$, have as large a correlation λ_1 as any other possible pair of functions, each composed of a linear combination of the original variables. Also, the "second" function pair have as large a correlation λ_2 as any other possible pair of functions, each being composed of a linear combination of the original variables *and* each being uncorrelated with both members of the first pair.

Either set of new variables can be predicted in a least-squares sense by the new variables in the other set. The prediction equations are

$$ \widehat{{}_n\mathbf{y}_q\mathbf{B}_r}={}_n\mathbf{x}_p\mathbf{A}_r\mathbf{\Lambda}_r, \tag{14} $$

and

$$ \widehat{{}_n\mathbf{x}_p\mathbf{A}_r}={}_n\mathbf{y}_q\mathbf{B}_r\mathbf{\Lambda}_r. \tag{15} $$

Also, the original variables in one set can be predicted in a least-squares sense by the new variables in the other set[3] by

$$ {}_n\hat{\mathbf{y}}_q={}_n\mathbf{x}_p\mathbf{A}_r\mathbf{\Lambda}_r\mathbf{B}_q'(\mathbf{B}_r\mathbf{B}_q')^{-1}={}_n\mathbf{x}_p\mathbf{A}_r\mathbf{\Lambda}_r\mathbf{B}_q'\mathbf{S}_{22q}. \tag{16} $$

In the case that $r=q$, (16) can be written as

$$ {}_n\hat{\mathbf{y}}_q={}_n\mathbf{x}_p\mathbf{A}_q\mathbf{\Lambda}_q\mathbf{B}_q^{-1}. \tag{17} $$

Eq. (16) contains the prediction equation for each of the y_i in terms of all of the x_i. One may want to relate one set of variables to the other set but involve only a portion of the correlations λ_i, perhaps those k correlations that are judged to be significantly different from zero. An equation corresponding to Eq. (16) can be written as

$$ {}_n\hat{\hat{\mathbf{y}}}_q={}_n\mathbf{x}_p\mathbf{A}_r\underline{\mathbf{\Lambda}}_r\mathbf{B}_q'\mathbf{S}_{22q}, \tag{18} $$

where ${}_r\underline{\mathbf{\Lambda}}_r$ has only k non-zero elements, the others having been set to zero. Eq. (18) has the effect of including a contribution from only those k columns of ${}_p\mathbf{A}_r$ and k rows of ${}_r\mathbf{B}_q'$ corresponding to the k non-zero correlations.

An error matrix ${}_n\mathbf{\varepsilon}_q$ can be defined as

$$ {}_n\mathbf{\varepsilon}_q={}_n\mathbf{y}_q-{}_n\hat{\mathbf{y}}_q. \tag{19} $$

Then the total variance of each variable y_i is given by the corresponding diagonal element of

$$ \frac{1}{n}({}_q\mathbf{y}_n'\mathbf{y}_q)=\frac{1}{n}({}_q\hat{\mathbf{y}}_n'+{}_q\mathbf{\varepsilon}_n')({}_n\mathbf{y}_q+{}_n\mathbf{\varepsilon}_q). \tag{20} $$

Substituting from Eqs. (16) and (1), recognizing that the predictors are uncorrelated with the errors, and simplifying, yields

$$ \frac{1}{n}({}_q\mathbf{y}_n'\mathbf{y}_q)={}_q\mathbf{S}_{22q}\mathbf{B}_r\mathbf{\Lambda}_r^2\mathbf{B}_q'\mathbf{S}_{22q}+\frac{1}{n}({}_q\mathbf{\varepsilon}_n'\mathbf{\varepsilon}_q). \tag{21} $$

In the event that $r=q$, (21) can be written as

$$ \frac{1}{n}({}_q\mathbf{y}_n'\mathbf{y}_q)={}_q(\mathbf{B}')_q^{-1}\mathbf{\Lambda}_q^2\mathbf{B}_q^{-1}+\frac{1}{n}({}_q\mathbf{\varepsilon}_n'\mathbf{\varepsilon}_q). \tag{22} $$

Each ith diagonal element of the first term on the right is the amount of variance of the corresponding y_i explained by the predictors and each ith diagonal element of the last term is the amount of variance of the corresponding y_i unexplained by the predictors. Division of a diagonal element of the first term on the right by the corresponding element of the term on the left gives the fraction of variance of that y_i which is explained. These q values are the diagonal elements of

$$ {}_q\mathbf{R}_q={}_q\mathbf{\sigma}_{22q}^{-1}\mathbf{S}_{22q}\mathbf{B}_r\mathbf{\Lambda}_r^2\mathbf{B}_q'\mathbf{S}_{22q}\mathbf{\sigma}_{22q}^{-1}, \tag{23} $$

where ${}_q\mathbf{\sigma}_{22q}$ is a diagonal matrix composed of the corresponding positive square roots of the diagonal elements of ${}_q\mathbf{S}_{22q}$.

The total explained variance (EV) and the fractional part $R_{y\cdot x}^2$ of the total variance (TV) explained are obtained by taking the trace (tr) as follows:

$$ \mathrm{EV}=\mathrm{tr}({}_q\mathbf{S}_{22q}\mathbf{B}_r\mathbf{\Lambda}_r^2\mathbf{B}_q'\mathbf{S}_{22q}), \tag{24} $$

$$ R_{y\cdot x}^2=\frac{\mathrm{tr}({}_q\mathbf{S}_{22q}\mathbf{B}_r\mathbf{\Lambda}_r^2\mathbf{B}_q'\mathbf{S}_{22q})}{\mathrm{tr}({}_q\mathbf{S}_{22q})}=\frac{\mathrm{tr}({}_q\mathbf{\sigma}_{22q}\mathbf{R}_q\mathbf{\sigma}_{22q})}{\mathrm{tr}({}_q\mathbf{S}_{22q})}. \tag{25} $$

Eqs. (23), (24), and (25) can be evaluated as a sum of r terms. For instance, Eq. (24) becomes

[3] Hereafter in this paper only the equations which arise from considering the y_i to be the predictand set are presented.

$$\mathrm{EV}=\mathrm{tr}\left({}_{q}\mathbf{S}_{22q}\mathbf{B}_{r}\begin{bmatrix}\lambda_1^2 & & & \\ & 0 & & 0 \\ & & 0 & \\ & 0 & & \ddots \\ & & & & 0\end{bmatrix}{}_{r}\mathbf{B}_{q}'\mathbf{S}_{22q}\right)$$

$$+\mathrm{tr}\left({}_{q}\mathbf{S}_{22q}\mathbf{B}_{r}\begin{bmatrix}0 & & & \\ & \lambda_2^2 & & 0 \\ & & 0 & \\ & 0 & & \ddots \\ & & & & 0\end{bmatrix}{}_{r}\mathbf{B}_{q}'\mathbf{S}_{22q}\right)+\cdots$$

$$+\mathrm{tr}\left({}_{q}\mathbf{S}_{22q}\mathbf{B}_{r}\begin{bmatrix}0 & & & \\ & 0 & & 0 \\ & & \ddots & \\ & 0 & & 0 \\ & & & & \lambda_r^2\end{bmatrix}{}_{r}\mathbf{B}_{q}'\mathbf{S}_{22q}\right). \quad (26)$$

The ith diagonal element in the jth term on the right before taking the trace, is the amount of variance of the corresponding y_i explained by the jth canonical function. Also, the jth trace on the right gives the total amount of variance of the y_i $(i=1, 2, \cdots, q)$ explained by the jth canonical function.

If Eq. (23) is expanded in a similar manner, the ith diagonal element in the jth term is the fractional amount of variance of the corresponding y_i explained by the jth canonical function. The jth term in the similar expansion of the right side of Eq. (25) is the fractional amount of the total variance of the y_i $(i=1, 2, \cdots, q)$ explained by the jth canonical function. It should be noted that ${}_{q}\mathbf{R}_{q}$ is invariant under linear transformations of scale of the predictors and predictands, and that EV and $R_{y\cdot x}^2$ are invariant under linear transformations of scale of the predictors but not of the predictands.

The prediction equations can be put in terms of the X_i and Y_i, if desired. For instance Eq. (17) becomes

$${}_{n}\hat{\mathbf{Y}}_{q}={}_{n}\mathbf{X}_{p}\mathbf{A}_{q}\boldsymbol{\Lambda}_{q}\mathbf{B}_{q}^{-1}-{}_{n}\bar{\mathbf{X}}_{p}\mathbf{A}_{q}\boldsymbol{\Lambda}_{q}\mathbf{B}_{q}^{-1}+{}_{n}\bar{\mathbf{Y}}_{q}. \quad (27)$$

3. Discriminant analysis formulation

Suppose that there exist n observations of each of p variables X_i $(i=1, 2, \cdots, p)$, and that these observations can be divided into G groups of sizes $n_1, n_2, \cdots, n_G$. The observations can be considered to represent points in a p-dimensional space where each is tagged with its respective group number. Also, the observations can be arranged to form the matrix ${}_{n}\mathbf{X}_{p}$. The p overall means are $\bar{X}_i$ and the deviations are given by $x_i=X_i-\bar{X}_i$. The individual group means of these deviations are given by $\bar{x}_{ij}$ $(i=1, 2, \cdots, p;\ j=1, 2, \cdots, G)$. New variables ${}_{n}\mathbf{x}_{p}\mathbf{V}_i$, $(i=1, 2, \cdots, r)$, where $r \leq$ the smaller of p and $G-1$, can be formed such that the values corresponding to a particular group tend to cluster together about a mean and also that that mean tends to be separated from other group means. The ${}_{p}\mathbf{V}_i$ are called discriminant functions.

Within groups, between groups, and total sum of squares matrices can be defined, respectively, as

$${}_{p}\mathbf{W}_{p}=({}_{p}\mathbf{x}_{n}'-{}_{p}\bar{\mathbf{x}}_{n}')({}_{n}\mathbf{x}_{p}-{}_{n}\bar{\mathbf{x}}_{p}), \quad (28)$$

$${}_{p}\mathbf{B}_{p}={}_{p}\bar{\mathbf{x}}_{n}'\bar{\mathbf{x}}_{p}, \quad (29)$$

$${}_{p}\mathbf{T}_{p}={}_{p}\mathbf{x}_{n}'\mathbf{x}_{p}={}_{p}\mathbf{W}_{p}+{}_{p}\mathbf{B}_{p}, \quad (30)$$

where the matrix ${}_{n}\bar{\mathbf{x}}_{p}$ is composed of the group means in the order that the groups are represented in ${}_{n}\mathbf{x}_{p}$.

The discriminant functions can be found by solving [see Bryan (1950)]

$$({}_{p}\mathbf{W}_{p}^{-1}\mathbf{B}_{p}-\mu_{ip}\mathbf{I}_{p})\,{}_{p}\mathbf{V}_i=\mathbf{0}, \quad (31)$$

in which the μ_i are solutions of

$$|\,{}_{p}\mathbf{W}_{p}^{-1}\mathbf{B}_{p}-\mu_{p}\mathbf{I}_{p}|=0. \quad (32)$$

Each μ_i is interpreted as the ratio of between-to-within-groups variance of the X_j (predictors) due to the ith discriminant function.

Now suppose $G-1\equiv q$ dummy variables Y_i $(i=1, 2, \cdots, q)$ are defined such that $Y_{ij}=1$ if the jth observation belongs to group i and $Y_{ij}=0$ if the jth observation does not belong to group i. A dummy variable corresponding to the Gth group is not defined since it would be redundant with the other $G-1$ groups and ${}_{G}\mathbf{S}_{22G}$ would be singular. Deviations from the mean are given by $y_{ij}=Y_{ij}-\bar{Y}_{ij}$ and can be put into the matrix ${}_{n}\mathbf{y}_{q}$. This predictand matrix and the predictor matrix ${}_{n}\mathbf{x}_{p}$ can now be used in the canonical correlation framework and all of the equations in Section 2 apply.

Tatsuoka (1955) and others have shown that the discriminant analysis solution [Eqs. (31) and (32)] is equivalent to the canonical correlation solution [Eqs. (5) and (6)], where

$$\mu_i=\frac{\lambda_i^2}{1-\lambda_i^2}, \quad (33)$$

and

$${}_{p}\mathbf{V}_i=C\,{}_{p}\mathbf{A}_i, \quad (34)$$

where C is an arbitrary constant which is necessary because if ${}_{p}\mathbf{V}_i$ is a solution of Eq. (31), $C\,{}_{p}\mathbf{V}_i$ is also, and no restriction was imposed on the variance of the discriminant functions. However, the discriminant analysis solution as usually defined stops with the computation of the ${}_{p}\mathbf{V}_i$ and μ_i and some other technique must be employed to find the actual estimates of group membership [for instance, see Miller (1962)], whereas Eq. (27) can be used to produce estimates of the $G-1$ binary predictands directly.

After estimates $\hat{Y}_{ij}$ $(i=1, 2, \cdots, G-1)$ are made, the estimate for the Gth group is

$$\hat{Y}_{Gj}=1-\sum_{i=1}^{G-1}\hat{Y}_{ij}. \quad (35)$$

Predictand group membership is designated by a "one" and non-membership is indicated by a "zero." Therefore, group estimates near "one" or estimates large with respect to the other estimates would indicate membership in that group.

Usually, $G-1=q$ is much less than p and it is more efficient to solve Eqs. (11), (12) and (13) rather than than Eqs. (5), (6) and (10) or Eqs. (31) and (32) unless special methods of solution are devised.

4. Multiple linear regression

Multiple linear regression is concerned with the estimation of one predictand by a linear combination of predictors and is a special case of canonical correlation. If $q=1$, the equations in Section 2 hold and are much simplified. In particular, Eq. (16), put in terms of the original variables X_i and Y, is the usual regression equation,

$$ {}_n\hat{\mathbf{Y}}_1={}_n\mathbf{X}_p\mathbf{S}_{11p}{}^{-1}\mathbf{S}_{12_1}-{}_n\bar{\mathbf{X}}_p\mathbf{S}_{11p}{}^{-1}\mathbf{S}_{12_1}+{}_n\bar{\mathbf{Y}}_1. \quad (36) $$

The regression estimates can be found separately for any number of predictands even though the predictands are linearly related. If G dummy variables denoting group membership are formed as discussed in Section 3 and used as predictands, the resulting regression equations are identical to those derived from the canonical correlation analysis and can be written

$$ {}_n\hat{\mathbf{Y}}_G={}_n\mathbf{X}_p\mathbf{S}_{11p}{}^{-1}\mathbf{S}_{12_G}-{}_n\bar{\mathbf{X}}_p\mathbf{S}_{11p}{}^{-1}\mathbf{S}_{12_G}+{}_n\bar{\mathbf{Y}}_G. \quad (37) $$

It may be sufficiently accurate for some purposes to consider the estimates $\hat{Y}_i$ $(i=1, 2, \cdots, G)$ to be the mean values of Y_i for each observable combination of predictor values. In this way the concept of $\hat{Y}_i$ being an unbiased estimate of the probability of Y_i is introduced. Also, for any observable combination of predictor values the sum of the probability estimates is

$$ \sum_{i=1}^{G}\hat{Y}_{ij}=\sum_{i=1}^{G}({}_1\mathbf{X}_p-{}_1\bar{\mathbf{X}}_p)_p\mathbf{S}_{11p}{}^{-1}\begin{bmatrix}\frac{1}{n}\sum_{j=1}^{n}x_{1j}y_{ij}\\ \frac{1}{n}\sum_{j=1}^{n}x_{2j}y_{ij}\\ \vdots\\ \frac{1}{n}\sum_{j=1}^{n}x_{pj}y_{ij}\end{bmatrix}+\sum_{i=1}^{G}\bar{Y}_{i\cdot} $$

$$ =\frac{1}{n}({}_1\mathbf{X}_p-{}_1\bar{\mathbf{X}}_p)_p\mathbf{S}_{11p}{}^{-1}\begin{bmatrix}\sum_{j=1}^{n}x_{1j}\sum_{i=1}^{G}y_{ij}\\ \sum_{j=1}^{n}x_{2j}\sum_{i=1}^{G}y_{ij}\\ \vdots\\ \sum_{j=1}^{n}x_{pj}\sum_{i=1}^{G}y_{ij}\end{bmatrix}+\sum_{i=1}^{G}\bar{Y}_{i\cdot} \quad (38) $$

Since

$$ \sum_{i=1}^{G}y_{ij}=\sum_{i=1}^{G}(Y_{ij}-\bar{Y}_{i\cdot})=\sum_{i=1}^{G}Y_{ij}-\sum_{i=1}^{G}\bar{Y}_{i\cdot}, \quad (39) $$

and

$$ \sum_{i=1}^{G}Y_{ij}=\sum_{i=1}^{G}\bar{Y}_{i\cdot}=1, \quad (40) $$

it is seen that the sum over all G groups of the coefficients of each predictor equals zero and that the sum of the probability estimates equals unity. However, this does not guarantee that the individual estimates $\hat{Y}_{ij}$ are bounded by zero and one [see Miller (1964)].

The fact that $\sum_{i=1}^{G}\hat{Y}_{ij}=1$ justifies Eq. (35). It is also interesting to note that the criterion for the determination of the regression equations, i.e.,

$$ \sum_{j=1}^{n}(Y_{ij}-\hat{Y}_{ij})^2 \text{ is a minimum} \quad (41) $$

also assures the minimum (best) P-Score [see Brier (1950)] on the dependent data obtainable by linear prediction equations and the predictors being used.

5. Measures of association between two sets of variables

Hooper (1959) discusses three measures of association between two sets of variables—Hotelling's (1936) vector alienation and vector-correlation coefficients and his proposed trace correlation. The following definitions can be made:

$$ \text{Vector correlation coefficient}=\left[\prod_{i=1}^{q}\lambda_i{}^2\right]^{\frac{1}{2}}, \quad (42) $$

$$ \text{Vector alienation coefficient}=\left[\prod_{i=1}^{q}(1-\lambda_i{}^2)\right]^{\frac{1}{2}}, \quad (43) $$

$$ \text{Trace correlation coefficient}\equiv\bar{r}=\left[\frac{1}{q}\sum_{i=1}^{q}\lambda_i{}^2\right]^{\frac{1}{2}}=\left[\frac{1}{q}\operatorname{tr}_q\Lambda_q{}^2\right]^{\frac{1}{2}}. \quad (44) $$

Hooper (1959) notes that the vector correlation coefficient has the undesirable property of being zero if there are not q non-zero canonical correlations. Also the vector correlation and vector alienation coefficients both tend to zero when q is large. He states that the trace correlation coefficient has "... none of these defects ..." and upon comparing

$$ \bar{r}^2=\frac{1}{q}\sum_{i=1}^{q}\lambda_i{}^2 \quad (45) $$

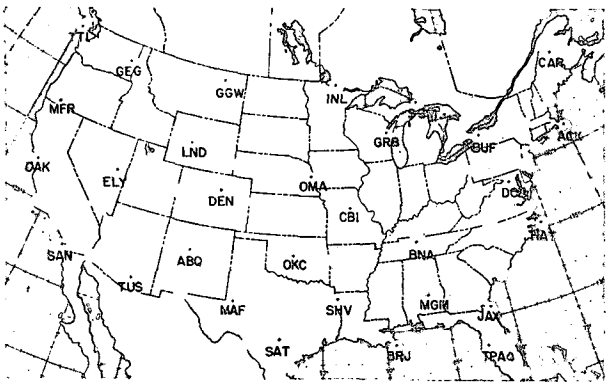

FIG. 1. The 30 stations used in the example.

with

$$1-\bar{r}^2=-\frac{1}{q}\sum_{i=1}^{q}(1-\lambda_i^2), \tag{46}$$

concludes that "... $\bar{r}^2$ can be naturally interpreted as that part of the total variance of the jointly dependent variables that is accounted for by the systematic part of the reduced form, and $1-\bar{r}^2$ as the unexplained part ..." However, it has been shown in this paper that Eq. (25) gives the fractional part of the total variance of the dependent variables accounted for by the predictors. It is proposed, therefore, that if a single measure of association is desired, $R_{y\cdot x}{}^2$ defined in Eq. (25) is probably as good as any other even though it is no more nor less than

$$R_{y\cdot x}{}^2=\frac{\sum_{i=1}^{q}\sigma_i^2 R^2_{y_i\cdot x_1\cdot x_2,\cdots,x_p}}{\sum_{i=1}^{q}\sigma_i^2}, \tag{47}$$

TABLE 1. The canonical correlations λ_i, canonical correlations squared λ_i^2, and fraction of total variance explained, EV_i/TV, by the most important 15 of the 30 canonical functions.

No. of function i	Canonical correlation λ_i	λ_i^2	EV_i/TV
1	0.94	0.88	0.231
2	0.93	0.86	0.173
3	0.91	0.82	0.097
4	0.90	0.81	0.094
5	0.87	0.76	0.066
6	0.82	0.67	0.023
7	0.79	0.62	0.020
8	0.76	0.58	0.012
9	0.72	0.51	0.013
10	0.67	0.44	0.004
11	0.64	0.41	0.005
12	0.59	0.34	0.002
13	0.55	0.30	0.002
14	0.49	0.24	0.003
15	0.46	0.21	0.001
$\sum_{i=1}^{6}$			0.683
$\sum_{i=1}^{30}$			0.749

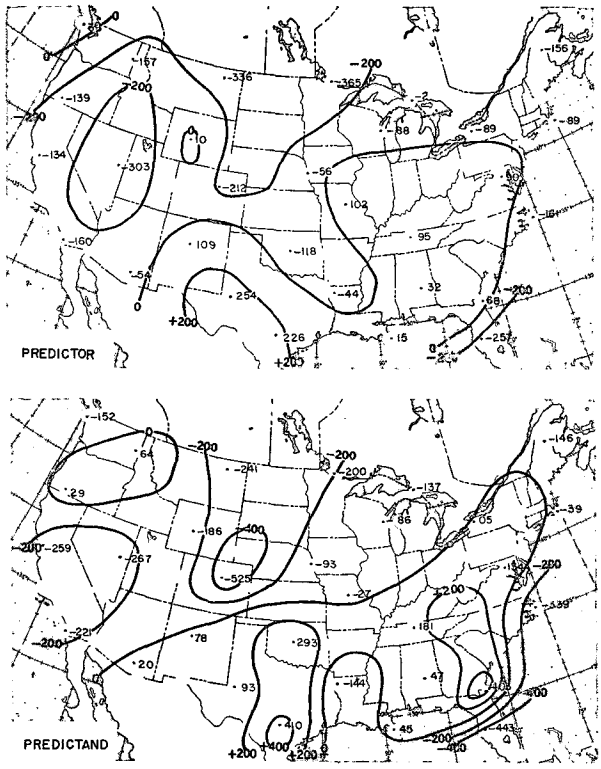

FIG. 2. The predictor and predictand functions $\times 10^5$ corresponding to λ_1.

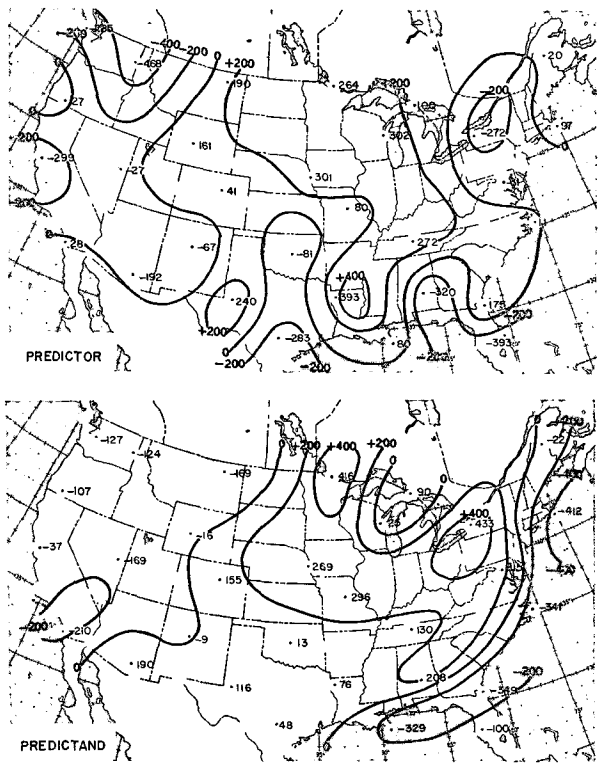

FIG. 3. The predictor and predictand functions $\times 10^5$ corresponding to λ_2.

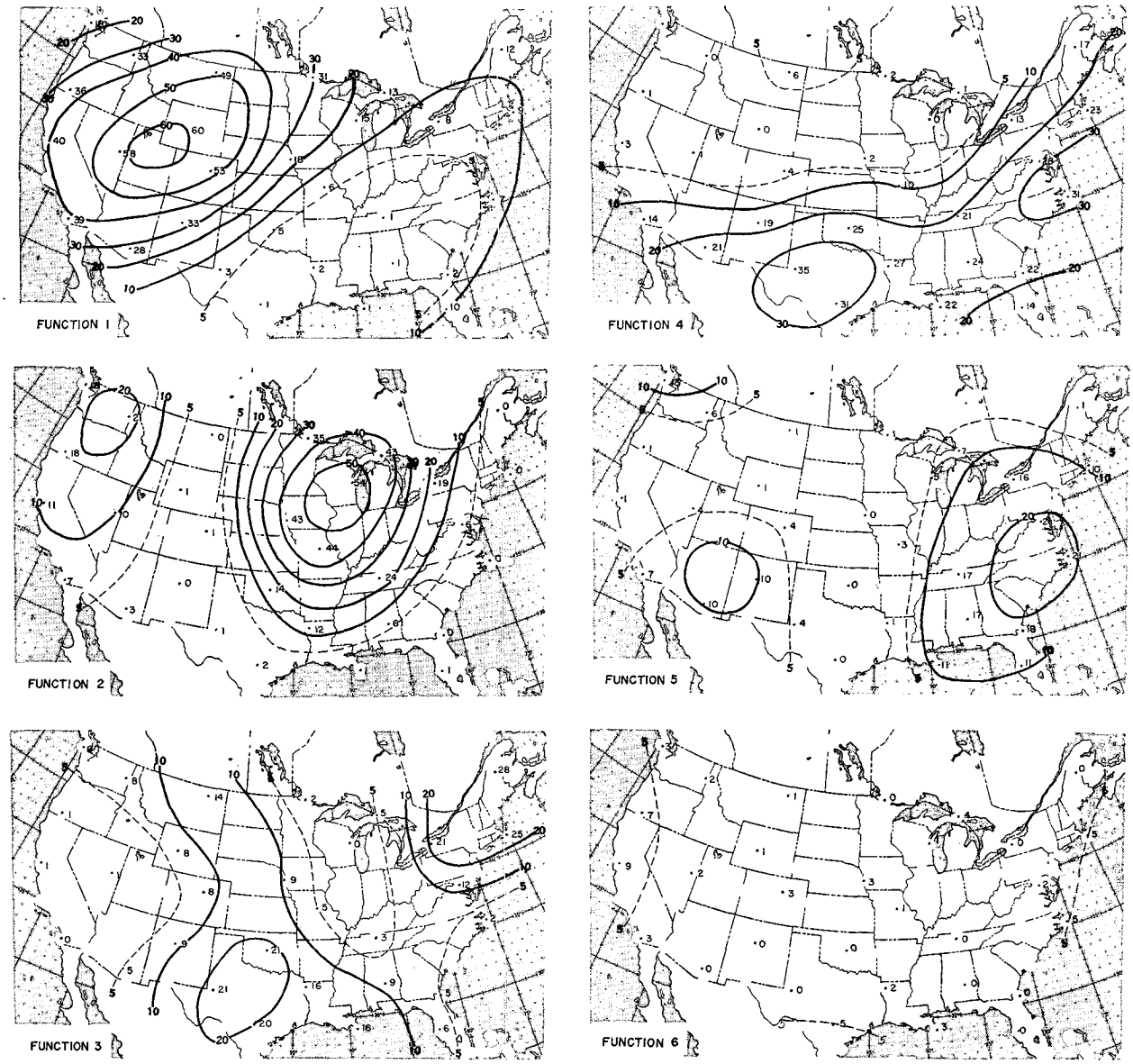

FIG. 4. The per cent of the variance of each predictand explained by each of the first 6 predictor functions.

where σ_i^2 is the variance of the ith predictand and $R^2_{y_i \cdot x_1 \cdot x_2, \cdots, x_p}$ is the reduction of variance of the ith predictand found by the usual multiple regression technique. (These reductions of variance and σ_i are also the diagonal elements of ${}_q\mathbf{R}_q$ and ${}_q\boldsymbol{\sigma}_{22q}$, respectively.) A suitable name for $R_{y \cdot x}$, the positive square root of $R_{y \cdot x}^2$, would be the "composite correlation coefficient."

As stated previously, $R_{y \cdot x}^2$ is not invariant to scale transformations of the predictands (unless the same linear transformation is imposed on each); however, this is not necessarily undesirable. If there is only a single predictand ($q=1$), then $R_{y \cdot x}^2$ is the square of the multiple correlation of that predictand with the p predictors.

6. An example

As an example of the canonical correlation technique, 5 years of 500-mb heights observed at 30 stations in the United States at 0000 GMT in June, July and August were related to the same variables 24 hr earlier. The stations used are shown in Fig. 1. The sample size was 455. The largest 15 of the resulting 30 canonical correlations are shown in Table 1.

The predictor and predictand functions corresponding to λ_1, and λ_2 are shown in Figs. 2 and 3; the coefficients are plotted at the station locations. These functions do not seem to correspond to any easily recognizable synoptic pattern, nor given the predictor functions

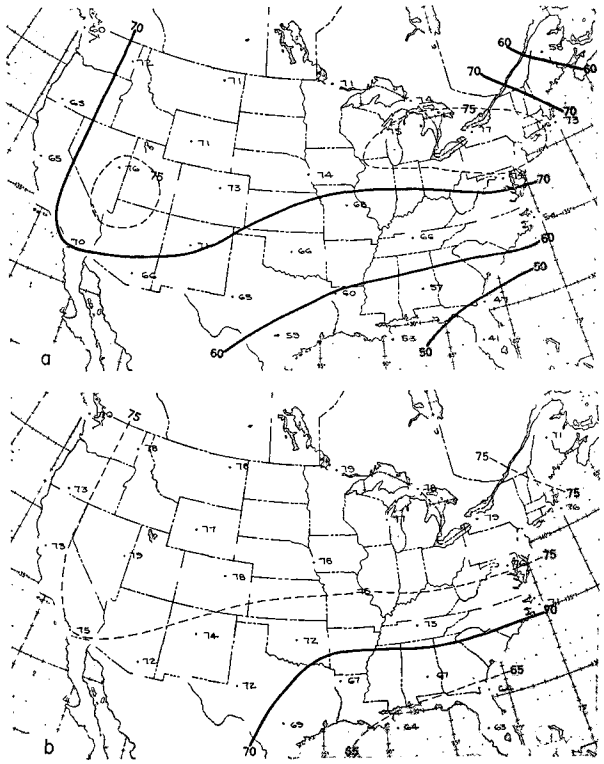

FIG. 5. The per cent of the variance of each predictand explained by the first 6 functions, a., and all 30 functions, b.

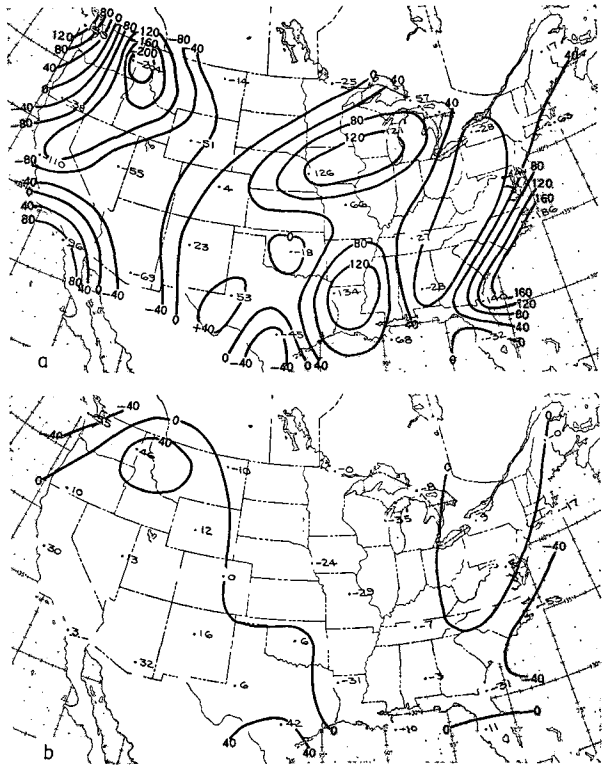

FIG. 6. The coefficients $\times 10^3$ in the regression equations for predicting the 500-mb height at Nashville, a., and Lander, b., derived from functions 2, 4 and 5.

would one readily expect the corresponding predictand functions.

The fractional part of the variance of each predictand explained by each of the first 6 predictor functions is shown in Fig. 4. These maps have well-defined patterns and it can be seen that a particular predictor function, even that corresponding to λ_1, is nearly useless for some predictands. The fractional part of the variance explained by the first 6 and all 30 predictor functions is depicted in Fig. 5. Fig. 5 shows that the first 6 functions explain most of the variance of which all 30 are capable.

In general, the heights at stations which have no "upstream" stations in the sample are less predictable than the others. Also, due to their less organized behavior, heights at southern stations are less predictable in terms of reduction of variance than those at northern stations. Finally, heights at those stations having few close neighbors are difficult to predict.

Figs. 4 and 5 show that functions 2, 4 and 5 together explain 62% of the variance of height at Nashville 24 hr later, while the other 27 functions increase this explained variance to only 73%. These same three functions explain only 2% of the variance of height at Lander. The coefficients in the regression equations derived from these functions for the predictands Nashville and Lander are shown in Fig. 6. The coefficients do not form a very smooth pattern although the larger positive values do tend to be near and slightly to the west of the predictand stations as would be expected.

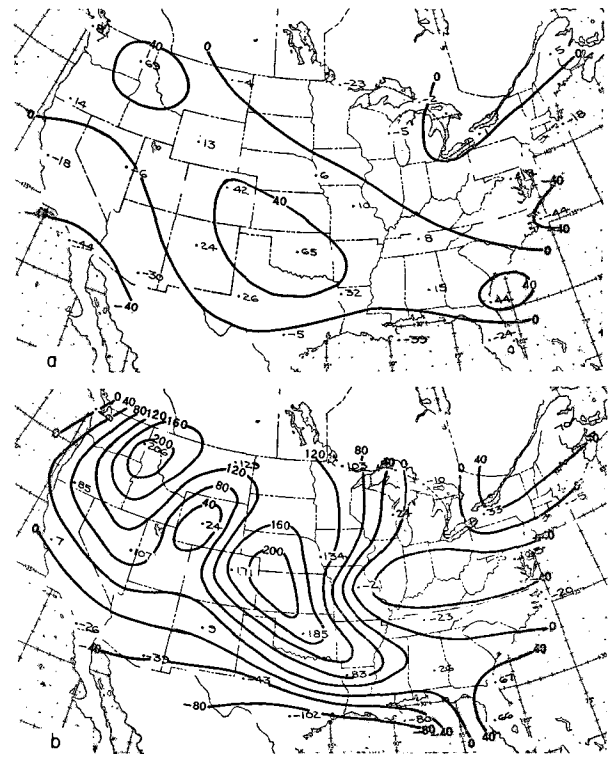

FIG. 7. The coefficients $\times 10^3$ in the regression equations for predicting the 500-mb height at Nashville, a., and Lander, b., derived from functions 1, 3 and 7.

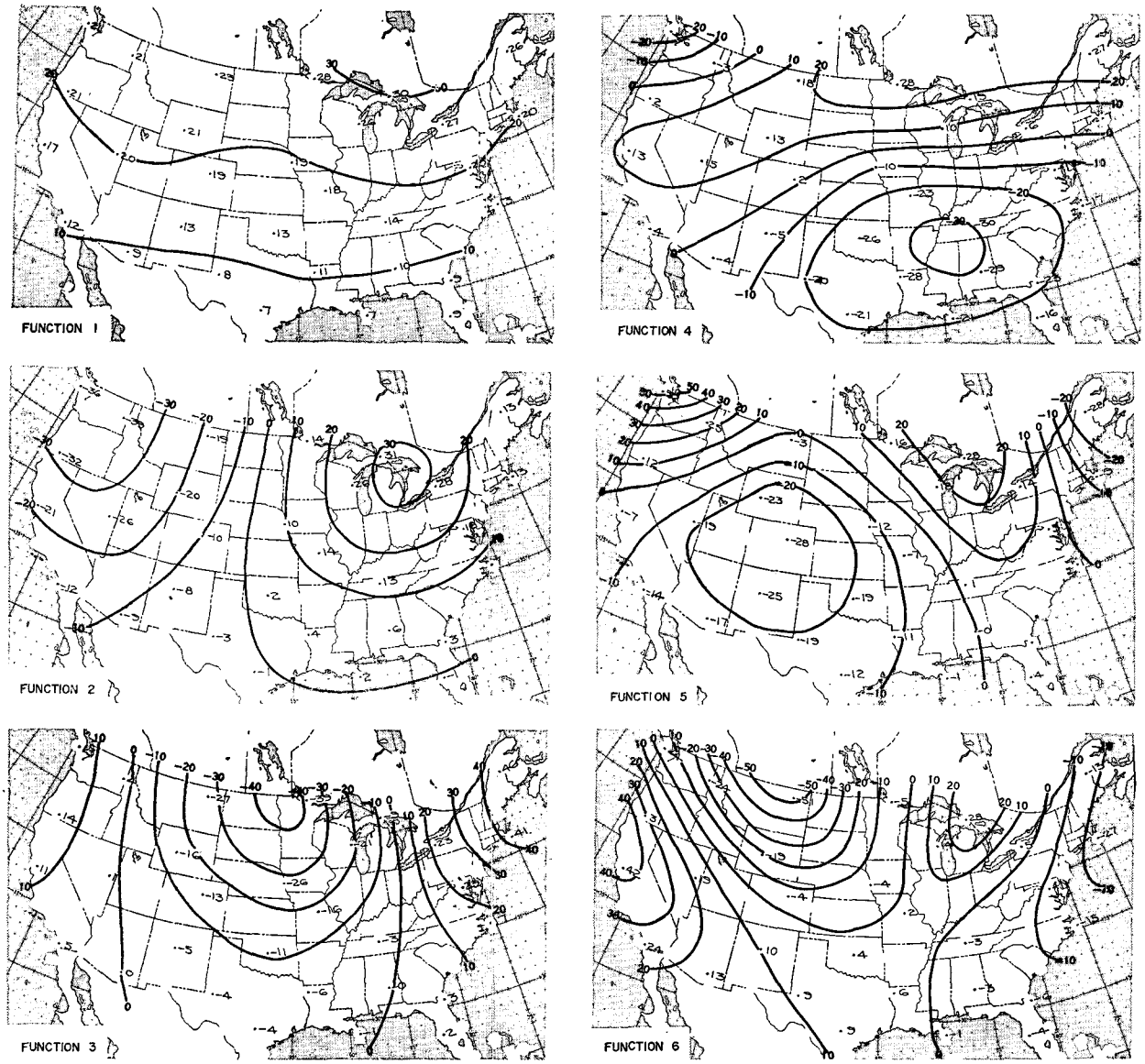

FIG. 8. The six most important principal components $\times 10^2$ of the predictand data.

The magnitudes of the coefficients are in general larger in the equation which explains a large portion of the variance of the predictand.

Functions 1, 3 and 7 together explain 71% of the variance of height at Lander, only 6% less than that afforded by all 30 functions, but they explain only 5% of the variance at Nashville. Again the magnitudes of the regression coefficients shown in Fig. 7 tend to be larger in the equation which explains a large portion of the variance of the predictand.

It is interesting to note that the predictor and predictand functions ${}_pA_i$ and ${}_qB_i$ (Figs. 2 and 3) do not exhibit the smooth patterns of the most important principal components of the predictand data which are shown in Fig. 8. (The principal components of predictor and predictand data are nearly the same in this example.) These principal components have the same characteristics as the empirical orthogonal functions (principal components) of sea-level pressure presented by Lorenz (1956).

Table 1 also contains λ_i^2, the fractional part of the variance of ${}_n\mathbf{Y}_q\mathbf{B}_i$ explained by ${}_n\mathbf{X}_p\mathbf{A}_i$ and vice versa, and EV_i/TV, the fractional part of the total variance of the predictands ${}_n\mathbf{Y}_q$ explained by the ith canonical predictor function. The canonical correlations decrease rather slowly as i increases but EV_i/TV decreases rapidly. It is possible for EV_i/TV to increase as i increases, as evidenced by the entries for $i=8$ and 9.

The data used here were highly correlated in time and 455 cases were evidently not sufficient to give very satisfactory results in terms of the patterns of the coefficients in the canonical functions and regression equa-

tions. (Existing significance tests give little guidance here since the assumptions underlying the tests are far from actuality. Lawley's (1959) test indicates the first 21 canonical correlations to be significant at the $2\frac{1}{2}\%$ level.) Given a much larger sample, one would expect the coefficients in the regression equations which explain a large portion of predictand variance to exhibit smoother patterns than those shown in Figs. 6 and 7. Since Fig. 6b and Fig. 7a represent regression equations which explain only a small portion of predictand variance and are probably the result of the random component in the data, one would not expect them to be meteorologically meaningful.

In many applications a predictor may occasionally be in error. If only a very small number of predictors are in the regression equation, the error may have a large and detrimental effect on the prediction. The more predictors there are in the equation the less an error in one of them will affect the prediction. In cases where there are several predictands, canonical correlation used very cautiously may give regression equations which are slightly better for operational use than those derived by the well known screening regression (Miller, 1962). Other factors, such as the problem of missing data, may overshadow this possible advantage.

Regression equations are not always to be preferred to discriminant (or canonical) functions. When the number of groups is large, analysis of the multi-dimensional discriminant space is difficult and involves additional assumptions. However if only 2, or perhaps 3, discriminant functions are important, hand analysis of scatter diagrams may yield probability estimates that surpass regression estimates; in this application, discriminant analysis will have reduced the problem from p dimensions to 2 or 3 dimensions.

REFERENCES

Anderson, T. W., 1958: *An Introduction to Multivariate Statistical Analysis.* New York, John Wiley & Sons, 374 pp.

Barnard, M., 1935: The secular variations of skull characters in four series of Egyptian skulls. *Ann. Eugenics,* **6**, 352–371.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **79**, 1–3.

Brown, G. W., 1947: Discriminant functions. *Ann. Math. Statist.,* **18**, 514–528.

Bryan, J. G., 1950: A method for the exact determination of the characteristic equation and latent vectors of a matrix with applications to the discriminant function for more than two groups. Ed. D. Dissertation, Harvard University, 290 pp.

Fisher, R. A., 1936: The use of multiple measurements in taxonomic problems. *Ann. Eugenics,* **7**, Part II, 179–188.

Hooper, J. W., 1959: Simultaneous equations and canonical correlation theory. *Econometrika,* **27**, 245–256.

Hotelling, H., 1936: Relations between two sets of variates. *Biometrika,* **28**, 139–142.

Lawley, D. N., 1959: Tests of significance in canonical analysis. *Biometrika,* **46**, 59–66.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Sci. Rept. No. 1, Statistical Forecasting Project, Massachusetts Institute of Technology, 49 pp.

Lund, I. A., 1955: Estimating the probability of a future event from dichotomously classified predictors. *Bull. Amer. Meteor. Soc.,* **36**, 325–328.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. *Meteor. Monogr.,* **4**, No. 25, 54 pp.

——, 1964: Regression estimation of event probabilities. Tech. Rept. No. 1, Contract Cwb-10704, The Travelers Research Center, 153 pp.

Tatsuoka, M. M., 1955: The relationship between canonical correlation and discriminant analysis; and a proposal for utilizing qualitative data in discriminant analysis. Cambridge, Educational Research Corporation, 47 pp.