CANONICAL CORRELATION ANALYSIS OF FORMULATION OPTIMIZATION EXPERIMENTS

N. R. Bohidar
Philadelphia College of Pharmacy and Science
Villanova University
1530 Bridal Path Road. Lansdale, PA. 19446
and
Norman R. Bohidar
University of Washington, Seattle, Washington

ABSTRACT

Formulation optimization experiments are primarily composed of two groups of variables, a set of independent variables and a set of dependent variables. Simultaneous consideration of all the variables in a single analysis is desirable since it provides an opportunity to study the interrelationships of all variables, independent as well as dependent at the same time and imparts an in-depth insight into the entire system as a whole. A multivariate statistical analysis, known as canonical correlation analysis, has indeed this capability. In addition, the analysis has the capacity of extracting the maximum possible correlation, called canonical correlation, between the variables of the two sets. The larger the value of the canonical correlation (0.90 or above), the higher is the predictability of one set from the other set. The analysis produces two composite canonical functions, one for each set. They can be used to streamline the subsequent search process associated with the full-fledged optimization analysis. The analysis also has the cardinal property to rank-order the variables in each set according to their relative contributions to the canonical prediction function, and to delineate the most important variable in each set. This information can be useful in monitoring the future performance of the formulation in a time-and-cost effective manner and in selecting variables for future experiments. All the relevant features of the analysis have been depicted in

this paper in the context of a mobile phase composition optimization experiment.

INTRODUCTION

Canonical correlation analysis succinctly depicts the latent structures of the interrelationship between a set of independent variables (X-variables) and a set of dependent variables (Y-variables) and, consequently it can appropriately be considered as a structural generalization of simple correlation analysis involving a single X-variable and a single Y-variable, and of multiple correlation analysis which involves several X-variables and a single Y-variable. However, the methods, mathematics, objectives and interpretations are uniquely distinct, in that the results of the analysis impart a global interpretation pertaining to the entire system consisting of a set of several interrelated variables. Since a formulation optimization experiment generates a set of several X-variables (2-5 excipients/process variables) and a set of several Y-variables (8-10 response variables), canonical correlation analysis is indeed aptly suitable for application to any pharmaceutical formulation optimization system for making simultaneous statistical inference about the interrelationships of the process and product variables considered.

To accomplish this task, canonical correlation analysis(1) establishes a linear function of the X-variables and a linear function of the Y-variables based on those coefficients for the X-variables and those coefficients for the Y-variables which maximize the correlation between the two linear functions noted above. In other words, canonical correlation analysis provides the maximum possible correlation between a set of X-variables and a set of Y-variables involving all variables simultaneously. For the purpose of recognition, the following terminologies are introduced: (i) The X-coefficients and Y-coefficients are called the canonical coefficients, (ii) the X-function and Y-function are called the canonical variates (composites or functions) and (iii) the correlation between the two canonical variates is called the canonical correlation which is the maximum correlation among all the correlations between the variables in the two sets. The analysis should be considered as an integral part of a formulation optimization analysis. The higher the magnitude of the canonical correlation, the greater is the chance of a successful optimization analysis. If the magnitude of the canonical correlation (maximum possible correlation) attains a value of say, 0.90 or above, it would essentially guarantee the attainment of a high resolution optimization search process associated with the M-SOOP

procedure(2,3,4). Because all the variables are in a single functional form, one would be able to predict from one set to the other. Predictions will lead to the detection of possible variable-ranges to be targeted for optimization. The relative magnitude of the canonical coefficients in each set would enable one to rank-order the variables in accordance of their importance thus providing a basis for identifying the key parameters associated with the process as well as the product variables based on the simultaneous consideration of all variables, a cardinal property of this analysis. Even though the analysis is computationally complex, the entire analysis can be accomplished by using only a few simple SAS(5) program statements provided in Table-I. Canonical correlation analysis must be carried out for each formulation optimization experiment in addition to the three analyses demonstrated in reference(2,3,4), all accomplished from the same data-set.

The primary purpose of this paper is (a) to determine the magnitude of the maximum correlation, the canonical correlation, (b) to assess the proportion of variation in the Y-set accounted for by the X-set variables, (c) to conduct a test of significance for the canonical correlation, (d) to determine the canonical coefficients for the X-variables, as well as the Y-variables, (e) to rank-order the coefficients in each set and (f) to interpret the results of the canonical correlation analysis in the context of an optimization analysis of a mobile phase composition experiment(2).

THEORY

Let the X-set containing the independent variables have p variables denoted by X_1, X_2, ---- X_p and let the Y-set containing the dependent variables have t variables denoted by Y_1, Y_2, --- Y_t, with $p < t$. Let the number of formulations considered for the study be equal to n. Let the respective linear functions of the X-variables and the Y-variables be expressed as,

$$W_i = a_1X_{i1} + a_2X_{i2} + \text{----} a_pX_{ip}, \quad i = 1, 2, \text{----} n$$

$$Z_i = b_1Y_{i1} + b_2Y_{i2} + \text{----} b_tY_{it}, \quad i = 1, 2, \text{----} n$$

where, W and Z represent the two canonical variates and a's and b's represent their respective canonical coefficients. Let W^* and Z^* denote the respective mean of W and Z. Since certain matrix properties will play an integral part of the maximization process, the expressions above are converted to a matrix/vector notation, as follows:

$\Sigma(W - W^*)^2 = A'S_{xx}A$, $A' = [a_1, a_2, \text{---}a_p]$, $\Sigma(Z - Z_*)^2 = B'S_{yy}B$, $B' = [b_1, \text{---}b_t]$ and $\Sigma(W - W^*)(Z - Z^*) = A'S_{xy}B$.

Note that, S_{xx} and S_{yy} are the matrices of sum of squares and sum of products of X variables and Y variables

respectively and S_{xy} is the matrix of sum of products between X and Y variables. Note that, S_{xx} and S_{yy} are two symmetric matrices of (p x p) and (t x t) dimensions, and S_{xy} is a matrix of (p x t) dimension. The entire matrix is of (p + t) x (p + t) dimension.

The product-moment correlation coefficient (ordinary correlation) of the two canonical variates is expressed as

$$R_{wz} = \Sigma(W_i - W^*)(Z_i - Z^*)/[\Sigma(W_i - W^*)^2]^{1/2}[\Sigma(Z_i - Z^*)^2]^{1/2}$$

Expressed in matrix notation, one has

$$R_{wz} = A'S_{xy}B/[A'S_{xx}A]^{1/2}[B'S_{yy}B]^{1/2}$$

The purpose here is to determine those values of a_i's and b_i's which maximize $R_{wz}(1)$. However, it is well known that the value of a correlation coefficient is invariant under scalar transformation. To overcome this invariance property of the correlation and to acquire an unique solution, it is necessary to impose the following universally accepted constraints,

$$A'S_{xx}A = 1 \quad \text{and} \quad B'S_{yy}B = 1$$

Now one proceeds with the maximization process by resorting to the Lagrange Multipliers method. The Lagrange Multiplier Function (LMF) is explicitly expressed as,

$$LMF = A'S_{xy}B - 1/2\theta_1(A'S_{xx}A - 1) - 1/2\theta_2(B'S_{yy}B - 1)$$

Taking partial derivatives of LMF with respect to A, B, θ_1 and θ_2, one obtains the following results,

(1) $dLMF/dA = S_{xy}B - \theta_1 S_{xx}A = 0$, (2) $dLMF/dB = S'_{xy}A - \theta_2 S_{yy}B = 0$, (3) $dLMF/d\theta_1 = A'S_{xx}A - 1 = 0$ and (4) $dLMF/d\theta_2 = B'S_{yy}B - 1 = 0$, where d stands for the partial derivative operator. For the solution of this set of equations, one begins with multiplying the first equation by A', and the second by B' and then imposing the constraints in the third and fourth equations. These operations reduce the system of equations to:

(1) $S_{xy}B - \theta S_{xx}A = 0$ and (2) $S_{yx}A - \theta S_{yy}B = 0$ where, $\theta_1 = \theta_2 = \theta = A'S_{xy}B$. Now multiplying the first equation by θ and the second by $S_{xy}S_{yy}^{-1}$, one has:

$\theta S_{xy}B - \theta^2 S_{xx}A = 0$ and $S_{xy}S_{yy}^{-1}S_{yx}A - \theta S_{xy}S_{yy}^{-1}S_{yy}B = 0$. Adding these two equations, one reaches the final form of the equation as,

$$(S_{xy}S_{yy}^{-1}S_{yx} - \theta^2 S_{xx})A = 0$$

For solving the above equation, one needs to evaluate the determinantal (characteristic) equation,

$$M_{aa} = \det(S_{xy}S_{yy}^{-1}S_{yx} - \theta^2 S_{xx}) = 0$$

where "det" stands for the determinant of the expression in the parenthesis. Since this represents a p-degree polynomial in θ, there are p roots, which are the eigen values. So the solution yields the eigen values, θ_1^2, θ_2^2, θ_3^2, -- θ_p^2 and their respective eigen vectors, A_1, A_2, A_3, ---A_p. The non-zero positive square root θ_1 of the eigen value θ_1^2 is called the canonical correlation between the two canonical variates, $W_1 = A'_1X$ and $Z_1 = B_1'Y$, which is the maximum correlation among the bivariate

correlations associated with the X-set and Y-set. The p elements of A_1 eigen vector constitute the p canonical coefficients, pertaining to the p X-variables. The other (p - 1) eigen values and their respective vectors can be defined in the similar manner. Now for the solution of B, one multiplies the second equation by Θ and the first by $S_{yx}S_{xx}^{-1}$, in the reduced set of equations numbered (1) and (2) above, and gets: $S_{yx}S_{xx}^{-1}S_{xy}B - \Theta S_{yx}A = 0$ and $\Theta S_{yx}A - \Theta^2 S_{yy}B = 0$. Adding these two equations leads to the results,

$$(S_{yx}S_{xx}^{-1}S_{xy} - \Theta^2 S_{yy})B = 0$$

The solution is accomplished by evaluating the determinantal equation:

$$M_{sb} = \det(S_{yx}S_{xx}^{-1}S_{xy} - \Theta^2 S_{yy}) = 0$$

the non-zero eigen values Θ_1^2, Θ_2^2, ---Θ_p^2 and their respective eigen vectors, B_1, B_2 ---B_p represent the solutions of the above system of equations. The non-zero positive square root Θ_1 of the eigen value Θ_1^2 is the same canonical correlation found from the determinantal equation, M_{sa}. The t elements of the eigen vector, B_1, derived from the eigen value Θ_1^2 constitute the t canonical coefficients pertaining to the t Y-variables in the Y-set. Note that, using M_{sa}, one can derive, $B_i = \Theta^{-1}{}_i S_{yy}^{-1}S_{yx}A_i$, i = 1,2,---p, (bypassing M_{sb}). The first eigen value Θ_1^2, the first canonical correlation, Θ_1 and the set of canonical coefficients associated with the first eigen vector are the most important parameters in this analysis.

Let R_{xx} be the (p x p) symmetric matrix of correlations between the X-variables, let R_{yy} be the (t x t) symmetric matrix of correlations between the Y-variables and let R_{xy} be the (p x t) matrix of correlations between the X-variables and Y-variables. Now,

$$R_{wz} = C'R_{xy}D/[C'R_{xx}C]^{1/2}[D'R_{yy}D]^{1/2}$$

where, C and D represent the vectors of standardized canonical coefficients, which are derived from the following pair of determinantal equations,

$$M_{rc} = \det(R_{xy}R_{yy}^{-1}R_{yx} - \Theta^2 R_{xx}) = 0 \quad \text{(for C)}$$

$$M_{rd} = \det(R_{yx}R_{xx}^{-1}R_{xy} - \Theta^2 R_{yy}) = 0 \quad \text{(for D)}$$

Note that the standardized canonical coefficients are easy to interpret. The following pair of determinantal equations represents another form of M_{rc} and M_{rd} equations.

$$M_{rdf} = \det(R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy} - \Theta^2 I) = 0 \quad \text{(for D vector)}$$

$$M_{rcf} = \det(R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx} - \Theta^2 I) = 0 \quad \text{(for C vector)}$$

For completeness, presented in the following is a pair of yet another form of the determinantal equations,

$$M_{saf} = \det(S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx} - \Theta^2 I) = 0 \quad \text{(for A vector)}$$

$$M_{sbf} = \det(S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy} - \Theta^2 I) = 0 \quad \text{(for B vector)}$$

Note that the solutions of $M_{saf} = M_{sa}$ and that of $M_{sbf} = M_{sb}$. All these different formats of the determinantal equations

are presented here only to reduce the amount of prevailing confusion that exists and to introduce the invariance properties of the partitioned matrices.

Now it would be appropriate to depict, in the following, those matrix properties that are crucial to the analysis, explicitly:

(a) The eigen values and their respective eigen vectors derived from the determinantal equations based either on the S-matrices (S_{xx}, S_{yy}, S_{xy}) or the R-matrices (R_{xx}, R_{yy}, R_{xy}) remain the central parameters of the analysis.

(b) For the derivation of the elements of the eigen vectors, pertaining to S-matrices, there are two equations, one for the eigen vector, A, (M_{sa}), whose elements serve as the canonical coefficients for the variables in the X-set, and the other for the eigen vector, B, (M_{sb}), whose elements serve as the canonical coefficients for the variables in the Y-set. These elements are called the unstandardized canonical coefficients because they are derived from the S-matrices.

(c) For the derivation of the elements of the eigen vectors pertaining to R-matrices, there are two equations, one for the eigen vector, C, (M_{rc}), whose elements serve as the canonical coefficients for the variables in the X-set, and the other for the eigen vector, D, (M_{rd}), whose elements serve as the canonical coefficients for the variables in the Y-set. These elements are called the standardized canonical coefficients because they are derived from the R-matrices.

(d) For a given equation, the solution provides as many eigen values as there are variables in the smaller of the two sets. Since the X-set has p variables and the Y-set has t variables, where p is less than or equal to t, the S-equations or the R-equations would yield only p non-zero eigen values.

(e) The p-eigen values are all positive (non-zero) and they are arranged in order of their magnitudes, θ^2_1, θ^2_2, --- θ^2_p. The non-zero positive square root of the largest eigen value, (θ_1) is called the canonical correlation coefficient.

(f) It is interesting to note that, all four equations (M_{sa}, M_{sb}, M_{rc} and M_{rd}) yield the same set of p eigen values arranged in order of their magnitudes. This last relationship shows that the magnitude of an eigen value remains unchanged whether one uses the variance-covariance matrix (S) or the correlation matrix (R).

(g) Associated with each eigen value, there is an unique eigen vector. For the purpose of derivation, each eigen value is inserted into each of the four equations, (M_{sa}, M_{sb}, M_{rc} and M_{rd}) to generate the elements of their respective eigen vectors. If θ_1 is inserted into the M_{sa} equation, one generates p elements of the eigen vector, A_1 and each element is assigned to one of the p variables in

the same order they are arranged in the X-set. Now, if Θ_1 is inserted into the M_{sb} equation, one generates the t elements of the eigen vector, B_1 and each element is assigned to one of the t variables in the same order they are arranged in the Y-set. Note that the magnitudes of the elements of the eigen vector A are not identical to that of C, nor are the elements of the eigen vector B identical to that of D.

(h) Consider in the following a symbolic representation of the matrix properties depicted above: Let the X-set contain 3 variables and the Y-set 6 variables (a typical optimization situation). Then only three non-zero eigen values are extracted from the equations, Θ^2_1, Θ^2_2 and Θ^2_3, which are numbered according to the descending order of their magnitudes. The positive square root of Θ^2_1, the largest eigen value, is the canonical correlation. The 9 x 9 [(3 + 6) x (3 + 6)] sum of squares and sum of products symmetric matrix is partitioned into the following four submatrices, presented with their respective dimensions, S_{xx}(3 x 3), S_{yy}(6 x 6), S_{xy}(3 x 6) and S_{yx}(6 x 3). The M_{sa} matrix is a 3 x 3 symmetric matrix with Θ^2_1, Θ^2_2 and Θ^2_3 as its eigen values and A_1, A_2 and A_3 as their respective eigen vectors, each containing 3 elements, the canonical coefficients. Based on Θ^2_1 and A_1, one constructs, $W_1 = a_1X_1 + a_2X_2 + a_3X_3$, the canonical variate for the X-set. The M_{sb} matrix is 6 x 6 symmetric matrix with Θ^2_1, Θ^2_2 and Θ^2_3 as its only non-zero eigen values (noting the fact that the rank of M_{sb} is only 3 and $\Theta^2_4 = \Theta^2_5 = \Theta^2_6 = 0$) and B_1, B_2 and B_3 as their respective eigen vectors, each containing 6 elements, the canonical coefficients. Based on Θ^2_1 and B_1, one constructs, $Z_1 = b_1Y_1 + b_2Y_2 + b_3Y_3 + b_4Y_4 + b_5Y_5 + b_6Y_6$. Note that Θ^2_1 is considered here as the predominant (significant) eigen value. Note also that there are two other pairs of linear functions, (W_2,Z_2) and (W_3,Z_3). Now one computes, for each formulation, a W_1-score and a Z_1-score generating n such pairs. The bivariate correlation of these pairs constitutes the canonical correlation. The trace of M_{sa} (the sum of the three diagonal elements) is equal to $\Theta^2_1 + \Theta^2_2 + \Theta^2_3$ and the trace of M_{sb} (the sum of the six diagonal elements) is equal to $\Theta^2_1 + \Theta^2_2 + \Theta^2_3$ also. Note that the last 3 eigen vector columns have identical elements. The M_{rc} and M_{rd} matrices yield the same three eigen values, Θ_1^2, Θ_2^2 and Θ_3^2, as obtained for the M_{sa} and M_{sb} matrices. The eigen vectors of M_{rc} denoted by C_1, C_2 and C_3 contain 3 elements each and that of M_{rd} denoted by D_1, D_2 and D_3 contain 6 elements each. The two canonical variates are expressed as, $W_1^* = c_1X_1^* + c_2X_2^* + c_3X_3^*$ and $Z_1^* = d_1Y_1^* + d_2Y_2^* + d_3Y_3^* + d_4Y_4^* + d_5Y_5^* + d_6Y_6^*$, where the original variables are expressed in their standardized form (X^* and Y^*).

(i) It is known that for a p x p symmetric matrix, the sum of the p eigen values is equal to the trace (sum of

the diagonal elements) of that matrix. However, this is not true for the S or R square matrix of (p+t) x (p+t) dimension. It is although true for the M_{sa}, M_{sb}, M_{rc} and M_{rd} matrices of rank p.

(j) The sum of squares of the elements of an eigen vector of a symmetric matrix adds to one. However, this is not true for the M_{sa}, M_{sb}, M_{rc} and M_{rd} matrices. Since the solution of Lagrangean multiplier equation is based on the constraints, $A'S_{11}A=1$, $B'S_{22}B = 1$, $C'R_{11}C = 1$ and $D'R_{22}D = 1$.

(k) The canonical coefficients are directly interpretable based on their respective relative magnitudes, in determining the importance of each variable in the X-set and the Y-set.

(l) The canonical coefficients are also used to construct a pair of scores (W and Z) for each formulation.

Univariate Structure Correlations(5): The development so far has explicitly depicted the role of canonical correlation, maximal eigen values, canonical coefficients, canonical variates as well as the associated matrix properties, in canonical correlation analysis. The formulas for these parameters are based on the simultaneous consideration of all the variables of the X-set and Y-set, which is obviously the appropriate multivariate approach to the analysis. The development depicted in the following, however, would concentrate primarily on the correlation of each variable in a set with the canonical variate of that set, involving one original variable and one canonical variate at a time. The formulas for the correlations are presented here for two specific reasons: (a) these results should only be considered as an adjunct to the multivariate results and (b) the SAS software program, which will be considered in the next section, do provide the results automatically on a routine basis. Note that, any interpretations drawn from the results must be considered in the context of an univariate approach. Note, corr(W,X) = $R_{xx}C$ and corr(Z,Y) = $R_{yy}D$ where corr denotes correlation.

Reciprocal Univariate Structure Correlation: Here one is interested in the correlation between the canonical variate of one set and the individual variables of the other set. Now corr(W,Y) = $(R_{yy}D)(R_c)$ and corr(Z,X) = $(R_{xx}C)(R_c)$ measuring the strength of each variable of one set accounted for by the canonical variate of the other set, where R_c is the square root of R_c^2, the maximal eigen value.

STATISTICAL TESTS OF SIGNIFICANCE IN CANONICAL CORRELATION ANALYSIS

This section is created to cover a comprehensive set of test statistics for conducting statistical tests to determine if the experimentally derived canonical

correlations are indeed statistically significant. There are four distinct test statistics, developed by four different authors, each accomplishing the same goal. Unfortunately, no one test statistic has been conclusively demonstrated to be universally superior or inferior. In other words, no single test is uniformly most powerful (that is, high probability of rejecting the null hypothesis of no perceptible canonical correlation, when the hypothesis is false) against all possible alternatives. Therefore it is proposed to present all four primary test statistics as well as the associated sequential and F-transformed test statistics. It is interesting to note that, each test statistic is a different function of the same eigen value derived from the matrix, $(R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx})$ or $(S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx})$. Therefore it is proposed to present not only the formulas of the various test statistics, but also the numerical results of the tests associated with the mobile phase optimization experiment noted in the introduction section(2). It should be noted here that this is the first time all these tests have been assembled in one place in an unified coherent manner.

To accomplish the computations of the test statistics, one needs the numerical values of (a) the eigen values, θ_1^2, θ_2^2,---- θ_p^2, (b) the number of the variables in the X-set, p, (c) the number of variables in the Y-set, t, and (d) the number of formulations considered in the study, n. For this study, one finds, (a) $\theta_1^2 = 0.925212$, $\theta_2^2 = 0.193195$ and $\theta_3^2 = 0.001612$, arranged in order of magnitude, (b)p = 3, (c)t = 3 and (d) n = 15. Presented in the following are the various tests:

I. Wilk-θ-test (6): Test statistic, in general,

$\theta = (1 - \theta_1^2)(1 - \theta_2^2)(1 - \theta_3^2)$ --- $(1 - \theta_p^2)$

Here, $\theta = (1 - \theta_1^2)(1 - \theta_2^2)(1 - \theta_3^2) = 0.060242$

(Note θ without a subscript or an exponent would denote the test statistic)

Rao's F-statistic (7) for Wilk-θ-test:

$F = (1 - \theta_1^b)(ms + 2d)/\theta_1^b 2r = 5.3223$

where, $b = s^{-1}$ and $s = [(p^2t^2 - 4)/(p^2+t^2 - 5)]^{1/2} = 2.4337$

$m = (n - 1) - (1/2)(p + t + 1) = 10.5$, $2r = pt = 9$,

$d = -(pt - 2)/4 = -1.75$, $b = s^{-1} = 0.41084$, degrees of freedom (d.f.) for the F-test, 2r = 9, ms + 2d = 22.05424, $(1 - \theta^b)/\theta^b = 2.17196$, (ms + 2d)/2r = 2.4505. Since the sample F-value (5.3225) exceeds the tabular F-value (2.34 for a = .05 and 3.35 for a = .01 for 9 and 22 d.f.) the first canonical correlation is highly significant ($p<<.01$). Note, throughout this section, "a" denotes the level of significance, 0.05 or 0.01).

II. Wilk-θ-test for the second canonical correlation:

$\theta = (1 - \theta_2^2)(1 - \theta_3^2) = 0.805504$

F-test for θ: $F = (1 - \theta^b)(ms + 2d)/\theta^b 2r = 0.5710$

$b = s^{-1}$ and $s = [(p-k)^2(t-k)^2-4]/[(p-k)^2 + (t-k)^2 - 5] = 2$

where k = the number of eigen values removed (here k = 1, p-k = 2 and t-k = 2), m = [(n - 1) - k] - (1/2)[(p - k) + (t - k) + 1] = 10.5, d = -[(p - k)(t - k) - 2]/4 = -0.5 or 2d = -1.0, b = s^{-1} = 0.5, d.f. for the F-test, 2r = (p - k)(t - k) = 4, ms + 2d = 20; $(1 - \Theta^b)/\Theta^b$ = 0.11421, (ms + 2d)/2r = 5. Since the sample F-value (0.5710) is less than the tabular F-value (2.87 for a = .05 for 4 and 20 d.f.) the second canonical correlation is not significant (p>0.05).

III. Wilk-Θ-test for the third canonical correlation:

$\Theta = (1 - \Theta_3^2) = 0.9984$

F-statistic = $(1 - \Theta^b)(ms + 2d)/\Theta^b 2r = 0.0178$

b = s^{-1}, s = $[(p-k)^2(t-k)^2 - 4]/[(p-k)^2 + (t-k)^2 - 5]$ = 1.0 where k = 2, p - k = 1, t - k = 1. m = same formula as in II = 10.5, d = 0.25, 2d = 0.5, d.f for the F-test, 2r = 1.0, ms + 2d = 11; $(1 - \Theta^b)/\Theta^b$ = 0.0016146, (ms + 2d)/2r = 11. Since the sample F-value (0.0178) is less than the tabular F-value (4.84 for a = .05 for 1 and 11 d.f.) the third canonical correlation is not significant (p>0.05).

IV. Bartlett's X^2chi (chi-square) statistic for Wilk-Θ-test(8):

(i) As before, $\Theta = (1 - \Theta_1^2)(1 - \Theta_2^2)(1 - \Theta_3^2)$ = 0.060242

Now, $X^2\text{chi} = -[(n - 1) - (1/2)(p + t + 1)]\text{Log}_e\Theta$

Since $\log_e(0.06024) = -2.80939$, n = 15, p = 3 and t = 3 X^2chi = 29.4986 with (pt) degrees of freedom = 9. Since the tabular X^2chi (with 9 d.f., 16.92 for a = .05 and 21.67 for a = .01) is less than the sample X^2chi, the first canonical correlation is highly significant (p<<.01). (ii) To conduct a test for the second canonical correlation, one computes the following:

$X^2\text{chi} = -[(n - 1) - (1/2)(p + t + 1)]\text{Log}_e\Theta$, where, now,

$\Theta = (1 - \Theta_2^2)(1 - \Theta_3^2)$ with (p - k)(t - k) d.f. (k=1). Note that only the value of $\log_e\Theta$ and the d.f. change here. Since the sample X^2chi (=2.2710) is less than the tabular X^2chi (9.49 for a = .05 with 4 d.f.), this canonical correlation is not significant (p > 0.05).

(iii) To conduct a test for the third canonical correlation, one computes,

$X^2\text{chi} = -[(n - 1) - (1/2)(p + t + 1)]\text{Log}_e\Theta$, where,

$\Theta = (1 - \Theta_3^2)$ with (p - k)(t - k) d.f. (k=2). Since the sample X^2chi (0.01694) is less than the tabular X^2chi value (3.84 for a = .05 with 1 d.f.) the third canonical correlation is not significant (p > 0.05).

V. Hotelling-Lawley Trace Test(9):

(i) Test for the first canonical correlation:

Let $H = \Sigma(\Theta_i^2/1-\Theta_i^2)$, i = 1,2,3 (H=12.6123), S=p(=3),

M = (1/2)[abs(p - t) - 1] = -0.5,

N = (1/2)[(n - 1 - t) - p -1] = 3.5, then

$F = [2H(SN + 1)]/S^2(2M + S + 1)]$ = 10.744 with

[S(2M + S + 1)](=9) and [2(SN + 1)] (=23) as the

d.f.'s. Since the sample F (=10.744) exceeds the tabular F (2.32 for a = .05 and 3.30 for a = .01 with 9 and 23 d.f.)., this correlation is highly significant ($p << 0.01$). Here, the abs = absolute value.

(ii) Test for the second canonical correlation: For this test, p and t must be reduced by one, since k=1. Thus,

$H = [\theta_2^2/1 - \theta_2^2] + [\theta_3^2/1 - \theta_3^2] = 0.24107$, S = 2, M = -0.5, N = 4.5 and F = 0.60268. Since the tabular F is 2.87 with 4 and 20 d.f., this correlation is not significant ($p > 0.05$).

(iii) Test for the third canonical correlation: For this test, p and t must be reduced by two, since k=2. Thus,

$H = [\theta_3^2/1 - \theta_3^2] = 0.001615$, S = 1, M = -0.5, N = 5.5 and F = 0.02002, which is not significant ($p > 0.05$), since the tabular F is 4.67 for a = 0.05 with 1 and 13 d.f.

VI. Pillai's Trace Test (10):

(i) Test for the first canonical correlation:

Let $P = \Sigma\theta_i^2$, (P=1.12), i = 1,2,3. Here, S, M and N retain the same definitions given in test V. Here,

$F = [(P)(2N + S + 1)]/[(S - P)(2M + S + 1)] = 2.1845$ with $[S(2M + S + 1)]$ (=9) and $[S(2N + S + 1)]$ (=33) as the two d.f.'s. Since the sample F (2.1845) exceeds the tabular F (2.1840 for a = .05), the correlation is significant ($p < 0.05$). This test is very conservative and is given here for completeness only.

(ii) Test for the second canonical correlation: For this test, p and t must be reduced by one, since k=1. Here

$P = \theta_2^2 + \theta_3^2 = 0.1948$ and F = 0.2769 which is not significant ($p > 0.05$) since the tabular F is 3.26 for a = 0.05 with 4 and 12 d.f. (iii) Test for the third canonical correlation: For this test, p and t must be reduced by two, since k=2. Here

$P = \theta_3^2 = 0.001612$ and F = 0.02099 which is not significant ($p > 0.05$), since the tabular F is 4.67 for a = 0.05 with 1 and 13 d.f.

VII. Roy's Greatest Characteristic Root Test (11):

(i) Test for the first canonical correlation:

Let $\theta_{max} = [\theta_1^2/1-\theta_1^2]$ (= 12.371129) and

$F_{upper} = [(\theta_1^2)(n - t - 1)]/[(1 - \theta_1^2)(t)]$ with n = 15 and t = 3. Since F_{upper} (= 45.36) exceeds the tabular F (3.59 for a = .05 and 6.22 for a = .01 with 3 and 11 d.f.), this correlation is very highly significant ($p << 0.01$). (ii) Test for the second canonical correlation: Here t must be reduced by one, since k=1. Here

$\theta_{max} = [\theta_2^2/1 - \theta_2^2] = 0.23946$ and $F_{upper} = 1.4367$ which is not significant ($p > 0.05$) since the tabular F is 3.88 for a = 0.05 with 2 and 12 d.f. (iii) Test for the third canonical correlation: For this test t must be reduced by two, since k=2. Here

$\theta_{max} = [\theta_3^2/1 - \theta_3^2] = 0.0016146$ and $F_{upper} = 0.02099$ which is not significant ($p > 0.05$) since the tabular F value is 4.67 for a = 0.05 with 1 and 13 d.f.

TABLE-I

SAS PROGRAM STATEMENTS FOR CANONICAL CORRELATION ANALYSIS

No.	Program Statement	No.	Program Statement
1	XXXX XXXX;	5	PROC CANCORR DATA = A ALL;
2	DATA A;	6	VAR X_1 X_2 X_3;
3	INFILE ACET;	7	WITH Y_1 Y_2 Y_3;
4	INPUT X_1 X_2 X_3 Y_1 Y_2 Y_3;	8	RUN;

Even though many formulas given above pertain to p = 3, they can easily be extended to any p. Note that, to accomplish these tests given above, one only needs the ordinary F or X^2chi tables universally available in any text book on statistical methods. Also note that, the tables for the distribution of the other tests given above are only sporadically available in the literature and they are usually not very extensive. Some tables such as Heck chart (table) for Roy's test and lower percentage points table for Wilk's-θ-test are available. However, the three degrees-of-freedom parameters used there have been defined especially for multivariate analysis of variance procedure only. Appropriate modifications are needed for their use in canonical correlation analysis.

Sometimes it may be of interest to construct 95% confidence limits for a canonical correlation (R_c) using Lawley's(12) asymptotic standard error (SE) formula, $SE(R_c = [(n - 1)^{-1}(1 - R_c^2)]^{1/2}$, which yields only an approximate result.

COMPUTATIONAL ASPECTS OF THE ANALYSIS

As evidenced by the deliberations of the preceding two sections, the computational burden associated with the analysis is enormous and the development of an in-house software for these matrix procedures is a formidable task. However, a short SAS (5) software program can accomplish the entire analysis instantly. Table-I contains the canonical correlation analysis program written in the SAS statistical computer language (5). The program has no restrictions on the number of variables in either set and

thus the statements, 4, 6 and 7 must be modified according to the number of variables to be processed. The program presented in Table-I is based on the specifications of the mobile phase optimization study containing three variables in each of the two sets, X and Y, and follows the same format presented in reference (2). The program calls for storing the data for these six variables (or however many variables involved) in columns of numbers in the data file named "ACET" (statement-3) which is created and stored prior to the execution of the program (OPT.SAS). Should the user prefer another name for that file, that name must be inserted instead. It should be noted here that, the same data file "ACET" could be used for all four programs listed in reference (2) in addition to this program without creating a new data file for each program, which is a tremendous time-saving device. Note also that statement-1 must be furnished by the computer department of the scientist's facility.

The computer output print-out consists of (a) mean and standard deviation of all six variables, (b) elements of R_{xx}, R_{yy} and R_{xy} matrices, (c) ordered canonical correlations (1,2 and 3), (d) ordered eigen values and the proportion of variation accounted for by each, (e) F tests for Wilk's Θ for each of the three canonical correlations, (f) F tests for Hotelling-Lawley trace, Pillai's trace and Roy's largest root, for only the first canonical correlation (note that, for the others use the formulas given in the previous section, (g) standardized canonical coefficients (eigen vectors) for each eigen value and (h) univariate structural correlations for the X-set, the Y-set and their reciprocals.

RESULTS, DISCUSSION AND INTERPRETATION

A canonical correlation analysis is conducted for a mobile phase composition optimization experiment, described in detail in reference (2). The motivation of such a study emanates from the fact that a successful separation of a pure drug substance from its degradation products leads to a successful control on the known contaminants. One of the primary functions of a Pharmaceutics Department is to develop appropriate procedures for these important chemical activities. Typically, the process involves the selection of a relevant set of solvents and modifiers, and of their respective concentrations for experimentally determining the optimum composition which meets the goals pertaining to a prescribed set of quantitatively assessed column performance parameters. The three variables of the Y-set comprise of (i) capacity factor(K')(Y_1), (ii) peak skew (Y_2) and (iii) HETP (Height equivalent theoretical plate)

(Y_3). Each of these column performance parameters is measured for each of the various combinations of the levels of the two organic modifiers, Methanol (X_2) and Acetonitrile (X_3) and buffer pH (X_1), which constitute the three variables for the X-set. Fifteen such combinations defining the various mobile phase compositions are generated based on an orthogonal central composite (augmented) full three factor factorial design with a center point. (The factorial structure of a three factor composite design can be derived from the table on page 159 in reference (4) and the exact concentrations of the X-variables can be found in reference (2)).

Simultaneous consideration of all six variables in a single analysis remains the most attractive feature of this statistical procedure whose results genuinely reflect all the available information in the experiment, enabling one to draw appropriate multivariate (rather than univariate) inferences and interpretations. The first canonical correlation is the maximum correlation between the variables of the two sets. In this case, the magnitudes of the eigen values are: $\theta_1^2 = 0.9252$, $\theta_2^2 = 0.1932$ and $\theta_3^2 = 0.001612$ and their respective canonical correlations are: $R_{c1} = 0.9619 = (0.9252)^{1/2}$, $R_{c2} = 0.4395 = (0.1932)^{1/2}$ and $R_{c3} = 0.0402 = (0.001612)^{1/2}$. Here the maximum correlation ($R_{c1} = 0.9619$) is indeed the maximum since the highest bivariate (regular) correlation between X and Y variables is only 0.819 (absolute magnitude); (see Table-II, intersection of Y_1-row and X_3-column). This confirms not only the derivations (theory section) but also the appropriateness of the multivariate analysis. Now one proceeds to determine if $R_{c1} = 0.9619$ is indeed statistically significant. In the section on statistical tests, it is clearly revealed that $R_{c1} = 0.9619$ is statistically significant ($p << 0.01$) based on each and every test depicted in that section. Furthermore, the tests show that the second and third canonical correlations are not statistically significant ($p > 0.05$) (see the section for details). Based on the significant canonical correlation, R_{c1}, the two canonical functions have the following forms, derived from the eigen value, $\theta_1^2 = 0.9252$,

$$W_i = -0.0072X_1 - 0.2831X_2 - 0.9591X_3 \text{ and}$$

$$Z_i = 0.5514Y_1 + 0.4599Y_2 + 0.1587Y_3$$

$i = 1,2,-------15$, where, X and Y must be expressed as standardized variables, and the numerical values represent their respective canonical coefficients.

The next step is to determine the extent to which the above two functions account for the six variables, much the same way one interprets the value of R^2 in optimization or multiple regression analysis. The magnitude of the first eigen value has the property of measuring the adequacy of the two functions, and in this

case, an impressive 92.5% (Θ_1^2 = 0.9252) of the information available in the six variables can adequately be represented by the two canonical functions. This indicates high potentiality for accurate predictability. Based on the two functions, 15 values for each canonical variate, W and Z, can be generated and the bivariate (ordinary) correlation between them is 0.9619, (which is numerically equal to R_{c1} (= $[\Theta_1^2]^{1/2}$)), implying a very strong correlation between the variates. If a regression line is fitted to W and Z data, it yields a slope of 0.9619 ($B_{z,w}$), an intercept of zero, and a canonical regression equation: Z = 0.9619W, with a R_{zw}^2 value of 0.9252, (=Θ_1^2), which is an excellent R^2-value. The utility of this will be described later.

Now that the two canonical functions are formulated, the next logical step is to interpret the absolute as well as the relative magnitudes of the canonical coefficients. The larger the absolute magnitude of a coefficient associated with a variable, the greater is the contribution of that variable to the canonical function. This way, the canonical coefficients not only have the property to rank order the variables, but also to delineate the most important variable based on the simultaneous consideration of all variables involved. There are three methods to accomplish this. Consider first the X-set, in which, C_1 = -0.0072, C_2 = -0.2831 and C_3 = -0.9591: (i) one merely rank order the absolute magnitudes by simple inspection, (ii) rank order the variables by expressing each coefficient as a percentage of the total, ($100C_i/\Sigma C_i$), X_1 = 0.57%, X_2 = 22.7% and X_3 = 76.8%, and (iii) Based on the Lagrange multiplier constraint, $C'R_{11}C = 1$, in which R_{11} is an identity matrix, the sum of squares of the coefficients is expressed as, ($100C_i^2/\Sigma C_i^2$), X_1 = 0.0%, X_2 = 8.0% and X_3 = 92.0%. All three methods clearly indicate that, X_3 contributes substantially to the X-canonical function. In other words, acetonitrile is exerting a significant effect on capacity factor and peak skew, relative to the other two X-variables, in this mobile phase system. It is also the most important variable since the following correlations are very high: corr(X_3,W) = -0.959, corr(X_3,Z) = -0.923, corr(X_3,Y_1) = -0.819, and corr(X_3,Y_2) = -0.816. The non-significant correlations associated with the other two variables are: corr(X_1,W) = -0.0072, corr(X_2,W) = -0.283, corr(X_1,Z) = -0.007, corr(X_2,Z) = -0.272, corr(X_1,Y_1) = 0.011, corr(X_2,Y_1) = -0.392, corr(X_1,Y_2) = -0.027 and corr(X_2,Y_2) = 0.029 (see Table II). It should be noted here that , since the X-set is an orthogonal system, the canonical coefficient of an X-variable is identically equal to the correlation value between that variable and the X-canonical variable (W). Note also that, when

TABLE-II

BIVARIATE CORRELATION VALUES BETWEEN VARIABLES

	X_1	X_2	X_3	Y_1	Y_2	Y_3
X_1	1.0	0.0	0.0	0.011	-0.027	-0.001
X_2	0.0	1.0	0.0	-0.392	0.029	-0.439
X_3	0.0	0.0	1.0	-0.819	-0.816	-0.604
Y_1	0.011	-0.392	-0.819	1.0	0.547	0.812
Y_2	-0.027	0.029	-0.816	0.547	1.0	0.272
Y_3	-0.001	-0.439	-0.604	0.812	0.272	1.0

comparing two correlations, one compares only their absolute magnitudes and not their directions (signs), in this context.

The interpretation of the canonical coefficients of the Y-set is the next consideration. The magnitudes are: $d_1 = 0.5514$, $d_2 = 0.4599$ and $d_3 = 0.1161$. Applying the first method, the rank order of the Y-variables is obvious.

For the second method, the numerical distribution is as follows: $Y_1 = 47.1\%$, $Y_2 = 39.3\%$ and $Y_3 = 13.6\%$. Since the Y-set is not an orthogonal system, (that is, there are non-zero correlations), the third method calculations are not as simple as that for the X-set, since $\Sigma d_i^2 = 1$. The Lagrange multiplier constraint is $D'R_{22}D = 1$, in which R_{22} is not a diagonal matrix. The matrix computations of the left hand side of the constraint equation yield: $Y_1 = 51.38\%$, $Y_2 = 37.01\%$ and $Y_3 = 11.61\%$. All three methods do concur with respect to the rank order of the variables and to the fact that Y_1 and Y_2 share the major portion (88%) of the contribution to the canonical function. In other words, capacity factor and peak skew emerge as the two significant variables directly influencing as well as being highly sensitive to the independent variables in this mobile phase system. They are also important variables, because of their high correlation values, as: $corr(Y_1,Z) = 0.932$, $corr(Y_2,Z) = 0.805$, $corr(Y_1,W) = 0.896$, $corr(Y_2,W) = 0.774$, $corr(Y_1,X_3) = -0.819$ and $corr(Y_2,X_3) = -0.816$. In this system the contribution of Y_3 is not as substantial as the other two Y-variables. Note that in this section "corr" implies correlation.

It should be clearly noted that for the purpose of interpretation the multivariate results, such as, Θ_1^2, R_{c1}, C_1 and d_1, are strongly emphasized and properly interpreted here. However, the univariate results, such as, corr(X_1,W), corr(Y_1,Z), corr(Y_1,W) and corr(X_1,Z), and corr(X_1,Y_1) are quoted here only to supplement the multivariate results.

Optimization Analysis and Canonical Correlation Analysis:

Now that the variables, X_3, Y_1 and Y_2 have emerged as the most significant contributors to their respective canonical variates, these variables would play an effective role in the subsequent ensuing optimization analysis. This forms the natural connection between the two analyses. Furthermore, these results clearly demonstrate the unique capacity of canonical correlation analysis for delineating the most significant variables, based on the simultaneous consideration of all variables involved. These selected set of variables could form a basis for a focused search process associated with the M-SOOP procedure of the optimization analysis. The mechanism for accomplishing this is as follows: (a) Based on the specific goal values of the Y-variables, a Z_o-value is generated from the Y-canonical variate ($Z_o = d_1Y_{1o} + d_2Y_{2o} + d_3Y_{3o}$), where Y_{1o}, Y_{2o} and Y_{3o} are the goal values. (b) Now using the canonical regression function ($Z = 0.9619W$) the Z_o-value is converted to a W_o-value, (c) These selected range of X-values derived from the W_o-value form the basis for the subsequent full fledged optimization analysis, associated with the M-SOOP procedure. Note: The results here pertain only to this study.

These selected set of few variables are primarily used for monitoring in a time-and-cost effective manner the future performance of the optimum composition or formulation and for discovering the various mechanisms governing the mobile phase or the pharmaceutical system.

It is strongly recommended that canonical correlation analysis should be conducted as a routine pre-optimization analysis for the various reasons depicted above.

ACKNOWLEDGEMENT

Deep gratitudes are due to Mrs. Barbara J. Tomlinson for her excellent talent in word-processing this manuscript with utmost rapidity and quality.

REFERENCES

1. H. Hotelling, Relations Between Two Sets of Variables. Biometrika, Vol.28, 321-377 (1936).
2. N.R. Bohidar, Pharmaceutical Formulation Optimization Using SAS. Drug Development and Industrial Pharmacy, Vol.17, No.3, 421-441 (1991).

3. N. R. Bohidar, Application of Optimization Techniques in Pharmaceutical Formulation-An Overview. Proceedings of the American Statistical Association. Biopharmaceutical Section. 6-13 (1984).
4. N. R. Bohidar and K. E. Peace, Pharmaceutical Formulation Development. Chapter IV. "Biopharmaceutical Statistics in Drug Development." Marcel Dekker, Inc. New York, N.Y. 149-229 (1988)
5. SAS Institute Inc. "SAS User's Guide: Statistics" Version 5 Edition. SAS Institute, Inc. Cary, NC(1985)
6. S.S. Wilks, Certain Generalization in the Analysis of Variance. Biometrika, Vol. 24, 471-494(1932).
7. C. R. Rao, An Asymptotic Expansion of the Distribution of Wilk's Criterion. Bull. Inter. Stat. Inst., Vol. 33, 177-180(1951).
8. M. S. Bartlett, Further Aspects of the Theory of Multiple Regression. Proc. Camb. Phil. Soc. Vol. 34, 33-40(1938).
9. C. R. Rao, "Linear Statistical Inference and its Application." John Wiley and Sons, New York (1965).
10. K. C. S. Pillai, Some New Test Criteria in Multivariate Analysis. Ann. Math. Stat., Vol. 26, 117-120(1955).
11. S. N. Roy, "Some Aspects of Multivariate Analysis." John Wiley and Sons, New York (1957).
12. D. N. Lawley, Test of Significance in Canonical Correlation Analysis, Biometrika, Vol. 46, 59-66(1959).