ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of DEVELOPMENT RESEARCH

International Journal of Development Research
Vol. 3, Issue, 9, pp.017-023, September, 2013

Full Length Research Article

MULTIVARIATE ANALYSIS OF CHEMICAL COMPONENTS OF TOBACCO LEAVES

[1,*]Ekezie Dan Dan, [1]Opara Jude and [2]Nwobi Anderson Chukwukailo

[1]Department of Statistics, Imo State University, PMB 2000, Owerri Nigeria
[2]Department of Statistics, Abia State Polytechnic, Aba Nigeria

ARTICLE INFO

Article History:
Received 28h June, 2013
Received in revised form
16th July, 2013
Accepted 15th August, 2013
Published online 19th September, 2013

Key words:
Canonical Correlations,
Canonical Variables,
Correlation Matrix,
Standardized Observation,
Likelihood Ratio.

ABSTRACT

This study examined the Multivariate analysis of Chemical Components of Tobacco Leaves using Canonical Correlation Analysis which seeks to identify and quantify the association between two sets of variables. The paper focused on using Canonical Correlation Analysis to analyze data on chemical components of 25 tobacco leaf samples. The multivariate data satisfied the normality assumption. The data for this study, which contains three criterion measures and six predictor variables, were analyzed using the "SAS" statistical software package. Based on the results obtained, and the hypotheses carried out, it was revealed that out of the three sample canonical correlations, the first two ($\hat{\rho}_1^* = (0.933, \hat{\rho}_2^* = 0.842)$ are significant, while the third one ($\hat{\rho}_3^*$ = 0.373) is insignificant. The analysis also revealed that the first sample canonical variate, $\hat{U}_1$, of the criterion measures is a "better" representative of its set than the first sample canonical variate, $\hat{U}_1$, of the predictor variables of its set.

INTRODUCTION

In many research settings, the social scientist encounters a phenomenon that is best described not in terms of a single criterion but, because of its complexity, in terms of a number of response measures (William and Matthew; 1984). In such cases, interest may center on the relationship between the set of criterion measures and the set of explanatory factors. In a manufacturing process, for instance, we might be concerned with the relationship between a set of organic chemical constituent variables, on the one hand, and various inorganic chemical constituent variables on the other hand, as it is applicable in this paper. In the business or economic fields, we might be interested in the relationship between a set of price indices and a set of production indices, with a view towards (say) predicting one from the other. The study of the relationship between a set of predictor variables and a set of response measures is known as canonical correlation analysis. Canonical correlation analysis seeks to identify and quantify the associations between two sets of variables (Johnson and Wichern; 1992).

Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. The idea is first to determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all parts uncorrelated with the initially selected pair. The process continues. The pairs of linear combinations are called the canonical variables, and their correlations are called canonical correlations. The canonical correlations measure the strength of association between the two sets of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pairs of canonical variables. With a growing number of large scales genomic data the focus these days have been in finding the relationship between two or more sets of variables. One of the classical methods that can be used in cases when we have two set of variables from the same subject is Canonical Correlation Analysis (CCA) but it lacks biological interpretation for situations in which each set of variables has more than thousands of variables. This issue was first addressed by Parkhomenko *et al.* (2009) who proposed a novel method for Sparse Canonical Correlation Analysis (SCCA).

***Corresponding author:** *Ekezie Dan Dan*
Department of Statistics, Imo State University, PMB 2000, Owerri Nigeria

Recently there are a few other proposed methods to find relationship between two sets of variables based on different penalty functions but there are very few comparative studies that have been done so far. A few of the proposed methods are Waaijenborg *et al.* (2008) who used SCCA to find relationships between the effect of copy number alterations on gene expression and progression of glioma, Witten and Tibshirani (2009) used SCCA to find association between gene expression and array comparative genome hybridization (CGH) measurements, Parkhomenko *et al.* (2009) and Waaijenborg *et al.* (2009) used SCCA technique to find correlation between Single-nucleotide polymorphism (SNP) and gene expression data, and Lee etal. (2011) used SCCA approach to find association between gene expression and proteomic data. SCCA was first introduced by Parkhomenko *et al.* (2009) in which a sparseness parameter controls how many variables will be included from each data set. The algorithm proposed by Witten *et al.* (2009) for computing Sparse CCA is similar to that of Waaijenborg *et al.* (2008). Waaijenborg *et al.* (2008) penalized the classical CCA as an iterative regression and then applied an elastic net penalty to find the canonical vectors. The elastic net is a combination of ridge regression and lasso. For more detail about ridge regression, see Hoerl (1962).

MATERIALS AND METHODS

The method of analysis used in this study is the Canonical Correlation Analysis. This paper shall focus on how to analyze a sample of 25 samples of tobacco leaf for organic and inorganic chemical constituents in a manufacturing company using the SAS Statistical Software Package.

Canonical Variates and Canonical Correlations

In this paper, we shall be interested in measures of association between two groups of variables. The first group of p variables is represented by the (p × 1) random vector $\mathbf{X}^{(1)}$. The second group of q variables is represented by the (q × 1) random vector $\mathbf{X}^{(2)}$. We assume, in the theoretical development, that $\mathbf{X}^{(1)}$ represents the smaller set, so that p ≤ q.

For the random vectors $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$, let

$$\left.\begin{array}{ll} E(\mathbf{X}^{(1)}) = \boldsymbol{\mu}^{(1)}; & Cov(\mathbf{X}^{(1)}) = \Sigma_{11} \\ E(\mathbf{X}^{(2)}) = \boldsymbol{\mu}^{(2)}; & Cov(\mathbf{X}^{(2)}) = \Sigma_{22} \\ Cov(\mathbf{X}^{(1)},\mathbf{X}^{(2)}) = \Sigma_{11} = \Sigma_{22}' & \end{array}\right\} \quad \ldots (1)$$

It will be convenient to consider $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ jointly, so, the random vector

$$\underset{((p+q)\times 1)}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \hline \mathbf{X}^{(2)} \end{bmatrix} = \begin{bmatrix} X_1^{(1)} \\ X_2^{(1)} \\ \vdots \\ X_p^{(1)} \\ \hline X_1^{(2)} \\ X_2^{(2)} \\ \vdots \\ X_q^{(2)} \end{bmatrix} \quad \cdots \quad (2)$$

has mean vector

$$\underset{((p+q)\times 1)}{\boldsymbol{\mu}} = E(\mathbf{X}) = \begin{bmatrix} E(\mathbf{X}^{(1)}) \\ \hline E(\mathbf{X}^{(2)}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \hline \boldsymbol{\mu}^{(2)} \end{bmatrix} \quad \cdots \quad (3)$$

and covariance matrix

$$\underset{(p+q)\times(p+q)}{\boldsymbol{\Sigma}} = \mathrm{E}(\mathbf{X}-\boldsymbol{\mu})\mathrm{E}(\mathbf{X}-\boldsymbol{\mu})'$$

$$= \left[\begin{array}{c|c} E(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})' & E(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})' \\ \hline E(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})(\mathbf{X}^{(1)}-\boldsymbol{\mu}^{(1)})' & E(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})(\mathbf{X}^{(2)}-\boldsymbol{\mu}^{(2)})' \end{array}\right]$$

$$\left[\begin{array}{c|c} \underset{(p\times p)}{\boldsymbol{\Sigma}_{11}} & \underset{(p\times q)}{\boldsymbol{\Sigma}_{12}} \\ \hline \underset{(q\times p)}{\boldsymbol{\Sigma}_{21}} & \underset{(q\times q)}{\boldsymbol{\Sigma}_{22}} \end{array}\right] \quad \cdots \quad (4)$$

The covariances between pairs of variables from different sets – one variable from $\mathbf{X}^{(1)}$, one variable from $\mathbf{X}^{(2)}$ – are contained in Σ_{12} or, equivalent, in Σ_{12}. That is, the pq elements of Σ_{12} measure the association between the two sets. When p and q are relatively large, interpreting the elements of Σ_{12} collectively is ordinarily hopeless (Johnson and Wichern; 1992). Moreover, it is often linear combinations of variables that are interesting and useful predictive or comparative purposes. The main task of canonical correlation analysis is to summarize the associations between the $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ sets in terms of a few carefully chosen covariance (or correlations) rather than the pq covariance in Σ_{12}.

Linear combinations $\mathbf{Z} = \mathbf{CX}$ have

$$\left.\begin{array}{l} \boldsymbol{\mu}_{\mathbf{z}} = E(\mathbf{Z}) = E(\mathbf{CX}) = \mathbf{C}\boldsymbol{\mu}_X \\ \Sigma_{\mathbf{Z}} = Cov(\mathbf{Z}) = Cov(\mathbf{CX}) = \mathbf{C}\Sigma_X\, C' \end{array}\right\} \quad \ldots (5)$$

and provide simple summary measures of a set of variables.

$$\text{Let} \quad \begin{array}{l} \mathbf{U} = \mathbf{a}'\mathbf{X}^{(1)} \\ \mathbf{V} = \mathbf{b}'\mathbf{X}^{(2)} \end{array} \quad \ldots (6)$$

for some pair of coefficient vectors **a** and **b**. Using Equations (5) and (6),

$$\left.\begin{array}{l} Var(\mathbf{U}) = \mathbf{a}'Cov(\mathbf{X}^{(1)})\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a} \\ Var(\mathbf{V}) = \mathbf{b}'Cov(\mathbf{X}^{(1)})\mathbf{b} = \mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b} \\ Cov(\mathbf{U},\mathbf{V}) = \mathbf{a}'Cov(\mathbf{X}^{(1)},\mathbf{X}^{(2)})\mathbf{b} = \mathbf{a}'\Sigma_{12}\mathbf{b} \end{array}\right\} \quad \ldots (7)$$

We shall seek coefficient vectors **a** and **b** such that

$$Corr(\mathbf{U},\mathbf{V}) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b}}} \quad \ldots (8)$$

is as large as possible. We then define:

The first pair of canonical variables are the pair of linear combinations $\mathbf{U}_1$, $\mathbf{V}_1$ having unit variances, which maximize the correlation in Equation (8);

The second pair of canonical variables are the linear combinations $\mathbf{U}_2$, $\mathbf{V}_2$ having unit variances, which maximize the correlation in Equation (8) among all choices which are uncorrelated with the first pair of canonical variables.

At the kth step

The kth pair of canonical variables are the linear combinations $\mathbf{U}_k$, $\mathbf{V}_k$ having unit variances, which maximize the correlation in Equation (8) among all choices uncorrelated with the previous k – 1 canonical variable pairs.

The correlation between the kth pair of canonical variables is called the kth canonical correlation. If the original variables are standardized with $\mathbf{Z}^{(1)}$ = $\left[\mathbf{Z}_1^{(1)}, \mathbf{Z}_2^{(2)}, \ldots, \mathbf{Z}_p^{(1)}\right]'$ and $\mathbf{Z}^{(2)} = \left[\mathbf{Z}_1^{(1)}, \mathbf{Z}_2^{(2)}, \ldots, \mathbf{Z}_q^{(1)}\right]'$ from first principles, the canonical varaites are of the form

$$\left.\begin{aligned} \mathbf{U}_k &= \mathbf{a}'_k\mathbf{Z}^{(1)} = \mathbf{e}'_k\boldsymbol{\rho}_{11}^{-1/2}\mathbf{Z}^{(1)} \\ \mathbf{V}_k &= \mathbf{b}'_k\mathbf{Z}^{(2)} = \mathbf{f}'_k\boldsymbol{\rho}_{22}^{-1/2}\mathbf{Z}^{(2)} \end{aligned}\right\} \qquad \ldots(9)$$

Here Cov($\mathbf{Z}^{(1)}$) = ρ_{11}, Cov($\mathbf{Z}^{(2)}$) = ρ_{22}, Cov($\mathbf{Z}^{(1)}$, $\mathbf{Z}^{(2)}$) = ρ_{12} = ρ_{21} and $\mathbf{e}_k$ and $\mathbf{f}_k$ are the eigenvectors of $\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}$ and $\boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{22}\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1/2}$, respectively. The canonical correlations, ρ_k^*, satisfy

$$Corr(\mathbf{U}_k, \mathbf{V}_k) = \boldsymbol{\rho}_k^*, \quad k = 1, 2, \ldots, p \qquad \ldots(10)$$

where $\rho_1^{*2} \geq \rho_2^{*2} \geq \ldots \geq \rho_p^{*2}$ are the nonzero eigenvalues of the matrix $\boldsymbol{\rho}_{11}^{-1/2}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}\boldsymbol{\rho}_{11}^{-1/2}$ (or equivalently, of $\boldsymbol{\rho}_{22}^{-1/2}\boldsymbol{\rho}_{22}\boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1/2}$).

It should be noted that:

$$a'_k(X^{(1)} - \mu^{(1)}) = a_{k1}(X_1^{(1)} - \mu_1^{(1)}) + a_{k2}(X_2^{(1)} - \mu_2^{(1)}) + \cdots + a_{kp}(X_p^{(1)} - \mu_p^{(1)})$$

$$= a_{k1}\sqrt{\sigma_{11}}\frac{(X_1^{(1)} - \mu_1^{(1)})}{\sqrt{\sigma_{11}}} + a_{k2}\sqrt{\sigma_{22}}\frac{(X_2^{(1)} - \mu_2^{(1)})}{\sqrt{\sigma_{22}}} + \ldots + a_{kp}\sqrt{\sigma_{pp}}\frac{(X_p^{(1)} - \mu_p^{(1)})}{\sqrt{\sigma_{pp}}}$$

where $\mathrm{Var}\left(X_i^{(1)}\right) = \sigma_{ii}$, i = 1, 2, …, p. Therefore, the canonical coefficients for the standardized variables, $Z_i^{(1)} = (X_i^{(1)} - \mu_i^{(1)})/\sqrt{\sigma_{ii}}$, are simply related to the canonical coefficients attached to the original variables $X_i^{(1)}$. Specifically, if a′$_k$ is the coefficient vector for the kth canonical variate U$_k$, then $a'_kV_{11}^{1/2}$ is the coefficient vector for the canonical variate constructed from the standardized variables Z$^{(1)}$. Here $V_{11}^{1/2}$ is the diagonal matrix with ith diagonal element $\sqrt{\sigma_{ii}}$. Similarly, $b'_kV_{22}^{1/2}$ is the coefficient vector for the canonical varaite constructed from the set of standardized variables Z$^{(2)}$. In this case $V_{22}^{1/2}$ is the diagonal matrix with ith diagonal element $\sqrt{\sigma_{ii}} = \sqrt{\mathrm{Var}(X_i^{(2)})}$. The canonical correlations are unchanged by the standardization. However, the choice of the coefficient vectors a$_k$, b$_k$ will not be unique if $\rho_k^* = \rho_{k+1}^{*2}$.

Identifying the Canonical Variables

Even though the canonical variables are artificial, they can often be "identified" in terms of the subject matter variables. This identification is often aided by computing the correlations between the canonical variates and the original variables. These correlations, however, must be interpreted with caution. They only provide univariate information in the sense that they do not indicate how the original variables contribute jointly to the canonical analyses. For this reason, many investigators prefer to assess the contributions of the original variables directly from the standardized coefficients in Equation (9).

Let $\underset{(p\times p)}{\mathbf{A}}$ = [a$_1$, a$_2$, …, a$_p$]′ and $\underset{(q\times q)}{\mathbf{B}} = [b_1, b_2, \ldots, b_q]'$, so

that the vectors of canonical variables are

$$\underset{(p\times 1)}{\mathbf{U}} = \mathbf{AX}^{(1)}, \quad \underset{(q\times 1)}{\mathbf{V}} = \mathbf{BX}^{(2)}, \qquad \ldots \qquad (11)$$

where we are primarily interested in the first p canonical variables in $\mathbf{V}$. Then

$$\mathrm{Cov}(\mathbf{U}, \mathbf{X}^{(1)}) = \mathrm{Cov}(\mathbf{AX}^{(1)}, \mathbf{X}^{(1)}) = \mathbf{A}\Sigma_{11} \qquad \ldots \qquad (12)$$

Because Var(U$_i$) = 1, $\mathrm{Corr}(U_i, X_k^{(1)})$ is obtained by dividing $\mathrm{Cov}(U_i, X_k^{(1)})$ by $\sqrt{\mathrm{var}(X_k^{(1)})} = \sigma_{kk}^{1/2}$. Equivalently, $\mathrm{Corr}(U_i, X_k^{(1)})$ = $\mathrm{Cov}(U_i, \sigma_{kk}^{1/2}X_k^{(1)})$. Introducing the (p × p) diagonal matrix $V_{kk}^{-1/2}$ with kth diagonal element $\sigma_{kk}^{-1/2}$, we have, in matrix terms,

$$\underset{(p\times p)}{\rho_{U,X^{(1)}}} = \mathrm{Corr}(U, X^{(1)}) = \mathrm{Cov}(U, V_{11}^{-1/2}X^{(1)}) = \mathrm{Cov}(AX^{(1)}, V_{11}^{-1/2}X^{(1)})$$

$$A\Sigma_{11}V_{11}^{-1/2}$$

Similar calculations for the pairs (U, X$^{(2)}$), (V, X$^{(2)}$) and (V, X$^{(1)}$) yield

$$\left.\begin{array}{ll} \underset{(p\times p)}{\rho_{U,X^{(1)}}} = A\Sigma_{11}V_{11}^{-1/2} & \underset{(q\times q)}{\rho_{V,X^{(2)}}} = A\Sigma_{22}V_{22}^{-1/2} \\ \underset{(p\times q)}{\rho_{U,X^{(1)}}} = A\Sigma_{12}V_{22}^{-1/2} & \underset{(q\times p)}{\rho_{V,X^{(2)}}} = A\Sigma_{21}V_{11}^{-1/2} \end{array}\right\} \qquad \ldots(13)$$

where $V_{22}^{-1/2}$ is the (q × q) diagonal matrix with ith diagonal element $\sqrt{\mathrm{var}(X_i^{(2)})}$.

Canonical variables derived from standardized variables are sometimes interpreted by computing the correlations.

$$\left.\begin{array}{ll}\rho_{\mathbf{U},\mathbf{Z}^{(1)}} = \mathbf{A}_z\rho_{11}; & \rho_{V,Z^{(2)}} = \mathbf{B}_z\rho_{22}\\ \rho_{\mathbf{U},\mathbf{Z}^{(2)}} = \mathbf{A}_z\rho_{12}; & \rho_{V,Z^{(1)}} = \mathbf{B}_z\rho_{21}\end{array}\right\} \quad \ldots(14)$$

where $\underset{(p\times p)}{A_z}$ and $\underset{(q\times q)}{B_z}$ are the matrices whose rows contain the canonical coefficients for the $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ sets, respectively. The correlations in the matrices displayed is Equation (14) have the same numerical values as those appearing in Equation (13), that is, $\rho_{\mathrm{U,X}^{(1)}} = \rho_{\mathrm{V,Z}^{(1)}}$ and so forth. This follows because, for example,

$$\rho_{\mathrm{U,X}^{(1)}} = \mathrm{A}\Sigma_{11}\mathrm{V}_{11}^{-1/2} = \mathrm{A}\mathrm{V}_{11}^{-1/2}\mathrm{V}_{11}^{-1/2}\,\Sigma_{11}\mathrm{V}_{11}^{-1/2} = \mathrm{A}_z\rho_{11} = \rho_{\mathrm{U,Z}^{(1)}}\,.$$

The correlations are unaffected by the standardization.

The Sample Canonical Variates and Sample Canonical Correlations

A random sample of n observations on each of the (p + q) variables $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$ can be assembled into the ((p + q) × n) data matrix

$$\mathbf{X} = \left[\begin{array}{c}\mathbf{X}^{(1)}\\ \hdashline \mathbf{X}^{(2)}\end{array}\right] = \left[\begin{array}{cccc} x_{11}^{(1)} & x_{12}^{(1)} & \cdots & x_{1n}^{(1)}\\ x_{21}^{(1)} & x_{22}^{(1)} & \cdots & x_{2n}^{(1)}\\ \vdots & \vdots & \ddots & \vdots\\ x_{p1}^{(1)} & x_{p2}^{(1)} & \cdots & x_{pn}^{(1)}\\ \hdashline x_{11}^{(2)} & x_{12}^{(2)} & \cdots & x_{1n}^{(2)}\\ x_{21}^{(2)} & x_{22}^{(2)} & \cdots & x_{2n}^{(2)}\\ \vdots & \vdots & \ddots & \vdots\\ x_{q1}^{(2)} & x_{q2}^{(2)} & \cdots & x_{qn}^{(2)}\end{array}\right] = \left[x_1, x_2, \ldots, x_n\right]$$

where $$x_j = \begin{bmatrix} x_j^{(1)}\\ x_j^{(2)}\end{bmatrix} \qquad \ldots \quad (15)$$

The vector of sample means can be organized as

$$\underset{(p+q)\times 1}{\bar{x}} = \left[\begin{array}{c}\bar{x}^{(1)}\\ \hdashline \bar{x}^{(2)}\end{array}\right] \text{ where } \left.\begin{array}{l}\bar{x}^{(1)} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} x_j^{(1)}\\ \bar{x}^{(2)} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} x_j^{(2)}\end{array}\right\} \quad \ldots(16)$$

Similarly, the sample covariance matrix can be arranged analogous to the representation in Equation (4). Thus

$$\underset{(p+q)(p+q)}{\mathbf{S}} = \left[\begin{array}{c:c}\underset{(p\times p)}{\mathbf{S}_{11}} & \underset{(p\times q)}{\mathbf{S}_{12}}\\ \hdashline \underset{(q\times p)}{\mathbf{S}_{21}} & \underset{(q\times q)}{\mathbf{S}_{22}}\end{array}\right]$$

where

$$\mathbf{S}_{kl} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j^{(k)} - \bar{x}^{(k)}\right)\left(X_j^{(l)} - \bar{x}^{(l)}\right)' \qquad \mathrm{k}, l = 1, 2, \qquad \ldots (17)$$

The linear combinations

$$\hat{U} = \hat{a}'x^{(1)}; \qquad \hat{\mathrm{V}} = \hat{\mathrm{b}}'\mathrm{x}^{(2)} \qquad \ldots \qquad (18)$$

have sample correlation

$$r_{\hat{U},\hat{V}} = \frac{\hat{\mathbf{a}}'\mathbf{S}_{12}\hat{\mathbf{b}}}{\sqrt{\hat{\mathbf{a}}'\mathbf{S}_{11}\hat{\mathbf{a}}}\sqrt{\hat{\mathbf{b}}'\mathbf{S}_{22}\hat{\mathbf{b}}}} \qquad \ldots \qquad (19)$$

The first pair of sample canonical variates is the pair of linear combinations $\hat{\mathrm{U}}_1, \hat{\mathrm{V}}_1$ having unit sample variances that maximize the ration in Equation (19).

In general: the kth pair of sample canonical variates is the pair of linear combinations $\hat{\mathrm{U}}_{\mathrm{k}}, \hat{\mathrm{V}}_{\mathrm{k}}$ having unit sample variances that maximize the ratio in Equation (19) among those linear combinations uncorrelated with the previous k – 1 sample canonical variates. The sample correlation between $\hat{\mathrm{U}}_{\mathrm{k}}$ and $\hat{\mathrm{V}}_{\mathrm{k}}$ is called the kth sample canonical correlation.

Data Presentation

The data used for this research was extracted from Neil H.T. (2002), Applied Multivariate Analysis, Exercises 4.3 page 216. A sample of 25 samples of tobacco leaf for organic and inorganic chemical constituents was used for the study. The dependent variables considered are defined as follows:

Y_1 : Rate of cigarette burn in inches per 1000 seconds
Y_2 : Percentage sugar in the leaf
Y_3 : Percentage nicotine

The fixed independent variables are defined as follows.

X_1: Percentage of Nitrogen
X_2: Percentage of Chlorine
X_3: Percentage of Potassium
X_4: Percentage of Phosphorus
X_5: Percentage of Calcium
X_6: Percentage of Magnesium

Table 1 shows the three dependent variables (Organic Chemical constituents) and six Explanatory variables (Inorganic Chemical constituents) of 25 samples of tobacco leaf.

Data Analysis

The organic chemical constituents, $\mathbf{X}^{(1)}$, and the inorganic chemical consttuents $\mathbf{X}^{(2)}$, were defined as:

$$\mathbf{X}^{(1)} = \begin{pmatrix} X_1^{(1)}\\ X_2^{(1)}\\ X_3^{(1)}\end{pmatrix} = \begin{pmatrix}\text{Rate of cigarette burn in inches per 1000 seconds}\\ \text{Percent sugar in the leaf}\\ \text{Percent nicotine}\end{pmatrix}$$

$$\mathbf{X}^{(2)} = \begin{pmatrix} X_1^{(2)}\\ X_2^{(2)}\\ X_3^{(2)}\\ X_4^{(2)}\\ X_5^{(2)}\\ X_6^{(2)}\end{pmatrix} = \begin{pmatrix}\text{Percentage of Nitrogen}\\ \text{Percentage of Chlorine}\\ \text{Percentage of Potassium}\\ \text{Percentage of Phosphorus}\\ \text{Percentage of Calcium}\\ \text{Percentage of Magnesium}\end{pmatrix}$$

Table 1: The Tobacco Data

Subject ID	Dependent variables			Independent variables					
	Y_1	Y_2	Y_3	X_1	X_2	X_3	X_4	X_5	X_6
1	1.55	20.05	1.38	2.02	2.90	2.17	0.51	3.47	0.91
2	1.63	12.58	2.64	2.62	2.78	1.72	0.5	4.57	1.25
3	1.66	18.56	1.56	2.08	2.68	2.40	0.43	3.52	0.82
4	1.52	18.56	2.22	2.20	3.17	2.06	0.52	3.69	0.97
5	1.70	14.02	2.85	2.38	2.52	2.18	0.42	4.01	1.12
6	1.68	15.64	1.24	2.03	2.56	2.57	0.44	2.79	0.82
7	1.78	14.52	2.86	2.87	2.67	2.64	0.5	3.92	1.06
8	1.57	18.52	2.18	1.88	2.58	2.22	0.49	3.58	1.01
9	1.60	17.84	1.65	1.93	2.26	2.15	0.56	3.57	0.92
10	1.52	13.38	3.28	2.57	1.74	1.64	0.51	4.38	1.22
11	1.68	17.55	1.56	1.95	2.15	2.48	0.48	3.28	0.81
12	1.74	17.97	2.00	2.03	2.00	2.38	0.50	3.31	0.98
13	1.93	14.66	2.88	2.50	2.07	2.32	0.48	3.72	1.04
14	1.77	17.31	1.36	1.72	2.24	2.25	0.52	3.10	0.78
15	1.94	14.32	2.66	2.53	1.74	2.64	0.50	3.48	0.93
16	1.83	15.05	2.43	1.90	1.46	1.97	0.46	3.48	0.9
17	2.09	15.47	2.42	2.18	0.74	2.46	0.48	3.16	0.86
18	1.72	16.85	2.16	2.16	2.84	2.36	0.49	3.68	0.95
19	1.49	17.42	2.12	2.14	3.30	2.04	0.48	3.28	1.06
20	1.52	18.55	1.87	1.98	2.90	2.16	0.48	3.56	0.84
21	1.64	18.74	2.10	1.89	2.82	2.04	0.53	3.56	1.02
22	1.40	14.79	2.21	2.07	2.79	2.15	0.52	3.49	1.04
23	1.78	18.86	2.00	2.08	3.14	2.60	0.50	3.30	0.80
24	1.93	15.62	2.26	2.21	2.81	2.18	0.44	4.16	0.92
25	1.53	18.56	2.14	2.00	3.16	2.22	0.51	3.37	1.07

Responses for variables $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ were recorded on a scale and then standardized. The sample correlation matrix based on 25 responses is:

$$\mathbf{R} = \left(\begin{array}{c:c} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \hdashline \mathbf{R}_{21} & \mathbf{R}_{22} \end{array}\right)$$

$$= \left(\begin{array}{ccc:cccccc} 1.000 & & & 0.226 & -0.523 & 0.487 & -0.320 & -0.085 & -0.313 \\ -0.320 & 1.000 & & -0.705 & 0.430 & 0.190 & 0.244 & -0.516 & -0.525 \\ 0.216 & -0.702 & 1.000 & 0.768 & -0.271 & -0.294 & -0.045 & 0.686 & 0.734 \\ 0.226 & -0.705 & 0.768 & 1.000 & & & & & \\ -0.623 & 0.430 & -0.271 & -0.089 & 1.000 & & & & \\ \hdashline 0.487 & 0.190 & -0.294 & -0.007 & -0.093 & 1.000 & & & \\ -0.320 & 0.244 & -0.045 & -0.112 & 0.074 & -0.205 & 1.000 & & \\ -0.085 & -0.516 & 0.686 & 0.604 & 0.095 & -0.583 & -0.009 & 1.000 & \\ -0.313 & -0.525 & 0.734 & 0.604 & 0.118 & -0.611 & 0.514 & 0.729 & 1.000 \end{array}\right)$$

The min(p, q) = min(3, 6) = 3 sample canonical correlations and the sample canonical variate coefficient vectors are displayed in Table 2.

For instance, the first sample canonical variate pair is

$$\hat{U}_1 = -0.095Z_1^{(1)} - 0.439Z_2^{(1)} + 0.665Z_3^{(1)}$$

$$\hat{V}_2 = -0.470Z_1^{(2)} - 0.355Z_2^{(2)} - 0.063Z_3^{(2)} - 0.124Z_4^{(2)} + 0.082Z_5^{(2)} + 0.479Z_6^{(2)}$$

with sample canonical correlation $\hat{\rho}_1^* = 0.933$. The results above were taken from the **SAS** Statistical software output shown in Appendix. To provide interpretation for $\hat{U}_1$ and $\hat{V}_1$, the sample correlations between $\hat{U}_1$ and its component variables and $\hat{V}_1$ and its component variables were computed. Also, we provide the sample correlations between variables in one set and the first sample canonical variate of the other set.

Again, the second sample canonical variate pair is

$$\hat{U}_1 = 0.990\ Z_1^{(1)} - 0.158\ Z_2^{(1)} - 0.342\ Z_3^{(1)}$$

$$\hat{V}_2 = 0.112Z_1^{(2)} - 0.626Z_2^{(2)} + 0.448Z_3^{(2)} - 0.175Z_4^{(2)} + 0.362Z_5^{(2)} - 0.537Z_6^{(2)}$$

with canonical correlations $\hat{\rho}_2^{x} = 0.842$

The sample correlations between $\hat{U}_2$ and its component variables and $\hat{V}_2$ and its component variables were computed, and presented in Table 4.

Estimating Proportions of Explained Sample Variance

Using the table of sample correlation coefficients presented in Table 3, we compute

$$R^2_{Z^{(1)}/\hat{U}_1} = \frac{1}{3}\sum_{k=1}^{3} \gamma^{2}_{\hat{U}_{1,Z_k^{(1)}}} = \frac{1}{3}\left[(0.189)^2 + (-0.876)^2 + (0.953)^2\right]$$
$$= 0.570$$

$$R^2_{Z^{(2)}/\hat{V}_1} = \frac{1}{6}\sum_{k=1}^{6} \gamma^{2}_{\hat{V}_{1,Z_k^{(2)}}} = \frac{1}{6}\left[(0.799)^2 + (-0.310)^2 + \ldots + (0.680)^2\right]$$
$$= 0.312$$

$$R^2_{Z^{(1)}/\hat{V}_1} = \frac{1}{3}\sum_{k=1}^{3} \gamma^{2}_{\hat{V}_{1,Z_k^{(1)}}} = \frac{1}{3}\left[(0.176)^2 + (-0.817)^2 + (0.88)^2\right]$$
$$= 0.496$$

$$R^2_{Z^{(2)}/\hat{U}_1} = \frac{1}{6}\sum_{k=1}^{6} \gamma^{2}_{\hat{U}_{1,Z_k^{(2)}}} = \frac{1}{6}\left[(0.857)^2 + (-0.332)^2 + \ldots + (0.802)^2\right]$$
$$= 0.359$$

The first sample canonical variate, $\hat{U}_1$, of the organic chemical constituents set accounts for 57% of the set's total sample variance. The next sample canonical variate, $\hat{V}_1$, of the organic chemical constituents set accounts for 49.6% of the set's total sample variance. The first sample canonical variates, $\hat{V}_1$ and $\hat{U}_1$, of the inorganic chemical constituents set explains 31.2% and 35.9% respectively of the set's total

Table 2: Canonical Variate Coefficients and Canonical Correlations

	Standardized variables					Standardized variables					
	$Z_1^{(1)}$	$Z_2^{(1)}$	$Z_3^{(1)}$	$\hat{\rho}_i^2$		$Z_1^{(2)}$	$Z_2^{(2)}$	$Z_3^{(2)}$	$Z_4^{(2)}$	$Z_5^{(2)}$	$Z_6^{(2)}$
$\hat{\mathbf{a}}_1'$:	-0.095	-0.439	0.665	0.933	$\hat{\mathbf{b}}_1'$:	0.470	-0.355	-0.063	-0.124	0.082	0.479
$\hat{\mathbf{a}}_2'$	0.990	-0.158	-0.342	0.842	$\hat{\mathbf{b}}_2'$	0.112	-0.626	0.448	-0.175	0.362	-0.537
$\hat{\mathbf{a}}_3'$	0.354	1.370	1.189	0.373	$\hat{\mathbf{b}}_3'$	-1.107	-0.143	1.497	0.551	1.552	0.442

Table 3: Sample Correlations between Original Variables and Canonical Variables

		Sample canonical variates				Sample canonical variates	
	$X^{(1)}$ variables	$\hat{U}_1$	$\hat{V}_1$		$X^{(2)}$ variables	$\hat{U}_1$	$\hat{V}_1$
1	Rate of cigarette burn in inches per 1000 seconds	0.189	0.176	1	Percentage of Nitrogen	0.857	0.799
2	Percent sugar in the leaf	-0.876	-0.817	2	Percentage of Chlorine	-0.332	-0.310
3	Percent nicotine	0.953	0.889	3	Percentage of Potassium	-0.349	-0.325
				4	Percentage of Phosphorus	-0.115	-0.107
				5	Percentage of Calcium	0.729	0.680
				6	Percentage of Magnesium	0.802	0.749

Table 4: Correlations between Original Variables and Canonical Variables

		Sample canonical variates				Sample canonical variates	
	$X^{(1)}$ variables	$\hat{U}_2$	$\hat{V}_2$		$X^{(2)}$ variables	$\hat{U}_2$	$\hat{V}_2$
1	Rate of cigarette burn in inches per 1000 seconds	0.967	0.814	1	Percentage of Nitrogen	0.086	0.073
2	Percent sugar in the leaf	-0.235	-0.198	2	Percentage of Chlorine	-0.704	-0.593
3	Percent nicotine	-0.017	-0.015	3	Percentage of Potassium	0.656	0.552
				4	Percentage of Phosphorus	-0.404	-0.340
				5	Percentage of Calcium	-0.331	-0.278
				6	Percentage of Magnesium	-0.568	-0.478

sample variance. We might infer that $\hat{U}_1$ of the organic chemical constituents is a "better" representative of its set than $\hat{U}_1$ of the inorganic chemical constituents is of its set.

Using the table of sample correlation coefficients presented in Table 4, we compute

$$R^2_{Z^{(1)}/\hat{U}_2} = \frac{1}{3}\sum_{k=1}^{3} \gamma^2_{\hat{U}_{2},Z_k^{(1)}} = 0.330$$

$$R^2_{Z^{(2)}/\hat{U}_2} = \frac{1}{6}\sum_{k=1}^{6} \gamma^2_{\hat{U}_{2},Z_k^{(2)}} = 0.255$$

Test of Significance of the Canonical Correlation

The first two canonical correlations, ρ_1^* and ρ_2^*, appear to be nonzero, small deviations from zero will show up as statistically significant. From a practical point of view, the third sample canonical correlation can probably be ignored since (i) it is reasonably small in magnitude and (ii) the corresponding canonical variate explains very little of the sample variation in the variable sets $X^{(1)}$ and $X^{(2)}$. Thus from the SAS output in Appendix, the p-values for both the first and second canonical correlation are small, implying that they are significant, while the third canonical correlation is insignificant because of the high p-value observed. The SAS output revealed that there is a relationship between the organic chemical constituents and the inorganic chemical constituents.

Conclusion

In light of the discussion in the analysis above, it is desirable to conclude that $\hat{U}_1$ of the organic chemical constituents is a "better" representative of its set than $\hat{U}_1$ of the inorganic chemical constituents is of its set. Again, it can be concluded that canonical relations exhibited by the organic chemical constituents-inorganic chemical constituents' data proved statistically significant in the first two canonical correlations, and statistically insignificant in the last (third) canonical correlation. Finally, we concluded that relationship exists between the organic chemical constituents and the inorganic chemical constituents of the Tobacco leaf samples.

REFERENCES

Hoerl, A.E. (1962): Application of Ridge Analysis to Regression Problems. Chemical Engineering Progress. 58, 54-59.

Johnson, R.A. and Wichern, D.W. (1992): Applied Multivariate Statistical Analysis, prentice Hall, Englewood Cliffs, New Jersey.

Johnson, R.A. and Wichern, D.W. (2007): Applied Multivariate Statistical Analysis, Sixth Edition. Pearson Prentice Hall, New Jersey.

Lee, W., Lee D, Lee Y., and Pawitan, Y. (2011): Sparse Canonical Covariance Analysis for High-throughput Data. Statistical Applications in Genetics and Molecular Biology: Vol. 10: Iss. 1, Article 30.

Lee, W., Lee, D, Lee Y., and Pawitan, Y. (2011): Sparse canonical covariance analysis. R package .http://www.meb.ki.se/yudpaw/.

Neil, H.T. (2002): Applied Multivariate Analysis. Springer-Verlag, New York Berlin Heidelberg.

Parkhomenko, E., Tritchler, D., and Beyene, J. (2009): Sparse Canonical Correlation Analysis with Application to Genomic Data Integration. Statistical Applications in Genetics and Molecular Biology: Vol. 8: Iss.1, Article 1.60.

Waaijenborg S., Verselewel. D.E., Witt, Hamer P.C., and Zwinderman, A.H. (2008): Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. Statistical Applications in Genetics and Molecular Biology: Volume 7: Issue 1, Article 3.

William, R.D. and Matthew, G. (1984): Multivariate Analysis, Methods and Applications. John Wiley & Sons, New York.

Witten, D.M. and Tibshirani, R. (2009) Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. Statistical Applications in Genetics and Molecular Biology: Vol. 8: Iss. 1, Article 28.

****