Multivariate Analysis: Greater Insights into Complex Systems

Kathleen M. Yeater,* Sara E. Duke, and Walter E. Riedell

ABSTRACT

Many agronomic researchers measure and collect multiple response variables in an effort to understand the more complex nature of the system being studied. Multivariate (MV) statistical methods encompass the simultaneous analysis of all random variables measured on each experimental or sampling unit. Many agronomic research systems studied are, by their very nature, MV; however, most analyses reported are univariate (analysis of one response at a time). The objective of this review is to outline a statistical foundation of applications of MV methods and techniques for the agronomic sciences. By utilizing two agronomic data sets, both typical in dimension and structure, we discuss three classes of MV techniques based on the research question and characteristics of the data: (i) hypothesis driven, such as MV analysis of variance; (ii) dimension reduction, such as principal components analysis; and (iii) classification and discrimination, which includes canonical discriminant analysis. Several advantages and disadvantages of the MV tools are explained. This review will provide researchers with a beginning framework of MV generalizations of univariate techniques, and methods that are unique to MV dimension analysis. It is important for researchers to capture the concept of variability within a MV data set to better understand the complex system.

The origins of applied multivariate (MV) statistical methods arose from the fields of chemistry, psychology, sociology, ecology, and climatology—all multifactor, complex systems with interactions and associations among variables and units (Johnson and Wichern, 2002; Quinn and Keough, 2002; Tabachnick and Fidell, 2007). Like these other systems, agricultural systems are complex, with multiple factors driving ecosystem functions and ecological communities, from soil microbiota to weed or vegetation communities and from soil texture to functional classification of landscapes (Digby and Kempton, 1987; Gauch, 1982; Kenkel, 2006; Kenkel et al., 2002).

In general, agricultural studies have been and continue to be evaluated within one class of statistical approaches: based on hypothesis-driven experiments, with one response variable analyzed at a time. Multivariate analysis refers to a broad category of methods used when multiple response variables are measured on a set of experimental units or sampling objects. Researchers intentionally measure multiple factors based on their subject matter knowledge of the system, presumably with the goal of revealing associations and interrelatedness between and within sets of variables to gain greater insight into the system’s function and behavior. Scientific inquiry has expanded far beyond the simple controlled study; data sets are becoming larger and more complex. Understanding the complexity of biotic and abiotic systems is better accomplished with MV thinking, allowing questions to be addressed that may otherwise be difficult to conceive by univariate approaches. The desired inferential outcome to a high-dimensional, data-rich study requires approaching analyses with applicable statistical methods, namely MV statistical methods.

Multivariate methods have been used and published in the crops, soils, and agronomic literature but are the exception rather than the rule. Riedell et al. (2009) and Badu-Apraku et al. (2011) revealed relationships among multiple environments, several agronomic traits, and grain yield utilizing canonical discriminant analysis (CDA), principal components analysis (PCA), and cluster analysis; Riedell et al. (2007) and Francisco et al. (2011) investigated various nutritive measures in grains and legumes exploring the results of canonical correlation analysis (CCA) and PCA; Schutter and Dick (2002) (PCA and CDA) and Lupwayi et al. (2011) (PCA) revealed interrelationships among soil community structures. Soil fertility and abiotic measures across temporal-spatial changes were research topics presented by Jaradat and Weyers (2011)

Abbreviations: CCA, canonical correlation analysis; CDA, canonical discriminant analysis; DA, discriminant analysis; MANOVA, multivariate analysis of variance; MV, multivariate; PC, principal component; PCA, principal components analysis; RV, random variable.
(partial least squares) and Mohammadi et al. (2010) (PCA, pattern analysis); and Dieleman et al. (2000a) utilized CCA to identify associations between site properties and abundance of weed species in agricultural fields and then in a second study (Dieleman et al., 2000b) addressed specific hypotheses generated from the initial analysis.

We have outlined a statistical foundation that describes three classes of MV methods and provides an introductory tutorial for the analyses of these methods using SAS software (SAS Institute, 2013). The focus is on discussion of the intuitive concepts of these methods rather than on their mathematical rigor. For greater mathematical development, see Johnson and Wichern (2002). The presentation of these classes of MV techniques is based on the characteristics of the data and the research goals: (i) hypothesis-driven comparison of treatment groups affected by experimental treatment structure utilizing MV analysis of variance (MANOVA); (ii) dimension reduction techniques such as PCA; and (iii) classification and discrimination techniques—CDA is presented here, using an approach similar to that used for the original presentation of the data by Riedell et al. (2009). Each of these methods is based on aspects of linear modeling concepts and the assumption of linear relationships among variables, where the familiar statistical representation is $Y = X\beta + \varepsilon$ for both analysis of variance (ANOVA) and regression models in the univariate paradigm, with one $Y$ variable and one or more $X$ variables.

Linear modeling techniques lend greater ease to solving complex problems because of linear matrix algebra mathematics. Mathematical notation used in MV linear algebra has some consistency; however, each author may have slight variations. Here, $p$ represents the number of response or outcome variables, with $Y$ in the equation above representing multiple responses. The $X$ represents the design matrix, which may contain categorical and/or continuous predictor variables or their corresponding design matrices. In MV analysis, it is common practice to use uppercase boldface letters for matrices and lowercase boldface letters for vectors. A common principal underlying MV methods is to reduce the dimensionality of a large data set by finding linear combinations of variables that optimally capture information (variability) in the data. Explaining variability is the basis for developing the theory of a system and can lead to predictive capabilities. Mathematically, linear combinations of variables, e.g., $(Y_1, Y_2, \ldots, Y_p)$ as in a multiple regression) can be expressed as

$$Z = c_1 Y_1 + c_2 Y_2 + \ldots + c_p Y_p = \sum_{j=1}^{p} c_j Y_j = c'Y$$

where $c'$ is a row vector (the transpose of a column vector) of coefficients and $Y$ is the matrix of the $p$ columns $\times n$ observations, each representing one of the $Y$ variables. These generated linear combinations are useful for the interpretation of some underlying dimension or structure that might predict important relationships associated with the system under study.

The above equation is similar conceptually to a multiple regression with many explanatory or “predictor” variables. The MV analysis linear combinations, or composites of variables, are interpreted as representing some underlying structure or dimension in the system that was not or could not be directly measured; therefore, they are called canonical variates, latent structures, or factors. Furthermore, there are multiple factors generated that cumulatively explain all of the variability in the data; however, only a few are used for interpretation based on the amount of variability explained by each factor. Thus, the latent structures that are produced reduce the dimension of the original data set, yet they capture the majority of the variability in the data. Further, they have the optimal characteristic of being independent of one another and represent an orthogonal axis in the multidimensional data space.

**STRUCTURE OF MULTIVARIATE DATA**

Random variables (RVs) are defined as measured responses, whether dependent outcomes or independent predictors, that exhibit variation due to conditions and uniqueness. The dimension of a MV data set is the number of independent observations or sampling units ($n$ rows) by the number of RVs measured ($p$ columns). An arbitrary data matrix $A$ is composed of column vectors representing variables $a_1$ to $a_p$, measured and rows representing each replicate observation, and each measured element is represented in a rectangular array of dimensions $n$ by $p$:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix}$$

Each RV (column) is summarized by its mean, but in a collective representation termed a mean vector, $\mu = \bar{a}_1, \ldots, \bar{a}_p$, and a square-symmetric $p \times p$ variance–covariance matrix:

$$\text{Cov}(A) = S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

where the diagonal elements are the variance of each variable and the off-diagonal elements are the covariance between variables.

The mean vector $\mu$, when plotted in multidimensional space, is termed the centroid of the data cloud. Because variance and covariance are dependent on the scale of the original variable, it is standard practice to use the correlation matrix as a mean-centered, variance-standardized version of the variance–covariance matrix:

$$\text{Cov}(A) = S = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

Recall that the correlation is defined as $r_{ij} = \text{cov}(a_i, a_j)/\sqrt{\text{var}(a_i)\text{var}(a_j)}$.

A generalized summary of the overall variance represented in the covariance matrix is the determinant. This single number is an index that shows how much the variables in the data set differ. For the correlation matrix, the determinant ranges from 0 to 1. If variables are strongly related, the variation among them is low (redundant information) and the determinant will be close to zero.
A goal of MV statistics is to account for sources of variability in high-dimensional space (three dimensions or higher). Multivariate methods involve some form of the variances and covariances. As mentioned above, using linear combinations or composites of the original variables is a method for describing the variability in a MV data set. This method of variance analysis is broadly called eigenanalysis—a mathematical discipline central to linear matrix algebra. Essentially, eigenanalysis is a method of redistributing the original total variance of variables through linear composites of variables that account for major directions of variation in the data. The result of eigenanalysis is a series of paired eigenvalues and eigenvectors. Each eigenvector defines a new axis of variability in the coordinate system (data space), and each eigenvalue measures the strength of the axis and is reported with the amount of variability it explains. This is the primary method for finding the latent structures.

**EXPERIMENTAL DATA USED FOR EXAMPLES**

Two agronomic MV data sets are used here to illustrate the methods discussed. For the MANOVA example, the data set is from a 2-yr experiment conducted to determine how leaf mineral nutrients (N, P, K, Ca, Mg, Zn, Fe, Mn, and Cu concentrations), chlorophyll content, and agronomic traits (leaf area index and yield) of oat (*Avena sativa* L.) responded to crop stresses imposed by aphid feeding treatments: greenbugs (*Schizaphis graminum* Rondani), Russian wheat aphids (*Diuraphis noxia* Mordvilko), bird cherry-oat aphids (*Rhopalosiphum padi* L.), an aphid-vectored virus (barley yellow dwarf virus, *Luteovirus*, PAV strain), and a control that received no aphid infestations or virus infections (Riedell et al., 2007). The experimental design was completely randomized for three replicates each year. For the purposes of our MANOVA example, there were 30 sample observations (*n* = 30) and 12 measured response variables (*p* = 12). Riedell et al. (2007) extended their MV analysis to include CCA, which revealed that certain leaf mineral nutrients were correlated with the agronomic traits, and therefore a large proportion of variation in the agronomic traits was explained by these mineral nutrients.

For the examples illustrating PCA and CDA, the data set is from an experiment conducted to determine how maize (*Zea mays* L.) would respond to monoculture (C–C), a 2-yr rotation (C–S) with soybean (*Glycine max* L. Merr.), or a 4-yr rotation (C–S–W/A–A) with soybean, wheat (*Triticum aestivum* L.), and alfalfa (*Medicago sativa* L.) under different N rates (Riedell et al., 2009). The study design evaluated three rates of N fertilizer (8.5 or 5.3 Mg/ha yield goal or no N) and three crop rotation (C–C, C–S, or C–S–W/A–A) treatment effects. The experimental design consisted of three replicate blocks of the three crop rotations. The whole-plot experimental units, rotations, were arranged in a randomized complete block design repeated for 2 yr. The three N rates were subplot treatments randomly assigned to each whole plot. In total, there were 54 sample observations (*n* = 54) and 28 measured response variables (*p* = 28). Response variables can be classified into three separate data sets on soil mineral (NO$_3$, P, K, Ca, Mg, Fe, Mn, and Zn concentrations) and plant nutrient composition (N, P, K, Ca, Mg, Fe, Mn, and Zn concentrations) and grain characterization (grain yield and grain composition [oil, starch and N, P, K, S, Ca, Mg, Fe, Mn, and Zn concentrations]).

**MULTIVARIATE APPLICATIONS AND INTERPRETATIONS**

**Multivariate Analysis of Variance**

The most fundamental analysis technique familiar to almost all life science and natural environment researchers is the ANOVA. In general, this procedure has become the cornerstone data analysis method of much of the agronomic and related fields, where the focus is on testing hypotheses regarding mean differences among treatments. In univariate analyses, like ANOVA or a *t*-test, each analysis is performed on a single response variable at a time, where the variance is partitioned into sums of squares and adjusted for degrees of freedom resulting in mean square errors (MSE). Sums of squares between “treatment” groups (SS$_b$/w) may explain a significant amount of the variation in the response variable relative to the within-groups sums of squares (SS$_w$/in, i.e., residual error) by testing the variance testing ratio of MSE$_b$/w/MSE$_w$/in with an *F* statistic. This is also characterized by the familiar univariate linear model $y = b_0 + b_1x + e$, where *y* is a single continuous response (dependent) variable vector and *x* is a single, fixed experimental (independent) treatment factor. This model can also represent a simple linear regression if *x* represents a continuous predictor variable. For the MV scenario, the MANOVA statistical linear model is represented as $y = XB + E$, where *Y* is the matrix of response variables, *X* is the design matrix of fixed explanatory variables, *B* is the matrix of parameter estimates, and *E* is the residual error variance–covariance matrix, assumed to be the MV normal probability distribution. Multivariate ANOVA utilizes two data matrices, one response or outcome matrix, *Y*, of dimension $n \times p$ and one explanatory or predictor data matrix *X*, of dimension $n \times m$ (where there are *m* hypothesis factors or explanatory classification variables and *n* observations).

When the research question pertains directly to the comparison of groups, MANOVA is the appropriate MV analysis approach. Multivariate ANOVA is recommended to evaluate the MV response profile in a controlled experimental study, to control for the chance of Type I error rate in the standard univariate analysis of each variable. (Recall that each univariate test evaluating the null hypothesis [H$_0$] of no treatment effect at a Type I significance level [such as at $\alpha = 0.05$] is based on a long-run average of rejecting H$_0$ incorrectly 5% of the time if all treatment means are truly equal; therefore, the more univariate H$_0$ tests performed, the greater the chance of making a Type I error within the whole study.) Multivariate ANOVA accounts for the shared information among response variables (correlations) simultaneously and thus can account for the joint “effect” of predictors on the joint MV response, which might otherwise be missed in separate univariate analyses (Brown et al., 2012). The comparison of mean vector responses (centroids) is the unit of comparison in a MANOVA, taking into account the relationships between the predictor and response variables.

The MANOVA model is a linear additive model that partitions the response data matrix variance into three matrices comprised of sums-of-squares and cross-product (SSCP) analogs to univariate ANOVA:

1. *H* matrix, analogous to SS$_b$/w, that corresponds to the effect hypotheses defined in the model
2. *E* matrix, the residual error matrix analogous to the SS$_w$/in
3. *T* matrix, the total variability matrix, analogous to corrected sums of squares, SS$_{total}$
These matrices are utilized to test whether the hypothesized effects are patterned in the higher dimensional space defined by the multiple outcomes.

There are four test statistics generally used in interpretation of a MANOVA that utilize the SSCP matrices in ratios of variances from different sources of variability (Table 1):

- Wilks’ lambda ($\lambda$) is the most commonly used test statistic and is the ratio of the generalized error or residual variance matrix $E$ to the generalized total variance matrix $T$. This test value is an expression of the amount of variability not explained by the proposed model of hypothesized factors; smaller values lead to rejection of the null hypothesis of no effect due to the fixed factors in the model.
- Pillai’s trace is the sum of the diagonal elements of the matrix containing the ratios of the variability explained by the hypothesized model to the total variability in the data; this is conceptually similar to the univariate $R^2$ statistic ($SS_b/SS_{total}$).
- Hotelling–Lawley trace is the sum of the diagonal elements of the matrix containing the ratios of the variability explained by the hypothesized model to the error (residual) variability in the data; this is conceptually similar to the univariate $F$ statistic, $MSE_b/MSE_w$.
- Roy’s greatest root is the largest eigenvalue from the $E^{-1}H$ matrix. Recall that the first and largest eigenvalue represents the maximum variance explained by the linear combination of the response variables.

In SAS, a MANOVA is performed using the GLM procedure and is similar to a completely randomized design notation for our example except that the model statement is defined with the right side, as would be done for any ANOVA design (see the appendix for more detail). In our example, the test results for model effect have been summarized for all the effects that were tested (Table 2). We chose to treat year as a fixed effect with the variable selection techniques that focus on model parsimony, i.e., selection of the minimum number of variables, in their original units, that “best” predicts a response outcome. Principal components analysis uses all of the variables in the data set and determines weights (coefficients, $e_i$ in the linear construct are the loading factors of the eigenvectors based on the contribution of variability and correlation to the axis or PC). The original observations are transformed into score values for each PC of significant importance:

$$PC_1 = e_1'X = e_{11}X_1 + e_{12}X_2 + \ldots + e_{1p}X_p$$
$$PC_2 = e_2'X = e_{21}X_1 + e_{22}X_2 + \ldots + e_{2p}X_p$$
$$\vdots$$
$$PC_p = e_p'X = e_{p1}X_1 + e_{p2}X_2 + \ldots + e_{pp}X_p$$

These new scores can then be used in a subsequent analysis such as regression or canonical correlation without the problems of collinearity and with substantially fewer predictors. Because they are composite variables, their direct interpretation may not be obvious—subject matter knowledge may help with interpretation; however, not all data sets lend themselves to naming the PC composite structure with a meaningful label.

### Table 1. Test statistics for testing hypotheses defined by the effects model in multivariate analysis of variance. The $H$ matrix corresponds to the effect hypotheses defined in the model, the $E$ matrix is the residual error matrix, and the $T$ matrix is the total variability matrix.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks’ lambda ($\lambda$)</td>
<td>$</td>
</tr>
<tr>
<td>Pillai’s trace</td>
<td>trace($HT^{-1}$)</td>
</tr>
<tr>
<td>Hotelling–Lawley trace</td>
<td>trace($E^{-1}H$)</td>
</tr>
<tr>
<td>Roy’s greatest root</td>
<td>largest eigenvalue $E^{-1}H$</td>
</tr>
</tbody>
</table>

### Table 2. The $P$ values for multivariate analysis of variance (MANOVA) tests of predictor effects on leaf mineral nutrients and agronomic traits from a study on response to crop stress on oat imposed by aphid feeding and an aphid-vectored virus conducted over 2 yr. The experimental materials were described by Riedell et al. (2007).

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Treatment</th>
<th>Year</th>
<th>Treatment $\times$ year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks’ lambda ($\lambda$)</td>
<td>0.0002</td>
<td>&lt;0.0001</td>
<td>0.0004</td>
</tr>
<tr>
<td>Pillai’s trace</td>
<td>0.0009</td>
<td>&lt;0.0001</td>
<td>0.0235</td>
</tr>
<tr>
<td>Hotelling–Lawley trace</td>
<td>0.0005</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Roy’s greatest root</td>
<td>0.0006</td>
<td>&lt;0.0001</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

### Principal Components Analysis

Principal components analysis is possibly the most widely used empirical MV statistical technique (Ping et al., 2004). It is a nonparametric method that extracts relevant information from a MV data set that may contain unapparent patterns or structure when evaluated by uni- or bivariate methods alone. This method empirically summarizes the covariance or correlation among variables into a few components. In PCA, there is no distinction between predictor and outcome variable matrices; there is just one set of variables (matrix). Principal components analysis is an eigenanalysis and is also known as singular value decomposition, with the goal of reducing the dimensionality of a data set containing a large number ($>4$) of RVs into a few linear constructs, called principal components (PCs), each representing a new axis (coordinate system) of major variability. The symmetric variance–covariance or correlation matrix is decomposed into eigenvalues and eigenvectors. Eigenvalue–eigenvector pairs are ordered largest to smallest according to the variance that each linear composite axis explains. Thus, $\lambda_1 > \lambda_2 >> \lambda_p$, when a covariance matrix is used in the eigenanalysis decomposition.

In PCA, there is not a model being tested but rather it is a method that re-expresses the data in a new coordinate system, hopefully of lower dimension than the original data set dimension. Principal components analysis differs conceptually from variable selection techniques that focus on model parsimony, i.e., selection of the minimum number of variables, in their original units, that “best” predicts a response outcome. Principal components analysis uses all of the variables in the data set and determines weights (coefficients, $e_i$ in the linear construct are the loading factors of the eigenvectors based on the contribution of variability and correlation to the axis or PC). The original observations are transformed into score values for each PC of significant importance:
Fig. 1. Scree plots of the three groups of data: (a) soil minerals, (b) plant nutrients, and (c) grain characterization and composition, as originally described by Riedell et al. (2009). The scree plot is a plot of the eigenvalues against the extracted component rank of variance explained for the analysis based in a matrix of correlations. The "elbow" bend of each plot visually identifies an eigenvalue close to 1. When the eigenvalues approach 1 or less, this suggests no more relevant correlation can be accounted for among the variables.
Table 3. Eigenvalues and proportion of variance for the first five principal components (PCs) of the principal components analysis of the eight soil variables collected and described by Riedell et al. (2009). Although there are eight PCs in total, only the first five are listed because only the first two or three PCs are likely to be considered useful for further evaluation.

<table>
<thead>
<tr>
<th>Principal component</th>
<th>Eigenvalue</th>
<th>Variance proportion</th>
<th>Cumulative variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC1</td>
<td>2.69250</td>
<td>0.3366</td>
<td>0.3366</td>
</tr>
<tr>
<td>PC2</td>
<td>1.18256</td>
<td>0.1478</td>
<td>0.4844</td>
</tr>
<tr>
<td>PC3</td>
<td>1.10193</td>
<td>0.1377</td>
<td>0.6221</td>
</tr>
<tr>
<td>PC4</td>
<td>0.97221</td>
<td>0.1215</td>
<td>0.7437</td>
</tr>
<tr>
<td>PC5</td>
<td>0.91104</td>
<td>0.1139</td>
<td>0.8575</td>
</tr>
</tbody>
</table>

In summary, the assumptions behind PCA utility include:
1. Linearity frames the idea of identifying new orthogonal axes for a data set as a linear combination.
2. Large variance exists in the data and is associated with meaningful structure. Principal components that account for large proportions of the variance contain useful structure.
3. A moderate degree of correlation exists between measured variables. Essentially PCA decorrelates a data set to remove second-order dependencies (i.e., the covariance structure among variables).

The complete data set we are using here, collected by Riedell et al. (2009), provides an excellent example of the effective use of PCA to reduce the dimension of each of three distinct pieces of information from the same system, namely, soil nutrients (eight variables measured), plant nutrients (eight variables measured), and grain quality and nutrients (12 variables measured). Each of these data sets can be reduced in dimension via PCA with minimal overall loss of information.

In SAS, the PRINCOMP procedure is used for the basic PCA (see the appendix). Using the ODS graphics plot functions from SAS, the resulting output generates aids in interpreting the number of significant PCs to retain for subsequent analysis. The scree plot is a plot of the eigenvalues against the extracted component rank of variance explained (Fig. 1). For correlation matrices, the rule of thumb is to visually identify the point where the eigenvalues make an “elbow” or turn from an initial decreasing curve to a somewhat flat (horizontal) curve. This elbow point is often associated with an eigenvalue close to 1 and K is the dominant weight in this PC. Upon review of the pairwise correlation among all variables, it was observed that K has relatively low correlation with all other soil variables; this lack of correlation supports the idea that K would dominate one PC (Jackson, 1991).

Table 4. Eigenvector elements that define the linear combination of variables, from singular value decomposition or eigenanalysis within principal component analysis, for the eight soil variables collected and described by Riedell et al. (2009). The variables with the largest influence for each principal component (PC1–PC5) are in bold.

<table>
<thead>
<tr>
<th>Element</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil NO₃</td>
<td>0.1460</td>
<td>–0.4949</td>
<td>0.4144</td>
<td>–0.1813</td>
<td>0.5414</td>
</tr>
<tr>
<td>Soil P</td>
<td>0.1638</td>
<td>0.3530</td>
<td>0.2674</td>
<td>–0.6105</td>
<td>0.2968</td>
</tr>
<tr>
<td>Soil K</td>
<td>–0.0985</td>
<td>0.0938</td>
<td>0.8425</td>
<td>0.1767</td>
<td>–0.3822</td>
</tr>
<tr>
<td>Soil Ca</td>
<td>–0.5062</td>
<td>0.1032</td>
<td>–0.1059</td>
<td>–0.1371</td>
<td>0.4077</td>
</tr>
<tr>
<td>Soil Mg</td>
<td>–0.3404</td>
<td>0.3923</td>
<td>0.1743</td>
<td>0.5191</td>
<td>0.4298</td>
</tr>
<tr>
<td>Soil Fe</td>
<td>0.4805</td>
<td>0.3119</td>
<td>–0.0651</td>
<td>0.3666</td>
<td>0.2823</td>
</tr>
<tr>
<td>Soil Mn</td>
<td>0.5645</td>
<td>–0.0461</td>
<td>0.0296</td>
<td>0.2114</td>
<td>0.1312</td>
</tr>
<tr>
<td>Soil Zn</td>
<td>0.1432</td>
<td>0.5979</td>
<td>–0.0068</td>
<td>–0.3093</td>
<td>–0.1567</td>
</tr>
</tbody>
</table>
functions, with a minimum loss of information. In CDA, an extension of MANOVA and closely related to CCA, the newly derived canonical variates summarize between-group variation and provide a simultaneous test describing which variables best account for group differences.

The goal of CDA is to test and describe the relationships among two or more groups based on a set of discriminating variables. Canonical DA involves deriving the linear combinations (i.e., canonical functions) of the variables that will discriminate the best (i.e., maximize the variation) among the predefined groups. For each sampled observation, a canonical score is calculated for each canonical variate, and the group centroid can be used to identify the most typical location of an observation from a given group. A comparison of group centroids indicates how far apart the groups are along the variate being tested. One available statistic to describe differences of canonical discriminant functions is a relative measure of the distance between the group centroids, Mahalanobis’ distance $D^2 = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)^T S^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)$, where the $S$ matrix is the inverse of the pooled variance matrix.

In the CDA example, the CANDISC procedure in SAS software was used to produce the presented results (see the appendix). The soil measures of NO$_3$, Ca, Fe, and Mn; plant measures of N, P, K, Ca, and Zn; and grain measures of N, P, K, S, and Mg were used to discriminate and classify among the crop rotations. Two canonical variates accounted for 100% of the among-groups variance (Table 5). Each canonical variate is the linear combination of the independently measured variables and is orthogonal to the others (Vaylay and van Santen, 2002). The significant ($P < 0.0001$ and 0.0070) canonical correlations between the crop rotations and the first and second canonical variates ($r_{c1} = 0.87$ and $r_{c2} = 0.69$) indicate that the canonical variates explain the differentiation of the rotation groups.

Canonical loadings measure the simple linear correlation between an original independent variable and the canonical variate. Thus, the canonical loading reflects the variance that the observed variables share with the canonical variate and can be interpreted as the relative contribution of each variable to each canonical variate function (Cruz-Castillo et al., 1994; Yeater et al., 2004). The first canonical function is dominated by large loadings from plant N and grain N, followed by soil NO$_3$ and a negative grain P loading (Table 5). The second canonical function is dominated by a large loading from plant Ca, followed by a negative grain P loading. These functions indicate that the crop rotations differ most in the uptake and utilization of N, followed by Ca and P.

The graphical representation of the canonical score information is useful in understanding the separation of a priori known group membership. The plot presented in Fig. 2 represents canonical structures of importance (discriminant canonical variables 1 and 2) on the $x$ and $y$ axes with centroids of each rotation group, and score values for each replicate observation presented in the grouped scatterplot. The meaning is inferred from the pattern of the correlation defined by the loading matrix (based on within-group correlations among predictors and the discriminant functions). The nature and meaning of the discriminant functions are not outwardly apparent from the plot alone because the linear combinations of predictors is maximally separating the groups based on observations not passing through the cloud of data points, as occurs in most linear techniques (like PCA). It is important to investigate the canonical discriminant variables and the canonical loadings of the measured variables for a more thorough interpretation.

The separation of the rotation groups is measured by the Mahalanobis distance, $D^2$ (Table 6). All pairwise distances between the centroids of crop rotation groups were significant ($P < 0.0042$). The two canonical variables for each observation of the experiment resulted in the C–C rotation with a centroid of ordinates $[-1.74, -0.93]$ (Fig. 2) and, along with the information provided by the canonical variates (Table 5), shows a strong inclination for the influence of the negative loading of grain P. The C–S–W/A–A rotation has larger positive values on the first canonical function axis and is more influenced by the large positive loadings of plant N, grain N, and soil N. The C–S rotation has larger positive values on the second canonical function axis and is therefore most influenced by the plant Ca loading.

**OTHER DATA EXPLORATION CONSIDERATIONS**

**Multivariate Normal Distribution**

The probability distribution applicable to the MV analysis methods presented here is the MV normal distribution. This probability density is a straightforward extension of the univariate normal density, to which it will reduce when $p = 1$. In notation, the multivariate normal distribution of a random vector $x$ can be written as $x \sim N(\mu, \Sigma)$, where $\mu$ is a vector of means and $\Sigma$ is a variance–covariance matrix. Important properties of the distribution include: homogeneous variance of conditional responses on predictors; common covariance across observations; and independent observations. Testing for MV normality can begin with a simple assessment of the univariate normality of each variable and progress to more

---

**Table 5.** The canonical loadings of the measured traits on the first two canonical variables for the study on crop rotations and N rates conducted over 2 yr, as described by Riedell et al. (2009). The canonical discriminant analysis example utilized four variables measured on soil, five variables from the plant material, and five variables measured from the grain material. The variables with the largest influence for each canonical variate are in bold.

<table>
<thead>
<tr>
<th>Measured variable</th>
<th>Canonical variate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Soil NO$_3$</td>
<td>0.3234</td>
</tr>
<tr>
<td>Soil Ca</td>
<td>-0.0436</td>
</tr>
<tr>
<td>Soil Fe</td>
<td>0.0672</td>
</tr>
<tr>
<td>Soil Mn</td>
<td>0.1628</td>
</tr>
<tr>
<td>Plant N</td>
<td>0.4029</td>
</tr>
<tr>
<td>Plant P</td>
<td>-0.1373</td>
</tr>
<tr>
<td>Plant K</td>
<td>-0.0386</td>
</tr>
<tr>
<td>Plant Ca</td>
<td>0.1566</td>
</tr>
<tr>
<td>Plant Zn</td>
<td>-0.0107</td>
</tr>
<tr>
<td>Grain N</td>
<td>0.5607</td>
</tr>
<tr>
<td>Grain P</td>
<td>-0.3280</td>
</tr>
<tr>
<td>Grain K</td>
<td>-0.1298</td>
</tr>
<tr>
<td>Grain S</td>
<td>0.2882</td>
</tr>
<tr>
<td>Grain Mg</td>
<td>0.0336</td>
</tr>
<tr>
<td>Canonical correlation</td>
<td>0.8721</td>
</tr>
<tr>
<td>$P$ level of significance</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Variance accounted for, %</td>
<td>77.50</td>
</tr>
</tbody>
</table>
advanced tools available through various software packages (e.g., the %multnorm macro in SAS [Khattree and Naik, 1999, 2000]). There is a vast amount of information available on the MV normal distribution, its properties, and the available evaluations (Khattree and Naik, 1999, 2000).

**Data Visualization**

Many software packages now have the capability of producing scatterplots of multiple variables simultaneously with the pairwise correlation for each pair of variables. We suggest this as a useful initial step in evaluating a MV data set: look at the correlation structures among all variables. This can easily become overwhelming and our eyes are drawn to the strongest linear relationship or possibly a strong nonlinear relationship. But one must ask, “are variables that show relationships with groups of variables redundant in the information they contain regarding the system under study?” When there is correlation among multiple variables, some of the variability explained may be shared or there may be portions of the covariation that is unique to a given variable.

Another approach is to use cluster analysis. There are many different algorithms by which to cluster variables and observations. However, the goal is to reduce the number of variables to a smaller number of clusters and to then use subject matter knowledge to decide which variables best represent each cluster, possibly leading to hypothesis generation and testing, prediction applications based on patterns for each cluster, and even outlier detection, where outliers are often identified as those observations “far away” from any cluster. This reduces some of the problems of collinearity among variables, leading to a more stable estimation of parameters in multiple regression.

**Table 6.** Pairwise squared distances between three crop-rotation classification groups, C–C (maize monoculture), C–S (2-yr rotation with soybean), and C–S–W/A–A (4-yr rotation with soybean, wheat, and alfalfa) under different N rates as calculated by the Mahalanobis distance, $D^2$. All distances between class groups are significant ($P < 0.0042$). The $D^2$ values are in the lower triangle, and the corresponding $F$ statistic (numerator df = 14, denominator df = 38) is in the upper triangle. The original study on crop rotations and N rates conducted over 2 yr was described by Riedell et al. (2009).

<table>
<thead>
<tr>
<th>Rotation</th>
<th>C–C</th>
<th>C–S</th>
<th>C–S–W/A–A</th>
</tr>
</thead>
<tbody>
<tr>
<td>C–C</td>
<td>2.94</td>
<td>8.23</td>
<td></td>
</tr>
<tr>
<td>C–S</td>
<td>6.13</td>
<td>5.52</td>
<td></td>
</tr>
<tr>
<td>C–S–W/A–A</td>
<td>17.18</td>
<td>11.51</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Grouped scatterplot of the three crop-rotation classification groups, C–C (maize monoculture), C–S (2-yr rotation with soybean), and C–S–W/A–A (4-yr rotation with soybean, wheat, and alfalfa), under different N rates on the two canonical discriminant functions. The centroid of each rotation class is marked by the abbreviation for the rotation. The original study on crop rotations and N rates conducted over 2 yr was described by Riedell et al. (2009).
Multivariate Approaches and Experiment-wise Error Rate

There are a plethora of tools available for exploring MV data to predict outcomes and identify patterns. The appropriate MV approach controls the inflation of experiment-wise (Type I) error rates by (i) considering the set of RVs in a multidimensional space and (ii) accounting for the relationships among the RVs as well as the relationships between explanatory variables and RVs (Truxillo, 2005). In MV statistical analysis methods, estimations of parameters are still functions of the mean and variance but also the covariances (correlations) among the variables. In a univariate multiple regression analysis, the correlation among the explanatory variables has been described as “ill-conditioned” and must be overcome to estimate stable parameter estimates (Draper and Smith, 1981; Schabenberger and Pierce, 2002). Accordingly, the assumptions of the underlying distributional properties correspond to the appropriate analysis of the data, such that we can make inferential statements about the statistics, estimated parameters, and model validation.

The primary advantage of MV models is that they incorporate the correlations between experimental outcomes whether they are longitudinal (repeated measures in time or space) or among different outcome variables measured on the same sampling unit (Bell et al., 2014; Demsar et al., 2012; Hirosawa et al., 1996). The experiment-wise error rate is controlled at the α level set by the researcher, where MV methods simultaneously test the treatment group centroids for differences in one test. A MANOVA null hypothesis 𝐻0: 𝑢1 = 𝑢2 = ... = 𝑢K (where the 𝑢 values are group mean vectors or centroids) tests for group equality vs. the alternative hypothesis that at least two of the group centroids are different. Following a significant MV Wilks’ 𝜆 test of the null hypothesis (i.e., rejection of 𝐻0), individual RV 𝐹 tests could be performed with the assurance of maintaining the experiment-wise error rate close to the α level set for the study. An alternative way to evaluate the influence of individual RVs on the separation of groups is to study the eigenvector associated with the largest eigenvalue of the E⁻¹H matrix. The coefficients in the eigenvector account for the presence of all other RVs in the study, and the larger the weight on an RV(s) the more contribution that RV(s) plays in discriminating between treatment groups (Rencher, 2010).

When each RV is evaluated individually in a univariate ANOVA 𝐹 test at a set α level, without consideration of MV relationships, then the experiment consists of 𝑝 hypothesis tests (given 𝑝 RVs), and the overall α level will increase beyond the nominal level initially intended. However, the realities of a system under study can sometimes be so complex that returning to univariate analyses can help to process all of the information; however, caution should be used when making lots of univariate tests because the probability of rejecting one or more univariate tests when 𝐻0 is actually true can range from 0.09 to 0.3 when 𝜂 = 0.05 for each univariate test (Rencher and Christensen, 2012).

Rules of Thumb for Sample Size Selection and Number of Variables

The number of observations in a study is directly related to power, where power is the ability to detect “real” differences when they exist (Kraemer and Thiemann, 1987). Effect size refers to the strength of the association between variables and/or the size of the difference estimated; the greater the effect size for a given sample size, the greater the power of the study and the smaller the sample required. Currently, all statistical tests that seek to determine differences are based on a sampling distribution; therefore, the number of observations is directly linked to the standard deviation of the sampling distribution. The greater the number of observations, the narrower the sampling distribution of the mean and the greater the chance that differences will be statistically detected (i.e., more power).

The general recommendation for the minimum sample size for MV analysis falls into two categories. One category states that the absolute number of observations (𝑛) is necessary, while the other says that the observation/variable ratio (𝑛/𝑝) is important. These recommendations have been reviewed in the literature for the application of factor analysis and PCA (Arrindell and van der Ende, 1985; Velicer and Fava, 1998; MacCallum et al., 2001; Osborne and Costello, 2004). The results of Osborne and Costello (2004) indicated an interaction between the two, where the optimum outcomes occurred in analyses with large 𝑛 and high ratios, agreeing with rules of thumb where 𝑛 should be at least 5𝑝 for DA and 𝑛 should be >10𝑝 for hypothesis-driven approaches like MANOVA.

Conclusions

Multivariate data and their corresponding analysis options consider complex, multidimensional relationships among random response variables compared with a univariate analysis that considers only a single response variable. However, the inherent higher dimensionality of MV analyses can be difficult to interpret. The practical implementation of these approaches may involve the inclusion of several univariate and MV analyses to better understand the relationships among variables and their relevance to the complexities that are being researched. It is important to recognize that more time and effort is required to understand the relationships of multidimensional data sets, but the outcome can be well worth the effort. In approaching a potential MV study, it is essential for proper examination of the research questions for the researcher to specify these questions a priori. Doing so enables the researcher to more efficiently design a study that properly examines the research questions. Especially in MV analysis, addressing adequate sample size and deciding on the appropriate measured variables will aid in optimizing information. Before data collection, it is also advantageous to understand and appropriately organize the data structure, as well to become familiar with the software tools and applications that will be utilized.
APPENDIX
SAS Code Examples for Generating Multivariate Analyses with Experimental Data

Description of SAS Program for MANOVA

The following SAS program example uses the MANOVA to measure differences in the leaf mineral nutrients (N, P, K, Ca, Mg, Zn, Fe, Mn, and Cu), chlorophyll (chl), leaf area index (lai), and grain yield (GY). The following statements invoke the GLM procedure to perform the MANOVA that generated the output as presented in Tables 1 and 2. Unlike in univariate ANOVA, all dependent response variables tested are included on the left side of the equation in the MODEL statement. The inclusion of a MANOVA statement, following the model statement, tells SAS to perform MV hypothesis tests for the model effects. All four MV tests are computed via “h=_all_” within the MANOVA statement. Exact F tests are calculated via the MSTAT=EXACT option. The SAS output includes univariate analysis (each variable analyzed alone) unless a <nouni> option (no univariate analysis) is used in the PROC GLM statement.

SAS Code Program for MANOVA Example

odds graphics on;
proc glm data=mvms.oat_aphid_study;
class Year Treatment Rep;
model N P K Ca Mg Zn Fe Mn Cu chl lai GY = Treatment Year Treatment*Year;
manova h = _all_ / mstat=exact summary;
run;
quit;
ods graphics off;

Description of SAS Program for Principal Component Analysis

This example program analyzes and generates the first five principal components for grain measures (yield, oil, starch, N, P, K, S, Ca, Mg, Fe, Mn, and Zn), plant measures (N, P, K, Ca, Mg, Fe, Mn, and Zn), and soil measures (NO3, P, K, Ca, Mg, Fe, Mn, and Zn) using the PRINCOMP procedure. First, each set of variables (grain, plant, or soil) was investigated with separate PCA assessments (data not shown). Then all variables were combined and data step programming was used to relabel and merge the corresponding output. The generated output is presented in Tables 3 and 4. The default plots are generated from the “plots=all” option and scree plot and corresponding variance plot provided as Fig. 1. The observations on plots are labeled with the “ID” statement calling for N input and crop rotation (variable name: in_rot).

SAS Code Program for Principal Component Analysis Example and Plots of Principal Components

odds graphics;
ods rtf file='C:\pca_grain.rtf';
proc princomp data=mvms.all_data_source_a out=mvms.pca_g n=5 plots=all;
id in_rot;
var Yield Oil Starch Grain_N Grain_P Grain_K Grain_S Grain_Ca Grain_Mg Grain_Fe Grain_Mn Grain_Zn;
run;
ods rtf close;

ods rtf file='C:\pca_plant.rtf';
proc princomp data=mvms.all_data_source_a out=mvms.pca_p n=5 plots=all;
id in_rot;
var Soil_NO3 Soil_P Soil_K Soil_Ca Soil_Mg Soil_Fe Soil_Mn Soil_Zn Che Plant_DW Plant_N Plant_P Plant_K Plant_Ca Plant_Mg Plant_Fe Plant_Mn Plant_Zn Yield Oil Starch Grain_N Grain_P Grain_K Grain_S Grain_Ca Grain_Mg Grain_Fe Grain_Mn Grain_Zn;
run;
ods rtf close;

ods rtf file='C:\pca_soil.rtf';
proc princomp data=mvms.all_data_source_a out=mvms.pca_s n=5 plots=all;
id in_rot;
var Soil_NO3 Soil_P Soil_K Soil_Ca Soil_Mg Soil_Fe Soil_Mn Soil_Zn Che Plant_DW Plant_N Plant_P Plant_K Plant_Ca Plant_Mg Plant_Fe Plant_Mn Plant_Zn Yield Oil Starch Grain_N Grain_P Grain_K Grain_S Grain_Ca Grain_Mg Grain_Fe Grain_Mn Grain_Zn;
run;
ods rtf close;

ods rtf file='C:\pca_all.rtf';
proc princomp data=mvms.all_data_source_a out=mvms.pca_all n=5 plots=all;
id in_rot;
var Soil_NO3 Soil_P Soil_K Soil_Ca Soil_Mg Soil_Fe Soil_Mn Soil_Zn Plant_DW Plant_N Plant_P Plant_K Plant_Ca Plant_Mg Plant_Fe Plant_Mn Plant_Zn Yield Oil Starch Grain_N Grain_P Grain_K Grain_S Grain_Ca Grain_Mg Grain_Fe Grain_Mn Grain_Zn;
run;
ods rtf close;
Plant_P Plant_K Plant_Ca Plant_Mg Plant_Fe Plant_Mn Plant_Zn Yield Oil Starch Grain_N Grain_P Grain_K Grain_S Grain_Ca Grain_Mg Grain_Fe Grain_Mn Grain_Zn;
set mvms.pcaG;
rename prin1=G_pca1 prin2=G_pca2 prin3=G_pca3;
run;

data mvms.all_princ;
merge mvms.G_pca mvms.PL_pca mvms.S_pca;
by obs year input rotation rep;
run;

description of sas program for canonical discriminant analysis

This example program of the CANDISC procedure for CDA created the output in Tables 5 and 6. The soil measures of NO3, Ca, Fe, and Mn, plant measures of N, P, K, Ca, and Zn, and grain measures of N, P, K, S, and Mg were used to discriminate and classify among the crop rotations (variable name: rotat). The DISTANCE option displays squared Mahalanobis distances between the class group means, F statistics, and the corresponding probabilities of greater squared Mahalanobis distances between the class group means. Additionally, SAS coding in the Graph Template Language using the TEMPLATE procedure and SGRENDER procedure is provided, which generated the plot of the first two canonical variables in Fig. 2.

SAS code program for canonical discriminant analysis example and plot of canonical variables

proc candisc data=mvms.all_data_source_a out=mvms.outcan
class rotation;
var Soil_NO3 Soil_Ca Soil_Fe Soil_Mn Plant_N Plant_P Plant_K Plant_Ca Plant_Zn Grain_N Grain_P Grain_K Grain_S Grain_Ca Grain_Mg Grain_Fe Grain_Mn Grain_Zn;
run;

proc template;
define statgraph sgdesign;
dynamic _CAN1=_CAN2=_ROTATION;
begingraph;
entrytitle halign=center 'Influence from Crop Rotation';
rowgutter=10 columngutter=10;
layout overlay / rowdatarange=data columndatarange=data
viewmax=5.0);
yaxisopts=(viewmin=-3.5 viewmax=4.25));
sfncolor='white'
scatterplot x=_CAN1 y=_CAN2 / group=_ROTATION name='scatter';
discregetlegend 'scatter' / opaque=0 halign=right valign=top displayclipped=true across=1 order=rowmajor location=inside;
endlayout;
endgraph;
end;
run;

proc sgrender data=WORK.OUTCAN template=sgdesign;
dynamic _CAN1='CAN1' _CAN2='CAN2' _ROTATION='ROTATION';
run;

references


Bell, A.R., J.L. Caviglia-Harris, and A.D. Cak. 2014. Characterizing land-use change over space and time: Applying principal components analysis in the Brazilian Legal Amazon. J. Land Use Sci. doi:10.1080/174743X.2013.832427


Agronomy Journal • Volume 107, Issue 2 • 2015 809


