

OUTLIERS AND REDUNDANT VARIABLES IN PRINCIPAL COMPONENT AND FACTOR ANALYSIS.

G. Mora-Aguilera, R. Kiltie¹, D. Nieto-Angel, and D. Téliz.

Centro de Fitopatología, CP. Montecillo, México; 56230.

¹Zoology Dept. University of Florida. Gainesville, Fl. 32611.

Abstract

Nine variables were measured to characterize each of 60 epidemics of papaya ringspot on papaya (*Carica papaya*) plantations in Veracruz, México. This matrix of data was explored with principal component and factor analysis based on the correlation matrix. Outliers were detected and evaluated as influential observations. Variables were screened to select those that contribute more to the overall variance. PROC FACTOR and PROC PRINCOMP of SAS/STAT were used and compared for these analysis. Outliers were identified by using standardization of the original matrix, and standardization of unrotated and rotated principal components scores. Varimax with PROC FACTOR of SAS/STAT was used for the later case. Plot of the data with respect to the first major principal components were also used to graphically detect outliers. Outliers were evaluated as influential observations by recomputing principal components scores with the outliers excluded. Redundant variables were identified by using minor principal components. Eigenvectors for the mayor principal components were utilized to obtain a biplot to also graphically detect redundant variables. Robustness of the different approaches for detecting outliers and redundant variables is discussed. Advantages by using PROC FACTOR and PROC PRINCOMP are discussed as well.

INTRODUCTION

The objectives of this paper are to detect and test outliers as influential observations and to reduce the number of the original variables. Researchers have accomplished these objectives by using several approaches such as major and minor principal components, multiple correlation and cluster analysis (Pack *et al*, 1988; Hawkins and Fatti, 1984; Jolliffe, 1973). Different types of rotation have been suggested to simplify structure of eigenvectors and improve the selection of important variables (Jolliffe, 1989; Manly, 1986). Graphical display of observations and variables by using principal components have also been suggested to detect outliers and define highly correlated variables that could be then eliminated (Pack *et al*, 1988; Gabriel, 1971; Bradu and Gabriel 1978). In this study, the central idea that was exploited was the definition of minor principal components (Hawkins and Fatti, 1984). A combination of methods suggested by Pack *et al* (1988) and Hawkins and Fatti (1984) and an extension of the biplot technique suggested by Gabriel (1971), were used to detect outlying and eliminate variables with low loss of explanatory capacity of the original matrix of data. In both cases varimax rotation was used and tested for improving the interpretation of eigenvectors. The probability of partial axis rotation is addressed in the context of the objectives of this study as a possibility to avoid spread of variance and loss of interpretation of components.

To achieve these objectives SAS was utilized (SAS Institute, 1988). Two procedures, PROC FACTOR and PROC PRINCOMP, were applied and compared in their features rather than in their computational resources. A matrix of real data consisting of 10 variables and 59 observations was used for this study.

MATERIAL AND METHODS

Variables measured. 59 epidemics of papaya ringspot virus, an important disease in papaya crop, were characterized with 10 variables. The rate of apparent infection was estimated with the Gompertz, monomolecular and logistic models (Mora-Aguilera *et al*, 1992). The last two models were standardized to the Gompertz (k_G) (Campbell and Madden, 1990). The scale (b) and shape (c) parameters of the Weibull model (Pennypacker *et al*, 1980) were estimated by the DUD method (Ralston and Jennrich, 1978). These variables were stimated with PROC GLM and PROC NONLIN procedure of SAS institute (1988). The area under disease progress curve (AUDPC) and the AUDPC standardized by time in days (AUDPC_s) were estimated by programming the midpoint rule in SAS (Campbell and Madden, 1990). The remaining variables were the initial and final disease incidence (Y_0 and Y_f respectively), time in days from transplant to initial incidence (X_0), time in days to reach 50 percent of the final incidence (T_{50}), and total duration of the epidemics in days (T_f).

Principal component analysis. Principal component analysis (PCA) on the corrected matrix was performed to detect and evaluate outliers and to reduce the original set of descriptive variables to a smaller number of principal components. These objectives were achieved through PROC PRINCOMP and PROC FACTOR of SAS (SAS institute, 1988). The relative efficiency of these procedures was compared. Minor principal components were identified by small variances ($\lambda < 0.7$) (Hawkins and Fatti, 1984; Jolliffe, 1982). Conversely major principal components were distinguished by relatively large variance ($\lambda > 0.7$).

Although PCA, automatically normalizes the variables when the correlation matrix is used, Y_o and Y_f were transformed with $\arcsin$ of $(Y_o)^{1/2}$ and $(Y_f)^{1/2}$ respectively with the objective of normalizing these variables (Lyman, 1988). This transformation was selected since Y_o and Y_f were measured as perceptual units (Lyman, 1988). This transformation was used only after outlier detection for reducing the set of variables.

Outlier and influential observation detection. Outliers were detected and tested as influential observations by extension procedures presented by Hawkins and Fatti (1984) and Pack *et al* (1988). Confidence intervals of approximately 95% for the original standardized variables (Hawkins and Fatti (1984), plots of standardized and non-standardized scores with respect to the major principal components, computation of non-rotated and varimax rotated standardized scores were used to detect outliers. An observation was considered an outlier if it has at least one standardized variable or component with an absolute value greater than 3.0, which is approximately equivalent to using a normal significance level of 0.05 (Lyman, 1988). Graphically, an outlier was identified if the plotted observation is clearly distant from the other observations (Pack, *et al* 1988). Outliers were evaluated as influential observations by recomputing principal components scores with outliers excluded individually (Pack, *et al* 1988) and in groups. An outlier was considered an influential observation if the recomputed eigenvalues or eigenvector coefficients change sufficiently to modify the number of principal components selected and/or redundant variables identified (Pack *et al*, 1988; Jolliffe 1989; Reymont, 1988).

Elimination of redundant variables. Principal component analysis was performed interactively to reduce the number of variables. Minor principal components was used as criteria to eliminate variables (Hawkins and Fatti, 1984). Previous to the first PCA interaction, the correlation matrix was examined to detect potential

collinearity between variables. Less uncorrelated variables were selected and sorted in two groups to avoid the effect of collinearity in the PCA estimation of the eigenvalues and eigenvector coefficients (Hawkins and Fatti, 1984). PCA was applied to each group of variables to eliminate variables strongly associated to minor principal components ($\lambda's < 0.7$) (Jolliffe, 1973; Hawkins and Fatti, 1984). A high eigenvector coefficient indicated this type of variable. A variable was associated with each of the minor principal components. Thus, the number of rejected variables was equal to the number of eigenvalues smaller than 0.70 (Jolliffe, 1973). In the second interaction, variables selected from each group in the first step were grouped and reduced again through PCA. PCA was also applied to the full set of variables. The reduction of variables followed the criteria already described. This approach was done to compare the effect of collinearity among variables in the PCA results.

Major principal components selected from PCA for the full set of variables were represented graphically by using a biplot display (Bradu and Gabriel, 1978; Gabriel, 1971). Pairs of coefficients, corresponding with the first two eigenvectors, were plotted to represent vectors (variables) projected onto the plane dimension. Long vectors indicate variables that are well represented by the plane and have relatively larger standard deviation than small vectors (Gabriel, 1972). Angles at or near to 180° or very small indicate respectively negative or positive high correlation between variables (Gabriel, 1971). Variables poorly represented by a plane and highly correlated were eliminated.

RESULTS AND DISCUSSION

Outlier detection. By using an empirical criterion of an absolute standardized value greater than 3.0 and applying three different approaches, observations 8, 10, 42 and 59 were implicated as outliers. The first section of Table 1 gives the standardized original scores of ten variables for outliers 10 and 59. These observations had scores higher than 3.0 for the variables b and Y_o respectively. The second and third section of Table 1 indicate standardized component scores with and without varimax rotation. In addition to observations 10 and 59, the observations 8 and 42 were selected by the former and latter method respectively.

Interestingly, a graphical display of the two first principal components as was suggested by Pack *et al* (1988), did not show any obvious outliers. However a plot of the second and third major principal components

TABLE 1. Original scores and component scores of observations which qualify as outliers along one or more components. Bold numbers determined such observations that qualified as outlier.

Standardized Scores from Original Matrix of Data										
OBSERVATION	K_G ¹	Y_0	X_0	Y_f	T_1	T_{30}	c	b	AUDPCS	AUDPC
10	-1.35	1.58	0.66	-2.16	0.01	2.04	-1.27	3.83	-0.59	-1.35
59	1.14	4.82	0.15	1.07	-0.77	-1.13	-0.15	-0.96	0.74	0.19
Components										
	1	2	3	4	5	6	7	8	9	10
Standardized Scores Computed with PCA without Varimax Rotation										
8	0.70	-0.28	-0.16	-0.67	3.04	0.87	1.52	-0.32	1.05	-0.45
10	-2.25	-0.08	1.71	1.59	-0.75	3.74	2.54	1.60	-2.14	0.66
59	1.29	0.06	2.55	3.83	0.75	-0.96	-0.48	0.46	0.31	0.39
Standardized Scores Computed with PCA-FACTOR and Varimax Rotation										
10	0.26	-0.83	-0.95	-0.05	1.63	5.13	1.43	0.77	2.33	-0.64
42	0.48	1.33	-1.37	-1.60	0.31	-1.19	4.08	-1.80	-0.41	0.25
59	-0.11	0.04	-0.15	-0.10	4.80	-1.11	-0.39	0.65	0.07	-0.37

¹ Variables as are described in Text.

showed a clear outlier corresponding to observation 4 (Fig. 1A). Similarly, a plot with the third and four mayor principal components showed a clear outlier corresponding to the observation 59. In this axes, outlier 4 was only moderately apparent (Plot no shown). These results indicated that an outlier does not necessarily appear in the first two axes. This results when an outlier has a high score on a variable loading heavily on later axes. In this case observations 4 and 59 have relatively high values for the variable X_0 and Y_0 respectively.

However, since this outlier 4 was not detected by any other approach (Table 1), and to eliminate the possibility that the plot of non-standardized scores used by Pack *et al* (1988) explain this difference, a plot of standardized scores of principal components 2 and 3 was constructed. Observation 4 was still detected outside of space occupied by the other observations but not clearly enough to be considered an outlier. Contrarily, observation 59 was still detected as outlier for the space defined by axes 3 and 4 (Plots not shown). Since standardization of original scores or components are needed to apply the principle of standard normality for

detecting outlier (Hawkins and Fatti, 1984), plots of standardized scores should be used to avoid dissimilar results with the former methods. Additionally, since scores are estimated using eigenvectors and these tend to have large loadings when they are associated to small eigenvalues (Hawkins and Fatti, 1984), large scores are expected for minor principal components. Thus a plot of standardized scores using the minor axes should also be explored in addition of plotting major axes as was suggested by Pack *et al*, 1988. This idea was demonstrated plotting standardized minor axes. Observation 10 and 42 were clear outliers for the space defined by the standardized minor principal components 6 and 7 (Fig. 1B). Notice that observation 10 is also indicated as outlier using a nongraphical approach in the second and third portion of table 1. Observation 42 is also indicated with the varimax rotation method.

These variables have high coefficients in the eigenvector corresponding to principal component 3 and 4 respectively (Table 2, outliers included). Thus any plot where principal component 3 or 4 was involved showed the observation 4 or 59 as an outlier although in different

TABLE 2. Eigenvalues and eigenvectors for the first four principal components obtained with the correlation matrix and PRINCOMP for the complete matrix of data and excluding the outliers 8, 10, 42 and 59.

	Components Including all Outliers					Components Excluding Outliers			
	1	2	3	4		1	2	3	4
Eigenvalues	4.3416	2.4191	1.3973	0.8026		4.4514	2.5667	1.3763	0.7511
k_a^1	0.402	-0.306	-0.088	0.096		0.387	-0.313	-0.138	0.076
Y_a	0.095	0.035	0.553	0.796		0.114	0.048	0.565	0.810
X_a	-0.120	-0.140	0.661	-0.501		-0.110	-0.166	0.679	-0.462
Y_c	0.450	0.004	-0.143	0.028		0.440	0.001	-0.151	0.040
T_c	-0.265	0.404	-0.235	0.032		-0.273	0.405	-0.202	0.143
T_{so}	-0.394	-0.037	-0.294	0.266		-0.413	-0.004	-0.260	0.200
c	0.200	-0.490	-0.175	-0.069		0.172	-0.506	-0.152	0.031
b	-0.429	-0.124	0.174	-0.007		-0.436	-0.141	0.144	-0.053
AUDPC _a	0.354	0.372	0.155	-0.123		0.358	0.354	0.145	-0.207
AUDPC	0.187	0.570	0.015	-0.107		0.189	0.551	0.027	-0.122

¹ Variables as are described in text.

proportion. Plot of axis 3 and principal component 2 showed the most clear outlier (Fig. 1A). These results indicate that several plots should be explored rather than just the initial pair specific combination of principal components.

Hawkins and Fatti (1984) suggest that standardized components with or without varimax rotation rather than the unstandardized original scores is a more powerful diagnostic for identifying outlying observations. These authors emphasize the importance of minor principal components to achieve this objective, especially when there are strong correlations among variables. Our results seem to corroborate this idea, since an additional observation (8) was found using the criteria of standardized scores and minor principal components. However, there is a problem that arises, when the varimax rotation is used. The number of minor principal components can be reduced as results of varimax rotation. Thus the space in which to look for outlying observations is reduced. This is explained because the variance is spread among the components leaving fewer components with small variance (Jolliffe, 1989). Our analysis exemplifies this problem. The total number of principal components were 10 since there were 10 original variables in the data matrix. The major principal components detected without varimax rotation were the first four axes ($\lambda > 0.7$) (Table 2) consequently the

minor principal components detected without rotation were the last six axes. This number was reduced to three after varimax rotation. If the minor axes were strictly used, any observation could be detected as outlier. The bottom section of table 1 shows that the scores that make the observations 10, 42, and 59 behave as outliers lay under the major principal components six, seven and five respectively. The eigenvalues accounted for these axes are 1.02, 0.83 and 1.03 in the same order.

Notice that when the axes were not rotated, the outlier 59 was detected also under a major axis (4th principal component, $\lambda = 0.80$). These results suggest that minor axes can be useful to detect outlying observations as was stated by Hawkins and Fatti (1984) but these are not necessarily restricted to this space. Moreover outliers detected in the major axes could be considered more important since they probably affect variables that could remain after removing redundant variables usually detected in the minor axes. Pack *et al* 1988, clearly addressed that outliers in components removed are of little interest. Thus the whole set of principal components should be explored to diagnose outliers. Varimax rotation or any other orthogonal approach could be useful if it simplifies the structure of eigenvectors (Hawkins and Fatti, 1984). This means that the variance indicated in an eigenvalue is almost entirely explained by a well defined variable. The well defined

TABLE 3. Correlation matrix of 10 standardized epidemiological variables. Variables are described in text. Bold numbers represent relatively high correlation between the variables involved.

	Y_0	Y_f	T_i	T_{50}	c	b	AUDPC _s	AUDPC	X_0
k_0	0.1257	0.8570	-0.7339	-0.5942	0.7144	-0.6492	0.3286	-0.0929	-0.2279
Y_0		0.1471	-0.1445	-0.1920	-0.0621	-0.0572	0.1810	0.0873	0.1226
Y_f			-0.4532	-0.6208	0.5158	-0.8313	0.5977	0.3911	-0.3538
T_i				0.4956	-0.5013	0.3079	-0.1771	0.3772	-0.1373
T_{50}					-0.2036	0.6858	-0.6694	-0.3613	-0.1078
c						-0.2220	-0.1673	-0.4297	0.0036
b							-0.6836	-0.4654	0.4075
AUDPC _s								0.8398	-0.1401
AUDPC									-0.2063

variable (s) are characterized by high coefficients (close to 1) whereas the other variables contribute almost nothing (coefficient close to 0) to the variance of their eigenvalue (Hawkins and Fatti, 1984). The higher coefficients after rotate explain why the highest outlying scores were found after varimax rotation in Table 1.

Since the influence of individual outliers is additive (Pack *et al*, 1988), to test the outliers 4, 8, 10, 42 and 59 as influential observations, each outlier was omitted from the original matrix one by one and in several combinations. Such combinations always included the observations 10 and 59 since those epidemics were consistently indicated as outliers by the three statistical approaches. PCA was recalculated with the correlation matrix and without rotation for each new set of data. Neither a single nor a combination of outliers changed substantially the variance of the major principal components. Also the coefficients of the eigenvectors associated with those components were not significantly changed. As an example of these finding, Table 2 shows results for the major principal components for the whole data and excluding the outliers 8, 10, 42 and 59. As it can be observed, the cut-off of components based in the size of their variance remain the same (4 components with $\lambda > 0.7$). Similarly, variables that had relatively high coefficients in any of the major axes for the complete matrix of data remain high for the reduced data. Since the number of principal components selected and their overall structure were unaffected by the presence of outliers it can be concluded that these epidemics can be considered as outliers but not as influential observations

(Pack *et al* 1988, Jolliffe 1989).

Reyment (1990), and Pack *et al* (1988), stated that an influential observation can not necessarily be detected by outlying in specific variables. Hawkins and Fatti (1984) point out that this situation can be obtained when none of the scores of an observation for a set of variables are individually aberrant but taken together they are. This idea suggests that looking for outliers by using high standardized scores that are outside of a normal distribution of a particular principal component or by plotting pair of axes does not necessarily explore all possible outliers. Thus the idea of influence in principal component is more complex than it might appear. Graphical approaches to represent a multivariate space of higher dimension than two could be useful. This idea is addressed later in this discussion.

Elimination of redundant variables. Minor principal components case. The reduction of the original set of variables was developed in several steps. In the first stage the correlation matrix of the standardized descriptive variables was analyzed (Table 3). It indicated several relationships between some variables. The highest correlations were between Y_f with b and k_0 ($r = -0.83$ and 0.86 respectively); AUDPC_s with AUDPC ($r = 0.84$) and k_0 with c and T_i ($r = 0.71$ and -0.74 respectively). Since the correlation between AUDPC_s and AUDPC was high as we expected and because the latter Variable is a less accurate measurement (Campbell and Madden, 1990), AUDPC was eliminated.

By using the correlation matrix, two sets of less uncorrelated variables were defined. The first set was constituted by k_G , $AUDPCs$, b and Y_0 ; the second set was T_0 , X_0 , c and Y_f . The eigenvalues and eigenvectors obtained with principal component analysis for these two sets of variables are indicated in Tables 4 and 5 respectively. By using the minor components approach variables b and k_G from the first set, and T_0 and Y_f from the second set were eliminated. The remaining variables explained 79.4% and 88.3% of the variance for the first and second set respectively. Eigenvalues 0.64 and 0.18 for the first group and 0.42 and 0.16 for the second group of variables defined the minor axes (Table 4 and 5). Notice that the number of minor components correspond with the number of eliminated variables. This should be done to get the full rank of a matrix (Gabriel, 1971).

In the third step, the pooled set of variables selected from the two sets were analyzed again by principal component analysis. Table 6 shows the eigenvalues and eigenvectors for this pool of variables. The variables T_0 was eliminated. The remaining variables $AUDPCs$, Y_0 , c and X_0 are involved in an equal number of principal components. Together they explain 95% of variability for the set of five variables data. The proportion of variance that these four variables could explain for the whole set of variables data (nine) could be 89.6%. This proportion of variance was estimated with principal components analysis for the whole set of variables (Table 7). Interestingly, when spurious variables were eliminated by using the minor principal components criteria, but applying PCA to the full set of variables, the variables $AUDPCs$, Y_0 , and X_0 were also selected as before. However, c was eliminated since it represented well the 8th axis (Table 7). This left Y_f as the other variable to join the variables selected to explain 89.6% of the total variance.

Although the suggestion of Jolliffe (1973) and Pack *et al* (16) for selecting principal components ($\lambda > 0.7$) led to with four nonredundant variables, one additional analysis could be performed to still reduce the number of variables. The elimination of variables could be based on a stricter selection of major principal components ($\lambda > 1.0$) (Hawkins and Fatti, 1984). However, it should be recognized that the fewer variables are selected the more variance remains unexplained. For example, selecting only eigenvalues higher than 1.0, three variables will be selected. These variables will explain only 80.1% of the total variance (whole set of variables) rather than 89.6%. Moreover, Jolliffe (1989), suggest that four variables is a reasonable number to be retained

TABLE 4. Eigenvalues and eigenvectors obtained with PROC PRINCOMP of the correlation matrix of four variables and 59 observations. Bold numbers were redundant variables.

	Components			
	1	2	3	4
λ	2.194	0.984	0.640	0.180
k_G	0.528	-0.092	0.737	0.410
$AUDPC_s$	0.551	0.026	-0.669	0.496
Y_0	0.155	0.977	0.077	-0.119
b	-0.626	0.188	0.051	0.754

TABLE 5. Eigenvalues and eigenvectors obtained with PROC PRINCOMP of the correlation matrix of five variables and 59 observations. Bold numbers were redundant variables.

	Components				
	1	2	3	4	5
λ	2.408	1.201	0.807	0.427	0.155
T_0	0.503	-0.289	-0.133	0.798	-0.083
T_{00}	0.482	-0.159	0.658	-0.197	0.518
X_0	0.051	0.880	-0.000	0.321	0.345
c	-0.451	0.025	0.727	0.379	-0.349
Y_f	-0.553	-0.340	-0.141	0.274	0.694

TABLE 6. Eigenvalues and eigenvectors obtained with PROC PRINCOMP of the correlation matrix of five variables selected by principal components analysis from two uncorrelated groups. Bold numbers were redundant variables.

	Components				
	1	2	3	4	5
λ	1.761	1.130	1.096	0.805	0.205
$AUDPC_s$	0.662	-0.311	-0.063	-0.174	0.655
Y_0	0.330	0.174	0.524	0.764	0.002
T_{00}	-0.671	-0.188	0.208	0.187	0.659
X_0	0.026	0.668	0.502	-0.509	0.203
c	0.007	0.625	-0.651	0.300	0.306

TABLE 7. Eigenvalues and eigenvectors obtained with PROC PRINCOMP of the correlation matrix of nine variables and 59 observations. Bold numbers were associated to redundant variables.

	Components								
	1	2	3	4	5	6	7	8	9
Eigenvalues	4.2487	1.6167	1.3517	0.8472	0.4365	0.2318	0.1450	0.0863	0.0358
k_G	0.442	0.257	-0.102	0.098	-0.076	0.030	0.082	0.450	0.708
Y_0	0.089	-0.080	0.515	0.835	0.086	-0.068	-0.013	-0.099	0.009
X_0	-0.115	0.289	0.667	-0.382	0.350	0.173	-0.354	0.152	0.091
Y_f	0.453	-0.050	-0.149	0.101	0.247	0.293	-0.129	0.498	-0.588
T_i	-0.315	-0.396	-0.214	0.063	0.750	-0.104	0.145	0.190	0.248
T_{30}	-0.383	0.128	-0.329	0.299	-0.074	0.622	-0.464	-0.058	0.160
c	0.252	0.578	-0.202	0.053	0.469	0.087	0.219	-0.527	-0.048
b	-0.413	0.285	0.144	0.070	-0.113	0.305	0.700	0.321	-0.145
AUDPC _s	0.308	-0.500	0.198	-0.187	-0.010	0.615	0.272	-0.307	0.180

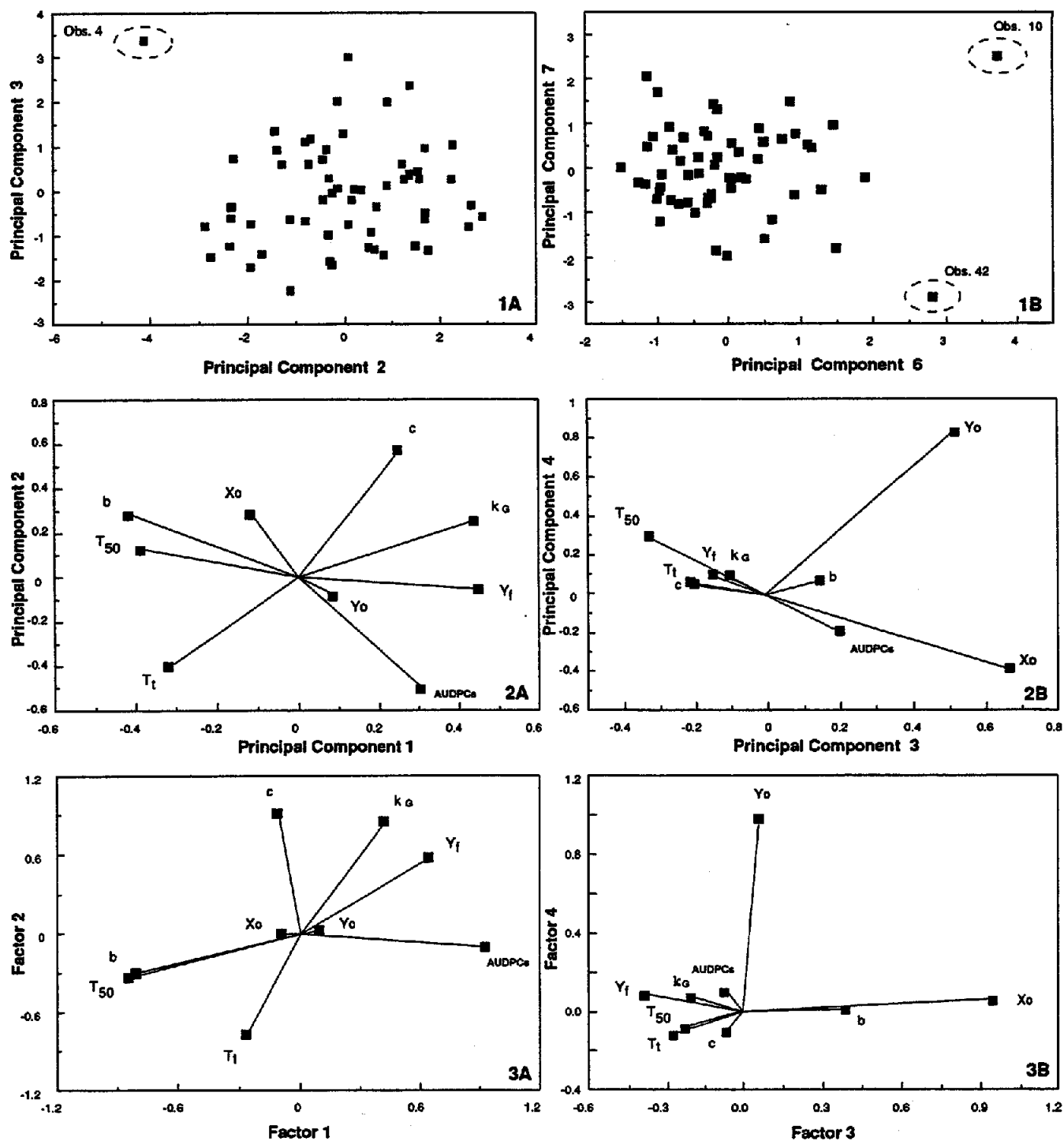
by using principal component analysis. Minor unrotated principal components were used in this study to identify redundant variables; however there is not a reason to compute principal components rotated to simplify eigenvectors structure (Hawkins and Fatti, 1984). Some problems arise with this approach. These will be discussed in the context of biplots below.

Elimination of redundant variables. Biplot case. The results of table 7 were used to display graphically the mayor axes. Four mayor eigenvalues were selected. This indicated that the matrix of data has a rank four. Therefore four variables are needed to describe its multivariate space. Theoretically, only a variable is needed to describe each subspace of rank one. Graphically, subspaces of rank 2 are well represented by a biplot (Gabriel, 1971; Bradu and Gabriel, 1978) and can be used to diagnostic correlations among variables and thus detect spurious variables. With this theoretical idea, biplots were made with only four mayor axes. The biplot made with the first two principal components represented a plane that account for 65.17% of the total data variability. Almost all this variability was explained in a additive manner by all the variables, except Y_0 and X_0 , since they were represented by the shorter vectors. This indicate that the longest vectors are variables that correspond best to the plane (Fig. 2A). Y_0 and X_0 (the

shorter vectors) were the poorest variables represented by this bidimensional space. Conversely, the biplot of axes 3 and 4 indicated Y_0 and X_0 as the best variable represented by this planar space (Fig. 2B).

This type of results were expected by Gabriel (1971) and Bradu and Gabriel (1978) because a biplot describe a plane and its vectors (variables) well represented in this space rather than a higher dimension. This explain why only some vectors are well represented in a particular combination of axes. Also it explain why the collinearity among variables observed in the correlation matrix (Table 4), were fairly well represented by these biplots. Notice that the biplots suggested some correlations which were not detected in the matrix of correlations, such as the negative correlation between Y_f and T_{30} indicated by vectors localized near to 180°. This problem is also observed in data presented by Anderson *et al.* (1990). Great part of their problem was because they display vectors only considering the first two axes. Thus they approximate their original matrix by a biplot of a matrix of rank two.

To correct the poor representation problem of these biplots, varimax rotation was apply to the first four axes. The minor principal components were not considered to avoid spread of variance and to loss



FIGURES 1 - 3. FIGURE 1. Plot of nonstandardized (1A) and standardized (1B) scores of 59 observations with respect to the principal component two and three, and six and seven respectively. The observations enclosed by dashed-circle were considered an outlier for the space defined by their axes. **FIGURE 2 and 3.** Approximation of correlations and relative importance of nine variables plotted in relation to major axes without (Fig. 2) and with varimax rotation (Fig. 3). Fig. 2A and Fig. 3A: Biplot display for the first and second principal components. Long vectors (lines) are variables well represented by the plane. Small vectors are not well represented in the bidimensional space. Small angles between vectors and vectors in the opposite direction indicate positive and negative correlation respectively. Description of variables is given in the text.

information on the most dominant individual source of variation for this axes (Jolliffe, 1989). Figure 3 shows that varimax rotation simplified the bidimensional space structure. The plane, represented by principal components 1 and 2 seem well explained by variables c and $AUDPCs$ (Figure 3A). Notice that these variables are almost perpendicular between them and to the axes. In general correlations between variables, such as b and T_{30} , T_1 and Y_f with k_G are now more concordant with those indicated in Table 3. The perpendicularity departure of these variables with respect to any axe indicate that these variables can be eliminated; at least from the space defined by axes 1 and 2. Similarly, the vectors of X_o and Y_o are still more reduced that before, reinforcing the idea that these variables does not belong to these space. The figure 3B indicate a more simple structure than the previous plot. Notice that Y_o and X_o are the longest vectors and are almost perpendicular to each other and to the axes. All the other variables are relatively small, suggesting that they do not have interpretability in this space and thus their correlations indicated in table 4 should not be taken into account.

Because $AUDPCs$ and c are fairly parallel to axes 1 and 2 respectively, and X_o and Y_o has the same behave for axes 3 and 4. We can conclude that each of these variables represent the linear subspace of their single axes to which they are parallel. Together these variables explain 89.61% of the matrix data variability and represent a four dimension multivariate space. These results agree with those obtained by using minor principal components of successive PCA starting with two low-correlated set of variables. Thus minor axes obtained with PCA for the whole matrix of data (9 variables), seems less appropriate to reject variables than those produced by partition of the data matrix or by biploting mayor axes, particularly after rotation. This will support the idea that strong correlations among variables will affect the PCA result (Hawkins and Fatti, 1984). However, these results are not conclusive and could require more study. Jolliffe (1973), demonstrated that different methods of discarding variables can select different set of parameters. However, he did not generate general conclusions from his research.

Biplots of mayor principal components were used to select the most important variables, however there is not a reason to produce biplots of reduced minor principal components. In this case the variables well represented in a plane are those that directly can be eliminated rather than to be selected. An advantage of this approach is that vectors of spurious variables are well represented in the

plane because their usual larger eigenvectors coefficients. Varimax rotation can also be apply only to this reduced set of minor axes to gain in interpretation, rather than to rotate the whole number of principal components as was suggested by Hawkins and Fatti, 1984. This approach avoid 'contamination' of the variance accounted for the minor principal components by the variance explained in the mayor axes (Jolliffe, 1989). However, FACTOR analysis performed with SAS does not have a direct option to rotate defined axes. Apparently, this problem can be overcome with BMDP package (Jolliffe, 1988). Interestingly, as was indicated by Gabriel (1974), 'threeplot' could also be used to diagnostic a matrix of data. The advantage of this plot is that a matrix of rank three is directly displayed. The same interpretation apply as before for this type of matrix. In the context of detecting multivariate outliers that are not clearly aberrant for a particular variable, this type of approach could be useful. Particularly because a tridimensional rather than a bidimensional space is explored. To achieve this purpose, plots of observations rather than variables vectors should be utilized.

Varimax rotation was useful in this analysis because it was possible simplify eigenvectors structure. However this can not be generalized to all types of matrices. Jolliffe (1989) indicated that rotation is suggested when the selected eigenvalues are close to each other. This evidence that the eigenvectors are not well defined. In this analysis the third and fourth eigenvalues were close to each other (1.35 and 0.84). However, even for the first and second eigenvalues, which appear not relatively close (4.24 and 1.61) the varimax rotation was useful to simplify structure. Hawkins and Fatti (1984), proposed that axes rotation is a powerful tool when there are strong correlations among variables. They point out that strong correlations are indicated by very small eigenvalues associated to the last principal components. Table 3 shows that correlations of variables were observed in this data. Thus there were small eigenvalues, such as 0.08 and 0.03, associated to the principal components eight and nine (Table 7).

PROC PRINCOMP vs PROC FACTOR. This analysis demonstrated that principal component analysis of the correlation matrix with *proc factor* can achieve the same goals than *proc princomp*. However for the objectives of this study *proc factor*. has three additional advantages: the scores are directly standardized, biplots can be obtained and axes can be rotated.

Princomp and *factor* procedures exhibit the same eigenvectors, eigenvalues and scores when the former

approach is executed with the option *std* to standardize the scores. When it is omitted only the eigenvalues and eigenvectors are the same. There are some differences between *Princomp* and *factor* in relation to computational performance. These differences are indicated SAS/STAT user's manual (1988). Since rotation of axes was found in this study as an effective tool for detecting outliers and removing variables, *proc factor* is recommended.

We wish to thank Researchers and Technicians of Research Center for the Development of the Humid Tropics at Paso San Juan, Veracruz, México for technical assistance in the data obtention.

LITERATURE CITED

- Anderson, W.F., Beute, M.K., Wynne, J.C., and Wongkaew, S. 1990. Statistical procedures for assessment of resistance in a multiple foliar disease complex of peanut. *Phytopathology* 80:1451-1459.
- Bradu, D. and Gabriel K.R. 1978. The biplot as a diagnostic tool for models of two-ways tables. *Thechnometrics* 20: 47-67.
- Campbell, C. L., and Madden, L. V. 1990. Introduction to Plant Disease Epidemiology. John Wiley Sons, New York. 532 pp.
- Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58: 453-467.
- Hawkins, D. M., and Fatti, L.P. 1984. Exploring multivariate data using the minor principal components. *The Statistician* 33: 325-338.
- Jolliffe, I. T. 1989. Rotation of III-defined principal components. *Appl. Statist.* 38: 139-147.
- Jolliffe, I. T. 1982. A note on the use of principal components in regression. *Appl. Statist.* 31:300-303.
- Jolliffe, I. T. 1973. Discarding variables in a principal component analysis. II: Real data. *Appl. Statist.* 22:21-31.
- Lyman, O. 1988. An Introduction to Statistical Methods and Data Analysis. 3ed. PWS CO. 855 p.
- Manly, B. F. J. 1986. Multivariate Statistical Methods. A Primer. Chapman and Hall. New Kork, NY. 159pp.
- Mora-Aguilera, G., Téliz, D., Campbell, C.L., and Avila, C. 1992. Temporal and spatial development of papaya ringspot in Veracruz, México. *Journal of Phytopathology* 136: 27-36.
- Pack, P., Jolliffe, I.T., and Morgan, J.T. 1988. Influential observation in principal component analysis: a case study. 15: 39-52.
- Pennypacker, S.P., Knoble, H.D., Antle, C.E., and Madden, L.V. 1980. A flexible model for studying plant disease progression. *Phytopathology* 70:232-235.
- Ralston, M. L., and Jennrich, R. I. 1978. DUD, a derive free algorithm for nonlinear least squares. *Technometrics* 20: 7-14.
- Reyment, R. A. 1988. Reification of classical multivariate statistical analysis in morphometry. Pp. 123-144 in *Proceeding of the Michigan Morphometrics workshop*. F.J. Rohlf and F.L. Bookstein, eds. University of Michigan Museum of Zoology. 380pp.
- SAS Institute, Inc. 1988. SAS/STAT User's guide. Release 6.03 edition. SAS Institute, Inc. Cary, NC. 1028 pp.