Descriptive Measures of Multivariate Scatter and Linear Dependence*

Multivariate Scatter and Linear Dependence

Daniel Peña† and Julio Rodríguez

Department of Statistics and Econometrics, Universidad Carlos III de Madrid, Spain‡
E-mail: dpena@est-econ.uc3m.es

In this paper we propose two new descriptive measures for Multivariate Data: The Effective Variance and the Effective Dependence. These measures have a direct geometric and statistical interpretation and can be used to compare groups with different number of variables. The contribution of these measures to understanding multivariate data is illustrated by several examples.

Key Words: correlation, principal components, variability.

1. INTRODUCTION

The trace and the determinant of the covariance matrix of a sample of multivariate data are often used as descriptive measures of multivariate variability. However, these measures cannot be used to compare the variability of sets of variables with different dimensions. The linear dependence between two variables is usually measured by the correlation coefficient, introduced by Galton and Pearson a century ago (see Rodgers and Nicewanders [14] for a brief history of this coefficient and 13 interpretations of its value). However, we do not have a simple measure of linear dependence among a set of variables that can be used as a standard descriptive measure in any dimension.

This paper proposes two new descriptive measures for multivariate data: The Effective Variance and the Effective Dependence. These measures have a direct geometric and statistical interpretation and can be used to compare groups with different numbers of variables. The paper is organized as follows. In the next

* This research has been sponsored by DGES (Spain) under project BEC 2000-167 and the Cátedra BBVA de Calidad.

† Correspondence may be sent to Prof. Daniel Peña, c/o Department of Statistics and Econometrics.

‡ email: dpena@est-econ.uc3m.es.

section we present some conditions that a useful measure of multivariate variability must satisfy. It is shown that neither the trace nor the determinant of the covariance matrix satisfy these conditions and the Effective Variance is proposed. In Section 3 we extend these conditions to a multivariate measure of linear relationship and the Effective Dependence is introduced. It is shown that the Effective Dependence can be used to estimate the number of principal components required to explain 90% of the data variability. Section 4 discusses the sample distributions of these measures. Section 5 illustrates their use in two examples.

2. A MEASURE OF MULTIVARIATE VARIABILITY

Let $\mathbf{X}$ be a p dimensional random variable with ¯nite covariance matrix $\mathbf{\Sigma}_{\mathbf{X}}$. We are interested in building a scalar measure of scatter $V(\mathbf{X})$ that summarizes in some optimal way the multivariate variability of the random variable. This measure should be useful for comparing the scatter of random variables of different dimension when they are measured on the same units. With this objective in mind, we want ¯rst that this measure of variability depends on the covariance matrix. Thus we are only taking into account linear relationships between the components of the $\mathbf{X}$. Second, given two vectors $\mathbf{X}$ and $\mathbf{Y}$ with covariance matrices $\mathbf{\Sigma}_{\mathbf{X}}$ and $\mathbf{\Sigma}_{\mathbf{Y}}$ we de¯ne the additional linear variability introduced by $\mathbf{Y}$ in the vector $\mathbf{Z}' = [\mathbf{X}'\mathbf{Y}']$ over the variability of $\mathbf{X}$ by

$$\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}} = \mathbf{\Sigma}_{\mathbf{Y}} - \mathbf{\Sigma}_{\mathbf{YX}}\mathbf{\Sigma}_{\mathbf{X}}{}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}. \tag{1}$$

Note that $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ is the covariance of the random variable $\mathbf{Y} - E(\mathbf{Y}) - \mathbf{B}(\mathbf{X} - E(\mathbf{X}))$, where $\mathbf{B} = \mathbf{\Sigma}_{\mathbf{X}}{}^{-1}\mathbf{\Sigma}_{\mathbf{XY}}$ and will be equal to the covariance of the random variable $\mathbf{Y}|\mathbf{X}$ only if $E(\mathbf{Y}|\mathbf{X})$ is linear on $\mathbf{X}$. If the linear variability introduced by $\mathbf{Y}$ exceeds the variability already present in $\mathbf{X}$, the variability of $\mathbf{Z}$ should also be greater than the variability in $\mathbf{X}$.

Thus, we establish that a useful scalar measure must satisfy the following properties:

(a) $V(\mathbf{X}) = g(\mathbf{\Sigma}_{\mathbf{X}})$. That is, the measure depends only of the covariance matrix.

(b) If X is scalar then $V(X) = var(X)$.

(c) If $\mathbf{Y} = \mathbf{QX}$ where $\mathbf{Q}$ is an orthogonal matrix, then $V(\mathbf{Y}) = V(\mathbf{X})$.

(d) If $\mathbf{Y} = \mathbf{BX} + \mathbf{C}$ where $\mathbf{B}$ is a non singular diagonal matrix and $\mathbf{C}$ a vector, then $V(\mathbf{Y}) = f^2(\mathbf{B})V(\mathbf{X})$.

(e) $V(\mathbf{X}) = 0$ if and only if $|\mathbf{\Sigma}_{\mathbf{X}}| = 0$.

(f) Let $\mathbf{Z}' = [\mathbf{X}'\ \mathbf{Y}']$ be a random vector of dimension $p + q$ where $\mathbf{X}$ and $\mathbf{Y}$ are random variables of dimension p and q respectively. Let us de¯ne the additional variability introduced by $\mathbf{Y}$ with respect to the one of $\mathbf{X}$, by $V(\mathbf{Y} : \mathbf{X}) = g(\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}})$, where $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X}}$ is given by (1). Then $V(\mathbf{Z}) \geq V(\mathbf{X})$ if and only if $V(\mathbf{Y} : \mathbf{X}) \geq V(\mathbf{X})$ and $V(\mathbf{Z}) \leq V(\mathbf{X})$ if and only if $V(\mathbf{Y} : \mathbf{X}) \leq V(\mathbf{X})$.

The two most often used measures to describe scatter about the mean in multivariate data are the *total variation*, Seber [15], given by $tr\,(\mathbf{\Sigma_X}) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$, and the *generalized variance*, Wilks [18], given by $|\mathbf{\Sigma_X}| = \lambda_1\lambda_2\cdots\lambda_p$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ are the eigenvalues of the covariance matrix $\mathbf{\Sigma_X}$. The former is often used as a measure of variation in principal components analysis and the latter plays an important role in maximum likelihood estimation and in model selection. It is straightforward to check that the total dispersion satisfies properties (a) to (c) and the generalized variance properties (a) to (e). Neither of them satisfies property (f): including an additional variable Y in a data set cannot decrease the trace, whereas it is well known that the determinant in dimension $p-1$, $|\mathbf{\Sigma}_{p-1}|$, and the determinant in dimension p, $|\mathbf{\Sigma}_p|$ are related by

$$|\mathbf{\Sigma}_p| = |\mathbf{\Sigma}_{p-1}|\,\sigma_p^2\left(1 - R^2_{p.1\cdots p-1}\right) \tag{2}$$

where σ_p^2 is the variance of the pth variable and $R^2_{p.1\cdots p-1}$ is the squared multiple correlation coefficient between the variable p and the variables $1,\ldots,p-1$. Thus, if we choose the determinant of the covariance matrix as a scalar measure of scatter, we can make $|\mathbf{\Sigma}_p|$ greater or smaller than $|\mathbf{\Sigma}_{p-1}|$ by choosing a variable Y such that $V(Y|\mathbf{X}) = \sigma_p^2\left(1 - R^2_{p.1\cdots p-1}\right)$ is greater or smaller than one.

The generalized variance is a measure of the hypervolume that the distribution of the random variables occupies in the space. Even if we standardize all the variables, we cannot compare generalized variances in set of different dimensions because, according to (2), this measure cannot increase by introducing new standardized variables. It is clear that with this hypervolume interpretation we cannot compare sets of different dimensions. An intuitive alternative is to use the average scatter in any direction. We propose the name *Effective Variance*, for the measure given by

$$V_e(\mathbf{X}) = |\mathbf{\Sigma_X}|^{1/p} = (\lambda_1\lambda_2\cdots\lambda_p)^{1/p} \tag{3}$$

that is, the geometrical mean of the univariate variances of the principal components of the data. It can also be interpreted as the length of the side of the hypercube whose volume is equal to the determinant of $\mathbf{\Sigma}_p$. Also, we can define the *Effective Standard Deviation* by

$$SD_e(\mathbf{X}) = \{V_e(\mathbf{X})\}^{1/2} = |\mathbf{\Sigma_X}|^{1/(2p)}.$$

It is straightforward to check that the Effective Variance satisfies properties (a) to (e). In order to check property (f) note that,

$$|\mathbf{\Sigma_Z}|^{1/(p+q)} = |\mathbf{\Sigma_X}|^{1/(p+q)}\left|\mathbf{\Sigma_{Y|X}}\right|^{1/(p+q)} \tag{4}$$

and, for instance, the condition $\left|\mathbf{\Sigma_{Y|X}}\right|^{1/q} \geq |\mathbf{\Sigma_X}|^{1/p}$ is equivalent to $\left|\mathbf{\Sigma_{Y|X}}\right|^{1/(p+q)} \geq |\mathbf{\Sigma_X}|^{q/(p(p+q))}$ which implies, by using (4), that $|\mathbf{\Sigma_Z}|^{1/(p+q)} \geq |\mathbf{\Sigma_X}|^{1/p}$.

From the properties of the geometric mean we have that

$$\lambda_p \leq V_e(\mathbf{X}) \leq \frac{1}{p}\sum_{i=1}^{p}\lambda_i$$

where λ_p is the minimum eigenvalue of $\mathbf{\Sigma_X}$.

Remark 1. Condition (f) implies that if we have two independent vectors, $\mathbf{X}$ and $\mathbf{Y}$, with the same measure of scatter, $V_e(\mathbf{X}) = V_e(\mathbf{Y})$, then $V_e(\mathbf{Z}) = V_e(\mathbf{X}) = V_e(\mathbf{Y})$.

Remark 2. Note that an alternative de¯nition for multivariate scatter is the Average Total Variation, $ATV(\mathbf{X}) = (1/p)tr(\mathbf{\Sigma_X})$, which does not satisfy properties (d) to (f). This measure does not take into account the covariance structure.

3. A MEASURE OF MULTIVARIATE LINEAR DEPENDENCE

The analysis in the previous section suggests a way to build a scalar measure of multivariate linear dependence that summarizes the linear relationships between the variables and can be applied in sets of different dimensions. We are interested in measures that are functions of the correlation matrix of the variables and, as before, we want to take into account linear relationships. We have to specify which properties a measure of dependence, $D(\mathbf{X})$, of a random vector $\mathbf{X}$, must have when the dimension of the vector is changed. Suppose that we increase its dimension by adding a new set of random variables $\mathbf{Y}$, to form the new vector $\mathbf{Z}' = [\mathbf{X}'\ \mathbf{Y}']$. Then the change of the dependence must depend on (i) the correlation matrix of the $\mathbf{Y}$ vector and (ii) the correlation between $\mathbf{X}$ and $\mathbf{Y}$ as measured by the matrix of cross correlations $\mathbf{R_{XY}}$. The additional correlation introduced by the $\mathbf{Y}$ variables can be measured by

$$\mathbf{R_{Y|X}} = \mathbf{R_Y}(\mathbf{I} - \mathbf{R_Y}^{-1}\mathbf{R_{YX}}\mathbf{R_X}^{-1}\mathbf{R_{XY}}), \tag{5}$$

which is the product of the correlation matrix of the $\mathbf{Y}$ variables and a correction term that depends on the canonical correlations between the vectors $\mathbf{Y}$ and $\mathbf{X}$.

Thus, we establish that the dependence measure, $D(\mathbf{X})$ must satisfy the following properties:

(a) $D(\mathbf{X}) = g(\mathbf{R_X})$. That is, the measure depends only of the correlation matrix.

(b) If X is scalar then $D(X) = 0$.

(c) If $\mathbf{Y} = \mathbf{QX}$ where $\mathbf{Q}$ is an orthogonal matrix then $D(\mathbf{Y}) = D(\mathbf{X})$.

(d) If $\mathbf{Y} = \mathbf{BX} + \mathbf{C}$ where $\mathbf{B}$ is a non singular diagonal matrix and $\mathbf{C}$ a vector then $D(\mathbf{Y}) = D(\mathbf{X})$.

(e) $0 \leq D(\mathbf{X}) \leq 1$, and $D(\mathbf{X}) = 1$ if and only if we can ¯nd a vector $\mathbf{a} \neq \mathbf{0}$ and $\mathbf{b}$ such that $\mathbf{a}'\mathbf{X} + \mathbf{b} = 0$. Also $D(\mathbf{X}) = 0$ if and only if $\mathbf{\Sigma_X}$ is diagonal.

(f) Let $\mathbf{Z}' = [\mathbf{X}'\,\mathbf{Y}']$ be a random vector of dimension $p+q$ where $\mathbf{X}$ and $\mathbf{Y}$ are random variables of dimension p and q respectively. We de¯ne the additional dependence as the additional correlation introduce by $\mathbf{Y}$ by $D(\mathbf{Y}:\mathbf{X}) = g(\mathbf{R}_{\mathbf{Y}|\mathbf{X}})$, where $\mathbf{R}_{\mathbf{Y}|\mathbf{X}}$ is given by (5). Then $D(\mathbf{Z}) \geq D(\mathbf{X})$ if and only if $D(\mathbf{Y}:\mathbf{X}) \geq D(\mathbf{X})$, and $D(\mathbf{Z}) \leq D(\mathbf{X})$ if and only if $D(\mathbf{Y}:\mathbf{X}) \leq D(\mathbf{X})$.

A standard measure of dependence in the bivariate case is ρ^2, the squared of the correlation coef¯cient. In the multivariate case, a possible generalization is $1-|\mathbf{R}_\mathbf{X}|$, where $\mathbf{R}_\mathbf{X}$ is the correlation matrix. This measure satis¯es properties (a) to (e), but again it is not appropriate for comparing the dependence structure between datasets with different numbers of variables. An alternative measure is $(1-|\mathbf{R}_\mathbf{X}|)^{1/p}$ which has the advantage that for $p=2$ it is equal to ρ, the linear correlation coef¯cient. However, this de¯nition does not satisfy property (f).

By analogy to the Effective Variance we de¯ne the*Effective Dependence* by

$$D_e(\mathbf{X}) = 1 - |\mathbf{R}_\mathbf{X}|^{1/p} \tag{6}$$

and it is easy to check that it satis¯es properties (a) to (e). Property (f) is obtained by noting that as in (4)

$$|\mathbf{R}_\mathbf{Z}|^{1/(p+q)} = |\mathbf{R}_\mathbf{X}|^{1/(p+q)} \left|\mathbf{R}_{\mathbf{Y}|\mathbf{X}}\right|^{1/(p+q)}$$

and the proof is the same as for the Effective Variance.

Remark 3. Condition (d) implies that the measures will be invariant if we change the sign of any elements of the vector $\mathbf{X}$. This is in agreement with condition (e) which implies that $D(\mathbf{X})$ is always positive.

Remark 4. The Effective Dependence for a bivariate random variable is $1-\sqrt{1-\rho^2}$ which is a useful measure of linear relationship. Let (y,x) be the components of the bivariate random vector, and $\sigma^2_{y|x} = \sigma^2_y(1-\rho^2)$. Then the Effective Dependence is $(\sigma_y - \sigma_{y/x})/\sigma_y$ and it directly provides the proportion of reduction in the standard deviation of the random variable due to the use of the linear information provided by the regressor. In the next section we will extend this idea for any dimension.

Remark 5. An alternative de¯nition for a dependence measure is$1-|\mathbf{R}_p|^{1/(p-1)}$ that in the bivariate case is equal to the squared correlation coef¯cient. For large p this measure will be very close to the Effective Dependence but we prefer the exponent $1/p$ for symmetry with respect to the Effective Variance.

3.1. Some properties of the Effective Dependence

Firstly, the Effective Dependence represents the average proportion of explained variability among the variables. To see this, note that by repeated use of (4), we can write

$$|\mathbf{R}_\mathbf{X}| = \left(1-R^2_{p.1\cdots p-1}\right)\left(1-R^2_{p-1.1\cdots p-2}\right)...\left(1-R^2_{2.1}\right). \tag{7}$$

where the ith term represents the proportion of unexplained variation in a regression between the $p-i+1$ variable and the variables $p-i, p-i-1, \ldots, 1$. As the squared correlation can always be interpreted as $R^2 = 1 - RSS/TSS$, where RSS is the residual sum of squares and TSS the total sum of squares and calling $RSS(i|1\cdots i-1)$ to the residual sum of squares in the regression of the ith variable on the $i-1, \ldots, 1$ and $TSS(i)$ to the total variability in this regression, we can write:

$$|\mathbf{R_X}|^{1/p} = \left\{ \frac{RSS(p|p-1,\cdots,1)\cdots RSS(2|1)RSS(1)}{TSS(p)\cdots TSS(2)TSS(1)} \right\}^{\frac{1}{p}} = \frac{\overline{RSS}}{\overline{TSS}},$$

where $RSS(1) = TSS(1)$ and $\overline{RSS}$ and $\overline{TSS}$ are the geometric means of the residual sum of squares and the total sum of squares of all the regressions. Note that this measure is invariant to any permutation of the variables. Thus, the Effective Dependence can be written as

$$D_e(\mathbf{X}) = 1 - \frac{\overline{RSS}}{\overline{TSS}}.$$

This interpretation also holds when the set of variables can be partitioned as $\mathbf{Z}' = [\mathbf{X}'\mathbf{Y}']$, where $\mathbf{X}$ has dimension p and $\mathbf{Y}$ has dimension q and suppose that $p \geq q$. We have

$$D_e(\mathbf{Z}) = 1 - \left\{ \left(1 - R^2_{xp.1\cdots p-1}\right) \cdots \left(1 - R^2_{x2.1}\right) \right\}^{\frac{1}{p+q}}$$
$$\left\{ \left(1 - R^2_{yq.1\cdots q-1}\right) \cdots \left(1 - R^2_{y2.1}\right) \right\}^{\frac{1}{p+q}} \left\{ \prod\nolimits_{i=1}^{l} \left(1 - r_i^2\right) \right\}^{\frac{1}{p+q}},$$

where $l = \min(p,q) = q$ and r_i^2 are the canonical correlation coef¯cients between the two sets of variables. This expression shows that if the two sets are uncorrelated, then $D_e(\mathbf{Z})$ is just the average of the internal dependence. When the two sets are correlated the Effective Dependence is an average of the internal dependence and the cross dependence as measured by the canonical correlation coef¯cients.

Note that the Effective Dependence satis¯es the following inequality

$$\frac{1}{p} \sum_{i=1}^{p} R^2_{i.1\cdots(i-1)} \leq D_e(\mathbf{X}) \leq 1,$$

obtained by noting that from (7),

$$1 - |\mathbf{R_X}|^{1/p} = 1 - \left[\prod_{i=1}^{p} (1 - R^2_{i.1\cdots(i-1)}) \right]^{(1/p)}$$

and by using the properties of the geometric mean. This inequality is satis¯ed for all permutations of the variables, since if $\mathbf{Y} = \mathbf{PX}$, where $\mathbf{P}$ is a permutation matrix, then $D_e(\mathbf{Y}) = D_e(\mathbf{X})$. Note that if $\mathbf{P}$ is a permutation matrix then $|\mathbf{P}| = \pm 1$ and $|\mathbf{\Sigma_Y}| = |\mathbf{P}|^2 |\mathbf{\Sigma_X}| = |\mathbf{\Sigma_X}|$ implies that $|\mathbf{R_Y}| = |\mathbf{R_X}|$.

Secondly, the Effective Dependence is a measure of the lack of sphericity of the standardized variables. Anderson (1984, p. 427) de¯nes sphericity as

$$\psi(\mathbf{\Sigma}_p) = \frac{|\mathbf{\Sigma}_p|^{1/p}}{(1/p)tr\left(\mathbf{\Sigma}_p\right)},$$

and he uses this measure for testing the hypothesis $H_0 : \mathbf{\Sigma} = \sigma^2\mathbf{I}$. If $\psi = 1$, then the geometric mean of the eigenvalues is equal to the arithmetic mean and all the variables are uncorrelated, and the shape of the data is a sphere. When ψ tends to zero, the data moves away from sphericity and when $\psi = 0$, we are in a lower dimension, and the ellipsoid is degenerate. For standardized variables $\psi(\mathbf{R}_p) = |\mathbf{R}_p|^{1/p}$, and

$$D_e(\mathbf{X}) = 1 - \psi(\mathbf{R}_p).$$

Thirdly, when all the off diagonal values of the correlation matrix are equal and the number of variables is large, the Effective Dependence will converge to this common correlation value.

To illustrate this property, suppose that the correlation matrix of a vector of p random variables has the simple structure

$$\mathbf{R}_p = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}.$$

Then the coef¯cient of determination in the regression of any variable with respect to the rest $R^2_{p.1\cdots p-1} \to \rho$ as $p \to \infty$. This result, shown by Mustonen [13], is a consequence of

$$\lim_{p\to\infty} \left(1 - R^2_{p.1\cdots p-1}\right) = (1-\rho) \lim_{p\to\infty} \left\{ \frac{1+(p-1)\rho}{1+(p-2)\rho} \right\} = (1-\rho)\,.$$

Then, it is easy to show that, for the generalized variance,

$$\lim_{p\to\infty} (1 - |\mathbf{R}_p|) = \lim_{p\to\infty} \left[1 - (1-\rho)^{p-1}\{1+(p-1)\rho\}\right] = 1 \qquad \forall \rho \in (0,1)\,,$$

whereas for the Effective Dependence,

$$\lim_{p\to\infty} D_e\left(\mathbf{X}\right) = \lim_{p\to\infty}\left(1 - \left|\mathbf{R}_p\right|^{1/(p-1)}\right) = \tag{8}$$
$$= \lim_{p\to\infty}\left[1-(1-\rho)\left\{1+(p-1)\,\rho\right)^{1/(p-1)}\right] =$$
$$= \rho \qquad \forall\rho\in(0,1)\,,$$

which provides an interesting interpretation of the correlation coef¯cient as the limiting average proportion of explain variability in this situation.

Finally, the Effective Dependence can be used to predict the number of principal components required to explained a given proportion of the variability of the data.

[FIGURE 1 ABOUT HERE]

One would expect that the larger the global correlation structure the smaller the number of principal components or factors needed to describe the linear properties of the observed data. A useful measure of linear dependence should inherit this property and we will show that the Effective Dependence is strongly related to the proportion of components needed to summarize the data (see Figure 1). Suppose that we have a sample of p standardized variables and let λ_i, $i = 1, ..., p$ be the eigenvalues of the correlation matrix of the data. We want to study the relationship between the D_e of the sample and the proportion of components, h/p, needed to explain 90% of the total variability. We carried out a simulation study by generating random correlation matrices of dimension p as follows: (1) the eigenvalues of the correlation matrix are drawn from a $Beta(\alpha, \beta)$ distribution, with α and β chosen from a grid in the interval $(0,3)^2$, obtaining 900 pairs of parameters (α_i, β_i). (2) The values are normalized so that their sum is p. For each ¯xed value p, we generated 900 matrices. This process was performed for $p = 40, 80, \ldots, 440$, so that 9900 correlation matrices were generated in total. For each one of these matrices, we calculate $\left(\frac{h}{p}\right) \in [0,1]$ and the D_e. We observed that the relation between $\left(\frac{h}{p}\right)$ and $D_e(\mathbf{X})$ is a sigmoid, but in the interval $D_e(\mathbf{X}) \in [0.1, 0.9]$ we can approximate it by the linear relation

$$\frac{h}{p} = 0.8230 - 0.492 D_e(\mathbf{X})$$

with $R^2 = 0.97$ (see Figure 1). This relationship can be approximated by

$$h = p(0.8 - 0.5 D_e(\mathbf{X})). \tag{9}$$

To illustrate this result, we present the analysis of Jeffers' [9] pine pitprops data, taken from Mardia et al. ([11] pp. $176-178$, $225-227$). This data set has 180 observations of pitprops cut from the Corsican pine tree. The data have 13 variables (X) measured on each prop. The Effective Dependence of these data is

$D_e(\mathbf{X}) = 0.563$. From equation (9) we obtain that $h = 0.519p$ and the estimated number of principal components explaining 90% of the total variability is 6.74. Thus we need 6 or 7 components. The eigenvalues of the correlation matrices are: 4.22, 2.38, 1.88, 1.11, 0.91, .82, .58, .44, .35, .19, .05, .04 and .04. For the ¯rst 6 components, the cumulative variability is 87.1% and if the seventh component is added, it is 91.5%. Therefore more than 90% of the cumulated variability is obtained considering the ¯rst 7 components. Jeffers took the ¯rst6 components in his analysis, because of their clear physical interpretation.

4. SAMPLE DISTRIBUTIONS

The sample distribution of the Effective Variance can be obtained from existing results on the generalized variance (see [2, 12]). The generalized variance is usually estimated by the sample generalized variance, $det(\mathbf{S}_p)$, where $\mathbf{S}_p$ is the sample covariance matrix with dimension $p \times p$. In the case of Effective Variance it is estimated by the pth root of the generalized sample variability. The following two lemmas derive the distribution of $(det(\mathbf{S}_p))^{1/p}$ when $\mathbf{S}_p$ is computed with a sample of size $N = n+1$, from the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_p)$ distribution. In this case $\mathbf{S}_p$ follows a Wishart distribution with n degrees of freedom and covariance matrix $(1/n)\boldsymbol{\Sigma}_p$, $W_p(n, (1/n)\boldsymbol{\Sigma}_p)$. The two lemmas characterize the asymptotic distribution and an approximation of the exact distribution by the sample Effective Variance.

LEMMA 4.1. *Let $\mathbf{S}_p$ be a $p \times p$ sample covariance matrix from the $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_p)$ with n degrees of freedom. Then*

$$\sqrt{n}\left(|\mathbf{S}_p|^{1/p} \,/\, |\boldsymbol{\Sigma}_p|^{1/p} - 1\right)$$

is asymptotically normally distributed with mean 0 and variance $2/p$.

Proof. The asymptotic distribution of the Effective Variance can be obtained from the asymptotic normality of the generalized variance. Anderson [2], shows that $\sqrt{n}\,(|\mathbf{S}_p| \,/\, |\boldsymbol{\Sigma}_p| - 1)$ is asymptotically normal with mean 0 and variance $2p$. Then, applying the $\delta-$method (see Ser©ing [16] p. 118)), for $g(x) = x^{1/p}$, it follows, that V_e is also asymptotically normally distributed, with mean 0 and variance $2/p$. ▮

LEMMA 4.2. *The exact distribution for the pth root of $|\mathbf{S}_p| \,/\, |\boldsymbol{\Sigma}_p|$ is*

$$|\mathbf{S}_p|^{1/p} \,/\, |\boldsymbol{\Sigma}_p|^{1/p} \sim \Gamma\left(\frac{p(n-p)}{2}, \frac{p(n-1)}{2}\left(1 - \frac{(p-1)(p-2)}{2n}\right)^{1/p}\right).$$

Proof. Using the results of Hoel [8] related to the exact distribution for the pth root of $|\mathbf{A}_p| \,/\, |\mathbf{\Sigma}_p|$, we have

$$|\mathbf{A}_p|^{1/p} \,/\, |\mathbf{\Sigma}_p|^{1/p} \sim \Gamma\left(\frac{p(n-p)}{2}, \frac{p}{2}c\right),$$

where

$$c = \frac{1}{2}\left(1 - \frac{(p-1)(p-2)}{2n}\right)^{1/p}$$

and $|\mathbf{A}_p|$ is $(n-1)^p\, |\mathbf{S}_p|$. Applying the property that, $X \sim \Gamma(\alpha, \beta) \rightarrow dX \sim \Gamma(\alpha, \beta/d)$, then

$$|\mathbf{S}_p|^{1/p} \,/\, |\mathbf{\Sigma}_p|^{1/p} \sim \Gamma\left(\frac{p(n-p)}{2}, \frac{p(n-1)}{2}\left(1 - \frac{(p-1)(p-2)}{2n}\right)^{1/p}\right).$$

∎

The exact distribution of $D_e(\mathbf{X})$ can be easily obtained from the exact distribution of $|\mathbf{R}_{\mathbf{X}}|$ given in Gupta and Rathie [7]. The asymptotic distribution of $-n\log|\mathbf{R}_X|$, under the hypothesis that $\mathbf{R}_{\mathbf{X}} = \mathbf{I}$, is a χ^2 with $p\,(p-1)\,/2$ degrees of freedom (see Box [3]). Thus, the asymptotic distribution of $npD_e(\mathbf{X})$ is a χ^2 with the same degrees of freedom.

5. EXAMPLES

To illustrate the information provided by for the Effective Variance and the Effective Dependence in a descriptive analysis of multivariate data, we apply them to two well known sets of data. The ¯rst is the Fisher Iris data, originally due to Anderson [1] and analyzed by Fisher [6] in his seminal paper on discriminant analysis. These data correspond to measures of three species of ©owers called, Iris Setosa, Iris Versicolor and Iris Virginica. There are 50 specimens of each species and four variables: Y_1=sepal length, Y_2=sepal width, Y_3=petal length and Y_4=petal width, all measured in cm.

[FIGURE 2 ABOUT HERE]

Figure 2 shows the scatterplot of the Iris data in the variables Y_1 and Y_4. The specimens for Setosa are squares, for Versicolor circles and for Virginica triangles. In ¯gure 2 two concentric circles centred in the mean of each group are plotted. The circle in solid line shows the observed scatter in the projected data and it has radius $2 \times SD_e(\mathbf{\Sigma}_{14}^{(i)})$, where $\mathbf{\Sigma}_{14}^{(i)}$ is the covariance matrix of variables Y_1 and Y_4 in the group i. The second circle, in dotted line, shows the real multivariate scatter and it has radius $2 \times SD_e(\mathbf{\Sigma}^{(i)})$, where $\mathbf{\Sigma}^{(i)}$ is the covariance matrix in the group i, for all variables. The similarity of the two circles indicates that the dispersion in the projected data is similar to the dispersion in the multivariate data.

Figure 2 shows that the circles are similar in the species Setosa, whereas in the two other groups, the multivariate dispersion is slightly inferior than the projected dispersion. If we compare the multivariate dispersion between the three groups, using the dotted circles, small differences in the dispersion between groups are observed.

[TABLE 1 ABOUT HERE]

In Table 1, some scatter measures for each group in the Iris data are shown. The ¯rst group of measures correspond to all variables and the second group to the projected data shown in the scatterplot in Figure 2. The total variability and the generalized variance do not provide a descriptive information to understand the data, and are not appropriate for comparing the variance in sets of different dimensions. The ATV and the new measure V_e provide information over the scatter in each group in units which are comparable in dimensions 4 and 2, corresponding to the multivariate dispersion and in the dispersion in the projected data. The ATV for each group is similar to the ATV for each group in the projected data, but this measure does not take into account the covariance structure of the variables in each group. The fourth row in Table 1 shows the V_e in each group for all variables. If we compare this V_e with the V_e for the groups in the projected data, we can see a clear resemblance in the species Setosa and small differences in the two other species. In the last row of each set of measures, the ratio between V_e and ATV is shown, which is the sphericity. Based on this measure for each group, we can observe that the sphericity is higher in the projected data than in the original data. Moreover, in the species Versicolor, the sphericity is smaller than in the rest of the groups. This descriptive analysis of Iris data shows differences, in form and scatter, between the covariance matrix in the groups. This conclusion coincides with the result shown by Krzanowski and Radley [10], over the difference in scatter in each species.

[FIGURE 3 ABOUT HERE]

To illustrate the information provided by the Effective Dependence we consider the data on air quality measurements in the New York metropolitan area from May 1, 1973 to September 30, 1973 from Cleveland et al. [4]. Only the $n = 111$ complete cases are considered here. This data set is obtained for studying the relationship between the variable Ozone concentration in parts per billion, X_1, with the variables Solar Radiation in Langleys (X_2), Wind Speed in miles/hour (X_3) and Temperature in degrees F (X_4). As the variables are measured in different units, we will study the standardized data. Figure 3 shows a scatterplot matrix of the standardized Ozone data. In each scatterplot we present three vectors with equal length. The angle between X_i and Z_1 shows the Effective Dependence between the variables X_i and X_j, in such a way that $\cos^2(\theta) = D_e((X_i, X_j))$, whereas the angle between X_i and Z shows the Effective Dependence among all variables, $(X_1, \ldots, X_4)$. If the angle between Z_1 and Z is small $D_e((X_i, X_j))$ for the projected data is similar to $D_e((X_1, \ldots, X_4))$ for the four variables. Figure 3 shows that in the projections (X_1, X_3) and (X_1, X_4) the angle between Z and Z_1 is small and we conclude that the linear relationship observed between the projected pairs of variables is similar to the average multivariate relationship. On the other hand, variables (X_1, X_2), (X_2, X_3), (X_2, X_4) and (X_3, X_4) show a weaker linear relationship than the average dependence in the data set.

[TABLE 2 ABOUT HERE]

The Effective Dependence for this data set is 0.27. The plot shows that this moderate value is due to the fact that only X_1 is strongly associated with the other variables, although this relationship is slightly nonlinear, (see Cook and Weisberg [5] and Velilla [17]). Table 2 illustrates the average position of the Effective Dependence with respect to the maximum and the minimum of the correlation and the determination coef¯cients.

Given the value of the Effective Dependence and applying the proposed rule (9), the number of principal components required to explain 90% of the variability for this data is, $h = 4(0.8 - 0.5 * 0.27) = 2.65$. Computing the principal components we obtain that the ¯rst two principal components explain 81.3%, whereas the ¯rst three explain 93.2%. This is in agreement with the proposed rule.

ACKNOWLEDGMENTS

We are very grateful to Mike Wiper for his help with the ¯nal draft, and to a referee for very helpful comments.

REFERENCES

1. E. Anderson. The irises of the Gaspe Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.
2. T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley, 2nd edition, 1984.
3. G. E. P. Box. A general distribution theory for a class of likelihood criteria. *Biometrika*, 36:317–346, 1949.
4. W. S. Cleveland, B. Kleiner, J. E. McRae, and R. E Warner, J. L. and. Pasceri. The Analysis of Ground-Level Ozone Data from New Jersey, New York, Connecticut, and Massachusetts: Data Quality Assessment and Temporal and Geographical Properties. Bell Laboratories Memorandum, 1975.
5. R. D. Cook and S. Weisberg. *An Introduction to Regression Graphics*. New York: John Wiley, 1994.
6. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
7. A. K. Gupta and P.N. Rathie. On the distribution of the determinant of sample correlation matrix from multivariate gaussian population. *Metron*, 1:43–56, 1983.
8. P. G. Hoel. A signi¯cance test for component analysis. *Annals of Mathematical Statistics*, 8:149–158, 1937.
9. J. N. R. Jeffers. Two case studies in the application of principal components analysis. *Appl. Statist.*, 16:225–236, 1967.
10. W. J. Krzanowski and Radley D. M. Nonparametric con¯dence and tolerance regions in canonical variate analysis. *Biometrics*, 45:1163–1173, 1989.

11. K. V. Mardia, J. T. Kent, and Bibby J. M. *Multivariate Analysis*. Academic Press, 1979.
12. R. B. Muirhead. *Aspect of multivariate statistical theory*. John Wiley & Sons, Inc., 1982.
13. S. Mustonen. A measure for total variability in multivariate normal distribution. *Computational Statistics & Data Analysis*, 23:321–334, 1997.
14. J. L. Rodgers and W. A. Nicewanders. Thirteen ways to look at the correlation coef¯cient. *The American Statistician*, 42:59–65, 1988.
15. G. A. F. Seber. *Multivariate Observations*. New York: John Wiley, 1984.
16. R. J. Ser©ing. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley, 1980.
17. S. Velilla. Assessing the number of linear components in a general regression problem. *Journal of the American Statistical Association*, 93:1088–1098, 1998.
18. S. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, 24:471–494, 1932.

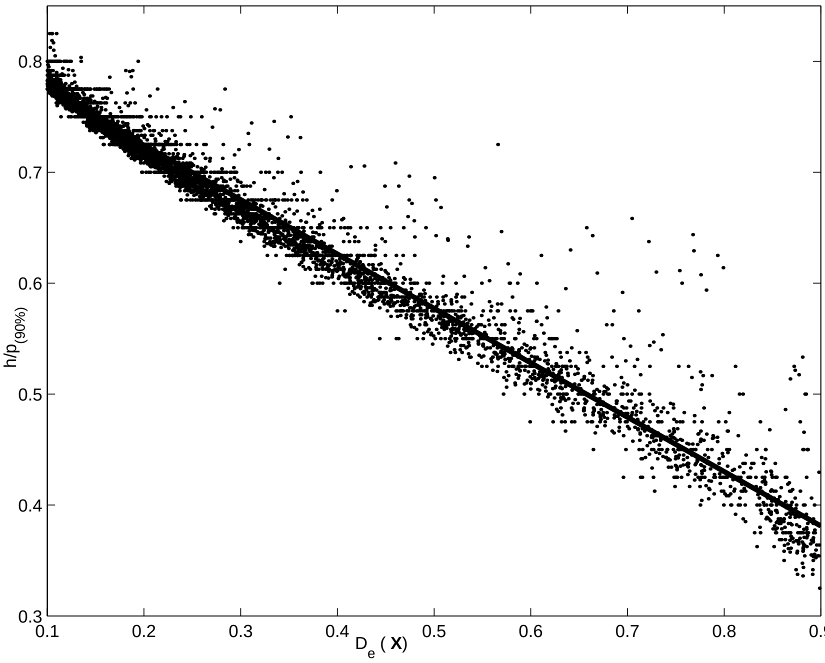

FIG. 1. Relationship between the proportion of the Principal Components which explain 90% of the total variability and the Effective Dependence in $[0.1, 0.9]$.

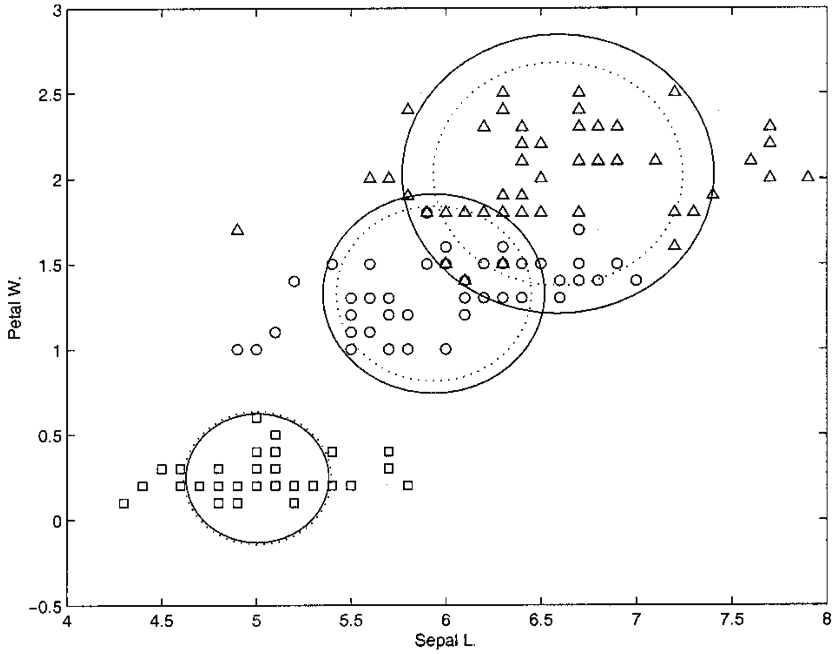

FIG. 2. Plot with the 3-groups for Fisher Iris data. The circle in solid line has radius $SD_e(\mathbf{\Sigma}^k)$ for each group, $k = 1, 2, 3$, and the circle in dotted line has radius $SD_e(\mathbf{\Sigma}_{14})$, where $\mathbf{\Sigma}_{14}$ is the covariance matrix between the variables X_1 and X_4 in the group i.

TABLE 1.

Descriptive measures of variance for each group in the Fisher Iris data.

	$Setosa$	$Versicolor$	$Virginica$
Measures of variability in all variables			
$TV(\mathbf{\Sigma}^{(i)})$	0.309	0.624	0.888
$GV(\mathbf{\Sigma}^{(i)})$	2.1×10^{-6}	1.9×10^{-5}	1.3×10^{-4}
$ATV(\mathbf{\Sigma}^{(i)})$	0.077	0.156	0.222
$V_e(\mathbf{\Sigma}^{(i)})$	0.038	0.066	0.107
$\psi(\mathbf{\Sigma}^{(i)})$	0.493	0.423	0.482
Measures of variability in projected data (Y_1, Y_4)			
$TV(\mathbf{\Sigma}_{14}^{(i)})$	0.135	0.305	0.480
$GV(\mathbf{\Sigma}_{14}^{(i)})$	0.001	0.007	0.028
$ATV(\mathbf{\Sigma}_{14}^{(i)})$	0.068	0.153	0.240
$V_e(\mathbf{\Sigma}_{14}^{(i)})$	0.036	0.085	0.168
$\psi(\mathbf{\Sigma}_{14}^{(i)})$	0.529	0.556	0.7

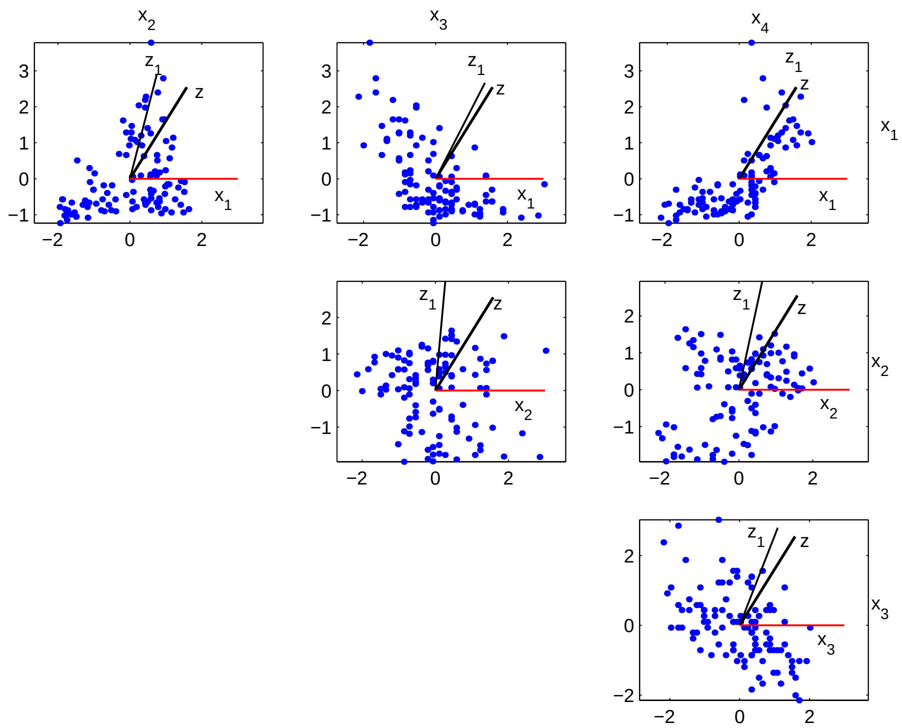

FIG. 3. Scatterplot matrix for the Ozone data. The angle between vectors Z_1 and X_i show the $D_e([X_i, X_j])$ and the angle between Z and X_i show the $D_e([X_1, \ldots, X_4])$.

TABLE 2.

Relative position of Effective Dependence in Ozone data

Dataset	min r^2	min R^2	$D_e(\mathbf{X})$	max r^2	max R^2
Ozone	0.121	0.140	0.270	0.475	0.605