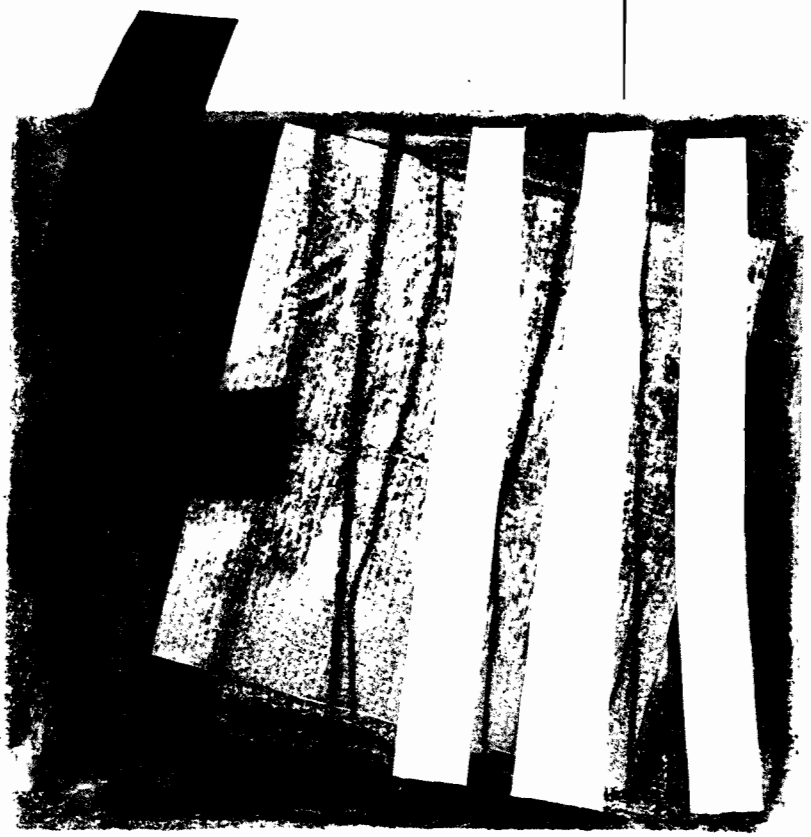
DESCRIPTIVE MEASURES OF MULTIVARIATE SCATTER AND LINEAR DEPENDENCE

Daniel Peña and Julio Rodríguez

00-50



WORKING PAPERS

Universidad Carlos III de Madrid

Working Paper 00-50
Statistics and Econometrics Series 22
September 2000

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

DESCRIPTIVE MEASURES OF MULTIVARIATE SCATTER AND LINEAR DEPENDENCE

Daniel Peña and Julio Rodríguez*

Abstract

In this paper we propose two new descriptive measures for Multivariate Data: The average variance and the Dependency or Average Square Correlation Coefficient. 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 for understanding multivariate data is illustrated in several examples.

Keywords: correlation; principal components; variability.

* Peña, Departamento de Estadistica y Econometría, Universidad Carlos III de Madrid, Spain, E-mail: dpena@est-econ.uc3m.es; Rodríguez, Departamento de Estadistica y Econometría, Universidad Carlos III de Madrid, Spain, E-mail: puerta@est-econ.uc3m.es

Descriptive Measures of Multivariate scatter and linear dependence

Daniel Peña and Julio Rodríguez
Universidad Carlos III de Madrid

Summary. In this paper we propose two new descriptive measures for Multivariate Data: The average variance and the Dependency or Average Square Correlation Coefficient. 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.

Keywords: 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 Nicewander (1988) 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 average variance and the Dependency or Average Square Correlation Coefficient. 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 section we present some conditions that a useful measure of multivariate variability must fulfil. It is shown that neither the trace nor the determinant of the covariance matrix verify these conditions and the average variance is suggested as a general descriptive measure of multivariate variability. In Section 3 we extend these conditions to a multivariate measure of linear relationship and the Dependency coefficient is introduced. It is shown that the Dependency can be used to estimate the proportion 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 X be a p dimensional random variable with finite covariance matrix Σ_x. We are interested in building a scalar measure of variability $V(X)$ that summarizes in some optimal way the multivariate variability of the random variable. This measure should be useful for comparing the variability of random variables of different dimension when they are measured

in the same units. With this objective in mind, we establish that a useful scalar measure must verify the following properties:

(a) $V(X) = g(\Sigma_x)$. That is, the measure must be a function only of the covariance matrix.
(b) If X is scalar then $V(X) = var(X)$.
(c) If $Y = QX$ where Q is an orthogonal matrix, then $V(Y) = V(X)$.
(d) If $Y = BX + C$ where B is a non singular diagonal matrix and C a vector, then $V(Y) = g^2(B)V(X)$.
(e) $V(X) = 0$ if and only if $|\Sigma_x| = 0$.
(f) Let $Z = [X\ Y]$ be a random vector of dimension $p + q$ where X and Y are random variables of dimension p and q respectively. Let us define the additional variability introduced by Y with respect to the one of X, by $V(\Sigma_{y/x})$, where $\Sigma_{y/x}$ is the covariance matrix of the random variable Y/X. Then $V(Z) \geq V(X)$ if and only if $V(\Sigma_{y/x}) \geq V(X)$ and $V(Z) \leq V(X)$ if and only if $V(\Sigma_{y/x}) \leq V(X)$.

The two most often used measures to describe scatter about the mean in multivariate data are the *total variation* (Seber, 1984), given by $tr\,(\Sigma_x) = \lambda_1 + \lambda_2 + \ldots + \lambda_p$, and the *generalized variance* (Wilks, 1932), given by $|\Sigma_x| = \lambda_1\lambda_2 \ldots \lambda_p$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ are the eigenvalues of the covariance matrix Σ_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 verifies properties (a) to (c) and the generalized variance properties (a) to (e). Neither of them verifies property (f): including an additional variable in a data set cannot decrease the trace whereas it is well known that the determinant in dimension $p-1$, $|\Sigma_{p-1}|$, and the determinant in dimension p, $|\Sigma_p|$ are related by

$$|\Sigma_p| = |\Sigma_{p-1}|\,\sigma_p^2\left(1 - R^2_{p.1\ldots p-1}\right) \tag{1}$$

where σ_p^2 is the variance of the p component and $R^2_{p.1\ldots p-1}$ 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 matrix covariance as a scalar measure of variables , we have that $|\Sigma_p|$ is greater or smaller than $|\Sigma_{p-1}|$ if $V(\Sigma_{y/x}) = \sigma_p^2\left(1 - R^2_{p.1\ldots p-1}\right)$ is greater or smaller than one. Introducing a new variable can increase or decrease the generalized variance without any connection between the additional variability, as measured by $\Sigma_{y/x}$, and the original variability, as measured by Σ_x.

These limitations are well known and some alternative measures of multivariate variability have been proposed. For instance, Mustonen (1997) proposed the measure

$$Mvar(\Sigma_p) = \max \sum_{i=1}^{p} \sigma^2_{i.i+1\ldots p} \tag{2}$$

where $\sigma^2_{i.i+1\ldots p}$ is the residual variance in the regression of the ith variable with respect to the variables $i+1, \ldots, p$, and the maximum is sought over all permutations of variables. This measure cannot decrease with p, because denoting by Σ_p the covariance matrix of a set of p variables, and Σ_{p-1} the covariance when one variable in the set is removed, then

$$Mvar(\Sigma_p) \geq Mvar(\Sigma_{p-1}),$$

where equality is only possible if the deleted variable is linearly dependent on the rest. It is easy to see that this measure satisfies properties (a) and (b) but does not satisfy properties (c) to (f). An additional disadvantage of $Mvar$ is that it is expensive computationally, because in order to find the maximum we have to examine the $p!$ possible arrangements of the variables.

The generalized variance is a measure of the hypervolume that the distribution of the random variables occupies in the space. If we avoid the effect of the units by standardizing all the variables we cannot compare generalized variances in set of different dimensions because according to (1), this measure cannot increase as the number of variables grows. It is clear that we need to get rid of this hypervolume interpretation if we want to compare sets of different dimensions. An intuitive alternative is to use the average scatter in any direction. We propose the name *Average Variance*, for the measure given by

$$AV(\Sigma_p) = |\Sigma_p|^{1/p} = (\lambda_1 \lambda_2 \ldots \lambda_p)^{1/p}, \tag{3}$$

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 Σ_p. Also, we can define the *Average Standard Deviation* by

$$ASD(\Sigma_p) = \{AV(\Sigma_p)\}^{1/2} = |\Sigma_p|^{1/2p}.$$

It is straightforward to check that the average variance verifies properties (a) to (e). In order to check property (f) note that we have

$$|\Sigma_z|^{1/p+q} = |\Sigma_x|^{1/p+q} \left|\Sigma_{y/x}\right|^{1/p+q} \tag{4}$$

and, for instance, the condition $\left|\Sigma_{y/x}\right|^{1/q} \geq |\Sigma_x|^{1/p}$ is equivalent to $\left|\Sigma_{y/x}\right|^{1/p+q} \geq |\Sigma_x|^{q/p(p+q)}$ which implies, by using (4), that $|\Sigma_z|^{1/p+q} \geq |\Sigma_x|^{1/p}$.

The Average Variance can be expressed as a function of the regression residual variances which appear in the measure proposed by Mustonen (2). We have that

$$AV(\Sigma_p) = \left(\Pi_{i=1}^{p} \sigma^2_{i.i+1\ldots p}\right)^{1/p} \tag{5}$$

and the Average Variance represents the geometric mean of the regression residual variances when each variable is predicted using all the remaining variables. When the variables are uncorrelated the *Average Variance* is the geometric mean of the individual variances. Notice that, in contract to Mustonen's measure, the *Average Variance* is invariant for any permutation of the variables.

From the properties of the geometric mean we have that

$$\lambda_p \leq AV(\Sigma_p) \leq \frac{1}{p} \sum_{i=1}^{p} \lambda_i$$

where λ_p is the minimum eigenvalue of Σ_p.

Remark. Note that an alternative definition for multivariate scatter, is the Average Total Variation, $ATV = (1/p)tr(\Sigma)$, which does not verify 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 dependency that summarizes the linear relationships between the variables and can be applied in sets of different dimensions. This measure, $D(X)$ must verify the following properties:

(a) $D(X) = g(R_x)$. That is, the measure must be a function only of the correlation matrix.
(b) If X is scalar then $D(X) = 1$.
(c) If $Y = QX$ where Q is an orthogonal matrix then $D(Y) = D(X)$.
(d) If $Y = BX + C$ where B is a non singular diagonal matrix and C a vector then $D(Y) = D(X)$.
(e) $0 \leq D(X) \leq 1$, and $D(X) = 1$ if and only if we can find a vector a such that $a'X = 0$. Also $D(X) = 0$ if and only if Σ_x is diagonal.
(f) If $Z = [X\ Y]$ where Y is a vector variable of dimension q and we define by $R_{y/x} = diag(\Sigma_{y/x})^{-1/2}\Sigma_{y/x}diag(\Sigma_{y/x})^{-1/2}$ the additional dependency introduced by Y. Then $D(Z) \geq D(X)$ if and only if $D(R_{y/x}) \geq D(R_x)$, and $D(Z) \leq D(X)$ if only if $D(R_{y/x}) \leq D(R_x)$.

The usual measure of dependence in the bivariate case is ρ, the correlation coefficient. In the multivariate case, a possible generalization is $1 - |R_p|$, where R_p is the correlation matrix. This measure satisfies properties (a) to (e), but again it is not appropriate for comparing the dependence structure between datasets with different numbers of variables. A possible generalization is $(1 - |R_p|)^{1/p}$ which has the advantage that for $p = 2$ it is equal to ρ, the linear correlation coefficient. However, this definition does not satisfy property (f). Note that by repeated use of (4), we can write

$$|R_p| = \left(1 - R^2_{p.1\ldots p-1}\right)\left(1 - R^2_{p-1.1\ldots p-2}\right)\ldots\left(1 - R^2_{2.1}\right). \tag{6}$$

and $|R_p|$ is the product of $p - 1$ terms, and the *ith* term represents the proportion of unexplained variation in a regression between the $p - i + 1$ variable and the variables $p - i$, $p - i - 1, ..., 1$. The average (obtained by the geometric mean) unexplained variation will be $|R_x|^{1/(p-1)}$ and we define the *Average Square Correlation Coefficient* or *Dependency* coefficient by :

$$D(R_x) = 1 - |R_x|^{1/(p-1)} \tag{7}$$

and the *Average Correlation Coefficient* (ACC) will be given by

$$\overline{\rho}(R_x) = \sqrt{1 - |R_x|^{1/(p-1)}}.$$

In the particular case $p = 2$, the ACC is equal to the standard Pearson correlation coefficient.

It is straightforward to show that the Dependency coefficient satisfies properties (a) to (e). To show that it also verifies property (f) suppose that $D(R_{y/x}) \geq D(R_x)$. Then we have $1 - |R_{y/x}|^{1/(q-1)} \geq 1 - |R_x|^{1/(p-1)}$. This implies $|R_x|^{(q-1)/(p-1)} \geq |R_{y/x}|$ and also $|R_x|^{(q-1)/(p-1)(p+q-2)} \geq |R_{y/x}|^{1/(p+q-2)}$. Then, $|R_x|^{1/(p-1)} \geq |R_x|^{1/(p+q-2)}|R_{y/x}|^{1/(p+q-2)}$ which implies by using (4) that $|R_x|^{1/(p-1)} \geq |R_z|^{1/(p+q-2)}$ and $D(Z) \geq D(X)$.

Note that the Dependency satisfies the following inequality

$$0 \leq D(X) \leq \frac{1}{p-1} \sum_{i=2}^{p} R^2_{i.1\cdots(i-1)}$$

for all the remaining variables.

Remark. An alternative definition for a dependency measure that also verifies properties (a) to (f) is the value $1 - |R_p|^{1/p}$. This measure can be interpreted as the geometric mean of p terms such that the last one is always equal to one. From this point of view we think that the dependency, which is the geometric mean of the $p-1$ terms of unexplained variability has a more clear interpretation. However, that for bivariate random variables both have some merits. The Dependency directly provides the squared correlation coefficient, whereas the measure $1 - |R_2|^{1/2}$ leads to $1 - \sqrt{1-\rho^2}$ which is also a useful measure of linear relationship. Let (y, x) be the components of the bivariate random vector, and $\sigma^2_{y/x} = \sigma^2_y(1-\rho^2)$. This measure 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. However, this intuitive explanation cannot be used for higher dimensions and therefore we think that the Dependency has a more straightforward interpretation as a descriptive measure of linear relationship.

3.1. Some properties of the Dependency coefficient

Firstly, the Dependency coefficient represents the average squared correlation among the variables.

To see this, note that the squared correlation can always be interpreted as

$$R^2 = 1 - \frac{RV}{TV}$$

where RV is the residual or unexplained variability (sum of squares) and TV the total variability. By (7) the Dependency can be written as

$$D(X) = 1 - \frac{RV(m)}{TV(m)}$$

where

$$\frac{RV(m)}{TV(m)} = \left\{ \frac{RV(p/1\cdots p-1)\cdots RV(2/1)}{TV(p)\cdots TV(2)} \right\}^{\frac{1}{p-1}}$$

is the geometric mean of the residual sum of squares divided by the total sum of squares of the regressions. Note that this measure in invariante to any permutation of the variables.

This interpretation holds when the set of variables can be split as $Z = (X, Y)$, where X has dimension p and Y has dimension q and suppose that $p \geq q$. Then it is well known that

$$|R_z| = |R_x|\,|R_y|\,\left|I - R_y^{-1} R_{yx} R_x^{-1} R_{xy}\right|$$

where the non null eigenvalues of the matrix $R_y^{-1}R_{yx}R_x^{-1}R_{xy}$ are the canonical correlation coefficients. Thus, we can write

$$D(Z) = 1 - \left\{\left(1 - R^2_{xp.1\ldots p-1}\right)\cdots\left(1 - R^2_{x2.1}\right)\right\}^{\frac{1}{p+q-1}}$$
$$\left\{\left(1 - R^2_{yq.1\ldots q-1}\right)\cdots\left(1 - R^2_{y2.1}\right)\right\}^{\frac{1}{p+q-1}}\left\{\prod_{i=1}^{l}\left(1 - r_i^2\right)\right\}^{\frac{1}{p+q-1}}$$

where $l = \min(p,q) = q$ and r_i^2 are the canonical correlation coefficients between the two sets of variables. This expression shows that if the two sets are uncorrelated, then $D(Z)$ is just the average correlation coefficient among all the variables. When the two sets are correlated the Dependency is an average of the internal dependence and the cross dependence as measured by the correlation coefficients.

Secondly, the Dependency is a measure of the lack of sphericity of the standardized variables.

Anderson (1984, p. 427) defines sphericity as

$$\psi(\Sigma_p) = \frac{|\Sigma_p|^{1/p}}{(1/p)tr\,(\Sigma_p)},$$

and he used this measure for testing the hypothesis $H_0 : \Sigma = \sigma^2\mathbf{I}$. If $\psi = 1$, then the geometric mean of the eigenvalues is equal to the arithmetic mean and all the variables are uncorrelated, then 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 $\psi(R_p) = |R_p|^{1/p}$, and the *Dependency* is

$$D(X) = 1 - \psi(R_p)^{p/(p-1)}.$$

(See Juan and Prieto (1995) for other non-sphericity measure to evaluate the performance of multivariate estimators with high breakdown point).

Thirdly, when all the off diagonal values of the correlation matrix are equal, the dependency is also equal 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

$$R_p = \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{bmatrix}.$$

Notice that given this structure, the coefficient of determination in the regression of any variable with respect to the rest is $R^2_{p.1\ldots p-1}$ and $R^2_{p.1\ldots p-1} \to \rho$ as $p \to \infty$. This result is shown by Mustonen (1997), and is a consequence of

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

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

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

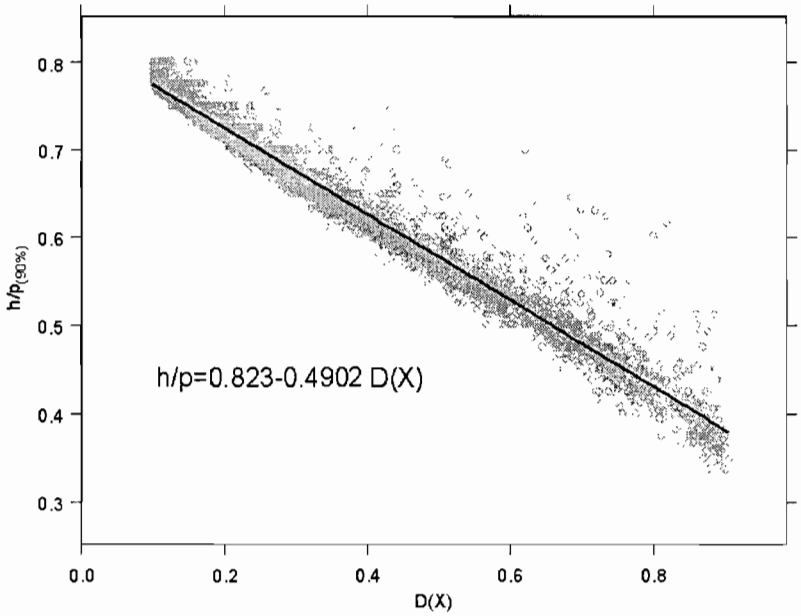


Fig. 1. Relationship between the proportion of the Principal Components which explain 90% of the total variability and the Dependency.

whereas for the Dependency coefficient,

$$\lim_{p\to\infty} D(X) = \lim_{p\to\infty} \left(1 - |R_p|^{1/(p-1)}\right) = \tag{8}$$
$$= \lim_{p\to\infty} 1 - (1-\rho)\left\{1 + (p-1)\rho\right)^{1/(p-1)} =$$
$$= \rho \qquad \forall \rho \in (0,1\}$$

which provides an interesting interpretation of the correlation coefficient as the limiting average proportion of explained variability in this situation.

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

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 Dependency should inherit this property and we will show that the dependency coefficient is strongly related to the proportion of components needed to summarize the data. 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 Dependency coefficient of the sample and the proportion of components, h/p, needed to explain 90% of the total variability. We have carried on a simulation study. We have generated random correlation matrices of dimension p as follows: (1) the eigenvalues of the correlation matrix have been generated as random values from a $Beta(\alpha, \beta)$ distribution where α and β were chosen from a mesh in the interval $(0, 3)^2$ obtaining 900 pairs of parameters, (α_i, β_i). (2) The values were normalized so that their sum is p. For each fixed value p, we generated 900 matrices. This process was performed for $p =$

$40, 80, \ldots, 440$ and 9900 correlation matrices were obtained. For each one of these matrices, we calculate $\left(\frac{h}{p}\right) \in [0,1]$ and the *Dependency* measure. We observed that the relation between $\left(\frac{h}{p}\right)$ and D is a sigmoid, but in the interval $D \in [0.1, 0.9]$ we can approximate it (see figure 1) by the linear relation

$$\frac{h}{p} = 0.8230 - 0.4902D(X)$$

with $R^2 = 0.967$. This relationship can be approximated by

$$h = p(0.8 - 0.5D(X)) \tag{9}$$

To illustrate this result, we present the analysis of Jeffers' (1967) pine piprops data, taken from Mardia et al. (1979, 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 *Dependency* of these data is $D(X) = 0.592$ and from equation (9) we obtain that $h = 0.503p$ and the estimated number of principal components for explain 90% of the total variability is 6.55. 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 first 6 components, this 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 first 7 components. Jeffers took the first 6 components in his analysis, because of their clear physical interpretation.

4. Sample Distributions

Given a sample, $(x_1, \ldots, x_N)$, with $x_i \in \mathbf{R}^p$, it is well known that the generalized sample variance

$$|S_p| = \left| \frac{1}{N-1} \sum (x_i - \overline{x})(x_i - \overline{x})' \right|$$

is a measure of the hypervolume occupied by the data in the following two sense. Firstly, if we consider the sample observations as N points in $\mathbf{R}^p$ and compute the volumes of all the different parallelotopes formed by using as principal edges p vectors of $x_1, \ldots, x_N$ as one set of endpoints and $\bar{x}$ as the other, the generalized variance is proportional to the sum of squares of these volumes and the factor of proportionality is N^{-p}. This interpretation is due to Anderson (1984 p. 263). Secondly, if we now consider the sample as p points in $\mathbf{R}^N$ the generalized variance is the volume of the parallelotope generated by the vectors. To illustrate this last interpretation, suppose a bivariate sample of size N and let us call $\widetilde{x}_1, \widetilde{x}_2 \in \mathbf{R}^N$ the $N \times 1$ vectors representing the variable in deviation to their means, that is $\widetilde{x}_1 = x_1 - \overline{x}_1\mathbf{1}$ and $\widetilde{x}_2 = x_2 - \overline{x}_2\mathbf{1}$, where $\mathbf{1} = (1, \ldots, 1)'$ and $\overline{x} = (\overline{x}_1, \overline{x}_2)'$ is the mean vector. Let $L_{\widetilde{x}_1}$ and $L_{\widetilde{x}_2}$ be the module of the vectors $\widetilde{x}_1$ and $\widetilde{x}_2$. The area of the parallelogram formed by these two vectors is $A = L_{\widetilde{x}_1} L_{\widetilde{x}_2} \sqrt{1 - \cos^2\theta_{12}}$ where θ_{12} is the angle between $\widetilde{x}_1$ and $\widetilde{x}_2$, and $\cos\theta_{12} = r_{12}$ is the correlation coefficient between the two variables. Let S_2 be the sample variance matrix, we also have $|S_2| = A^2/(N-1)^2$. Suppose now that we add a third variable, $\widetilde{x}_3$. We can easily compute the volume of the parallelotope generated by the vectors $\widetilde{x}_1$, $\widetilde{x}_2$, $\widetilde{x}_3$ by multiplying the area of the base by the height, where the base is

formed with the vectors $\tilde{x}_1$ and $\tilde{x}_2$, $(n-1)\,|S_2|^{1/2}$, and the height is $s_3 \sin\theta_{3.12}$, where $\theta_{3.12}$ is the angle between the variable $\tilde{x}_3$ and the plane defined by $\tilde{x}_1$ and $\tilde{x}_2$. Thus we have that the volume is $Vol = (n-1)^{3/2}\,|S_2|^{1/2}\, s_3\sqrt{(1-r^2_{3.12})}$ and $|S_3| = Vol^2/(N-1)^3$. In general, we have $|S_p| = Vol_p^2/(N-1))^p$ where Vol_p is the hypervolume in p dimensions. Thus, the average sample variance in p dimensions, $|S_p|^{1/p}$ is proportional to the length of the side of the hypercube with hypervolume equal to $|S_p|$.

The sample distribution of the Average Variability can be obtained from existent results on the generalized variance (see Anderson (1984) and Muirhead (1982)). The generalized variance is usually estimated by the sample generalized variance, $det(S_p)$, where S_p is the sample covariance matrix with dimension $p \times p$. In the case of Average Variability it is estimate by the pth root of the generalized sample variability. The following two lemmas provide the distribution of $(det(S_p))^{1/p}$ when S_p is computed with a sample of size $N = n+1$, from the $N_p(\mu, \Sigma)$ distribution. In this case S_p follows a Wishart distribution with n degrees of freedom and covariance matrix $(1/n)\Sigma_p$, $W_p(n, (1/n)\Sigma_p)$. We state the two lemmas in order to characterize the asymptotic distribution and an approximation of the exact distribution by the sample *Average Variance.*

LEMMA 1. *Let S_p be a $p \times p$ sample covariance matrix with n degrees of freedom. Then*

$$\sqrt{n}\left(|S_p|^{1/p} / |\Sigma_p|^{1/p} - 1\right)$$

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

PROOF. The asymptotic distribution of AV can be obtained from the asymptotic normality of the generalized variance. Anderson (1984), shows that $\sqrt{n}\,(|S_p| / |\Sigma_p| - 1)$ is asymptotically normal with mean 0 and variance $2p$. Then, applying the $\delta-$method (see Serfling (1980 *p.* 118)), for $g(x) = x^{1/p}$, it follows, that the AV is also asymptotically normally distributed with mean 0 and variance $2/p$.

LEMMA 2. *The 'exact' distribution for the pth root of $|S_p| / |\Sigma_p|$ is*

$$|S_p|^{1/p} / |\Sigma_p|^{1/p} \sim \Gamma\left(\frac{p(n-p)}{2}, \frac{p(n-1)}{2}\left(1 - \frac{(p-1)(p-2)}{2n}\right)^{1/p}\right).$$

PROOF. Using the results of Hoel (1937) related to the 'exact' density for the pth root of $|A_p| / |\Sigma_p|$, we have

$$|S_p| / |\Sigma_p| \sim \frac{c^{\frac{1}{2}p(n-p)} y^{\frac{1}{2}p(n-p)-1} e^{-cy}}{\Gamma(\frac{1}{2}p(n-p))}$$

where

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

and $|A_p|$ is $(n-1)^p\,|S_p|$. Applying the property that, $X \sim \Gamma(\alpha, \beta) \rightarrow dX \sim \Gamma(\alpha, \beta/d)$, then

$$|S_p|^{1/p} / |\Sigma_p|^{1/p} \sim \Gamma\left(\frac{p(n-p)}{2}, \frac{p(n-1)}{2}\left(1 - \frac{(p-1)(p-2)}{2n}\right)^{1/p}\right).$$

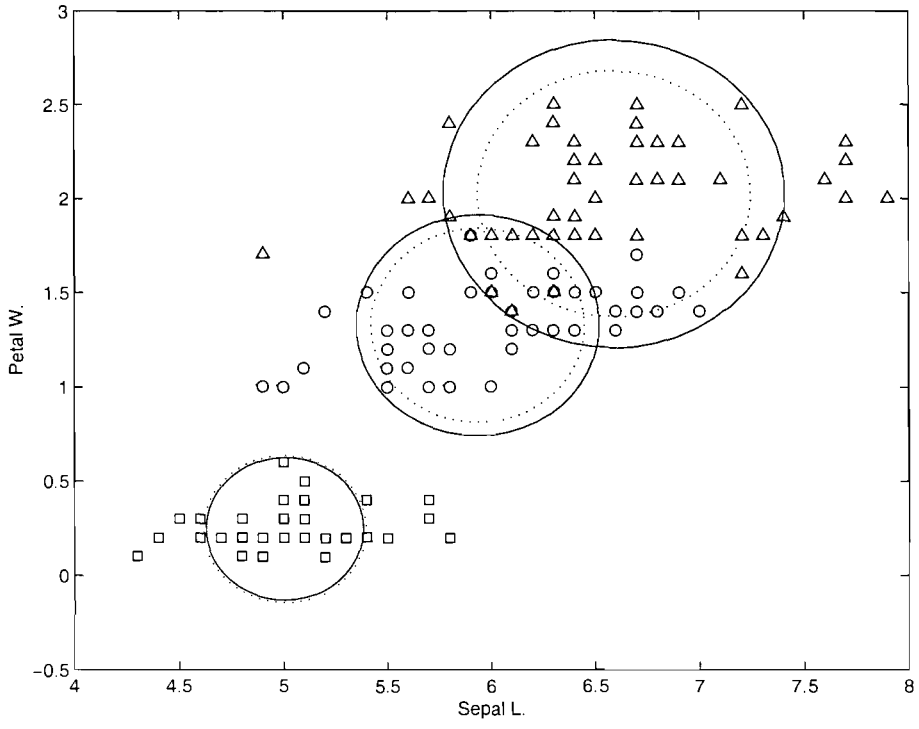


Fig. 2. Plot with the 3-groups for Fisher Iris data. The circle in solid line has radius $ASD(\Sigma^k)$ for each group, $k = 1, 2, 3$, and the circle in dotted line has radius $ASD(\Sigma_{14})$, where Σ_{14} is the covariance matrix between the variables X_1 and X_4 in the group i.

Regarding the distribution of $D(R_x)$, the exact distribution can be easily obtained from the exact distribution of $|R_x|$ given in Gupta and Rathie (1983). The asymptotic distribution of $-n \log |R_x|$ is a χ^2 with $p(p-1)/2$ degrees of freedom (see Box 1949). Thus, the asymptotic distribution of $n(p-1)D(R_x)$ is a χ^2 with the same degrees of freedom.

5. Examples

To illustrate the information provided be for the Average Variance and the Dependency in a descriptive analysis of multivariate data, two known sets of data are presented. The first of these datasets is the Fisher Iris data, originally due to Anderson (1935) and analyzed by Fisher (1936) in his seminal paper on discriminant analysis. These data correspond to measures of three species of flowers called, Iris Setosa, Iris Versicolor and Iris Virginica. There are 50 specimens of each specie 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 shows the projection of the Iris data in the variables Y_1 and Y_4. The specimens for Setosa are squares, for Versicolor circles and triangles for Virginica. In figure 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 $2 \times ASD(\Sigma_{14}^{(i)})$, where $\Sigma_{14}^{(i)}$

Table 1. Descriptive measures of variance for each group in the Fisher Iris data.

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

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 $2 \times ASD(\Sigma^{(i)})$, where $\Sigma^{(i)}$ is the covariance matrix in the group i, for all variables. The similarity in both circles indicates that the dispersion in the projected data is similar to the dispersion in the multivariate data. Figure 2 shows that both 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.

In Table 1, some dispersion measures for each group in the Iris data are shown. The first measures correspond to all variables and the second to the projected data shown 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 two new measures, ATV and AV 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 AV in each group for all variables. If we compare this AV with the AV 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 AV 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, (1989), over the difference in dispersion in each species.

To illustrate the information provided by the Dependency measure we consider the data on air quality measurements in the New York metropolitan area from May 1, 1973 to September 30, 1973 from Chambers et al. (1975). Only the $n = 111$ complete cases are considered here. The data are ordered in time, but gaps of up to 10 days may exist between

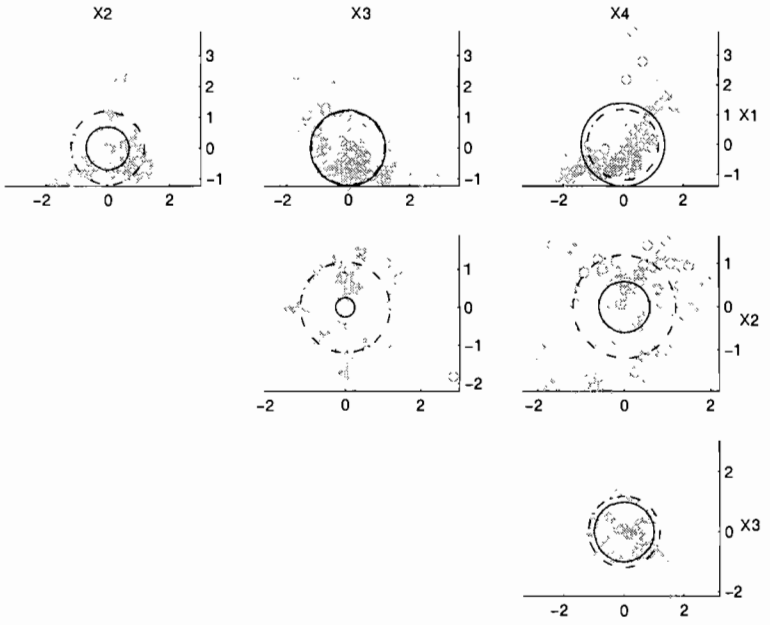


Fig. 3. Scatterplot matrix for the Ozone data. The circle in solid line shows the Dependency in the projected data and the circle in dotted line shows the Dependency in the data.

Table 2. Relative position of Dependency in Ozone data

Dataset	min r	max R	$D(R_p)^{1/2}$	max r	max R
Ozone	0.348	0.374	0.594	0.698	0.778

readings. This data set is obtained for studying the relationship between the variable Ozone concentration in part 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 represent the standardized data. Figure 3 shows a scatterplot of the standardized Ozone data. In each plot we present two circles showing the Dependency computed from all the variables (dotted line) and the Dependency computed for the couple of variables represented in the scatterplot (solid line). These circles have radius $2 \times \overline{\rho}(R)$, see (3), and $2 \times \overline{\rho}(R_{ij})$, respectively, where R and R_{ij} are the correlation matrix for all variables and for variables X_i and X_j. In Figure 3 we observe that both circles are similar in the projections (X_1, X_3), (X_1, X_4) and (X_3, X_4). We conclude that the linear relationship between these pair of variables is similar to the average multivariate relationship. On the other hand, variables (X_1, X_2), (X_2, X_3) and (X_2, X_4) show a weaker linear relationship than the average in the data set.

The *Dependency* for this data set is $0.594^2 = 0.35$. The plot shows that this moderate value is due to the fact that only X_1 shows a high relation with respect to the other variables, although this relationship is slightly no linear, (see Cook and Weisberg (1994) and Velilla (1998)). Table 2 illustrates the average position of Dependency with respect to the maximum and the minimum of the correlation and the determination coefficients.

Given the value of the Dependency coefficient and applying the proposed rule (9), we

obtain that the number of principal components required to explain 90% of the variability is for this data, $h = 4(0.8 - 0.5 * 0.35) = 2.51$. Computing the principal components we obtain that the first two principal components explain 81.3% whereas the first three explain 93.2%. This is in agreement with the proposed rule.

Acknowledgments

This research has been sponsored by DGES (Spain) under project PB-96-0111 and the Cátedra BBVA de Calidad . We are very grateful to Mike Wiper for his help with the final draft.

References

Anderson, E. (1935) The irises of the Gaspe peninsula. *Bulletin of the American Iris Society*, **59**, 2-5.

Anderson, T. W. (1984) *An Introduction to Multivariate Statistical Analysis.* New York: John Wiley. (2nd. Ed.)

Box, G. E. P. (1949) Ageneral distribution theory for a class of likelihood criteria. *Biometrika*, **36**, 317-346.

Cleveland, W. S., Kleiner, B, McRae, J. E., Warner, J. L. and. Pasceri, R. E (1975) 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.

Cook, R. D. and Weisberg, S. (1994) *An Introduction to Regression Graphics.* New York: John Wiley.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179-188.

Gupta, A. K. and Rathie, P.N. (1983) On the distribution of the determinant of sample correlation matrix from multivariate Gaussian population. *Metron*, **1**, 43-56.

Hoel, P. G. (1937) A significance test for component analysis.*Annals of Mathematical Statistics*, **8**, 149-158.

Jeffers, J. N. R. (1967) Two case studies in the application of principal components analysis. *Appl. Statist.*, **16**, 225-236.

Juan, J. and Prieto F. J. (1995) A subsampling method for the computation of multivariate estimators with high breakdown point. *Computational and Graphical Statistics*, **4**, 319-334.

Krzanowski, W. J. and Radley D. M (1989) Nomparametric Confidence and Tolerance Regions in Canonical Variate Analysis. *Biometrics*, **45**, 1163-1173.

Mardia, K. V., Kent, J. T. and Bibby J. M. (1979) *Multivariate Analysis.* Academic Press.

Manly, B. F. J. (1986) *Multivariate Statistical Methods: A Primer.* New York: Chapman & Hall, 11.

Muller, K. E. (1982) Understanding canonical correlation through the general linear model and principal components. *The American Statistician*, **36**, 342-354.

Mustonen, S. (1997) A measure for total variability in multivariate normal distribution. *Computational Statistics & Data Analysis*, **23**, 321-334.

Rodgers, J. L. and Nicewanders, W. A. (1988) Thirteen ways to look at the correlation coefficient. *The American Statistician*, **42**, 59-65.

Seber, G. A. F. (1984) *Multivariate Observations*. New York: John Wiley.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*. New York: John Wiley.

Velilla, S. (1998) Assessing the Number of Linear Components in a General Regression Problem. *Journal of the American Statistical Association*, **93**, 1088-1098.

Wilks, S. S. (1932) Certain generalizations in the analysis of variance. *Biometrika*, **24**, 471-494.