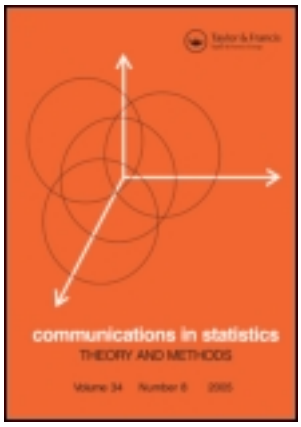

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

Distance-Based Measures of Association with Applications in Relating Hyperspectral Images

Carles M. Cuadras [a], Silvia Valero [b], Daniel Cuadras [c], Philippe Salembier [d] & Jocelyn Chanussot [b]

[a] Department d'Estadística, Universitat de Barcelona, Barcelona, Spain

[b] Grenoble Institute of Technology (INPG), GIPSA-Lab, Grenoble, France

[c] Servei d'Assessoria Estadística, Institut d'Investigacio Biomèdica de Bellvitge, Hospitalet, Spain

[d] Department de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunga, Barcelona, Spain

Available online: 18 May 2012

To cite this article: Carles M. Cuadras, Silvia Valero, Daniel Cuadras, Philippe Salembier & Jocelyn Chanussot (2012): Distance-Based Measures of Association with Applications in Relating Hyperspectral Images, Communications in Statistics - Theory and Methods, 41:13-14, 2342-2355

To link to this article: http://dx.doi.org/10.1080/03610926.2012.654880

Communications in Statistics—Theory and Methods, 41: 2342–2355, 2012

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2012.654880

Distance-Based Measures of Association with Applications in Relating Hyperspectral Images

CARLES M. CUADRAS,[1] SILVIA VALERO,[2] DANIEL CUADRAS,[3] PHILIPPE SALEMBIER,[4] AND JOCELYN CHANUSSOT[2]

[1]Department d'Estadística, Universitat de Barcelona, Barcelona, Spain
[2]Grenoble Institute of Technology (INPG), GIPSA-Lab, Grenoble, France
[3]Servei d'Assessoria Estadística, Institut d'Investigacio Biomèdica de Bellvitge, Hospitalet, Spain
[4]Department de Teoria del Senyal i Comunicacions, Universitat Politècnica de Catalunga, Barcelona, Spain

We propose a distance-based method to relate two data sets. We define and study some measures of multivariate association based on distances between observations. The proposed approach can be used to deal with general data sets (e.g., observations on continuous, categorical or mixed variables). An application, using Hellinger distance, provides the relationships between two regions of hyperspectral images.

Keywords Binary partition tree; Canonical correlations; Hellinger distance; Metric multidimensional scaling; Wilks lambda.

Mathematics Subject Classification Primary 62H20; Secondary 62H35.

1. Introduction

Several coefficients have been proposed to measure the relationships between two data sets taken on the same individuals. The data sets are often represented by two sets of variables, which in practice can be identified with the columns of two quantitative data matrices **X**, **Y**, with the same number of rows. Then some measures of multivariate association, as extensions of Pearson correlation coefficient, can be used. Most measures are based on canonical correlations, as the first canonical correlation, first proposed by Hotelling (1936), or the average of the squared canonical correlations. This class of symmetric measures has important applications. Ecology is an example where environmental data is related to species abundance. In genetics, the relationship between environmental variables and genetic frequencies plays an important role. In biometry, the user is interested in

Received November 29, 2010; Accepted December 28, 2011
Address correspondence to Carles M. Cuadras, Dept. d'Estadistica, Universitat de Barcelona, Díagonal 645, Barcelona 08028, Spain; E-mail: ccuadras@ub.edu

relating some physical characteristics of individuals to the same characteristics in their offspring. In psychology, it is important to relate physical characteristics to mental tests. There are examples dealing with classic types of multivariate data in Manly (1986), Rao (1952), and Rencher (1995); see Cramer and Nicewander (1979) for a complete repertoire of multivariate measures of association.

Nowadays, the sources of data are much more complex. In genomics we have many base pairs, microarrays, etc., and we may seek relationships between genotype and phenotypes of interest. Often in medical studies the variables are of mixed type (continuous, categorical, nominal), and cannot be treated as quantitative variables. In image processing, multivariate images are captured by hyperspectral remote sensors, containing, for each position, the solar radiation reflected by a material at different wavelengths. Then, in order to partition a multivariate image in similar regions, we should relate these regions by using an association measure.

Dealing with quantitative variables and using conventional coefficients may not be appropriate in these frameworks, because we can have more variables than observations, or the quantification may be artificial. The information can alternatively be given by a similarity or dissimilarity matrix. From this matrix and via Metric Multidimensional Scaling (MMDS), we can obtain principal coordinates providing two matrices $\mathbf{X}$, $\mathbf{Y}$, and next apply the proposed association measures. This distance-based approach, originated in Cuadras (1989) and Cuadras and Arenas (1990), has been used as a tool in prediction and multivariate analysis; see Amat et al. (1998), Bartkowiak and Jakimiec (1994), Boj et al. (2007), and Esteve et al. (2009). The procedure proposed here extends McArdle and Anderson (2001), Wessel and Schork (2006), and Zapala and Schork (2006), to relate a quantitative variable to some mixed variables by using distance-based regression.

In this article, we propose several ways of constructing association measures, based on appropriate statistical models. These results reduce to the measures proposed in Cramer and Nicewander (1979) when the distance is Euclidean.

2. A Motivating Application

In image processing, by varying the wavelength, any material reflects and absorbs the solar radiation differently. This radiance (number of photons) is registered by hyperspectral sensors, which collect multivariate discrete images in a series of contiguous wavelength bands, providing the spectral curves, which can distinguish between materials. In order to partition a multivariate image in regions belonging to different materials, we need to compare these regions objectively, by first transforming them into frequency data matrices and then converting into matrices with probability distributions in the rows.

Figure 1 shows an example of such interpretation. Figure 1(a) shows a hyperspectral region formed by three spectral curves belonging to the same material. Figure 1(b) shows how to describe (for one specific wavelength λ_i) a region by means of a probability distribution f_{λ_i}. Then, by varying the wavelength, the hyperspectral region is described by a set of probability distributions according to the observed radiance values of this region.

The study of hyperspectral image as a set of regions has been addressed by the so-called Binary Partition Tree (BPT) image representation; see Salembier and Garrido (2000). BPT is a set of regions structured in a hierarchical tree, as shown in Fig. 2. In this example, it is quite apparent how each BPT node corresponds to a

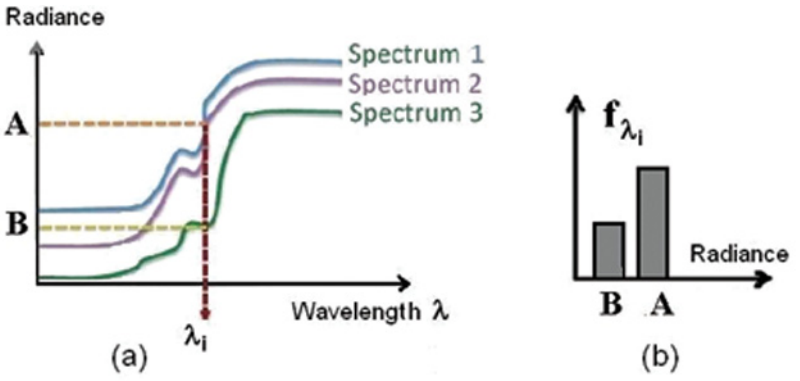

Figure 1. (a) Three spectral curves describing a region. For each value of the wavelength (horizontal axis) there are three radiance values (vertical axis). (b) Instead of taking only the mean, we consider the statistical distribution of these three values. (color figure available online.)

region of the image at a given level. BPT is constructed by merging the most similar adjacent regions in an iterative process. For instance, in the first iteration, R_3 and R_4 are merged together to form a parent node R_6. In order to construct the BPT, it is necessary to use a similarity measure to compare the adjacent regions. This measure has to deal with the type of data shown in Fig. 1.

As an illustration, a real data example is presented in Table 1. This (partial) data matrix contains, for each spectrum, the number of photons or radiances observed at a given wavelength λ_i. The simplest solution to describe the spectra is to take the mean values in the columns of Table 1, i.e., to consider the spectra mean of the region. Instead, we consider all possible radiance values, therefore this table

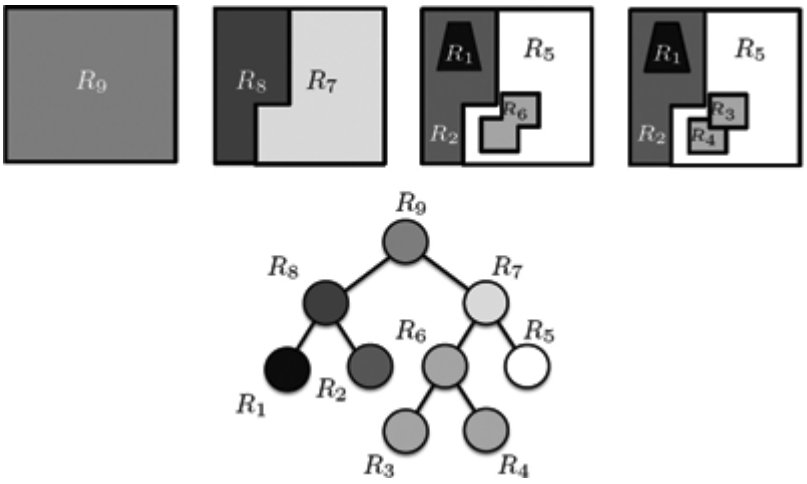

Figure 2. Example of Binary Partition Tree. Four pairs of regions forming an image are merged by following an iterative algorithm to construct the tree representation. In each step, the similarity is measured by comparing several probability distributions.

Table 1
Radiance values (number of photons) registered for four wavelengths (WL) and six spectral curves (top). Grouped distribution of the radiance values for each wavelength (bottom)

		Spectral values						
		S_1	S_2	S_3	S_4	S_5	S_6	$\cdots$
WL	λ_1	634	979	851	991	823	843	$\cdots$
	$\lambda 2$	861	1055	1061	1035	835	874	$\cdots$
	$\lambda 3$	1104	1030	1173	1101	1073	1103	$\cdots$
	$\lambda 4$	1100	1110	1193	1098	1257	1237	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

		Radiance values (class intervals)					
		601–700	701–800	801–900	901–1000	1001–1100	$\cdots$
WL	λ_1	0.125	0	0.375	0.250	0	$\cdots$
	$\lambda 2$	0	0	0.375	0	0.375	$\cdots$
	$\lambda 3$	0	0	0	0	0.375	$\cdots$
	$\lambda 4$	0	0	0	0	0.250	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

is converted into a probability data matrix, with rows containing the empirical distribution of the radiances for each λ_i.

Our aim is to compare two probability data matrices, obtained from two hyperspectral regions, as initial step to perform the algorithm to get the BPT. Of course, this high level but invisible image (in contrast with a visible ordinary image) must be studied by using mathematical and computational techniques.

3. MMDS and the Angle Between Subspaces

Let Ω be a finite set with n objects or individuals. Let $\delta_{ij} = \delta_{ji} \geq \delta_{ii} = 0$ a distance or dissimilarity function between pairs of individuals in Ω. This gives an $n \times n$ distance matrix $\Delta_x = (\delta_{ij})$. We suppose that this distance matrix is Euclidean, i.e., there exists a configuration $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$, with coordinates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, $i = 1, \ldots, n$, such that $\delta_{ij}^2 = (\mathbf{x}_i - \mathbf{x}_j)'(\mathbf{x}_i - \mathbf{x}_j)$. Thus, the coordinates of Ω constitute an $n \times p$ matrix $\mathbf{X} = (x_{ij})$ such that the distance between two rows i and j equals the original distance δ_{ij}.

If $\mathbf{I}_n$ is the identity matrix and $\mathbf{1}_n$ is the column vector of ones, metric multidimensional scaling (MMDS) is a well-known way of obtaining $\mathbf{X}$ from Δ_x. First, we find the $n \times n$ matrices $\mathbf{A} = -\frac{1}{2}\Delta_x * \Delta_x$ and $\mathbf{G}_x = \mathbf{H}_c\mathbf{A}\mathbf{H}_c$, where $\Delta_x * \Delta_x = (\delta_{ij}^2)$ and $\mathbf{H}_c = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}_n'$ is the centering matrix. Next we compute the spectral decomposition $\mathbf{G}_x = \mathbf{U}\Lambda_x^2\mathbf{U}'$, which provides the matrix of coordinates $\mathbf{X} = \mathbf{U}\Lambda_x$. Notice that Δ_x is a Euclidean distance matrix iff $\mathbf{G}_x$ is positive semi-definite matrix.

Assuming the eigenvalues in Λ_x^2 arranged in descending order, matrices $\mathbf{X}$ and $\mathbf{U}$ contain the principal and standard coordinates, respectively, of the n individuals

with respect to distance δ. The aim of MMDS is to represent the n individuals in reduced dimension (usually 2), by taking the first principal coordinates. However our interest here is to relate these coordinates to a second data set.

For a second data set consisting of observations on the same n individuals, we may consider another distance matrix Δ_y and find $\mathbf{G}_y = \mathbf{V}\Lambda_y^2\mathbf{V}'$ by using the same procedure. If the eigenvalues in Λ_y^2 are also arranged in descending order, the principal coordinates are $\mathbf{Y} = \mathbf{V}\Lambda_y$. With these coordinates, the relationship between both data sets reduces to the relationship between the centered matrices $\mathbf{X}(n \times p)$ and $\mathbf{Y}(n \times q)$.

In Cuadras (2008) the following multivariate association measure between $\mathbf{X}$ and $\mathbf{Y}$ is proposed:

$$\eta(\mathbf{X}, \mathbf{Y}) = \sqrt{\det(\mathbf{U}'\mathbf{V}\mathbf{V}'\mathbf{U})} = \sqrt{\det(\mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V})},$$

which satisfies $0 \leq \eta(\mathbf{X}, \mathbf{Y}) = \eta(\mathbf{Y}, \mathbf{X}) \leq 1$, and reduces to the multiple correlation coefficient when $\mathbf{Y}$ is univariate ($q = 1$). Since $\mathbf{U}'\mathbf{V}$ is a Gram matrix, η can be interpreted as the cosine of the angle between two subspaces expanded by $\mathbf{U}$ and $\mathbf{V}$; see Jiang (1996).

4. Measures Based on Multivariate Regression

If we consider the columns of $\mathbf{X}$ and $\mathbf{Y}$, as predictor and response variables, respectively, a standard way to relate them is by multivariate linear regression

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \Xi,$$

where $\mathbf{B}$ is a $p \times q$ matrix of parameters and Ξ is a $n \times q$ matrix of errors. The least-squares estimator of $\mathbf{B}$ is $\widehat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and the prediction matrix is $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}} = \mathbf{H}_t\mathbf{Y}$ where $\mathbf{H}_t = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat matrix.

4.1. *F test*

Clearly, there is no relationship if $\mathbf{B} = \mathbf{0}$. Assuming $\mathbf{X}, \mathbf{Y}$ centered, an appropriate statistic for testing this null hypothesis is based on

$$F_* = \mathrm{tr}(\mathbf{Y}'\mathbf{H}_t\mathbf{H}_t\mathbf{Y})/\mathrm{tr}[\mathbf{Y}'(\mathbf{I} - \mathbf{H}_t)\mathbf{Y}].$$

Suppose that $\mathbf{X}, \mathbf{Y}$ have been obtained by MMDS from two distance matrices. Then $\mathbf{X} = \mathbf{U}\Lambda_x$ and $\mathbf{H}_t = \mathbf{U}\mathbf{U}'$. As $\mathbf{H}_t = \mathbf{H}_t^2$ we have $\mathrm{tr}(\mathbf{Y}'\mathbf{H}_t\mathbf{Y}) = \mathrm{tr}(\mathbf{H}_t\mathbf{Y}\mathbf{Y}'\mathbf{H}_t) = \mathrm{tr}(\mathbf{H}_t\mathbf{G}_y\mathbf{H}_t)$, and similarly $\mathrm{tr}[(\mathbf{Y}'(\mathbf{I} - \mathbf{H}_t)\mathbf{Y}] = \mathrm{tr}[(\mathbf{I} - \mathbf{H}_t)\mathbf{G}_y(\mathbf{I} - \mathbf{H}_t)]$. Therefore, the ratio F_* can be formulated in terms of distances:

$$\begin{aligned} F_* &= \frac{\mathrm{tr}(\mathbf{H}_t\mathbf{G}_y\mathbf{H}_t)}{\mathrm{tr}[(\mathbf{I} - \mathbf{H}_t)\mathbf{G}_y(\mathbf{I} - \mathbf{H}_t)]} \\ &= \frac{\mathrm{tr}(\mathbf{G}_x^-\mathbf{G}_x\mathbf{G}_y\mathbf{G}_x^-\mathbf{G}_x)}{\mathrm{tr}[(\mathbf{I} - \mathbf{G}_x^-\mathbf{G}_x)\mathbf{G}_y(\mathbf{I} - \mathbf{G}_x^-\mathbf{G}_x)]}, \end{aligned}$$

where $\mathbf{G}_x^- = \mathbf{U}\Lambda_x^{-2}\mathbf{U}'$ is a g-inverse of $\mathbf{G}_x$, i.e., $\mathbf{G}_x\mathbf{G}_x^-\mathbf{G}_x = \mathbf{G}_x$.

By taking $F = F_* \times (n - p - 1)/(p + 1)$ we can invoke the F test when $q = 1$ and the only column of $\mathbf{Y}$ comes from a normal population. The F test is still justified when the rows of $\mathbf{Y}$ are multinormal with covariance matrix $\Sigma = \sigma^2\mathbf{I}$. For general data, testing $\mathbf{B} = \mathbf{0}$ can be performed by a permutation test.

To perform this test, we keep $\mathbf{Y}$ fixed, then find the $n!$ permutations of the rows of $\mathbf{X}$ and obtain the permutation distribution of F_*. There will be evidence against $\mathbf{B} = \mathbf{0}$ if the observed F_* is in the extreme tail. If n is large, we may choose at random (with repetition) a subset of $n!$ permutations; for an example with mixed data; see Cuadras (2011).

Tests based on F_* when only $\mathbf{Y}$ comes from a distance, have been used by McArdle and Anderson (2001) in relating ecological data, and Wessel and Schork (2006) in large-scale multilocus association studies. Here this test has been adapted to two distance matrices. However, this $\mathbf{F}$ approach has three drawbacks. First, it depends on $\mathbf{G}_y = \mathbf{V}\Lambda_y^2\mathbf{V}'$, i.e., on the diagonal matrix Λ_y^2, whose entries are proportional to the variances of the columns of $\mathbf{Y}$. Second, if F_* is significant, we accept dependence but we do not know the degree of association between both data sets. Finally, F_* is non symmetric in $\mathbf{X}$ and $\mathbf{Y}$.

4.2. *Wilks' Measure of Association*

There are alternative criteria for testing $\mathbf{B} = \mathbf{0}$ in the multivariate linear regression model $\mathbf{Y} = \mathbf{X}\mathbf{B} + \Xi$, which provide symmetric measures of multivariate association.

Let $\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}$ and $\mathbf{H} = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}$ the "error matrix" and the "hypothesis matrix," respectively, with $\mathbf{Y} = \mathbf{V}\Lambda_y$ and $\widehat{\mathbf{Y}} = \mathbf{H}_t\mathbf{Y} = \mathbf{U}\mathbf{U}'\mathbf{V}\Lambda_y$. We then have $\mathbf{E} = \Lambda_y(\mathbf{I} - \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V})\Lambda_y$ and $\mathbf{E} + \mathbf{H} = \mathbf{Y}'\mathbf{Y} = \Lambda_y\mathbf{V}'\mathbf{V}\Lambda_y = \Lambda_y^2$.

Likelihood ratio criterion or Wilk's lambda is well known in multivariate analysis; see Mardia et al. (1979). Wilk's lambda is

$$\begin{aligned} W &= \det(\mathbf{E})/\det(\mathbf{E} + \mathbf{H}) \\ &= \det(\mathbf{I} - \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}). \end{aligned}$$

W does not depend on Λ_y.

The canonical correlations r_i between $\mathbf{X}$ and $\mathbf{Y}$ satisfy the eigenequation

$$\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{v}_i = r_i^2\mathbf{Y}'\mathbf{Y}\mathbf{v}_i$$

i.e.,

$$\mathbf{H}\mathbf{v}_i = r_i^2(\mathbf{E} + \mathbf{H})\mathbf{v}_i,$$

where $\mathbf{v}_i$ is the corresponding eigenvector. This implies $\mathbf{E}\mathbf{v}_i = (1 - r_i^2)(\mathbf{E} + \mathbf{H})\mathbf{v}_i$. Therefore, W can be expressed in terms of canonical correlations

$$W = \prod_{i=1}^{s}(1 - r_i^2),$$

where $s = \min(p, q)$. Hence,

$$A_W = 1 - W = 1 - \prod_{i=1}^{s}(1 - r_i^2) \tag{1}$$

is an association measure which is 0 if $\mathbf{X}, \mathbf{Y}$ are independent, and 1 if $\mathbf{X}, \mathbf{Y}$ are linearly dependent. As A_W can be very close to 1 for large data sets, it is necessary to choose in advance the number of principal coordinates p and q, which determine the number s of canonical correlations. A proposal is presented in Sec. 6.

4.3. *More Association Measures*

For testing $\mathbf{B} = \mathbf{0}$, we may also employ other criteria, which also provide measures of association. For example, if $\mathbf{H}\mathbf{v}_i = \lambda_i \mathbf{E}\mathbf{v}_i$ gives the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$, and $\lambda_1 = \max(\mathbf{v}'\mathbf{H}\mathbf{v}/\mathbf{v}'\mathbf{E}\mathbf{v})$, then Roy R, Lawley-Hotelling U and Pillai's criterion V are given by

$$R = \lambda_1/(1+\lambda_1),$$
$$U = \mathrm{tr}[\mathbf{E}^{-1}\mathbf{H}] = \sum_{i=1}^{p} \lambda_i,$$
$$V = \mathrm{tr}[(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}] = \sum_{i=1}^{p} \lambda_i/(1+\lambda_i).$$

If $r_i, i = 1, \ldots, s$, are the canonical correlations, then $\lambda_i = r_i^2/(1-r_i^2)$ and the three criteria provide measures of multivariate association, $A_R = r_1^2$, $A_{LH} = (U/s)/(1+U/s)$ and $A_P = V/s$, also based on canonical correlations; see Table 2. For the derivation and sampling distribution of R, U, V and W under multinormality, see Anderson (2003).

We can obtain other measures based on generalized multiple correlation $\det(\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy})/\det(\mathbf{S}_{yy})$, the vectorial correlation $\mathrm{tr}(\mathbf{S}_{xy}\mathbf{S}_{yx})/\sqrt{\mathrm{tr}(\mathbf{S}_{xx}^2)\mathrm{tr}(\mathbf{S}_{yy}^2)}$, see Escoufier (1973), Rencher (1995), and the Procrustes statistic

$$P^2 = 1 - [\mathrm{tr}(\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X})^{1/2}]^2/[\mathrm{tr}(\mathbf{X}'\mathbf{X})\mathrm{tr}(\mathbf{Y}'\mathbf{Y})];$$

see Cox and Cox (2001). The distance-based version of these coefficients are

$$A_{HC} = \prod_{i=1}^{s} r_i^2, \quad A_P = \left(\sum_{i=1}^{s} r_i^2\right) \Big/ s \quad \text{and} \quad A_{PR} = \left(\sum_{i=1}^{s} r_i\right)^2 \Big/ s^2,$$

respectively.

Cramer and Nicewander (1979) gave geometrical arguments in terms of lengths of vectors and volumes of parallelotopes, to propose the measures $A_{CN1} = A_{HC}^{1/s}$, $A_{CN2} = 1 - W^{1/s}$. They also proposed the average $A_P = (\sum_{i=1}^{s} r_i^2)/s$, which also arises from Pillai's criterion. However, in general, coefficients A_{CN1}, A_{CN2} and A_P do not increase with the dimensionality s, the number of canonical correlations considered.

These measures are described and justified in Cuadras (2011). Table 2 reports all measures in terms of the association matrix $\mathbf{A} = \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}$, as well as $\mathbf{A}_c = \mathbf{I} - \mathbf{A}$. Thus $W = \det(\mathbf{A}_c)$ and $A_P = \mathrm{tr}(\mathbf{A})/s$.

Note that we are interested in relating two data sets rather than testing statistical hypotheses, but the above tests help us build association measures.

Table 2
Eight symmetric measures of multivariate association between two data sets based on distances

Measure of association	Expression in terms of $\mathbf{A} = \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}$, $\mathbf{A}_c = \mathbf{I} - \mathbf{A}$	Expression in terms of canonical correlations and $\lambda_i = r_i^2/(1 - r_i^2)$
A_R	First eigenvalue of $\mathbf{A}$	r_1^2
A_{HC}	$\det(\mathbf{A})$	$\prod_{i=1}^{s} r_i^2$
A_W	$1 - \det(\mathbf{A}_c)$	$1 - \prod_{i=1}^{s}(1 - r_i^2)$
A_{LH}	$[\mathrm{tr}(\mathbf{A}_c^{-1}\mathbf{A})/s]/[1 + \mathrm{tr}(\mathbf{A}_c^{-1}\mathbf{A})/s]$	$\frac{(\sum_{i=1}^{s} \lambda_i)/s}{1+(\sum_{i=1}^{s} \lambda_i)/s}$
A_P	$\mathrm{tr}(\mathbf{A})/s$	$(\sum_{i=1}^{s} r_i^2)/s$
A_{PR}	$[\mathrm{tr}(\mathbf{A}^{1/2})]^2/s^2$	$(\sum_{i=1}^{s} r_i)^2/s^2$
A_{CN1}	$[\det(\mathbf{A})]^{1/s}$	$(\prod_{i=1}^{s} r_i^2)^{1/s}$
A_{CN2}	$1 - [\det(\mathbf{A}_c)]^{1/s}$	$1 - [\prod_{i=1}^{s}(1 - r_i^2)]^{1/s}$

5. Choosing the Association Measure

It is worth noting that all association measures in Table 2 reduce to the squared multiple correlation coefficient when $\mathbf{Y}$ is univariate. In general, these measures are different. Since we have seven measures, a criterion to choose one is necessary.

As it is proved in Cramer and Nicewander (1979) and Cuadras (2011), the above measures of association can be ordered as follows:

$$A_{HC} \le r_s^2 \le A_{CN1} \le A_{PR} \le A_P \le A_{CN2} \le A_{LH} \le A_R \le A_W,$$

where $A_R = r_1^2$ and r_s^2 are the largest and smallest squared canonical correlations. Because of this order relationship, in practice A_{HC} can be quite small and A_W almost one.

We choose A_W because it is the largest coefficient, it depends on the s canonical correlations and it does not take small values if s is high, a property which does not hold for other coefficients. Thus, A_W works well with hyperspectral data in order to compare hyperspectral regions.

6. How Many Dimensions?

Choosing the number of (principal) dimensions is an important aspect in most techniques of multivariate analysis. Often the maximum dimension is limited by the number of variables p. However, when we work with distances, this maximum dimension can be larger, even can reach $n - 1$. In MMDS the number of dimensions considered in graphical representation, based on the percentage of variability accounted for by the first dimensions. This quality measure can be generalized; see Graffelman (2001). We propose a criterion, which extends a sequence defined in Cuadras et al. (1996).

First we fix two maximum dimensions K, L suggested by the data, see below. Let $\mathbf{u}_i, i = 1, \ldots, K$, $\mathbf{v}_j, j = 1, \ldots, L$, be the first K, L columns of $\mathbf{U}$ and $\mathbf{V}$, respectively. We define the sequence

$$c_{kl} = \frac{\mathbf{1}_k'[\Lambda_x^2(\mathbf{U}'\mathbf{V} * \mathbf{U}'\mathbf{V})\Lambda_y^2]\mathbf{1}_l}{\mathbf{1}_K'[\Lambda_x^2(\mathbf{U}'\mathbf{V} * \mathbf{U}'\mathbf{V})\Lambda_y^2]\mathbf{1}_L} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\lambda_{ix}^2(\mathbf{u}_i'\mathbf{v}_j)^2\lambda_{jy}^2}{\sum_{i=1}^{K}\sum_{j=1}^{L}\lambda_{ix}^2(\mathbf{u}_i'\mathbf{v}_j)^2\lambda_{jy}^2}, \quad k, l = 1, \ldots, K, L, \quad (2)$$

where $\mathbf{1}_k = (1, \overset{(k}{\ldots}, 1, 0, \overset{(n-k}{\ldots}, 0)'$ and λ_{ix}^2, λ_{jy}^2 are the eigenvalues (also called inertias) of $\mathbf{G}_x = \mathbf{U}\Lambda_x^2\mathbf{U}'$, $\mathbf{G}_y = \mathbf{V}\Lambda_y^2\mathbf{V}'$, respectively. These eigenvalues are proportional to the variances of the corresponding principal axes. Here, $*$ denotes element wise multiplication. Note that $\mathbf{u}_i'\mathbf{v}_j$ is just the correlation coefficient between the ith and jth principal coordinates obtained from Δ_x and Δ_y, respectively. Thus, the numerator in c_{kl} is a weighted average of the relationships between principal axes. Clearly,

$$0 < c_{11} \leq \cdots \leq c_{kl} \leq \cdots \leq c_{k'l'} \leq \cdots \leq c_{KL} = 1, \quad \text{if } k \leq k', \ l \leq l'.$$

We should choose dimension $s = \min(k, l)$ if $100 \times c_{kl}$ is high, for example, 90%.

As for the maximum dimensions, if we take K_1, L_1 initially, the correct dimensions should be K, L if $c_{K+1,L+1} \simeq \cdots \simeq c_{K_1,L_1} = 1$, or very close to 1; see two numerical examples in Sec. 8.

In practice, $(\mathbf{u}_i'\mathbf{v}_j)^2$ decreases as i and j increase, but a dimension $i < s$ of the first data set may be highly correlated with a (removed) dimension $j > s$ of the second data set. As discussed in Cuadras (1993), this quirk could be interpreted in the sense that this ith principal dimension depends on "noise," rather than the main variability of the second data set.

7. Comparing Regions of Hyperspectral Images

7.1. *Frequency Table*

One hyperspectral image is divided into regions, the initial ones being pixels. Each data matrix, representing a region, is obtained by considering a set of spectral curves depicting the radiation (number of photons) at different wavelengths. For example, spectrum S_1 (see Table 1) reflects 634, 861, 1104, 1100 photons at wavelengths $\lambda_1, \lambda_2, \lambda_3, \lambda_4$. The distribution of the radiances at wavelength λ_1 is then computed, and similarly for the other wavelengths, as is described in Table 1 (bottom).

Accordingly, this information (counts of photons) is transformed into a data matrix, where the columns correspond to radiances (number of photons) and the rows to wavelengths. Thus, each row represents the observed statistical distribution of the radiance for a given wavelength.

In general, given a set of S spectra belonging to a hyperspectral region of an image, we must consider the table:

		Radiance values (photons)					
		1	2	$\cdots$	j	$\cdots$	p
Wave-lengths	λ_1	f_{11}	f_{12}	$\cdots$	f_{1j}	$\cdots$	f_{1p}
	λ_2	f_{21}	f_{22}	$\cdots$	f_{2j}	$\cdots$	f_{2p}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	λ_n	f_{n1}	f_{n2}	$\cdots$	f_{nj}	$\cdots$	f_{np}

where f_{ij} is the frequency of the radiance j at wavelength λ_i, that is, f_{ij} spectra reflects j photons at this wavelength. (Note that we can also have class intervals $I_1, I_2, \ldots, I_p$ of radiances rather than specific values.) For each row, the sum of the frequencies is n. Both n and p can take large values. If $\mathbf{F} = (f_{ij})$, the data is represented by an $n \times p$ matrix $\mathbf{P} = \mathbf{F}/n$, with non negative entries p_{ij} such that

$$\sum_{j=1}^{p} p_{ij} = 1,$$

i.e., $\mathbf{P}\mathbf{1}_p = \mathbf{1}_n$. Since the n wavelengths represent discrete contiguous values, the distribution in a row is very similar to the distribution in neighboring rows. To measure this proximity, we use Hellinger distance between rows:

$$\delta^2_{ii'} = \sum_{j=1}^{p} (\sqrt{p_{ij}} - \sqrt{p_{i'j}})^2 = 2\left(1 - \sum_{j=1}^{p} \sqrt{p_{ij}}\sqrt{p_{i'j}}\right).$$

This gives an $n \times n$ (squared) distance matrix:

$$\Delta * \Delta = 2(\mathbf{1}_n\mathbf{1}'_n - \sqrt{\mathbf{P}}\sqrt{\mathbf{P}'}),$$

where $\sqrt{\mathbf{P}} = (\sqrt{p_{ij}})$. As $\mathbf{H}_c\mathbf{1}_n = \mathbf{0}$, $\mathbf{1}'_n\mathbf{H}_c = \mathbf{0}'$, where $\mathbf{H}_c$ is the centering matrix, we have

$$\mathbf{H}_c\sqrt{\mathbf{P}}\sqrt{\mathbf{P}'}\mathbf{H}_c = \mathbf{U}\Lambda^2\mathbf{U}',$$

and the principal and standard coordinates of the n wavelengths are the rows of $\mathbf{X} = \mathbf{U}\Lambda$ and $\mathbf{U}$, respectively. We can also find $\mathbf{U}$ from the SVD

$$(\sqrt{\mathbf{P}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n\sqrt{\mathbf{P}}) = \mathbf{U}\Lambda\mathbf{T}'.$$

7.2. *Image Comparison*

Suppose that we have another hyperspectral image region of the same size, which provides the $n \times p$ matrix $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{1}_p = \mathbf{1}_n$. Following the same procedure, we obtain the standard coordinates $\mathbf{V}$ from the SVD

$$(\sqrt{\mathbf{Q}} - \frac{1}{n}\mathbf{1}_n\mathbf{1}'_n\sqrt{\mathbf{Q}}) = \mathbf{V}\Lambda\mathbf{W}'.$$

To relate both regions we find the association between $\mathbf{P}$ and $\mathbf{Q}$ by computing the matrix $\mathbf{A} = \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}$. Finally we use Wilks association measure (1) $A_W = 1 - \det(\mathbf{I} - \mathbf{A})$, which approaches 1 when $\mathbf{P}$ is very similar to $\mathbf{Q}$.

7.3. *Distance Choice*

Note that, by using distances and MMDS, we can address the case $p > n$. But, what distance? We choose Hellinger distance between two probability densities for several reasons. First, it has a simple geometric interpretation and provides a closed

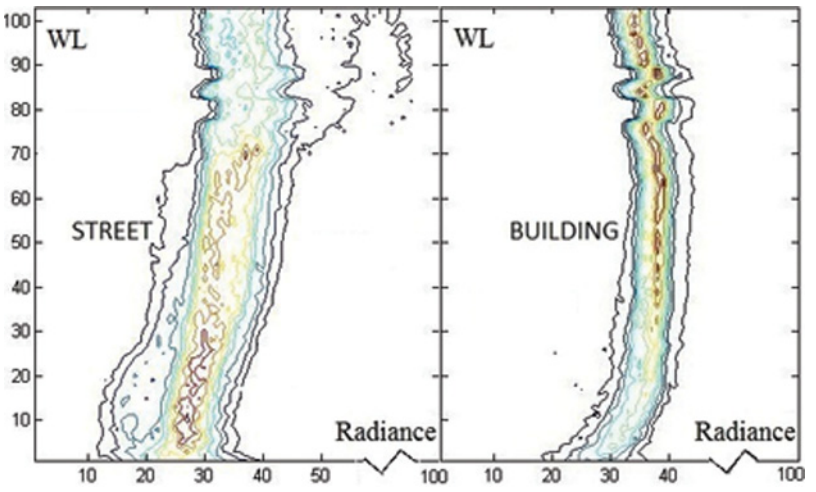

Figure 3. Contour plots of the probability matrices obtained from the regions of two spectral images taken from a street and a building. The white lateral parts mean that these matrices contain many zeros. The radiance values are grouped in 100 intervals. Wilks association measure is 0.8157. (color figure available online.)

formula for obtaining the principal coordinates. Second, it gives values similar to other distances as chi-square and log-ratio, which belong to a parametric family including Hellinger; see Cuadras and Cuadras (2006). Third, as discussed in Cuadras et al. (2006) and Rencher (1995), Hellinger distance is more appropriate when we have multinomial populations. This is the case of the rows of **F**. It is worth noting that while log-ratio (Aitchison) distance is suitable for compositional data, it can not be used here because **F** and **P** contain many zeros; see Fig. 3. Finally, Hellinger distance δ locally coincides with Bhattacharyya distance $\arccos(\delta)$, the information metric under a multinomial statistical model.

8. Two Examples

We present two examples illustrating the relation between two contiguous regions of hyperspectral curves. These images have been obtained from Pavia University (Pavia, Italy).

Firstly, we consider two data sets, Tree1 and Tree2, obtained from landscapes containing trees. The data used here consists of $n = 103$ wavelengths and $p = 200$ distinct class intervals (with length 35) of radiance values, providing two matrices of order 103×200. We take initially $K_1 = L_1 = 6$. Then (2) gives

$$c_{11} = 0.937 < c_{22} = 0.938 < c_{33} = 0.980 < c_{44} = 0.983 < c_{55} = 0.991 < c_{66} = 1.$$

This result suggests the choice $K = L = 3$. Again from (2) we obtain the $[100c_{kl}]$ table

$$\begin{bmatrix} 95.6 & 95.6 & 95.8 \\ 95.6 & 95.7 & 99.4 \\ 95.9 & 96.1 & 100 \end{bmatrix}.$$

We take $s = 1$ and the association measure is

$$A_W(\text{Tree1, Tree2}) = 0.9510.$$

As A_W is close to 1, both data sets represent similar trees, belonging to the same cluster, i.e., the same class of material. Therefore, these regions are merged. Note that $p > n$.

Secondly, we consider two data sets, Building and Street, obtained from the same city. Now we have $n = 103$ wavelengths and $p = 100$ class intervals (with length 65) of radiance values. Figure 3 shows the contour plots of the corresponding matrices $\mathbf{P}, \mathbf{Q}$. Initially, we choose $K_1 = L_1 = 6$. Then (2) gives

$$c_{11} = 0.681 < c_{22} = 0.930 < c_{33} = 0.987 < c_{44} = 0.995 < c_{55} = 0.999 < c_{66} = 1.$$

We directly take $s = 2$ dimensions; see Fig. 4. The association measure is

$$A_W(\text{Building, Street}) = 0.8157,$$

indicating that both data sets are relatively dissimilar, representing regions of urban objects belonging to different clusters, i.e., different classes of material. Therefore these regions are not merged.

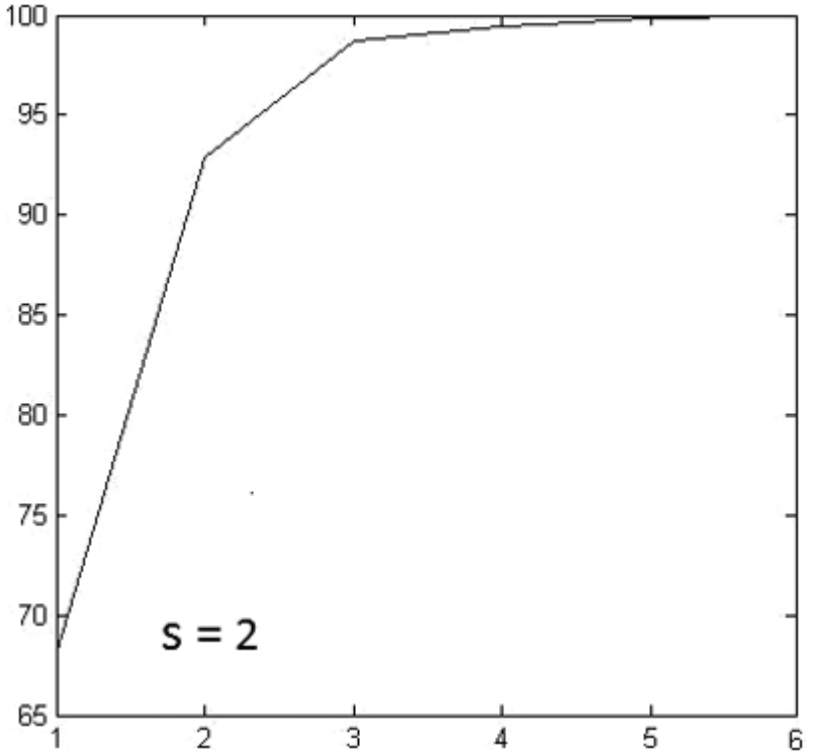

Figure 4. Plot of $100 \times c(i, i)$ vs. dimension i, see Eq. (2), for the images taken from a street and a building, indicating two canonical dimensions.

In image processing, when the partition of the wavelength band is very fine or the spectral resolution of the sensor increases, the dimensions n and p of the data matrix can be much larger. However, increasing the dimensionality from moderate to high, may not change the results much.

9. Discussion

The association measures presented here are obtained from distances, and therefore can be used for general data sets. For example, we can deal with mixed data, by using a distance based on Gower's similarity coefficient and relating matrices of principal coordinates; see Cuadras (2011).

We can relate compositional data sets (the rows sum up to 1), by using Hellinger distance. These sets come from regions of hyperspectral images, encoded as matrices with probability distributions in the rows, making Hellinger distance a suitable choice because there are many zeros. Moreover, the number of radiances p can be higher than the number of wavelengths n. Then, if we use a standard approach, we may find both the first canonical correlation and A_W equal to one. This trivial solution is avoided by using Hellinger distance and computing coordinates.

Other association measures may be used. However, because the order relationships $0 \leq A_{HC} \leq \cdots \leq A_W \leq 1$, we can find values A_{HC} close to zero and A_W close to one. As reported by Rencher (1995), measures A_W, A_{CN2}, A_{LH} agree in general, but A_{HC}, A_P may not indicate the same level of association. Of course, a single measure may not fully inform about the comparison between two data sets. However, to construct the Binary Partition Tree, we must summarize all of the relationships between two hyperspectral regions. We recommend the use of A_W for two reasons: it is the largest measure of association and, in general, does not take too small values if the dimensionality increases. So it works well with hyperspectral data.

Acknowledgments

Work supported in part by MEC (Spain) grant MTM2008-00642. Thanks are due to two anonymous referees for useful comments.

References

Amat, L., Robert, D., Besalú, E., Carbó-Dorca, R. (1998). Molecular quantum similarity measures tuned 3D QSAR: An antitumoral family validation study. *Chem. Inform. Comput. Sci.* 38:624–631.

Anderson, T. W. (2003). *An Introduction to Multivariate Analysis*. 3d ed. New York: Wiley.

Bartkowiak, A., Jakimiec, M. (1994). Distance-based regression in prediction of solar flare activity. *Qüestiió* 18:7–38.

Boj, E., Claramunt, M. M., Fortiana, J. (2007). Selection of predictors in distance-based regression. *Commun. Statist. B Simul. Computat.* 36:87–98.

Cox, T. V., Cox, M. A. A. (2001). *Multidimensional Scaling*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

Cramer, E. M., Nicewander, W. A. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika* 44:43–54.

Cuadras, C. M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In: Dodge, Y., ed. *Statistical Data Analysis Inference*. Amsterdam: Elsevier Science Publishers B. V. (North–Holland), pp. 459–473.

Cuadras, C. M. (1993). Interpreting an inequality in multiple regression. *Amer. Statistician* 47:256–258.

Cuadras, C. M., Arenas, C. (1990). A distance based regression model for prediction with mixed data. *Commun. Statist. A Theor. Meth.* 19:2261–2279.

Cuadras, C. M., Arenas, C., Fortiana, J. (1996). Some computational aspects of a distance-based model for prediction. *Commun. Statist. B Simul. Computat.* 25:593–609.

Cuadras, C. M., Cuadras, D. (2006). A parametric approach to correspondence analysis. *Linear Alg. Applic.* 417:64–74.

Cuadras, C. M., Cuadras, D., Greenacre, M. (2006). A comparison of different methods for representing categorical data. *Commun. Statist. B Simul. Computat.* 35:447–459.

Cuadras, C. M. (2008). Distance-based multisample tests for multivariate data. In: Arnold, B. C., Balakrishnan, N., Sarabia, J. M., Mínguez, R., eds. *Advances in Mathematical Statistical Modeling*. Boston: Birkhauser, pp. 61–71.

Cuadras, C. M. (2011). Distance-based approach in multivariate association. In: Ingrassia, S., Rocci, R., Vichi, M., eds. *New Perspectives in Statistical Modeling Data Analysis*. Berlin: Springer, pp. 535–542.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* 29:751–760.

Esteve, A., Boj, E., Fortiana, J. (2009). Interaction terms in distance-based regression. *Commun. Statist. A Theor. Meth.* 38:3498–3509.

Graffelman, J. (2001). Quality statistics in canonical correspondence analysis. *Environmetrics* 12:485–497.

Hotelling, H. (1936). Relations between two sets of variants. *Biometrika* 28:321–377.

Jiang, S. (1996). Angles between Euclidean subspaces. *Geometriae Dedicata* 63:113–121.

McArdle, B. H., Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance based redundancy analysis. *Ecology* 82:290–297.

Manly, B. F. J. (1986). *Multivariate Statistical Methods: a Primer*. London, New York: Chapman and Hall.

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

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: Wiley.

Rao, C. R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiió* 19:23–63.

Rencher, A. V. (1995). *Methods of Multivariate Analysis*. New York: Wiley.

Salembier, P., Garrido, L. (2000). Binary partition tree as an efficient representation for image processing, segmentation and information retrieval. *IEEE Trans. Image Process.* 9:561–576.

Wessel, J., Schork, N. J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *Amer. J. Hum. Genet.* 79:792–806.

Zapala, M. A, Schork, N. J. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci. USA* 103:19430–19435.