Multivariate association of compositional data matrices with applications in comparing hyperspectral images

C. M. CUADRAS[1] and S. VALERO[2]

[1]Departament d'Estadistica - Universitat de Barcelona, Barcelona, Spain
[2]GIPSA-lab, Signal & Image Dept. - Grenoble Institute of Technology, Grenoble, France

1 Introduction

It is well-known in image processing that, by varying the wavelength, any material reflects and absorbs in a different way the solar radiation. This 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 which are previously modelled by using compositional data matrices, where the entries in each row is a statistical discrete distribution of the radiance values (columns). These rows correspond to distinct but contiguos wavelengths. Thus the distribution in a row is very similar to the distribution in close rows. To measure this proximity, we use Hellinger distance between rows, which provides a distance matrix.

Given two hyperspectral regions of an image providing two compositional data matrices, we obtain the corresponding distance matrices and, by using metric multidimensional scaling, we compute two sets of principal coordinates, which are related by a multivariate association measure based on canonical correlations.

We ilustrate this approach comparing some multivariate regions of images captured by hyperspectral remote sensors.

2 Transforming hyperspectral images to compositional data matrices

Any data matrix, representing a region, is obtained by considering a set of spectral curves depicting the radiation (number of photons) at different wavelengths. As a simple illustration, suppose 3 spectra corresponding to 3 pixels:

	Wavelength		
	$780nm$	$50\mu m$	$1mm$
S_1	22	12	80
S_2	20	12	89
S_3	18	12	89

Each spectrum reflects photons, for instance S_1 reflects 22 at wavelength $780nm$, and S_3 reflects 89 at wavelength $1mm$. This information is transformed into a data matrix, where the columns correspond to radiance (number of photons) and the rows to wavelengths. Any row sums up to 1 and represents the observed statistical distribution of the radiance for a given wavelength. The above data gives:

	Radiance																				
	1	$\cdots$	12	$\cdots$	18	$\cdots$	20	$\cdots$	22	$\cdots$	80	$\cdots$	89	$\cdots$	100
$780nm$	0	$\cdots$	0	$\cdots$	1/3	$\cdots$	1/3	$\cdots$	1/3	$\cdots$	0	$\cdots$	0	$\cdots$	0
$50\mu m$	0	$\cdots$	1	$\cdots$	0	$\cdots$	0	$\cdots$	0	$\cdots$	0	$\cdots$	0	$\cdots$	0
$1mm$	0	$\cdots$	0	$\cdots$	0	$\cdots$	0	$\cdots$	0	$\cdots$	1/3	$\cdots$	2/3	$\cdots$	0

In general, given a set of spectra belonging to a hyperspectral region of an image, we have a continuous range of wavelengths, conveniently discretized in n equidistant values in order to obtain

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

a finite radiance table. Thus we have n wavelengths and p radiances and the transformed data is represented by a $n \times p$ matrix $\mathbf{P}$, with non negative entries p_{ij} such that

$$\sum_{j=1}^{p} p_{ij} = 1, \qquad i = 1, \cdots, n,$$

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 close rows. To measure this proximity, we may use Aitchison's distance. However, each row of $\mathbf{P}$ contains a relatively large sequence of zeros and our aim is not to represent the data in a low dimensional space (Aitchison and Greenacre, 2002). Hellinger distance between rows

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

may be more suitable for our study and can be handled easily by using matrix algebra. This gives a $n \times n$ (squared) distance matrix:

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

where $\sqrt{\mathbf{P}} = (\sqrt{p_{ij}})$.

Next, we use metric scaling to find the principal coordinates for $\Delta^{(2)}$, i.e., the spectral decomposition $\mathbf{H}(-\frac{1}{2}\Delta^{(2)})\mathbf{H} = \mathbf{U}\Lambda^2\mathbf{U}'$, where Λ is diagonal and $\mathbf{H} = \mathbf{I}_n - (1/n)\mathbf{1}_n\mathbf{1}'_n$ is the centring matrix. As $\mathbf{H}\mathbf{1}'_n = \mathbf{1}_n\mathbf{H} = \mathbf{0}$, we have

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

and the principal and standard coordinates of the n wavelengths are the rows of $\mathbf{X} = \mathbf{U}\Lambda$ and $\mathbf{U}$, respectively. As it has been noted above, we are not interested in representing the n wavelengths in low dimension. Our aim is to compare two regions of an image.

3 Comparing regions of hyperspectral images

Given two hyperspectral regions of an image providing two data matrices $\mathbf{P}$ and $\mathbf{Q}$, from $\mathbf{H}\sqrt{\mathbf{P}}\sqrt{\mathbf{P}'}\mathbf{H} = \mathbf{U}\Lambda_x^2\mathbf{U}'$, and $\mathbf{H}\sqrt{\mathbf{Q}}\sqrt{\mathbf{Q}'}\mathbf{H} = \mathbf{V}\Lambda_y^2\mathbf{V}'$, we obtain the matrices of standard coordinates $\mathbf{U}, \mathbf{V}$ and the association matrix $\mathbf{A} = \mathbf{V}'\mathbf{U}\mathbf{U}'\mathbf{V}$. Notice that the entries $\mathbf{u}'_i\mathbf{v}_j$ in the $p \times q$ matrix $\mathbf{U}'\mathbf{V}$ are the correlation coefficient between the $i-$th and $j-$th principal coordinates obtained from $\mathbf{P}$ and $\mathbf{Q}$, respectively. We choose the so-called Wilks multivariate association measure

$$A_W = 1 - \det(\mathbf{I} - \mathbf{A}) = 1 - \prod_{i=1}^{s}(1 - r_i^2),$$

where $r_i, i = 1, \ldots, s \leq \min(p, q)$ are the first s canonical correlations between $\mathbf{X}$ and $\mathbf{Y}$. This measure satisfies $0 \leq A_W \leq 1$ and comes from considering a multivariate regression model relating $\mathbf{X}$ and $\mathbf{Y}$. See Cramer and Nicewander (1979) and Cuadras (2008).

4 Choosing the predictive dimensions

In general, since the column dimensions p, q of $\mathbf{P}, \mathbf{Q}$ are quite large, the coefficient A_W could be very close to 1. To determine how many dimensions (or coordinates) K and L should be taken from $\mathbf{X}$, and $\mathbf{Y}$, respectively, where $K < p$, $L < q$, we choose the (provisional) values K, L suggested by the data. Next, by extending a coefficient defined in Cuadras *et al.* (1996), we propose the sequence

$$c_{kl} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\lambda_{ix}^2(\mathbf{u}'_i\mathbf{v}_j)^2\lambda_{jy}^2}{\sum_{i=1}^{K}\sum_{j=1}^{L}\lambda_{ix}^2(\mathbf{u}'_i\mathbf{v}_j)^2\lambda_{jy}^2}, \quad k, l = 1, \ldots, K, L, \tag{1}$$

where λ_{ix}^2, λ_{jy}^2 are eigenvalues of $\mathbf{H}\sqrt{\mathbf{P}}\sqrt{\mathbf{P}'}\mathbf{H}$ and $\mathbf{H}\sqrt{\mathbf{Q}}\sqrt{\mathbf{Q}'}\mathbf{H}$. Thus the numerator in c_{kl} is a weighted average of the relationships between principal axes. Clearly

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

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

5 Two examples

Firstly, we consider two data sets, Tree1 and Tree2, captured from landscapes containing trees. The data used here consists of $n = 103$ wavelengths and $p = 200$ radiance values, providing two matrices of order 103×200. We choose $K = L = 3$ and from (1) we obtain the $[100c_{kl}]$ table

$$\begin{bmatrix} 96.2 & 96.32 & 96.4 \\ 96.3 & 96.4 & 99.4 \\ 96.5 & 96.8 & 100 \end{bmatrix}.$$

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.

Secondly, we consider two data sets, Building and Street, captured from a city. Now we have $n = 103$ wavelengths and $p = 100$ radiance values. We also choose $K = L = 3$ and obtain the $[100c_{kl}]$ table

$$\begin{bmatrix} 63.9 & 81.9 & 82.3 \\ 75.1 & 93.4 & 95.1 \\ 79.6 & 98.2 & 100 \end{bmatrix}.$$

We take $s = 2$ dimensions. The association measure is

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

indicating that both data sets are relatively dissimilar, representing urban objects belonging to different clusters, i.e., different classes of material.

References

Aitchison, J., Greenacre, M. (2002) Biplots of compositional data. *Applied Statistics*, **51**, 375-392.

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

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

Cuadras, C. M., Arenas, A., Fortiana, J. (1996) Some computational aspects of a distance-based model for prediction. *Communications in Statistics: Simulation and Computation*, **25**, 593-609.