Discriminant Correspondence Analysis

Hervé Abdi[1]

1 Overview

As the name indicates, discriminant correspondence analysis (DCA) is an extension of discriminant analysis (DA) and correspondence analysis (CA). Like discriminant analysis, the goal of DCA is to categorize observations in pre-defined groups, and like correspondence analysis, it is used with nominal variables.

The main idea behind DCA is to represent each group by the sum of its observations and to perform a simple CA on the groups by variables matrix. The original observations are then projected as supplementary elements and each observation is assigned to the closest group. The comparison between the *a priori* and the *a posteriori* classifications can be used to assess the quality of the discrimination. A similar procedure can be used to assign new observations to categories. The stability of the analysis can be evaluated using cross-validation techniques such as jackknifing or bootstrapping.

[1]In: Neil Salkind (Ed.) (2007). *Encyclopedia of Measurement and Statistics.* Thousand Oaks (CA): Sage.
Address correspondence to: Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utd.edu/~herve

2 An example

It is commonly thought that the taste of wines depends upon their origin. As an illustration we have sampled 12 wines coming from 3 different origins (4 wines per origin) and asked a professional taster (unaware of the origin of the wines) to rate these wines on 5 scales. The scores of the taster were then transformed into binary codes to form an indicator matrix (as in multiple correspondence analysis). For example, a score of 2 on the "Fruity" scale would be coded by the following pattern of 3 binary values: 0 1 0. An additional unknown wine was also evaluated by the taster with the goal of predicting its origin from the ratings. The data are given in Table 1.

3 Notations

There are K groups, each group comprising I_k observations and the sum of the I_k's is equal to I which is the total number of observations. For convenience, we assume that the observations constitute the rows of the data matrix, and that the variables are the columns. There are J variables. The $I \times J$ data matrix is denoted $\mathbf{X}$. The *indicator* matrix is an $I \times K$ matrix denoted $\mathbf{Y}$ in which a value of 1 indicates that the row belongs to the group represented by the column and a value of 0 indicates that it does not. The $K \times J$ matrix denoted $\mathbf{N}$, is called the "group matrix," it stores the total of the variables for each category. For our example, we find that:

$$\mathbf{N} = \mathbf{Y}^{\mathsf{T}}\mathbf{X} = \begin{bmatrix} 3 & 1 & 0 & 0 & 1 & 3 & 0 & 2 & 2 & 2 & 2 & 0 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 2 & 1 & 2 & 1 & 1 & 0 & 1 & 3 & 1 & 1 & 1 & 1 \\ 0 & 1 & 3 & 3 & 1 & 0 & 1 & 1 & 2 & 3 & 1 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{1}$$

Performing CA on the group matrix $\mathbf{N}$ provides two sets of factor scores: one for the groups (denoted $\mathbf{F}$) and one for the variables (denoted $\mathbf{G}$). These factor scores are, in general scaled such that their variance is equal to the eigenvalue associated with the factor.

The grand total of the table is noted N, and the first step of the analysis is to compute the probability matrix $\mathbf{Z} = N^{-1}\mathbf{N}$. We

Table 1: Data the 3 region wines example: 12 wines from 3 different regions are rated on 5 descriptors. A value of 1 indicates that the wine possesses the given value of the variable. The wine *W?* is an unknown wine treated as a supplementary observation.

		Woody			Fruity			Sweet			Alcohol			Hedonic			
Wine	Region	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	4
1	1 Loire	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0
2	1 Loire	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0
3	1 Loire	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0
4	1 Loire	1	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1
$\sum$	1 Loire	3	1	0	0	1	3	0	2	2	2	2	0	1	1	1	1
5	2 Rhône	1	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0
6	2 Rhône	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0
7	2 Rhône	0	0	1	0	1	0	0	1	0	0	1	0	1	0	0	0
8	2 Rhône	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1
$\sum$	2 Rhône	1	2	1	1	2	1	2	1	1	0	1	3	1	1	1	1
9	3 Beaujolais	0	0	1	1	0	0	0	0	1	1	0	0	1	0	0	0
10	3 Beaujolais	0	1	0	1	0	0	0	0	1	1	0	0	0	1	0	0
11	3 Beaujolais	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1
12	3 Beaujolais	0	0	1	1	0	0	1	0	0	1	0	0	0	0	1	0
$\sum$	3 Beaujolais	0	1	3	3	1	0	1	1	2	3	1	0	1	1	1	1
W?	?	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0

denote $\mathbf{r}$ the vector of the row totals of $\mathbf{Z}$, (*i.e.*, $\mathbf{r} = \mathbf{Z1}$, with $\mathbf{1}$ being a conformable vector of 1's) $\mathbf{c}$ the vector of the column totals, and $\mathbf{D_c} = \text{diag}\{\mathbf{c}\}$, $\mathbf{D_r} = \text{diag}\{\mathbf{r}\}$. The factor scores are obtained from the following singular value decomposition:

$$\mathbf{D_r}^{-\frac{1}{2}}\left(\mathbf{Z} - \mathbf{rc}^{\mathsf{T}}\right)\mathbf{D_c}^{-\frac{1}{2}} = \mathbf{P\Delta Q}^{\mathsf{T}} . \tag{2}$$

($\mathbf{\Delta}$ is the diagonal matrix of the *singular* values, and $\mathbf{\Lambda} = \mathbf{\Delta}^2$ is the matrix of the *eigenvalues*). The row and (respectively) column factor scores are obtained as

$$\mathbf{F} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{P\Delta} \qquad \text{and} \qquad \mathbf{G} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{Q\Delta} . \tag{3}$$

The squared (χ^2) distances from the rows and columns to their respective barycenters are obtained as

$$\mathbf{d_r} = \text{diag}\left\{\mathbf{FF}^{\mathsf{T}}\right\} \qquad \text{and} \qquad \mathbf{d_c} = \text{diag}\left\{\mathbf{GG}^{\mathsf{T}}\right\} . \tag{4}$$

The squared *cosines* between row i and factor ℓ and column j and factor ℓ are obtained respectively as:

$$o_{i,\ell} = \frac{f_{i,\ell}^2}{d_{r,i}^2} \qquad \text{and} \qquad o_{j,\ell} = \frac{g_{j,\ell}^2}{d_{c,j}^2} . \tag{5}$$

(with $d_{r,i}^2$, and $d_{c,j}^2$, being respectively the i-th element of $\mathbf{d_r}$ and the j-th element of $\mathbf{d_c}$). Squared cosines help locating the factors important for a given observation. The *contributions* of row i to factor ℓ and of column j to factor ℓ are obtained respectively as:

$$t_{i,\ell} = \frac{f_{i,\ell}^2}{\lambda_\ell} \qquad \text{and} \qquad t_{j,\ell} = \frac{g_{j,\ell}^2}{\lambda_\ell} . \tag{6}$$

Contributions help locating the observations important for a given factor.

Supplementary or illustrative elements can be projected onto the factors using the so called *transition* formula. Specifically, let $\mathbf{i}_{\text{sup}}^{\mathsf{T}}$ being an illustrative row and $\mathbf{j}_{\text{sup}}$ being an illustrative column to be projected. Their coordinates $\mathbf{f}_{\text{sup}}$ and $\mathbf{g}_{\text{sup}}$ are obtained as:

$$\mathbf{f}_{\text{sup}} = \left(\mathbf{i}_{\text{sup}}^{\mathsf{T}}\mathbf{1}\right)^{-1}\mathbf{i}_{\text{sup}}^{\mathsf{T}}\mathbf{G\Delta}^{-1} \text{ and } \mathbf{g}_{\text{sup}} = \left(\mathbf{j}_{\text{sup}}^{\mathsf{T}}\mathbf{1}\right)^{-1}\mathbf{j}_{\text{sup}}^{\mathsf{T}}\mathbf{F\Delta}^{-1} . \tag{7}$$

[note that the scalar terms $\left(\mathbf{i}_{sup}^{T}\mathbf{1}\right)^{-1}$ and $\left(\mathbf{j}_{sup}^{T}\mathbf{1}\right)^{-1}$ are used to insure that the sum of the elements of $\mathbf{i}_{sup}$ or $\mathbf{j}_{sup}$ is equal to one, if this is already the case, these terms are superfluous].

After the analysis has been performed on the groups, the original observations are projected as supplementary elements and their factor scores are stored in a matrix denoted $\mathbf{F}_{sup}$. To compute these scores, first compute the matrix of row profiles

$$\mathbf{R} = \left(\text{diag}\left\{\mathbf{X1}\right\}\right)^{-1}\mathbf{X} \tag{8}$$

and then apply Equation 7 to obtain

$$\mathbf{F}_{sup} = \mathbf{R}\mathbf{G}\boldsymbol{\Delta}^{-1}\ . \tag{9}$$

The Euclidean distance between the observations and the groups computed from the factor scores is equal to the χ^2-distance between their row profiles. The $I \times K$ distance matrix between observations and groups is computed as

$$\mathbf{D} = \mathbf{s}_{sup}\mathbf{1}^{T} + \mathbf{1}\mathbf{s}^{T} - 2\mathbf{F}_{sup}\mathbf{F}^{T} \tag{10}$$

with

$$\mathbf{s}_{sup} = \text{diag}\left\{\mathbf{F}_{sup}\mathbf{F}_{sup}^{T}\right\} \text{ and } \mathbf{s} = \text{diag}\left\{\mathbf{F}\mathbf{F}^{T}\right\}\ . \tag{11}$$

Each observation is then assigned to the closest group.

3.1 Model Evaluation

The quality of the discrimination can be evaluated as a fixed effect model or as a random effect model. For the fixed effect model, the correct classifications are compared to the assignments obtained from Equation 10. The fixed effect model evaluates the quality of the classification on the sample used to build the model.

The random effect model evaluates the quality of the classification on new observations. Typically, this step is performed using cross-validation techniques such as jackknifing or bootstrapping.

Table 2: Factor scores, squared cosines, and contributions for the variables (J-set). Contributions corresponding to negative scores are in italic.

			Woody			Fruity			Sweet			Alcohol			Hedonic			
			1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	4
Axis	λ	%	Factor Scores															
1	.251	55	.93	−.05	−.88	−.88	−.05	.93	−.51	.33	.04	−.14	.33	−.20	0	0	0	0
2	.201	44	−.04	.35	−.31	−.31	.35	−.04	.64	−.13	−.28	−.74	−.13	1.40	0	0	0	0
Axis			Squared Cosines															
1			.998	.021	.892	.892	.021	.998	.384	.864	.021	.035	.864	.021	0	0	0	0
2			.002	.979	.108	.108	.979	.002	.616	.137	.979	.965	.137	.979	0	0	0	0
Axis			Contributions															
1			.231	*.001*	*.207*	*.207*	*.001*	.231	*.051*	.029	.001	*.007*	.029	*.008*	0	0	0	0
2			*.0006*	.0405	*.0313*	*.0313*	.0405	*.0006*	.1019	*.0056*	*.0324*	*.2235*	*.0056*	.4860	0	0	0	0

Table 3: Factor scores, squared cosines, and contributions for the regions, and the supplementary rows being the wines from the region and the mysterious wine (W?). Contributions corresponding to negative scores are in italic.

				Loire Wines					Rhône Wines					Beaujolais Wines				
			Region					Region					Region					
Axis	λ	%	Loire	1	2	3	4	Rhône	1	2	3	4	Beaujolais	1	2	3	4	W?
Factor Scores																		
1	251	55	0.66	0.82	0.50	0.43	0.89	−0.10	0.07	−0.66	−0.11	0.29	−0.56	−0.74	−0.41	−0.11	−0.96	1.01
2	201	44	−0.23	−0.42	−0.05	−0.25	−0.22	0.63	1.05	0.93	−0.10	0.64	−0.39	−0.73	−0.43	−0.10	−0.32	−0.15
Squared Cosines																		
1			.89	.79	.99	.75	.94	.03	.00	.33	.56	.17	.67	.51	.47	.56	.90	.98
2			.11	.21	.01	.25	.06	.97	1.00	.67	.44	.83	.33	.49	.53	.44	.10	.02
Contributions																		
1			.58	.	.	.	.	*.01*	.	.	.	.	*.41*	.	.	.	.	.
2			*.09*	.	.	.	.	.65	.	.	.	.	*.26*	.	.	.	.	.

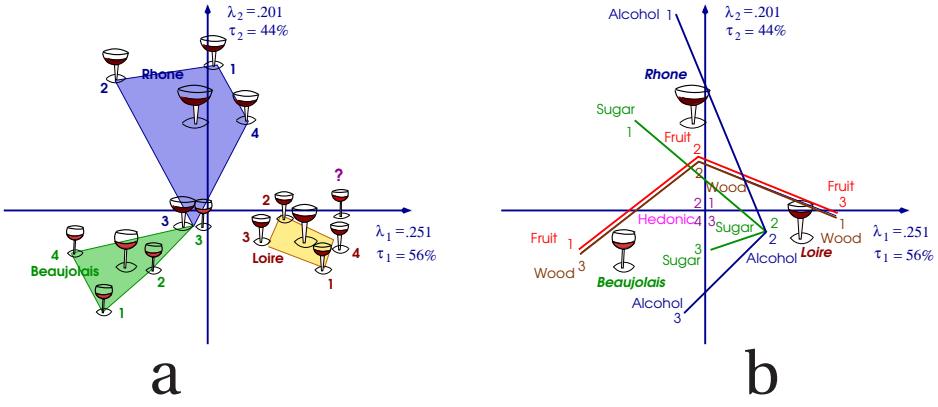

Figure 1: Discriminant Correspondence Analysis. Projections on the first 2 dimensions. (a) The I set: rows (*i.e.,* wines), The wines are projected as supplementary elements, Wine ? is an unknown wine. (b) The J set: columns (*i.e.,* descriptors). The wines categories have also been projected for ease of interpretation. Both figures have the same scale (some projection points have been slightly moved to increase readability). (Projections from Tables 2 and 3).

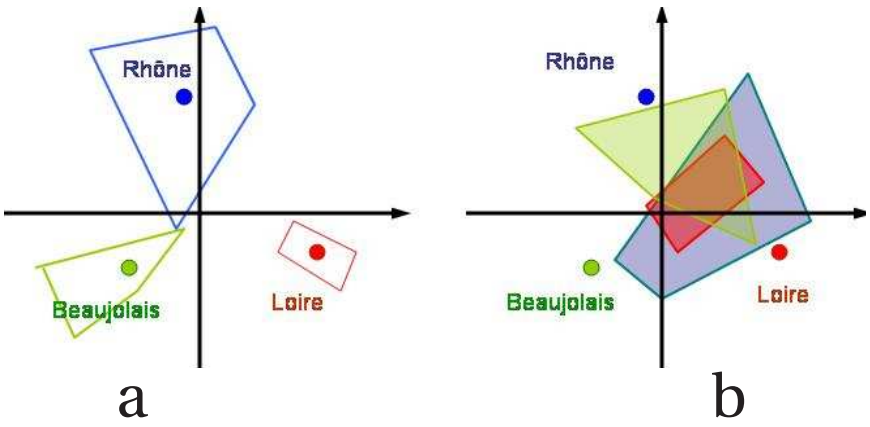

Figure 2: Discriminant Correspondence Analysis. Projections on the first 2 dimensions. (a) Fixed effect model. The three regions and the convex envelop for the wines. (b) Random effect model. The jackknifed wines have been projected back onto the fixed effect solution. The convex envelop shows that the random effect categories have a larger variability and have moved.

4 Results

Tables 2 and 3 give the results of the analysis and Figure 1 displays them. The fixed effect quality of the model is evaluated by the following confusion matrix:

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 1 & 4 \end{bmatrix}. \tag{12}$$

In this matrix, the rows are the assigned groups and the columns are the real groups. For example, out of 5 wines assigned to the wine region Beaujolais (Group 3), one wine was in fact from the Rhône region (Group 2) and 4 wines were from Beaujolais. The overall quality can be computed from the diagonal of the matrix. Here we find that 11 $(4+3+4)$ wines out of 12 were correctly classified.

A jackknife procedure was used in order to evaluate the generalization capacity of the analysis to new wines (*i.e.,* this corresponds to a random effect analysis). Each wine was in turn taken out of the sample, a DCA was performed on the remaining sample of 11 wines, and the wine taken out was assigned to the closest group. This gave the following confusion matrix:

$$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}. \tag{13}$$

As expected, the performance of the model as a random effect is less impressive than as a fixed effect model. Now only 6 $(2+2+2)$ wines out of 12 are correctly classified.

The differences between the fixed and the random effect models are illustrated in Figure 2 where the jackknifed wines have been projected onto the fixed effect solution (using metric multidimensional scaling, see entry). The quality of the model can be evaluated by drawing the convex envelop of each category. For the fixed effect model, the centers of gravity of the convex envelops are the categories and this illustrates that DCA is a least square estimation technique. For the random effect model, the degradation

of performance is due to a larger variance (the areas of the convex envelops are larger) and to a rotation of the envelops (the convex envelops are no longer centered on the category centers of gravity).

References

[1] Abdi, H. (2003). Multivariate analysis. In M. Lewis-Beck, A. Bryman, & T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks: Sage.

[2] Clausen, S.E. (1998). *Applied correspondence analysis*. Thousand Oaks (CA): Sage.

[3] Greenacre, M.J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

[4] Greenacre, M.J. (1993). *Correspondence analysis in practice*. London: Academic Press.

[5] Weller S.C., & Romney, A.K. (1990). *Metric scaling: Correspondence analysis*. Thousand Oaks (CA): Sage.