BARYCENTRIC DISCRIMINANT ANALYSIS

Hervé Abdi[*] *& Lynne Williams*

Barycentric discriminant analysis (BADA) generalizes discriminant analysis, and like discriminant analysis, it is performed when measurements made on some observations are combined to assign these observations, or *new* observations, to a priori defined categories. For example, BADA can be used (a) to assign people to a given diagnostic group (e.g., patients with Alzheimer's disease, patients with other dementia, or people aging without dementia) on the basis of brain imaging data or psychological tests (here the a priori categories are the clinical groups), (b) to assign wines to a region of production on the basis of several physical and chemical measurements (here the a priori categories are the regions of production), (c) to use brain scans taken on a given participant to determine what type of object (e.g., a face, a cat, a chair) was watched by the participant when the scans were taken (here the a priori categories are the types of object), or (d) to use DNA measurements to predict whether a person is at risk for a given health problem (here the a priori categories are the types of health problem).

BADA is more general than standard discriminant analysis because it can be used in cases for which discriminant analysis cannot be used. This is the case, for example, when there are more variables than observations, when the predictors are colinear, or when the measurements are categorical.

BADA is a class of methods that all rely on the same principle: Each category of interest is represented by the *barycenter* of its observations (i.e., the weighted average of the observations of a given category; the barycenter is also called the *center of gravity* or *center of mass*), and a generalized principal components analysis (GPCA) is performed on the category by variable matrix. This analysis gives a set of discriminant factor scores for the categories and another set of factor scores for the variables. The original observations are then projected onto the category

[*] **In Bruce Frey (Ed.), *The SAGE Encyclopedia of Research Design.* Thousand Oaks, CA: Sage. 2022**

Address Correspondence to Hervé Abdi,
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75080—3021 USA
E-mail: `herve@utdallas.edu` `https://personal.utdallas.edu/~herve.`

factor space, providing a set of factor scores for the observations. The distance of each observation to the set of categories is computed from the factor scores, and each observation is assigned to the closest category. The *a priori* and *a posteriori* category assignments are compared to assess the quality of the discriminant procedure. The prediction for the observations that were used to compute the barycenters is called the *fixed-effect* prediction. The fixed-effect performance is evaluated by counting the number of correct and incorrect assignments and storing these numbers in a confusion matrix. Another index of the performance of the fixed-effect model—equivalent to a squared coefficient of correlation—is the ratio of the category variance to the sum of the category variance and the variance of the observations within each category. This coefficient is denoted R^2 and is interpreted as the proportion of variance of the observations explained by the categories or as the proportion of the variance explained by the discriminant model. The performance of the fixed-effect model can also be represented graphically as a *toleranc*e ellipsoid that encompasses a given proportion (say 95%) of the observations. The overlap between the tolerance ellipsoids of two categories is proportional to the number of misclassifications between these two categories.

New observations can also be projected onto the discriminant factor space, and they can be assigned to the closest category. When the actual assignment of these observations is not known, the model can be used to *predict* category membership. The model is then called a *random* model (as opposed to the fixed model). An obvious problem, then, is to evaluate the quality of the prediction for new observations. Ideally, the performance of the random-effect model is evaluated by counting the number of correct and incorrect classifications for new observations and computing a confusion matrix based on these new observations. However, it is not always practical or even feasible to obtain new observations, and therefore the random-effect performance is often evaluated using computational cross-validation techniques such as the *Leave One Out* (LOO) or the *Bootstrap.* For example, a leave one out approach can be used by which each observation is taken out of the set, in turn, and predicted from the model built on all the other observations. The predicted observations are then projected in the space of the fixed-effect discriminant scores. This can also be represented graphically as a *prediction* ellipsoid. A prediction ellipsoid encompasses a given proportion (say 95%) of the new observations. The overlap between the prediction ellipsoids of two categories is proportional to the number of misclassifications of new observations between these two categories.

The stability of the discriminant model can be assessed by a cross-validation model such as the Bootstrap. In this procedure, multiple sets of observations are generated by sampling with replacement from the original set of observations, and the category barycenters are computed from each of these sets. These barycenters are then projected onto the discriminant factor scores. The variability of the barycenters can be represented graphically as a *confidenc*e ellipsoid that encompasses a given proportion (say 95%) of the barycenters. When the confidence intervals of two categories do not overlap, these two categories are *significantly* different.

In summary, BADA is a GPCA performed on the category barycenters. GPCA encompasses various techniques, such as correspondence analysis, biplot, Hellinger distance analysis, discriminant analysis, and canonical variate analysis. For each specific type of GPCA, there is a corresponding version of BADA. For example, when the GPCA is correspondence analysis, this gives the most well-known version of BADA: discriminant correspondence analysis (DICA). Because BADA is based on GPCA, it can also analyze data tables obtained by the concatenation of blocks (i.e., subtables). In this case, the importance (often called the *contribution*) of each block to the overall discrimination can also be evaluated and represented as a graph.

Hervé Abdi and Lynne J. Williams

See also Bootstrapping; Canonical Correlation Analysis; Correspondence Analysis; Discriminant Analysis; Jackknife; Matrix Algebra; Principal Components Analysis

Further Readings

Abdi, H. (2007). Discriminant correspondence analysis. In N.J. Salkind (Ed.): *Encyclopedia of Measurement and Statistics.* Thousand Oaks (CA): Sage. pp. 270–275.

Abdi, H., Williams, L.J., Beaton, D., Posamentier, M., Harris, T.S., Krishnan, A., & Devous, M.D. (2012). Analysis of regional cerebral blood flow data to discriminate among Alzheimer's disease, fronto-temporal dementia, and elderly controls: A multi-block barycentric discriminant analysis (MUBADA) methodology. *Journal of Alzheimer Disease*, **31**, s189–s201.

Beaton, D., Dunlop, J., ADNI, & Abdi, H. (2016). Partial Least Squares-Correspondence Analysis: A framework to simultaneously analyze behavioral and genetic data. *Psychological Methods*, **21**, 621–651.

Abdi H., Williams, L., & Béra, M. (2018). Barycentric discriminant analysis (BADA). In R. Alhajj and J. Rokne (Eds.), *Encyclopedia of Social Networks and Mining (2nd Edition).* New York: Springer Verlag.