Electronic Journal of Applied Statistical Analysis
EJASA (2012), Electron. J. App. Stat. Anal., Vol. 5, Issue 3, 438 – 444
e-ISSN 2070-5948, DOI 10.1285/i20705948v5n3p438
© 2012 Università del Salento – http://siba-ese.unile.it/index.php/ejasa/index

DISSIMILARITY PROFILE ANALYSIS: A CASE STUDY FROM ITALIAN UNIVERSITIES

Nadia Solaro[*]

Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Italy

Received 02 August 2012; Accepted 12 November 2012
Available online 16 November 2012

Abstract: *Dissimilarity profile analysis (DPA) is introduced as an explorative tool for object-oriented data analysis techniques that address the problem of latent dimension extraction by using proximity measures. Potentialities of DPA are shown within a case study from Italian universities, where undergraduate courses are examined with respect to students' enrolment, career, and degree attainment.*

Keywords: *Dimensionality reduction, SMACOF multidimensional scaling*

1. Introduction

Proximities, namely nonnegative measures of pairwise similarity/dissimilarity between objects (subjects, units, items, stimuli, etc.), are often used in many areas of research, whenever comparisons among units in analysis are of primary concern [1]. They are input data of object-oriented multivariate analysis (MVA) techniques, which address the problem of data dimensionality reduction by taking into account closeness of objects in a multidimensional space. Although proximities are the basic structure of object-oriented-type analyses, they rarely are pre-processed for explorative purposes. On the other hand, discovering diversity patterns and potential data abnormalities could be of considerable usefulness. The literature seems lacking in these kinds of contributions. Nonetheless, performing explorative analyses on a proximity matrix could help interpret results derived from the application of an object-oriented MVA technique, as well as disclose anomalies that might undermine statistical analyses.

The analysis tool developed here, called Dissimilarity Profile Analysis (DPA), takes the main idea from Profile Analysis for quantitative data matrices [2]. DPA turns out to be effective in the overall analysis of pairwise comparisons among objects, summing up all informative content of

[*] E-mail: nadia.solaro@unimib.it

variables. Moreover, DPA does not require the knowledge of a starting data matrix so that proximity matrices can be explored directly, without recovering object coordinates. Potentialities of DPA are illustrated within a case study on undergraduate courses of Italian universities, which are compared with respect to students' enrolment, career, and degree attainment.

2. Dissimilarity profile analysis

Profile Analysis (PA) is a technique that can be applied to different data structures of MVA. Usually, it represents the basic approach for analyzing contingency tables, where the word profile stands for row or column conditional frequencies of given categories. When PA is applied to multidimensional data matrices $\mathbf{Y}$ ($n \times p$), with p variables (columns) observed on n objects (rows), the term profile simply denotes the n row-vectors ($1 \times p$) of observations y_{ij}, ($i = 1,\ldots,n$; $j = 1,\ldots,p$). If, in addition, variables are all quantitative with the same scale of measurement, or are standardized, profiles admit a geometric representation called profile plot, where the p values of each profile are plotted against the labels of variables taken in arbitrary order. Profile plot reveals itself as a powerful visualization tool for comparisons, because similarities or differences among profiles are immediately apparent over variables. Moreover, the nature of differences between two profiles can be delved into by comparing them in terms of so-called overall level, scatter, and shape, which constitute the three key components of PA [2, chap. 10].

Dissimilarity Profile Analysis (DPA) is designed as a PA for dissimilarities, which are proximity measures expressing the degree of diversity among pairs of objects [1], [2]. Let $\boldsymbol{\Delta}$ be a ($n \times n$) symmetric dissimilarity matrix with dissimilarities δ_{ir} as elements ($i,r = 1,\ldots,n$). Diagonal δ_{ii}s, being self-dissimilarities, are all equal to zero. With the aim of extending PA to dissimilarities, we define as Dissimilarity Profiles (DPs) the n row-vectors ($1 \times n$) of dissimilarities δ_{ir} forming matrix $\boldsymbol{\Delta}$. Although DPs admit a geometric representation fairly similar to PA, there are several elements of distinction. First, DPs are plotted against the labels of objects, rather than variables. Second, given that self-dissimilarities are comprised in DPs, in a DP plot every trajectory falls down to zero in correspondence to the object on the x-axis to which it refers.

The three components: level, scatter, and shape, can also be defined for DPA. Given two generic DPs *i* and *r*, we have: (1) level of *i*-th DP: $\bar{\delta}_{i.} = \frac{1}{n}\sum_{l=1}^{n}\delta_{il}$, which is the average distance of object *i* with respect to the others; (2) scatter of *i*-th DP: $v_{i.}^2 = \sum_{l=1}^{n}(\delta_{il} - \bar{\delta}_{i.})^2$, expressing the variation of DP *i* around its level; (3) shape of *i*-th and *r*-th DPs: $q_{ir} = v_{ir} / v_{i.}v_{r.}$, where: $v_{ir} = \sum_{l=1}^{n}(\delta_{il} - \bar{\delta}_{i.})(\delta_{rl} - \bar{\delta}_{r.})$, which indicates whether DPs *i* and *r* share an analogous diversity pattern, i.e. whether they differ from the other units in a similar fashion. As in PA [2], level, scatter, and shape can be proved to be strictly related if distances among pairs of DPs are measured by the square of Euclidean distance: $d_{ir}^2 = \sum_{l=1}^{n}(\delta_{il} - \delta_{rl})^2 = \sum_{\substack{l=1 \\ l \neq i \neq r}}^{n}(\delta_{il} - \delta_{rl})^2 + 2\delta_{ir}^2$. In such a case, the following relation holds:

$$d_{ir}^2 = n(\bar{\delta}_{i.} - \bar{\delta}_{r.})^2 + (v_{i.} - v_{r.})^2 + 2v_{i.}v_{r.}(1 - q_{ir}), \qquad (i,r = 1,\ldots,n). \qquad (1)$$

Relative contributions of each component can be computed by simply dividing each additive term in formula (1) by the square of Euclidean distance. In such a way, it is possible to assess the relative importance of level, scatter, and shape in explaining the observed diversity patterns.

3. DPA of groups of Italian university courses

Potentialities of DPA as an explorative tool are illustrated within a case study concerning Italian universities. Data are drawn from the database of the University Education Survey, coordinated every year by the Office of Statistics of the Italian Ministry of University and Research (MIUR) [3]. By law, Italian universities are required to supply to MIUR all information pertaining to undergraduate courses (*corsi di laurea triennali*), and postgraduate courses (*corsi di laurea specialistici/magistrali*), in terms of enrolled students, students' career, passed exams, dropouts, and the attainment of a university qualification (3-year bachelor's degree, 2-year master's degree, respectively). Specialisation schools, first- and second-level vocational master courses, and doctorates are also involved in the survey.

The study here proposed focuses specifically on the undergraduate courses (UCs) of 3-year legal duration, introduced in Italy with the "3+2" reform (DM n. 509 3/11/1999 and DM n. 270 22/10/2004). All the relevant information about the number of enrolled and first-year students, *fuori corso* (beyond prescribed time, BPT) students (i.e. students not completing studies within the legal duration), composition by gender, achieved University Formative Credits (UFCs – *crediti formativi*), transfers and dropouts, number of graduates (i.e., number of students attaining a bachelor's degree), BPT graduates, and graduates with the highest marks (101 – 110 with honours), are drawn from the last available data collection, i.e. the academic year 2010/2011. Then, with the purpose of comparing the different types of formative careers, data have been aggregated over all Italian (public and private) universities and referred to the fifteen disciplinary areas according to which UCs are classified by MIUR [4]: Agricultural, Architecture, Chemical-Pharmaceutical, Economic-Statistical, Physical Education, Geo-biological, Legal, Engineering, Educational, Literary and Arts, Linguistic, Medicine, Political-Social, Psychological, and Scientific. Regarding variables, the set of university indicators reported in detail in Table 1 is computed for each disciplinary group. Next, the indicators have been standardized, and analyses carried out in three main steps: (1) PA is applied to standardized indicators to detect the main differences among profiles of UC groups over the considered indicators; (2) DPA is then carried out to highlight diversity patterns over all UC groups. Dissimilarities between UC groups are measured with Euclidean distance; (3) the MDS method known as SMACOF [5] is finally applied to synthesise the original indicators with a small number of unobservable dimensions. All routines for DPA have been implemented, and analyses carried out with the R software [6].

Results of PA are displayed in Figure 1. By observing the peaks of the profile plot, several remarks are worth making. The Legal group of UCs has the highest standardized scores on percentages of BPT students (p.BPT), students not attaining UFCs (p.stud.no.ufc), and BPT graduates (p.grad.BPT). Conversely, the Medicine group has the lowest scores on these indicators, as well as the lowest dropout rate (drop.rate), and the highest percentage of graduates (p.grad). The Geo-biological, Chemical-Pharmaceutical, Legal, Agricultural, and Scientific groups are characterized by the highest levels of dropout, whereas the Literary and Arts group has the highest percentage of graduates with the highest marks (p.grad.h.m).

Table 1. Description of university indicators.

Label	Description
en.stud	total number of students enrolled at undergraduate courses in 2010/11
p.Fstud	percentage of female students (over enrolled students)
p.BPT	percentage of students enrolled beyond prescribed time (*fuori corso*) (over enrolled students)
p.1st.y	percentage of first-year students (over enrolled students)
p.1st.y.h.m	percentage of first-year students with the highest marks (90−100) at the exit from secondary school (over first-year students)
p.transf	percentage of students' transfers from one university to another (over enrolled students)
p.1st.y.lic	percentage of first-year students coming from *liceo* (senior secondary school)
drop.rate	dropout rate, i.e. percentage of first-year students in 2009/10 not enrolled at the second year in 2010/11 (over first-year students in 2009/10)
p.stud.no.ufc	percentage of students enrolled in 2009/10 not achieving UFCs in 2010
p.grad	percentage of students attaining a bachelor's degree in 2010 (over students enrolled in 2006/07)
p.grad.BPT	percentage of students graduated beyond prescribed time (*fuori corso*)
p.grad.h.m	percentage of graduates with the highest marks (101−110 with honours)

As regards first-year students, the Geo-biological, Literary and Arts, Chemical-Pharmaceutical, Engineering, Linguistic, and Scientific groups have the highest percentages of students coming from *liceo* (p.1st.y.lic.), i.e. secondary schools specialized in humanities, modern languages, or science, while the Engineering and Scientific groups are characterized by the highest percentages of students with the highest marks upon exiting from secondary school (p.1st.y.h.m). Finally, the Political-Social group tends to have an intermediate position.

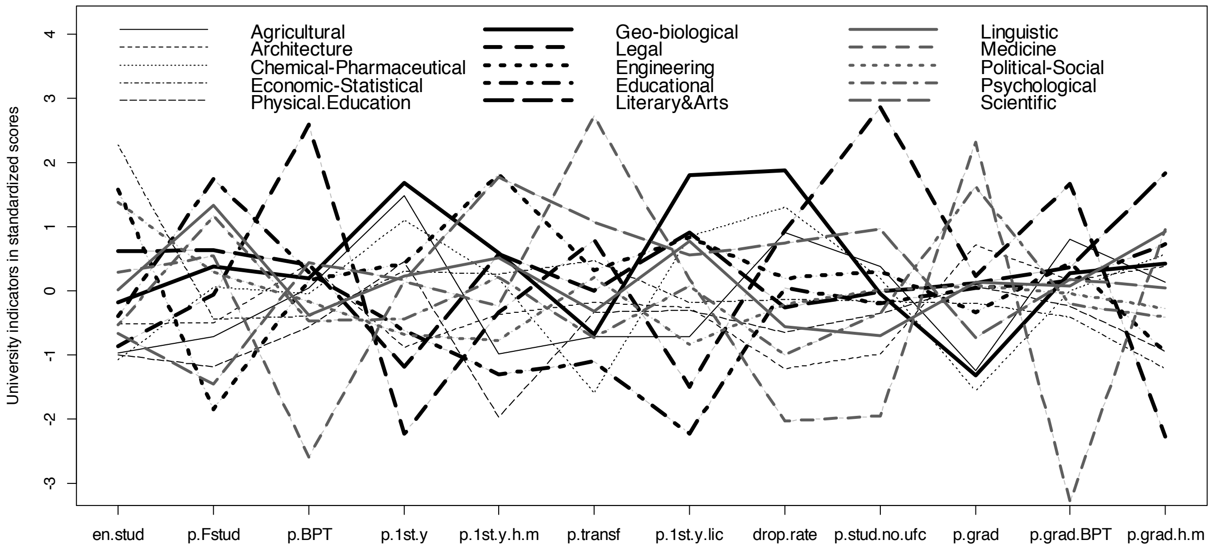

Figure 1. Profile plot of disciplinary groups of Italian undergraduate courses – Academic Year 2010/2011.

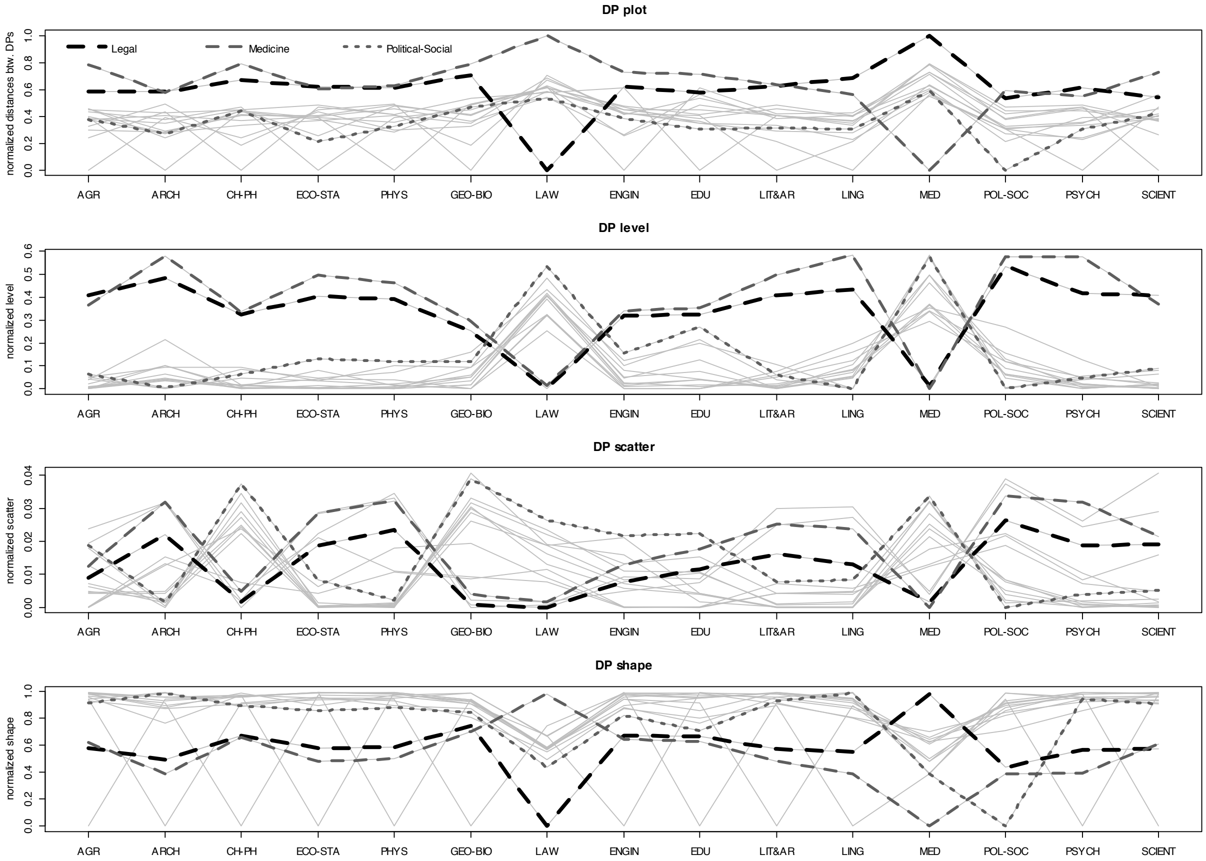

Figure 2. Dissimilarity profile plot of (normalized) Euclidean distances between disciplinary groups of undergraduate courses (1st panel). Decomposition in level (2nd panel), scatter (3rd panel), and shape (4th panel).

DPA is applied with the objective of detecting UC groups with similar diversity profiles, i.e. that are likewise different (or distant) from other groups of UCs. For each UC group, the DP is formed by Euclidean distances computed with respect to each other group. Results of DPA are reported in Figure 2, where the Legal (LAW), Medicine (MED), and Political-Social (POL-SOC) groups are picked out for illustrative purposes. DP plot (1st panel) shows that LAW and MED are equally distant from the other groups (their trajectories overlap), but they are very different from each other (their trajectories at LAW and MED tickmarks drastically separate). This suggests that LAW and MED could be represented as opposite, extreme points in a multidimensional space. Subsequent analyses carried out with respect to the three components (2nd to 4th panel) highlight that in the comparison with each other group, LAW and MED share very similar distance level, scatter, and shape patterns (trajectories tend to overlap), and that differences in level are the main factor explaining diversity. In turn, LAW and MED differ between them exclusively for the shape (their trajectories at LAW and MED tickmarks coincide for level and scatter while separate for shape). This strengthens the interpretation that LAW and MED are UC groups in an opposite condition, though very far from the other groups. Conversely, POL-SOC group has a fairly low trajectory in the DP plot, suggesting that it is similar to most UC groups. Moreover, at LAW and MED tickmarks, POL-SOC is located in an intermediate position, thus adducing evidence that POL-SOC could be a central point in a multidimensional space. The main factor distinguishing POL-SOC from the other UC groups is the shape (the trajectory in the

DP shape plot is almost always at a high level) while differences in level are quite negligible, with the exception of LAW and MED, for which the DP level plot shows the highest peaks.

Finally, SMACOF MDS [5] is applied in order to extract unobservable variables as syntheses of the original university indicators and verify if the above given remarks are confirmed by a dimensionality reduction analysis. Following an explorative study, a configuration in three dimensions has been judged as a satisfactory fit with input data (i.e. Euclidean distances between UC groups). Goodness-of-fit statistics "normalized metric stress" σ_N (the smaller, the better) and dispersion accounted for (DAF) are equal, respectively, to σ_N = 0.0146 and DAF = 1 – σ_N = 0.9852. DAF value, in particular, indicates that, on the whole, 98.52% of observed Euclidean distances are reproduced by the three-dimensional configuration while the first two dimensions alone account for 92.52%.

Figure 3 reports the main achieved results. The correlation circle in the left-hand panel proves that dimension 1 can be interpreted as an indicator of university students' failures, given that it is highly positively correlated with dropouts, percentages of students not achieving UFCs, BPT students, and BPT graduates, and highly negatively correlated with percentage of graduates. Dimension 2 represents attraction capacity and excellence of UC groups, given the high positive correlations with percentages of first-year students coming from *liceo* and/or with the highest marks at the exit from secondary school and of graduates with the highest marks. Dimension 3 (here omitted) represents size of UC groups (enrolled students and transfers) and composition by gender. The right-hand panel of Figure 3 displays the scatterplot of the first two dimensions, a map of UC groups with coordinates given by the scores of the two compound indicators "students' failures" and "attraction capacity and excellence". As can be clearly seen, the map broadly supports interpretations obtained from PA and DPA. In particular, LAW and MED are therein represented as extreme points, almost opposite (although they have a similar score on dimension 3), while POL-SOC is in the central part of the map (this holds for dimension 3, also).

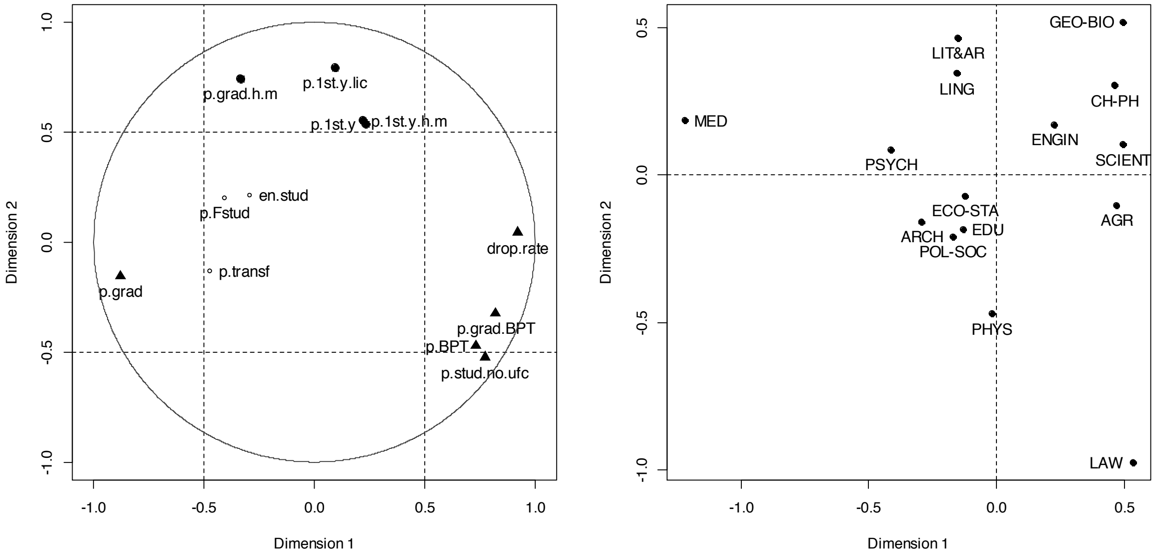

Figure 3. SMACOF MDS configuration of points. Left-hand panel: correlation circle of the first two dimensions. Right-hand panel: two-dimensional map of UC groups.

In conclusion, DPA is designed as an analysis tool for exploring proximity matrices. It aims at detecting the main characteristics of diversity patterns within a set of objects. In the study here proposed, input proximities are Euclidean distances, therefore, measured at ratio-scale level. In principle, however, DPA could be extended to proximities measured at any level of measurement.

References

[1]. Everitt, B.S., Rabe-Hesketh, S. (1997). *The Analysis of Proximity Data*. Kendall's Library of Statistics, 4. London: Arnold.

[2]. Jobson, J.D. (1992). *Applied Multivariate Data Analysis. Volume II: Categorical and Multivariate Methods*. New York: Springer-Verlag.

[3]. Ministero dell'Università e della Ricerca - Ufficio di Statistica. Indagine sull'Istruzione Universitaria (2012). http://statistica.miur.it/normal.aspx?link=datiuniv

[4]. Ministero dell'Università e della Ricerca (2011). *L'Università in cifre 2009/2010*. Appendice. http://statistica.miur.it/normal.aspx?link=pubblicazioni

[5]. de Leeuw, J., Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31, 3, 1–30.

[6]. R Development Core Team (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna. http://www.R-project.org