Statistica Applicata – ITALIAN JOURNAL OF APPLIED STATISTICS

EDITORIAL TEAM

A.S.A CONTACTS
Principal Contact
Francesco Palumbo (Editor in Chief)
editor@sa-ijas.org

Support Contact
Domenico Vistocco (Editorial Manager)
ijas@sa-ijas.org

JOURNAL WEBPAGE
https://www.sa-ijas.org/ojs/index.php/sa-ijas

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising:Cleup sc, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: info@cleup.it.
Rules for manuscript submission: https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions
Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.

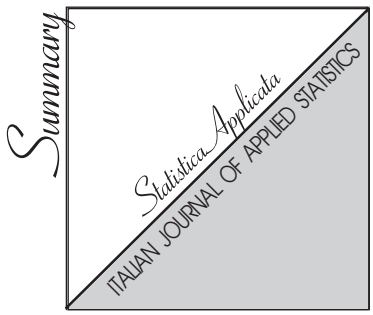
Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (http/web.it.nctu.ew/~qtqm/)

SINERGIE – Italian Journal of Management

Pubblished: November 2023

Cleup sc
'Coop. Libraria Editrice Università di Padova'
via G. Belzoni, 118/3 – Padova Italy
Phone +39 049 8753496 Fax +39 049 9865390
info@cleup.it – www.cleup.it – www.facebook.com/cleup



Vol. 35, Number 3

doi.org/10.26398/IJAS.0035-012

SPECIAL ISSUE IN MEMORY OF SIMONA BALBI - EDITORIAL

Jörg Blasius
Department of Political Science and Sociology, University of Bonn, Bonn, Germany

Luigi Fabbris
Tolomeo Studi e Ricerche, Padua and Treviso, Italy

Michael Greenacre
Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

Germana Scepi, Maria Spano
Department of Economics and Statistics, Federico II University of Naples, Naples, Italy

This special issue is dedicated to Simona Balbi, full professor of Statistics at the University Federico II of Naples, who prematurely died in February 2018.

The issue takes its cue from the scientific meeting in memory of Simona Balbi, "Statistics and Data Science", organized in February 2019 by the Department of "Scienze Economiche e Statistiche" at the Federico II University of Naples. This event aimed to bring together people who have known Simona Balbi and shared part of their scientific path with her but also, often, a bond of friendship and affection. The morning was dedicated to speeches by some foreign colleagues (Michael Greenacre, Jörg Blasius, Ludovic Lebart, Mirelle Summa, Gilbert Saporta, Mónica Bécue). The afternoon included a session of personal testimonies, followed by a session of scientific contributions by Italian colleagues who, with Simona Balbi, have developed part of their research paths (Natale Carlo Lauro, Luigi Fabbris, Corrado Crocetta, Furio Camillo, Rosanna Verde, Francesco Palumbo, Michelangelo Misuraca, Roberta Siciliano,

Vincenzo Esposito Vinzi, Giuseppe Giordano, Gabriella Grassia, Salvatore Ingrassia, Germana Scepi, and others).

Contributions regarding three main fields of Simona's scientific research were presented at the conference: *a) the social area, and in particular the education system; b) correspondence analysis and related methods;* and, finally*, c) textual data analysis.*

Simona was aware that statistical methodology is a tool to solve real problems, so any problem unsolved by statistics required a methodological improvement. From 2002 to 2012, she participated in various research projects in which the effectiveness of the higher education system was measured through indicators involving the matching between the education received by university graduates and the expectations of the labour market in Italy. In these projects, financed by the Italian Ministry of Education, Research and University, she and the research group at Naples University studied how to fuse databases at various levels of granularity (graduates, courses, universities, local labor markets) for making comparisons feasible.

Moreover, she organized three scientific meetings on higher education effectiveness at the Federico II University of Naples, in which she presented various papers, jointly with colleagues, and was a guest editor of two books (in Italian) on: "Jobs and graduates' competencies" and "Quantitative representation of an effective educational process that may improve best practices". She had the capacity to softly involve colleagues and young scholars in this field of applied research, being able to fit the content purposes with the statistical methods' potentiality.

The first paper of this special issue, "*Counting the Poor in Italy and EU*" by Luigi Fabbris can be placed in the first area of Simona's scientific research. In this paper, the author examines some approaches to poverty and discusses the properties of poverty measures. The paper suggests an interesting correspondence between poverty measurement approaches and intervention purposes. It is highlighted that the choice of a suitable approach to poverty is based on the pertinence of the properties of the statistical measures with respect to the policies and actions to overcome poverty.

Simona wrote many papers and organized a lot of conferences on the second thematic area: correspondence analysis and related methods. The editors remember the World Congress of Sociology in Bielefeld, 18-23 July 1994, and

the second of the conferences on correspondence analysis, called Visualization of Categorical Data, which took place in Cologne, May 17-19, 1995. Simona attended all the conferences which came to be called CARME (Correspondence Analysis and Related Methods). She did not only participate, but she also organized the seventh CARME conference, with Jörg Blasius and Michael Greenacre, which took place in Naples from September 20-23, 2015 (http://www.carme-n.org/carme2015). There is a video of the meeting on the YouTube channel, https://www.youtube.com/watch?v=WxQd1eA4fnw

Simona also organized the RC33 conference (Research Committee on Logic and Methodology of the International Sociological Association). This meeting takes place every four years and has about 500 participants. The Naples meeting was held September 1-5, 2008, and was one of the largest ever, with 88 sessions and over 500 papers. As in the previous two and subsequent conferences, the CARME theme was strongly represented with several sessions, including Automatic Textual Analysis (organized by Simona Balbi), Biplots (John Gower), Correspondence Analysis and Related Methods (Jörg Blasius), Geometric Data Analysis (Brigitte Le Roux), and Multidimensional Scaling (Mark de Rooij), to mention just a few.

The paper titled "*From Plain to Sparse Correspondence Analysis: A Generalized SVD Approach"* by Hervé Abdi, Vincent Guillemot, Ruiping Liu, Ndèye Niang, Gilbert Saporta, Ju-Chi Yu can be placed in the Simona's second research area. The authors propose an extension of the sparse correspondence analysis method developed by Liu et al. (2023) by adding a new global algorithm. This algorithm allows us the simultaneous optimization of the dimensionality of the sparsified space and the sparsification parameters of the rows and columns of the data. The paper discusses the properties of the new version of sparse CA estimates. The method is applied to interpret the relationships in a big textual data set—obtained from the Project Gutenberg (Gerlach and Font-Clos, 2020) — compiling common words used in 100 books each from five book categories: Biographies, Love stories, Mystery, Philosophy and Science Fiction. The results show that sparse correspondence analysis simplifies the interpretation of large tables by highlighting important categories and obtaining simple successive dimensions in the spirit of the simple structure of factor analysis.

Finally, the research topic that accompanied Simona throughout her academic career and that probably fascinated her the most was the statistical analysis of textual data.

Simona was a pioneer in this field, producing various scientific contributions focusing on both methodological and applicative aspects of different text mining tasks. Her early work focused on the use of factorial methods, from studying the stability of configurations obtained with non-symmetrical correspondence analysis to the use of Procrustes techniques for the analysis of multilingual corpora. In some of her work, she addressed topics specific to natural language processing, such as word sense disambiguation, proposing the joint use of correspondence analysis and network analysis tools, and the analysis of complex lexical structures through symbolic data analysis methods.

The application areas where she tested her methodological proposals were the most diverse, from classic open-ended survey responses to newspaper articles, annual reports, job postings, and finally to her latest work where her interest shifted to the growing need to analyze data from social media.

Since 2002, she was a member of the permanent scientific committee of JADT, *Journées internationales d'Analyse Statistique des Données Textuelles*, the biennial conference which has constantly gained importance since its first occurrence in 1992, open to all scholars and researchers working in the field of textual data analysis, including natural language processing and lexicography, text mining, information science, computational linguistics, sociolinguistics, analysis of political discourse and content analysis.

In July 2022, the conference was held in Naples, on the proposal of her colleagues Massimo Aria, Giuseppe Giordano, Michelangelo Misuraca, Germana Scepi, and Maria Spano, with the support of the Vadistat association founded by Simona herself in 2008 to spread statistical culture in the scientific, educational, and institutional fields that today continues its activity with the name "Vadistat for Simona Balbi".

Two articles on this issue can be placed in the textual data analysis area: the first paper is titled "Multilingual textual data: an approach through multiple factor analysis" by Belchin Kostov, Ramon Alvarez-Esteban, Mónica Bécue-

Bertaut and Francois Husson. The authors propose a new approach, the multiple factor analysis for generalised aggregate lexical tables (MFA-GALT), for analysing open-ended questions answered in different languages, in the case of multilingual surveys. MFA-GALT is performed in two steps, exactly like a classic MFA. First, each sub-table is analyzed separately while, in the second step, a global factorial analysis is performed on all sets of multiple tables. The properties and the graphical representations of this new approach are also shown with the application to data collected by a railway company on satisfaction of passengers regarding its night trains.

The second paper in this issue deals with textual data is titled "*Visualization of Textual Data: A Complement to Authorship attribution*" by Ludovic Lebart. In textual data analysis, authorship attribution is precisely a leading case of statistical decision. The author analyses a large corpus of 50 French novels of the 20th century, by comparing descriptive (or unsupervised) methods with confirmatory (or supervised) methods. It is shown that additive trees applied to the coordinates of a preliminary correspondence analysis can provide a useful strategy for describing, comparing, understanding this peculiar data and their relationships.

This issue is dedicated by the guest editors to Simona Balbi and, in the conclusion, we report some of the main papers of Simona in the three previously described research areas.

Simona was a dear friend for all of us and to all who met her, with her quiet manner and sense of humor that will always be remembered, not only her academic work and significant contribution to the field of statistics.

The Guest Editors
Jörg Blasius
Luigi Fabbris
Michael Greenacre
Germana Scepi
Maria Spano

MAIN REFERENCES OF SIMONA'S CONTRIBUTIONS

a) Social area:

Lauro, N., Balbi, S., Mola, F., Perna, A., and Scepi, G. (1991). Indagine sui laureati in Economia e Commercio di Napoli negli anni 1986-1989.

Balbi, S., Lauro, N. C., and Scepi, G. (1994). A multiway data analysis technique for comparing surveys. *Methodologica*, 3, 79-90.

Lauro, N., Balbi, S., and Scepi, G. (1994). The analysis of repeated surveys on Italian manufacturing enterprises: a multidimensional approach. In *Technique and Uses of Enterprise Panels-Proceedings of the First Eurostat International Workshop on Techniques of Enterprise Panels. Eurostat.*

Balbi, S., Balzano, S., and Bruzzese, D. (2002). A conceptual apporach to edit and imputation in repeated surveys. In *Compstat* 2002 (Vol. 400).

Balbi, S., and Grassia, M.G. (2003). Meccanismi di accesso al mercato del lavoro degli studenti di Economia a Napoli, Profiling dei laureati attraverso tre indagini ripetute. In *Transizione Università-Lavoro: la definizione delle competenze,* 97-110. Cleup.

Balbi, S., and Grassia, M.G. (2007). Profiling and labour market accessibility for the graduates in economics at Naples University. *Effectiveness of University Education in Italy: Employability, Competences, Human Capital*, 345-356.

Balbi, S., Grassia, M.G., Nappo, D., Tortora, C., and Triunfo, N. (2010). Come misurare l'efficacia di un Corso di Laurea. In *La rappresentazione quantitativa del processo universitario che genera efficacia e attiva il miglioramento*, 17-32. CLEUP.

Aria, M., Balbi, S., and Piscitelli, A. (2017). Profili ideali di laureati per lavorare nel turismo. Indagine sulle strutture alberghiere di Napoli. In *Scienza e coscienza a 1000 euro al mese. Neolaureati e mercato del lavoro.* Università degli Studi di Padova.

b) *Correspondence analysis and related methods:*

Balbi, S. (1992). On stability in non-symmetrical correspondence analysis using bootstrap. *Statistica Applicata*, 4(4), 543-552.

Lauro, N., Balbi, S., and Scepi, G. (1993). Multidimensional data analysis and experimental design. In *Contributed Papers of 49th ISI Session* (Vol. 1).

Lauro, N., Scepi, G., and Balbi, S. (1993). Empirical confidence regions for multidimensional control charts. In *Contributed Papers 49th ISI Session* (Vol. 2).

Balbi, S. (1994). Influence and stability in non symmetrical correspondence analysis. *Metron*, 52, 111-128.

Balbi, S. (1998). Graphical displays in nonsymmetrical correspondence analysis. In *Visualization of Categorical Data*, 297-309. Academic Press.

c) Textual data analysis:

Balbi, S. (1996). Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms. In *Proceedings of the JADT: 3es JADT, Journées Internationales d'Analyse Statistique des Données Textuelles, CISU, Roma*, *2*, 5-12.

Balbi, S. (1998). Textual data analysis for open-questions in repeated surveys. In *Advances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98) Università "La Sapienza", Rome,* 449-456. Springer Berlin Heidelberg.

Balbi, S., and Giordano, G. (2001). A factorial technique for analysing textual data with external information. In *Advances in Classification and Data Analysis,* 169-176. Springer Berlin Heidelberg.

Bolasco, S., Verde, R., and Balbi, S. (2002). Outils de text mining pour l'analyse de structures lexicales à éléments variables. In *Proceedings of the JADT: 6es Journes Internationales d'Analyse Statistique des Données Textuelles, St. Malo,* 197-208.

Balbi, S., Bolasco, S., and Verde, R. (2002). Text mining on elementary forms in complex lexical structures. In *Proceedings of the JADT*: *6es Journées Internationales d'Analyse Statistique des Données Textuelles, St. Malo*, 89-100.

Balbi, S., and Di Meglio, E. (2004). A text mining strategy based on local contexts of words. In *Proceedings of the JADT: 7es Journées Internationales d'Analyse Statistique des Données Textuelles, Louvain,* 4, 79-87.

Balbi, S., and Di Meglio, E. (2004). Una strategia di text mining basata su regole di associazione. In *Applicazioni di analisi statistica dei dati testuali*, 29-40. Casa Editrice Università La Sapienza.

Balbi, S., and Di Meglio, E. (2004). Contributions of textual data analysis to text retrieval. In *Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago*, 511-520. Berlin, Heidelberg: Springer Berlin Heidelberg.

Balbi, S., and Misuraca, M. (2005). Pesi e metriche nell'analisi dei dati testuali. *Quaderni di Statistica*, 7, 55-68.

Balbi, S., and Misuraca, M. (2006). Procrustes techniques for text mining. *In Data Analysis, Classification and the Forward Search: Proceedings of the Meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, University of Parma*, 227-234. Springer Berlin Heidelberg.

Balbi, S., and Misuraca, M. (2010). A doubly projected analysis for lexical tables. *Advances in Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks*, 13-19.

Balbi, S., Infante, G., and Misuraca, M. (2008). Conjoint analysis with textual external information. In *Proceedings of the JADT 2008: 9es Journées Internationales d'Analyse Statistique des Données Textuelles, Lyon,* 1, 129-136.

Balbi, S., Infante, G., and Misuraca, M. (2009). Il text mining per l'individuazione dell'offerta universitaria di competenze nel terzo settore. *formazione e lavoro*, 95-106.

Balbi, S., and Misuraca, M. (2010). A doubly projected analysis for lexical tables. Advances in *Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks*, 13-19.

Balbi, S. (2010). Beyond the curse of multidimensionality: High dimensional clustering in text mining. *Italian Journal of Applied Statistics*, 22(1), 53-63.

Balbi, S., Crocetta, C., Romano, M. F., Zaccarin, S., and Zavarrone, E. (2011). Competences and professional options of the Italian graduates: Results from the textual analysis of the degree course information data. In *Statistical Methods for the Evaluation of University Systems*, 195-207. Physica-Verlag HD.

Balbi, S., Stawinoga, A., and Triunfo, N. (2012). Text mining tools for extracting knowledge from firms annual reports. In *Proceedings of the*

JADT 2012: 11th International Conference on Statistical Analysis of Textual Data, Vol. 2012, p. 11es).

Balbi, S., and Triunfo, N. (2012). Statistical tools in the joint analysis of closed and open-ended questions. In *Survey Data Collection and Integration,* 61-72. Berlin, Heidelberg: Springer Berlin Heidelberg.

Balbi, S., and Stawinoga, A. (2013). Mining the ambiguity: Correspondence and network analysis for discovering word sense. In *Proceedings of the Conference of the Italian Statistical Society*.

Balbi, S., and Stawinoga, A. (2014). Textual data analysis tools for word sense disambiguation. In *Proceedings of the JADT: 12es Journées Internationales d'Analyse Statistique des Données Textuelles, Paris* (1), 57-66. INALCO Sorbonne Nouvelle.

Stawinoga, A., Balbi, S., and Scepi, G. (2016). Network tools for the analysis of brand image. *Italian Journal of Applied Statistics*, 26, 37-48.

Balbi, S., Misuraca, M., and Spano, M. (2016). A cosine-based validation measure for document clustering. In D. Mayaffre, C. Poudat, L. Vanni, V. Magri, P. Follette (eds.), *Statistical Analysis of Textual Data. Proceedings of 13th International Conference* (JADT16), Presses Fac Imprimeur, Nice, vol. 1, 65-74.

Balbi, S., Misuraca, M., and Scepi, G. (2017). A polarity-based strategy for ranking social media reviews. In *SIS 2017. Statistics and Data Science: New Challenges, New Generations. Proceedings of the Conference of the Italian Statistical Society*, 95-102. Firenze University Press.

doi.org/10.26398/IJAS.0035-013

COUNTING THE POOR IN ITALY AND THE EU

Luigi Fabbris[1]
Tolomeo Studi e Ricerche, Padua and Treviso, Italy

Abstract. *In this paper, we survey some popular methods for measuring poverty in a community. We include a method for detecting 'relative' versus 'absolute' poverty as well as 'extreme' poverty. We also consider alternative ways to measure poverty: material deprivation, at risk of poverty or social exclusion and variants of these methods. The analysis shows that each method has its own technical and logical properties that make it appropriate for use with specific informative targets, demonstrating that poverty indicators should not be used as if they all represent the same concept. Our analysis ends with suggestions concerning plausible relationships between measurement methods and the social purposes of the measures.*
Keywords*: Poverty measures; Relative poverty; Absolute poverty; Extreme poverty; Material deprivation; Social exclusion*

1. INTRODUCTION

Poverty is an intuitive concept that can be defined as lacking the financial resources to satisfy certain individual or community needs. However, when asked to pinpoint who in a community is poor, one may answer that a homeless person is certainly poor or that certain deprived villages in a faraway country are poor. Thus, everybody can identify situations of extreme poverty, but even an expert may find it puzzling to classify other less poor groups, which constitute the large majority of the poor.

An insufficient amount of money is the main reference for measuring poverty. However, poverty is also not having access to school and not knowing how to read. Poverty is not having a job; it is fear of what will be in the future, living one day at a time. Definitely, poverty is a complex and fuzzy concept that requires well-articulated reasoning.

[1] Luigi Fabbris, luigi.fabbris@unipd.it

In what follows, we specify the properties of a set of poverty measures with respect to the possible uses of the measure itself. Specifically, we define the technical and rational features that can make a poverty measure more appropriate than other competitive measures within a social-political scope.

The remainder of this paper is organised as follows. In Section 2, we present various approaches for measuring poverty. Only measures possessing properties that are relevant to social studies and official statistics are considered to conduct a systematic review of the data available on the subject matter. Section 3 links the inner properties of these measures to information that is relevant for social policy decisions. Section 4 highlights the social policy aims that each poverty measure could satisfy. Section 5 concludes the paper.

2. POVERTY MEASUREMENT APPROACHES

Relative poverty

Relative poverty is a general concept, because a person or family is considered poor when compared to a reference population: one is considered relatively poor if one has an income lower than a significant part of the population currently living in the same territory (Townsend, 1954). The reference to income makes the poverty concept flexible enough for people to choose if, when and how to achieve or integrate the goods and services they need but cannot access because of a lack of money.

In Italy, the 'relatively poor' are those whose income is below the poverty line, which is defined as 60% of the median income of the population in the area. Istat, the Italian Institute of Statistics, annually publishes two poverty thresholds—also called lines—and correspondingly, two poverty rates: one for individuals and another for families below the lines. With reference to families, a poverty line refers to a conventional family of two people, but corrective measures for larger families are applied. See Section 3.1 for more details.

The European Union (EU) statistical office, Eurostat (https://ec.europa.eu/eurostat/), has suggested a new relative measure, *at risk of poverty*, aimed at subrogating the relative poverty measure. The at-risk-of-poverty rate is the percentage of persons in the total population who

are at risk of poverty because their equivalised disposable income,[2] as calculated after social transfers, is below a certain threshold.

Specifically, the at-risk-of-poverty rate is computed for within-EU comparisons by counting the households possessing an income below 60% of the median equivalised income disposable to households in the same country before all social transfers. Each household member is considered to have the same equivalised income. When comparing the poverty thresholds of different EU member states, the thresholds are standardised according to purchasing power standards that, when controlling for differences in price levels between countries, convert different national currencies into a common expenditure currency.

The reference population for both the Istat and Eurostat measures consists of all persons living in private households. Thus, persons sleeping in collective households, in institutions or rough, are generally ignored in general statistics. To compute the at-risk-of-poverty rate, Eurostat uses data from the EU-SILC (Statistics on Income and Living Conditions) survey.[3]

Absolute poverty

Absolute poverty is an alternative concept: 'absolutely poor' is defined as a person or a family with insufficient resources to live with dignity in a given area at a given time. This inspiring concept refers to the minimum amount of money necessary for a family of a given size and composition to achieve a 'basket' of goods and services qualifying a decent lifestyle in the area. The absolute poverty threshold, which is expressed in monetary terms, is immediately operative because the prices of items composing the

[2] Income is to be equivalised to redistribute it within the household. The equivalised disposable income is calculated in three steps: 1) all monetary incomes received from any source by each member of a household are added up; 2) to reflect the differences in a household's size and composition, the total (net) household income is divided by the number of 'equivalent adults', using the so-called OECD-modified equivalence scale, which gives a weight to all members of the household; and finally, 3) the resulting figure, the equivalised disposable income, is attributed equally to each member of the household (Atkinson et al., 2017).

[3] https://ec.europa.eu/eurostat/statistics explained/index.php/EU_statistics_on_income_and_living_conditions_ (EU-SILC)_methodology_-_monetary_poverty

basket can be summed and allow social control on both the demand and supply sides of income.

Defining the basket is crucial to defining the threshold. A basket is a priori defined in monetary terms in all its components and may vary in space and time in parallel with the varying concept of a decent life. The concept of what makes life 'decent' may also vary in different population groups, for instance, people leaving alone versus couples with children.

This relativity could be seen as analogous to that of relative poverty, though in a poverty measurement survey, the number of absolutely poor people is computed by counting those who are below the a priori given poverty line. Instead, the relatively poor are those below the line computed after the collection of the other residents' income data.

In Italy, Istat (2009) has published measures of absolute poverty since 2005. The Italian basket includes goods and services that an expert committee considers essential for ensuring an acceptable minimum standard of living for a household with given characteristics residing in Italy. For instance, it includes having decent accommodations, two decent meals a day, the possibility of autonomous transport, access to health and education services, and so on, that a family would need. Instead of the necessary basket, the expert committee could directly evaluate the amount of money that is sufficient for a decent life.

The European Commission has not computed a similar measure for EU countries[4] (see Section 3.2 for more details).

Material deprivation

Another way of measuring poverty is through *material deprivation*. The basic idea is that if one does not possess certain goods and services that are relevant to living standards, then that person is considered poor (Townsend, 1979). In response to a study by Guio (2009), the EU defined a set of nine basic goods or services any European citizen should access (European Commission, 2010):

[4] In this paper, we circumscribe our analysis to EU countries, ignoring on purpose out-of-Europe comparisons because the measures of poverty strikingly differ in developing compared with developed countries (United Nations, World Bank).

- Five types of economic strain a household could not afford: (1) covering unexpected expenses; (2) taking a one-week holiday away from home in a year; (3) paying arrears (mortgage or rent, utility bills or hire purchase instalments); (4) eating a meal with meat, chicken or fish every second day; and (5) keeping the home adequately warm.
- Four types of durables the household could not afford (if it wanted to): (6) having a washing machine; (7) owning a colour television; (8) having a telephone; and (9) possessing a personal car.

The material deprivation and absolute poverty approaches are similar in that both imply the definition of a basket of goods and services while periodically updating the basket. Yet the definition of a basket of goods and services as well as the poverty threshold received criticism (among others, Bradshaw and Mayhew, 2010). Therefore, the EU appointed another commission to update the list of goods and services. This commission (Guio et al., 2017) studied those indicators describing the quality of housing and stated that the housing indicators could be added to the other deprivation indicators because they were rather independent of each other. In other words, the housing quality items define a second dimension of deprivation that is not included in the economic ones involved in the previous indicators of economic strain and durable goods.

In 2017, an updated EU list was issued that contains six items from the previous list (numbers 1–5 and 9) and seven new ones (https://www.poverty.ac.uk/world/european-union-2017). The new items concern the following:

- (1) Replacing worn-out clothes with new ones; (2) having two pairs of properly fitting shoes; (3) spending a small amount of money each week on oneself; (4) enjoying regular leisure activities; (5) getting together with friends/family for a drink/meal at least monthly; (6) having an internet connection; and (7) replacing worn-out furniture.

Poverty and social exclusion

The EU publishes another measure that includes both poverty and social exclusion: the *at-risk-of-poverty or social exclusion rate*. This measure is derived from the union intersection of the relative poverty rate, material

deprivation rate and unemployment rate of a community at a given time. It aims to evaluate the extent of *social exclusion*, which is a wider concept than poverty.

The union-intersection rule implies that the three composing rates may overlap. Thus, similar to all measures of poverty, this rate includes people who are at high risks of social exclusion, of being poor, deprived and unoccupied and those who are only deprived and unoccupied or facing another combination of difficulties.

Extreme poverty

Extreme poverty is another poverty concept. An 'extremely poor' person can be defined as a person who is close to the bottom line of income distribution and constitutes a social group whose necessities require urgent help. [5]

It is easy to identify the *homeless* as extremely poor, but we could enlarge this concept to include all those who use public services that are intended for the homeless (soup kitchens, shower stalls, public dormitories, etc.) or to families persistently and severely deprived and below the poverty line. We refer to the possibility of introducing another analytic dimension: the *persistence of poverty*, which implies that people who are below the poverty line for consecutive time spans are poorer than those who fall into poverty at a single point in time who or just occasionally do so and then recover (Bane and Ellwood, 1986; Whelan et al., 2002; Aaberge and Mogstad, 2007).

Therefore, we should also study *chronic poverty*, a concept that refers to persons or families that systematically or for long periods of time are in poverty (Fabbris and Sguotti, 2013). A person may be chronically poor either in a relative or absolute sense. The EU-SILC survey can

[5] The concept of extreme poverty crosses with that of absolute poverty. The experts of the United Nations (1995) had mainly the developing countries in mind when they stated, '*Absolute poverty is a condition characterized by severe deprivation of basic human needs, including food, safe drinking water, sanitation facilities, health, shelter, education and information. It depends not only on income but access to services*'. Also the World Bank's approach to poverty refers to a similar minimalistic principle (Ravallion et al., 2008).

generate a persistent poverty measure over a four-year period (OECD, 2008; European Commission, 2010).

Despite its urgency, the extreme poverty concept is rarely ascertained by national and European statistical offices. In Section 3.7, we examine various measures of this concept.

Psychological poverty

For completeness, we also introduce a subjective or *psychological* approach to poverty. The basic idea of this approach is that the poor are those who perceive themselves as poor. This approach is very different from those previously introduced. The reference is not income or a basket of items but rather a subjective perception: one's interpretation of the relevance (quantity, quality, persistence, etc.) of one's own income or basket availability either in absolute terms or with respect to others' incomes or baskets.

This approach, which could be relevant if used in tandem with an objective measure of poverty, requires specific research and therefore will not be dealt with in the following.

3. PROPERTIES OF POVERTY MEASURES

The title of the present paper refers to 'counting' the poor. Counting refers to the possibility of classifying each person or family as poor or non-poor. This classification is relevant if we refer to the possibility of intervening in and solving individual poverty problems.

Alternatively, interventions to fight poverty may refer to the general population or the totality of a social group. The poverty rate inherent to a community or social group can be expressed in terms of the probability that poverty will affect the concerned population or the percentage of people likely affected by poverty in that community/group. On one hand, this approach does not count the poor; on the other hand, it is in line with the hypothesis that policies are carried out at the broad community rather than individual level.

The properties of the poverty measures that are examined in what follows relate to their possible use. The distinctive properties of relative poverty are presented in Section 3.1, absolute poverty in Section 3.2, the

intensity of poverty in Section 3.3, material deprivation in Section 3.4, at risk of poverty in Section 3.5, at risk of poverty or social exclusion in Section 3.6 and extreme poverty in Section 3.7.

3.1. RELATIVE-POVERTY RATE

The peculiar property of the relative measure of poverty is that it depends on the median of income distribution. The median is a rather steady-in-time centrality parameter and, with reference to income, represents what the median citizen earns. Therefore, the following considerations apply:

- The median value can be considered the most representative value of the income distribution at hand because a possible increase in the median value means that a large percentage of the population, not just a few extreme income earners, had an income increase.[6]
- A time series of median values is usually quite steady and the poverty line consistently flat over time, making the latter a fair reference for mean-run policies for income integration.
- The poverty rate can be estimated not only directly through the data collected with an income survey but also by combining the parameters of the income distribution, which are fairly stable over time. The latter possibility particularly favours an early estimation of thresholds and rates and academic studies.

The main concern regarding a relative measure of poverty is that it depends on the shape of the distribution, which entails the following:

- The poverty rate is insensitive to constant increments of income spread over the population, because if everybody becomes richer, the poverty line shifts accordingly and the poverty rate remains unchanged. This is because the shape of the income distribution is similar throughout the world and over time. For the same reason, the poverty rate does not change if the whole population becomes proportionally poorer. Moreover, during economic downturns, when many low- and middle-income families lose income and

[6] This may not be the case if we consider the mean, instead of the median, as a centrality parameter. For instance, a mean increase could be obtained because either only the better-off had an income increase or only the poor became poorer. Instead, if the central part of the income distribution for two comparable years did not change, the two medians would be the same.

better-off families maintain or improve their standard of living, the relative poverty rate could paradoxically decrease (see, for instance, the years 2008–2011 in Figure 1).

- The relative poverty rate is almost constant over time, no matter the country; therefore, the rates of two countries do not mirror the difference in terms of income but instead in terms of 'relative poor people' of the countries. A poor person in one country could have an income twice as high as a poor person in another country, provided the median citizen in the former country has an income twice as high as the median citizen in the latter country.[7]
- The definition of the threshold at 60% of the median income is conventional. It has been fixed at 60% for the national country thresholds to provide a common reference. Indeed, the 60% proportion was defined as an EU reference. To make worldwide comparisons, the OECD, among others, fixes the poverty line at 50% of the country's median income (https://data.oecd.org/natincome/net-national-income.htm). Some European countries (Croatia, France, Germany, Latvia, Lithuania, Luxembourg, Malta, Romania, Spain and the UK) use, for national statistics on poverty, lower relative thresholds at 40% or 50% of the median equivalent income.
- In Italy, the relative poverty rate of families was steady at around 11% from 2005 to 2011 and showed higher values from 2011 on. In 2017, the rate rose to 15.6% and then dropped for two consecutive years. Instead, the individual poverty rate remained steady at around 13% until 2011 and was then artificially deflated in 2014 by a computational change before finally sharply increasing from 2017 on (Figure 1).
- The absolute poverty rate of families varied within a short range at or right above 4% until 2010; then it started increasing with the rise of absolute poverty in Italy, reaching 6.9% in 2017 and 7.5% in 2021. If the system had remained the same as before 2014, the 2021 rate would have been close to or above 10% (Figure 1).

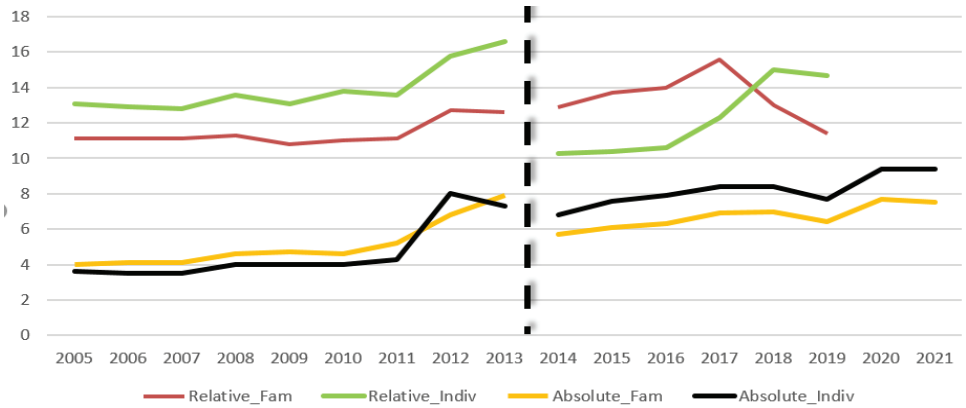
[7] A symptomatic example is reported by the European Commission (2010): the relative poverty threshold for a couple with two children in Estonia in 2008 was, in terms of purchase parity, 9,770 euros per year, and in the UK, it was 24,380 euros per year. Nevertheless, the at-risk-of-poverty rate in both countries was 19%.

- In line with Sen (1983), the relative rate is a difficult measure for people to understand. The precariousness of this rate is clearer if the time series of relative and absolute rates are compared: during tough economic times, the absolute poverty rate reasonably increases, while the relative rate may not grow accordingly and could be overcome by the absolute one.
- Income may not be as good an indicator of command over resources as expenditure, not least because it does not take into account the capacity to borrow, saving use, gifts, social assistance[8] and the value of family production for its own consumption.

The uncertainty about the poverty line induced Istat to define two more measures involving people or families close to the line, either above or below it. Persons 20% above the standard line—that is, between 60% and 72% of the median income—are considered 'almost poor', while those 20% below the line are considered 'scarcely poor' (Istat, 2017). Units above and below the almost poor are certainly non-poor and certainly poor, respectively. These two additional measures make sense only if they are published together with the pertinent poverty line.

[8] Several countries have linked their national poverty threshold to their minimum income (social assistance scheme) or other benefits. This situation may interfere with the definition of a threshold and could suggest the necessity to refer to disposable income, which is the basis of the at-risk-of-poverty rate, instead of just income.

Figure 1. Relative- and absolute-poverty rates in Italy according to type of population, 2005–2017 ***(in 2014, the estimation system changed[9])***



3.2. ABSOLUTE-POVERTY RATE

In a given country, for a series of absolute poverty rates to be comparable, the low-income families of that country should share the same basket of goods and services over a certain timespan. This is realistic in the short term. A long-term comparison, however, requires a regular update of the basket. This can be done by either consulting an expert commission and/or by adopting a time-dependent formula and parametrically updating the threshold values of the sets under comparison.

Scapin (2015) successfully surveyed a panel of experts (local administrators, politicians, charity representatives and academics) to obtain their views regarding the different needs and thresholds of various groups of families. In addition, Istat set a parametric function linking the needs of social groups living in differently developed regions. The poverty threshold can be interactively computed as a combination of four parameters: geographical area, municipality size, family composition and year.[10] More

[9] Until 2013, the reference data came from the survey on family consumption; since 2014, the data have come from the survey on family expenditure (Istat, Statistiche Report, https://www.istat.it). Istat estimated the new system rates for some years before the system change: the differences are relevant for the family absolute rates (2011: 4.3%; 2012: 5.6%; 2013: 6.3%) and dramatic for the individual relative rates (2011: 9.9%; 2012: 10.8%; 2013: 10.4%).

[10] See: https://www.istat.it/it/dati-analisi-e-prodotti/contenuti-interattivi/soglia-di-poverta.

complex functions could be created if we hypothesise that other characteristics are relevant to poverty.

In principle, the absolute-poverty rates of two Western countries that share the same basket of goods and services could also be compared. However, a common basket implies that the two countries also agree on the political use of the threshold(s). In Europe, it is difficult to imagine such an agreement arising from a Eurostat framework.

Absolute rates are relevant for within-country analysis. The regional partition of the country into regions—or the identification of family groups with different thresholds—are matters of social-political intervention if national or local governments or charity organisations are willing to intervene in the more deprived regions. Absolute rates are relevant for within-country analyses. The partition of a country into regions—or the identification of family groups with different thresholds—is a matter of social-political intervention if the national or local governments or charitable organisations are willing to intervene in the more deprived regions.

Once a family's income level is ascertained, the way of computing absolute poverty allows for analysing the possible subsidies to reach the threshold of the group to which the family belongs in that year.

3.3. INTENSITY OF POVERTY

Poverty rates involving a simple count of units below the poverty line are insensitive to what happens below the line. This is why Istat also computes the *poverty gap rate*, which is also called the *intensity of poverty* (http://www.istat.it/dati/catalogo/20090422_00/), representing how poor the poor are. It estimates the relative amount of money needed for all people below the poverty line to reach that line. The higher the intensity, the poorer the poor are.

In Italy, the intensity rate has been stable for a long time at 20–24%. It was 20.9% in 2017 and 23.8% in 2019 but has varied according to geographical partitions and types of families. For a local administrator, it is a useful tool to pinpoint the areas and population groups whose incomes are lower.

In symbols, let us consider the income, Y, of a given country, whose median, $Me(Y)$, is the income possessed by the country's central income

earner. The poverty threshold, *T(Y)*, which in our case is defined as *T(Y) = 0.6 Me(Y)*, enables the identification of N^* ($N^* < N$) units not exceeding the threshold. N and N^* can refer to both individuals and families. The relative poverty rate, R_1, is the proportion of units below the poverty threshold:

$$R_1 = \frac{N^*}{N},$$

while the poverty gap, R_2, is the budget necessary for all units below the poverty line to reach the line:

$$G(Y) = \sum_{i}^{N*} T(Y) - y_i \; = N^* T(Y) - \sum_{i}^{N*} y_i \;\; (i = 1, \dots, N^*),$$

where y_i is the income observed at unit i ($i = 1, \dots, N$) and the summation applies to just the first N^* units ordered according to income.

The amount *G(Y)* can be relativised by dividing it by its maximum, $T(Y)N^*$, giving the intensity of the poverty rate:

$$R_2 = G(Y)/Max(G(Y)) = G(Y)/[T(Y)N^*],$$

which varies between 0 and 1.

Even an individual intensity of poverty could then be computed:

$$R_{2i} = 1 - \frac{y_i}{T(Y)} \;\; (i = 1, \dots, N^*).$$

Similarly, it is possible to compute an (*absolute*) *poverty gap rate*, calculated as the amount of money needed for people below the absolute poverty line to reach the line and thus not be considered poor. In this case, the income threshold for a household in group h is T_h ($h = 1, \dots, H$), to which the household income, Y_i, refers. Unit i is in poverty if its equivalised income is below the threshold of the group of households to which that unit belongs, that is to say, if $Y_i - T_h < 0$.

The absolute poverty rate, R_3, is the proportion of units below the absolute poverty threshold, N^{*a}, and is computed in the same way as R_1—that is, $R_3 = N^{*a}/N$—and the poverty gap estimate is the budget necessary for all units below the poverty line to reach their threshold. The intensity of absolute poverty is computed using the same formula as for relative poverty.

3.4. MATERIAL DEPRIVATION RATE

The material deprivation rate can be used to compare countries, provided that the basket of goods and services is equivalent for all involved countries. This is possible if the basket is 'normative' in the sense that a European household is labelled as materially deprived if it does not access a common standard of goods and services.

The peculiarity of the material deprivation approach is that the basket functions as a physical standard. Bradshaw and Mayhew (2010) argued that for the EU policy to eradicate social exclusion, the challenge is to raise the living standards of the poor in poorer countries, and to achieve this, the EU should adopt at least an absolute-type indicator. The material deprivation basket shares with the absolute poverty basket the characteristic that items are physical entities, with the difference being that those in the latter basket are evaluated in monetary terms.[11]

Pooling deprivation indicators into a single indicator implies defining a composite index. Indeed, Guio's (2009) analyses support the idea that the economic strain and durable indicators of the basket could be treated as a composite deprivation index. The set of items involving the quality of housing, which the basket that Eurostat revised in 2017 contains, enriches the composite index.

Experts have disputed the definition of the deprivation threshold. The larger the basket, the more uncertain the threshold. One of the criticisms of the deprivation approach is that not having some items could be a lifestyle choice of someone who is perfectly capable of purchasing these items. Therefore, the count should be limited to the items a person cannot afford. Other items may be of secondary priority in the household budget, which people may plan to acquire after some time. One of the reasons why a colour television was removed from the 2017 item list is that

[11] The experience of the deprivation threshold computation (European Commission, 2010) suggests that in richer countries, a substantial proportion who are defined as 'poor' being below the at-risk-of-poverty threshold are lacking no deprivation items and state they do not have difficulty making ends meet. Moreover, Bradshaw and Mayhew (2010) have created an exercise of crossing the 60% at-risk-of-poverty rates with the 4+ material deprivation rate based on 2008 EU-SILC data for EU member countries, finding that for most countries, the two rates are highly consistent, hence highlighting a common latent factor; however, some countries (Latvia, Spain, UK, Italy, Greece and Hungary) show a second dimension uncorrelated with the previous one.

people could have technical alternatives to it, such as projectors or phones, or items may be possessed but broken. Other items, such as getting together with friends/family for a drink/meal at least monthly, could be an aspect of an inward-looking lifestyle.

Before 2017, there were two thresholds: one for *deprivation*, which was evaluated by three missing items, and another for *severe deprivation*, which was evaluated by four or more missing items. After 2017, the *material and social deprivation* rate became the proportion of the population experiencing an enforced lack of at least five out of thirteen deprivation items and *severe deprivation* involving seven or more items (https://ec.europa.eu/social/BlobServlet?docId=19228&langId=en).

A problem with the deprivation rate is that the instability of the basket and its multidimensionality. This suggests that the basket requires periodic updating and that alternative ways of evaluating the data other than merely counting the missing items should be studied (Whelan et al., 2008).[12]

Another problem is that, for policy purposes, the items must be transformed into monetary values. Generally, policy cannot intervene at the level of item supply, but it can and usually does intervene by providing income. For the indicator to be a useful tool with which to intervene at the individual or group level, a threshold defining a household as poor *and* deprived could be adopted. This rationale has been adopted by some European countries (among others, Ireland).

In symbols, the estimation of a material deprivation rate is a sensible operation only if the B ($B = 13$) items at unit i ($i = 1, \ldots, N$) of the N-sized population represent a common underlying construct, θ, which Eurostat calls 'material deprivation'. Of course, some units of the population may not be deprived, while others will have degrees of deprivation.

Let us consider unit i with a deprivation level θ_i and a vector of dichotomous items $\boldsymbol{Y}_i$ representing θ, where $Y_{ij} = 1$ if household i possesses item j, and 0 otherwise. The count of the items that the household possesses,

[12] An item could be weighted with the proportion of households that do have it (European Commission, 2010). The effect of this would be to give more weight to the lack of an item in a small minority of households. The underlying justification for this is that because most people have it, lacking it means greater deprivation.

$y_i = \sum_j^B Y_{ij}$, is an empirical score sufficient for estimating the unit deprivation level θ_i.

The deprivation threshold, *T(Y)*, is the minimum number of missing items in a deprived household. If household *i* (*i* = 1, …, *N*) has a number of missing items Θ_i equal to or larger than $T(Y^d)$ it is considered deprived. So, the deprivation threshold enables the identification of N^{*d} $(N^{*d} < N)$ deprived households in the population. The material deprivation rate, $R_4 = N^{*d}/N$, is the proportion of households at or above the deprivation threshold:

$$N^{*d} = \sum_i^N G(Y_i) \ \ where\ G(Y_i) = 1 \ \ if\ T(Y) \geq y_i\ and\ 0\ otherwise.$$

To estimate R_4, instead of simply counting the (not) possessed items, which implies that items are independent and equally important, we could assume that items have different weights with respect to the underlying one-dimensional poverty construct for that population and that the estimate of θ for unit *i*, θ_i, is a weighted combination of the items measured at unit i, y_{ij} (Walker, 2015):

$$\widehat{\theta_i} = \sum_j^B \hat{\beta}_j y_{ij} \ \ (i = 1, \ldots, N),$$

where $\hat{\beta}_j$ is the weight assigned to item *j* (*j* = 1, …, *B*). One possibility is that deprivation items are rated according to severity by a selected panel of experts.[13] Another is that item weights are estimated through a factor analysis of preference data after statistical standardisation.

If more than one dimension underlies the deprivation items—as has generally been taken for granted since the 2017 reform—a composite indicator could be constructed that summarises the underlying factors. The construction of a composite indicator requires further insights into the nature of the deprivation items.

[13] The importance of consulting experts in assessing the exchangeability of items is well known to scholars. According to Ravallion (2011), 'those with a stake in the outcomes will almost certainly be in a better position to determine what weights to apply than the analyst calibrating a measure of poverty'.

3.5. AT-RISK-OF-POVERTY RATE

The at-risk-of-poverty rate is a new Eurostat proposal to subrogate the relative poverty rate. It is computed for both individuals and households. The difference between this measure and the traditional one, as computed by Istat, is relevant. With reference to 2017, the new measure for Italy is 20.3%, which is comparable with an Istat relative poverty estimate of 12.3%.

To gain a sense of the representativeness of Eurostat estimates, the German at-risk-of-poverty rate for 2017 was 16.1% and the Greek rate was 20.2%.[14] Moreover, in Italy, over the past 10 years, the minimum was 18.4% in 2008 and the maximum was 20.6% in 2016. We leave it to experts to evaluate how realistic it is that the rate of poverty in Italy is about the same as in Germany.

We can conclude that, for a given country, this poverty measure varies moderately over time, similar to the relative poverty rate, but is much higher than the relative poverty rate. In addition, at a given time point, the differences between country levels may be small even if the country levels are high. The little between-country differentiation does not help the European Commission assign countries resources to fight poverty in a way that is proportional to the effective needs of these countries.

To measure the poverty threshold of EU countries, 60% of the median is the standard. The at-risk-of-poverty rate is the measure on which the EU particularly relies. However, it is not free from criticism:

- This measure is not easy for the general population or technical poverty experts to understand. Its computational refinement makes its meaning vague and detached from poverty intervention. This reduces its power as a poverty measure (Bradshaw and Mayhew, 2010).
- It refers to disposable income instead of living standards or expenditures. The income-based concept ignores the capacity to borrow, dissaving, gifts and the value of home production. Indeed, the capacity to borrow and use savings depends on the duration of economic difficulties and the exceptionality of the household's

[14] See: https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tessi010&plugin=1.

situation. Gifts, charity and considerable family support depend on the household's social umbrella. Home production is very important in rural areas and in families with older adults. Thus, despite its refined definition of 'equivalised disposable income before all social transfers excluding pensions that are below the at-risk-of-poverty thresholds calculated after social transfers' (https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:At-risk-of-poverty_rate), it is too narrow a concept to describe the command capacity regarding the household's resources.

- The reference threshold as 60% of income is arbitrary, although this criticism applies to all measures based on indirect measures of social uneasiness. Moreover, the EU indicator uses an OECD equivalence scale to adjust income to household needs with respect to family composition, while the OECD itself has abandoned this scale and adopted an equivalence scale based on the square root of the number of people in the household, which is believed to be more science based (OECD, 2008).

Eurostat also computes a *persistent at-risk-of-poverty* rate, which covers persons who have been living in private households for four years and who have been on the EU-SILC panel for all four relevant years.

In statistical terms, *mutatis mutandis*, all that has been mentioned for the relative poverty rate is valid for the at-risk-of-poverty rate. The difference is in the denomination of the basic variable, which is equivalised income for the former rate and equivalised disposable income before all social transfers, and so forth, for the latter rate.

3.6. AT-RISK-OF-POVERTY OR SOCIAL EXCLUSION RATE

The union-intersection of three indicators of household uneasiness—poverty (formerly, relative poverty rate, currently, at-risk-of-poverty rate), deprivation (the severe material deprivation rate) and unemployment (jobless rate)—should amount to the risk level of social exclusion of the households of an area at a reference time (Eurostat, 2020). This risk is estimated by the *at-risk-of-poverty or social exclusion* rate. Figure 2 shows the union intersection among the three basic estimates for the EU.

Figure 2. Millions of persons at risk of poverty or social exclusion in the EU, 2017

(*Source:Eurostat,https://twitter.com/EU_Social/status/1052508790774480896/photo/1*)

The rate is the most comprehensive of the measures described in Section 3. Thus, it may be considered a measure of social exclusion including poverty.

Some rates allow for better understanding what the indicator is aimed at highlighting[15]: 22.5% of the EU population (in absolute terms: 112.9 million people) was at risk of poverty or social exclusion in 2017, and this figure comes from merging the 16.9% of the population at risk of poverty, 9.3% of the population aged 0–59 years living in households with very low work intensity and 6.9% severely materially deprived population. The analogous figures for Italy were 28.9%, 20.3%, 11.8% and 10.1%, respectively. It is evident that the overall rate construction is dominated by the at-risk-of-poverty rate, which overlaps with both material deprivation

[15] The data in Figure 2 are in millions of persons, while the comments refer to friendlier rates. The peculiar construction of the overall rate should be noted: it is the union intersection of three rates (corresponding to the total surface covered by three bubbles) whose denominators are heterogeneous. In fact, the low work intensity rate is computed as a proportion of the population aged 0–59 years living in households, while the other two rates are computed as a proportion of the whole population.

and unemployment, but material deprivation and unemployment alone each add a quota of social exclusion.

To better understand what the indicator shows, let us compare the at-risk-of-poverty and social exclusion rates of Italy and Germany in 2017. Italy's rate is 28.9% and Germany's is 19%, but the difference between the at-risk-of-poverty rates of the two countries is much less: 20.3% for Italy and 16.1% for Germany. These data show that the at-risk-of-poverty rate does not differentiate countries that differ in terms of the material deprivation and employment of their populations.

The eclecticism of the social exclusion principle has produced several variants that constitute targets of Europe 2020 and 2030 strategies. To promote social inclusion, in particular through reducing poverty, EU leaders have called for further work to be undertaken on appropriate indicators of this target, covering the dimensions of relative poverty, material deprivation and a more dynamic aspect among the following: labour market exclusion, poverty anchored at a point in time and in-work poverty.

3.7. EXTREME POVERTY RATE

Those who are *extremely poor*, according to the OECD (2008) and the European Commission (2010), are persons and families that are persistently and severely deprived *and* below the poverty line. Yet there is no agreement in the EU regarding the estimation procedure of the extreme poverty rate in a country. The European Commission (2010) has limited its recommendations to either basing the estimate on deprivation indicators alone or constructing a composite indicator on the basis of the overlap between deprivation indicators and living on an income below a budget standard threshold.

Let us first examine the *homeless* phenomenon, which is the most extreme condition of the extreme poverty group. A homeless person is someone without a house to sleep in, though the full definition varies according to where these people sleep. Homeless people can sleep in the street or in buildings not designed for human habitation but also in public dormitories or other communal facilities, in temporary accommodation in a hotel or guesthouse or in accommodation temporarily provided by friends

or relatives.[16] Eurostat (2004) reported other possibilities stemming from a European survey on key witnesses of the homeless phenomenon at the national level. The variety of national concepts is so large that, despite a rough estimate of the EU homeless rate at or below 0.5%, in Germany, that estimate is above 20%.

Some countries have tried to count the homeless in their national censuses, though with uncertain results.[17] Country statistics are also collected and systematised by the European Observatory on Homelessness, which has published two statistical updates on the extent and profile of the homeless in the EU (Edgar, 2009; Busch-Geertsema et al., 2014). However, there are no periodic European statistics on the homeless, nor are single countries' statistics made comparable within this framework (see Edgar, 2009; Stephens et al., 2010).

Aiming to find a minimum common multiplier for defining and then measuring the national homeless phenomenon, Eurostat (2004) has suggested to overcome the current definition of homeless, which refers to an individual status, and elaborate indicators of the wider concept of 'insecure housing conditions and homelessness', which refers to housing conditions. Furthermore, Edgar (2009) and Amore et al. (2011) proposed adopting measures of 'at risk of homelessness' or 'housing exclusion', instead of a homelessness rate.

Nevertheless, we propose to evaluate the homelessness phenomenon by estimating the number of people who sleep either rough or in public dormitories, because these people are in more dramatic conditions. Indeed, people sleeping in shelters or rough areas can be seen as a very socially relevant problem. Even if limited in number, this group is not to be confused in statistical terms with other people who can be the

[16] The way people without their own home sleep allows for distinguishing between the roofless and houseless. A person is *roofless* if they live in a shelter, hotel, hostel of other type of institution or temporary accommodation paid for through social welfare benefits; they are *houseless* if they live in temporary accommodations for the homeless (Busch-Geertsema et al., 2014).

[17] As an example, the rolling French census of 2011 listed 16,339 homeless in municipalities with more than 10,000 inhabitants. According to Busch-Geertsema et al. (2014), this figure is an underestimate because INSEE/INED surveys on French homelessness have indicated some 86,000 homeless people in 2001 and 141,500 in 2012.

target of general socioeconomic policies as the homelessness phenomenon requires specific policies.

Moreover, the interviews conducted by Martini et al. (2007) evinced that such deep poverty is a non-return condition—that is, no interviewed person was able to return to 'normality'. For this type of studies, *normal life is the threshold*. Other studies (Culhane et al., 1994) have instead shown that point-in-time counts of homeless people tend to underestimate the probability of exiting that condition.

To estimate the homeless population in the daytime, the so-called oasis method could be applied. This entails counting those who attend certain sites (soup kitchens, shower stalls, public dormitories, centres for clothing distribution, etc.) to address their primary biological needs. This method consists of sampling the sites of the concerned area and counting how many people frequent them at a given point in time. Alternatively, data on service users can be collected either from service personnel or service registers.[18] Given the stationary condition of this population group, it is possible both to keep a record of service use and combine register-based data with survey data.

The simultaneous count requires many contemporaneous observers to avoid double counting and the risk of non-poor or people occasionally in need being confused with the poor. Moreover, the oasis method may ignore the homeless and living rough who, during the observation period, did not frequent the sampled structures, the Roma and other mobile groups and unregistered or ethnic minority people hosted in institutions, prisons, hospitals, hostels or camps. Moreover, for linguistic or social reasons, these people, when contacted, tend to elude surveys (CPRC, 2001).

The oasis method could also be applied in the street because everybody sleeps somewhere at night. However, for a street survey, it may be necessary to apply area sampling, which is a more complex estimation

[18] There are registers of people receiving support from charitable organisations and local authorities. Busch-Geertsema et al. (2014) reported that a nationwide survey on homelessness has been conducted in Denmark since 2007. These national counts are realised by asking all local service providers and authorities who are in contact with or have knowledge of homeless people to fill out a short individual questionnaire for each homeless person during a 'count week'. The survey covers homeless shelters, addiction treatment centres, psychiatric facilities, municipal social centres, job centres and social drop-in cafes; there is a high response rate from local service providers.

method. Area sampling implies knowing in advance the sites where people sleep, guessing the density of homeless people at each site and then optimally sampling the areas with a probability proportional to the homeless frequency. The sites include not only the streets, station areas, the ground floor of directional buildings and abandoned houses or establishments but also other closed-off sites that volunteers and police know are frequented by homeless people. Then the sampled areas are visited early in the night by squads of observers to count or interview homeless individuals.

Of course, this sampling technique may be dangerous for observers; therefore, any squad of observers should include at least one person who is involved in on-site homeless care. As a matter of fact, homeless individuals—some of whom include 'classical' vagabonds, including people with substance abuse problems, previously imprisoned people, individuals with long-term, multifaceted psychosocial vulnerabilities and irregular immigrants (Fabbris, 2005; Edgar, 2009; Istat, 2015; Benjaminsen, 2016)—either flee as soon as the data collection squad enters the site or refuse to speak with them. However, counting and even interviewing the homeless is feasible. In Italy, such a survey was conducted for the Veneto region (Fabbris, 2007) and it may be reproduced in other local contexts and at the national level.

A method similar to the oasis method was adopted by Istat (2015) for its second sample survey[19] of the homeless. The survey counted people frequenting Italian charitable structures to eat or sleep. Istat estimated a total of 50,724 homeless people in Italy in 2014 (https://www.istat.it/it/files//2015/12/Persone_senza_dimora.pdf). The proportion of surveyed people corresponds to approximately 0.2% of the resident population, a figure close to that of the homeless rate of other European countries.

[19] The 2014 survey of people living in extreme poverty was carried out by a joint effort of Istat, the Welfare Ministry, the Italian Federation of organisations for the homeless (Fio.PSD) and Caritas for Italy. A previous survey held in 2011 estimated 47,684 homeless. The survey included the main Italian municipalities, provincial capitals with more than 30,000 inhabitants and all municipalities in the hinterlands of major towns and cities. A special weighting procedure based on information about the repeated use of services was used to control for double counting (Istat, 2015).

4. CORRESPONDENCE BETWEEN STATISTICAL PROPERTIES AND AIMS OF POVERTY MEASURES

We consider the following aims of poverty measures:

1. The possibility of individual interventions, for both on persons and families, targeted at subsidising a below-threshold household income.
2. The possibility of intervening in within-country areas or at-risk groups. This possibility refers to the definition of a normative policy at the regional or social group level.
3. The possibility of making comparisons between different EU countries.
4. Other aims.

The results of the merging of computational approaches and possible aims are described in Table 1. A cross in the cell of the table represents full correspondence between the approach and aim.

Table 1. Correspondence between poverty measurement approaches and intervention purposes

Measurement approach	Individual intervention	Areas or groups at risk	National, local intervention	Between-country comparison
Relative poverty	****	X		
Intensity of poverty	X	X		
Persistent poverty	X	X		X*
Absolute poverty	X	X	X	
Material deprivation	X***	X		X
Homelessness	X**		X**	
At risk of poverty	****	X		X*
At risk of poverty or social exclusion				X

(*) The comparability between EU countries depends on sensibility of the rate with respect to the country's poverty level. (**) The homelessness phenomenon is so socially relevant that each single case should be followed; statistics should be collected to focus the attention on the phenomenon. (***) For an individual intervention, deprivation has to be transformed into monetary values. (****) A correspondence is virtually possible but could be biased.

The synopsis shows that the relative and absolute poverty rates, as well as the poverty intensity rate, are adequate for community and group interventions. The absolute poverty rate can also be the informative basis for national or local policies intended to help persons and households overcome structural difficulties.

The at-risk-of-poverty rate and analogously targeted persistent poverty rate, as proposed by Eurostat, are adequate for highlighting those areas or population groups at risk of poverty and partially also those EU countries that rank low or high in the poverty rate. Regarding the relative poverty threshold, the European Commission (https://ec.europa.eu/social/main.jsp?catId=89&langId=it&newsId=982&furtherNews=yes) recognises the following:

> *While justified in many ways, presents some weaknesses and, especially does not properly reflect the real living conditions of EU citizen. Living under the poverty threshold in richer countries does not involve the same difficulties as living under the poverty threshold in the poorest ones. The at-risk-of-poverty threshold is also very low in some of the poorer countries. For example, in Romania, the threshold is €1.71 per day per person.*

The relative poverty indicator could also be a reference for individual interventions, but the thresholds require further insights.

The material deprivation rate can be reasonably broken down by areas and social groups and is adequate for making comparisons between countries in terms of deprivation. Instead, it does not offer enough information for an individual intervention unless a suitable monetary transformation of deprived items is defined.

Statistics on the homeless could be useful for both individual intervention and understanding the dimension of the phenomenon at the local or national levels.

5. CONCLUSION

In the current paper, we have discussed approaches for measuring poverty. Poverty was conceived as a social syndrome, varying in intensity and persistence, and associated with a lack of income and other personal and social problems, in particular housing and employment. We claim that the rationale for choosing a suitable approach is based on the pertinence of the

properties of statistical measures with respect to the policies and actions to overcome poverty.

We determined that two measures currently in use at Istat—the absolute poverty and intensity of poverty rates—are informative for local and national intervention purposes. The absolute poverty rate is particularly suitable for interventions at the individual, geographical and social group levels.

Instead, the statistical measures defined at the EU level, in particular the at-risk-of-poverty rate, which was introduced to identify poverty at the local and national levels, emulate the Istat relative poverty rate but without producing adequate information to intervene. If an indicator shows that all countries are similarly poor, it should be better defined.

Concerning EU measures, we have examined three rates: the at-risk-of-poverty, material deprivation and at-risk-of-poverty or social exclusion rates. Here, the material deprivation rationale is appropriate for individual interventions, though some permanent 'Rosetta stone' is necessary to translate item deprivation into an income need. Indeed, the dimensions of deprivation can be plural, item and dimension can be weighted, and currency parities for time and space comparisons require further research. Thus, the subjectivities of the at-risk-of-poverty rate and material deprivation rate, which together constitute two-thirds of the at-risk-of-poverty or social exclusion rate, makes the latter rate problematic. A shift in Eurostat's attitude about this would certainly be welcome.

Our analysis leaves the following issues open:

- If we were asked to state our preference between a measure of poverty that highlights local people in extreme difficulties, which could enable regional or national institutions to politically intervene, and a measure suitable for comparing countries with respect to poverty, we would answer that it depends on which territorial level the intervention is expected to target. If the local authorities, which stand shoulder to shoulder with the poor and are resilient to changes in social preferences, are prone to intervening for local poverty alleviation, an absolute measure of poverty or a monetary transformation of material deprivation seems adequate. The absolute poverty approach is applied in Italy, a country in which the largest part of social interventions is realised directly by municipalities and

charitable organisations. If the relevant intervention occurs at the national level, again, an absolute poverty measure can help. Indeed, economic and social policies on a large territorial scale could be combined with targeting the poor and active participation in local initiatives and civil society groups' engagement (Craig and Porter, 2003). Moreover, knowing which EU countries are poor may help only if the European Commission wishes to help the poorer countries. Therefore, more cogent indicators, in particular an extreme poverty measure, could pinpoint particularly poor countries or areas (see the next bullet point). Finally, the estimation of the distance between a poor individual and a national or international median level—which is the basis of all relative poverty or social exclusion measures—is nothing but an exercise in curiosity.

- The complete eradication of poverty in Europe is a far target. In any society, a quota of poor people is physiological in the sense that it is possible for a person or family to face social difficulties for some time because of health, welfare or social inequalities, the labour market, the political system, criminality diffusion or other social diseases. Think, for instance, of the diffusion of gambling among adults, which can suddenly deplete individuals' or families' resources. Therefore, if we conceive of a poverty rate as a gauge of the inequality of a society, something relevant for academic debate or to solicit social compassion, we can use any measure proposed by Eurostat, with the consequence that the larger the rates, the more the poverty problem seems overwhelmingly difficult to solve. If a rate or threshold is conceived of as an informative tool for intervention, a radical review of the current Eurostat measures is needed. The European Commission decided to measure poverty as a combination of a lack of income and deprivation to highlight areas where poverty is endemic in the EU. This may be a way to proceed, provided the baskets and thresholds are adequately defined and the parity between income and deprivation is made explicit.
- The homeless comprise a subgroup of poor individuals who are socially discomfiting and erased in official statistics. We roughly know this population's size, with rare exceptions. In Italy, Istat has attempted some surveys on the homeless, though a more attentive sample survey would be needed to highlight their necessities and

potential to switch to normality. They differ from other poor people in terms of both characteristics and needs. Therefore, at least every few years, a survey of the homeless population would be useful.

References

Aaberge, R. and Mogstad, M. (2007). *On the Definition and Measurement of Chronic Poverty.* IZA Discussion Paper No. 2659. Bonn: IZA.

Amore, K., Baker, M. and Howden-Chapman, P. (2011). The ETHOS definition and classification of homelessness: An analysis. *European Journal of Homelessness,* 5(2): 19-37.

Atkinson, A.B., Leventi, C., Nolan, B., Sutherland, H. and Tasseva, I. (2017). Reducing poverty and inequality through tax-benefit reform and the minimum wage: The UK as a case-study. *The Journal of Economic Inequality*, 15(4): 303-323.

Bane, M.J. and Ellwood, D.T. (1986). Slipping into and out of poverty: The dynamics of spells. The *Journal of Human Resources*, 21(1): 1-23.

Benjaminsen, L. (2016). Homelessness in a Scandinavian welfare state: The risk of shelter use in the Danish adult population. *Urban Studies*, 53(10): 2041-2063.

Bradshaw, J. and Mayhew, E. (2010) Understanding extreme poverty in the European Union. *European Journal of Homelessness*, 4: 171-186.

Bradshaw, J., Middleton, S., Davis, A., Oldfield, N., Smith, N., Cusworth, L. and Williams, J. (2008). *A Minimum Income Standard for Britain: What People Think.* York: Joseph Rowntree Foundation.

Busch-Geertsema, V., Benjaminsen, L., Filipovič Hrast, M. and Pleace, N. (2014) *Extent and Profile of Homelessness in European Member States: A Statistical Update.* Brussels: FEANTSA.

CPRC (2001). *CPRC Methods Toolbox*. Chronic Poverty Research Centre (www.chronicpoverty.org/page/toolbox).

Craig, D. and Porter, D. (2003). Poverty reduction strategy papers: A new convergence. *World Development*, 31(1): 53-69.

Culhane, D.P., Dejowski, E.F., Ibañez J., Needham, E.N. and Macchia, I. (1994). Public shelter admission rates in Philadelphia and New York City: The implications of turnover for sheltered population count. *Housing Policy Debate*, 5(2): 107–140.

Edgar, W. (2009). *European Review of Statistics on Homelessness.* Brussels: FEANTSA.

European Commission (2010). *The Measurement of Extreme Poverty in the European Union*. Brussels: European Union.

Eurostat (2004). *The Production of Data on Homeless and Housing Deprivation in the European Union: Survey and Proposals*, Luxembourg: Office for Official Publications of the European Communities (https://ec.europa.eu/eurostat/documents/3888793/5832745/KS-CC-04-008-EN.PDF.pdf/2a7f26b4-4a10-4f05-a43a-52786a114279?t=1414779149000).

Eurostat (2020). *Glossary: At Risk of Poverty or Social Exclusion (AROPE)* (https://ec.europa.eu/eurostat/statistics-xplained/index.php/Glossary:At_risk_of_poverty_or_social_exclusion_(AROPE)).

Fabbris, L. (2005). La ricerca. In: Azienda ULSS 16 – Osservatorio per la tutela e promozione della persona, *Presenze nascoste. Viaggio nelle estreme povertà in Veneto*, Regione del Veneto-Giunta Regionale, ULSS 16, Veneto sociale, Padova: 31-66.

Fabbris, L. (2007). Methodology for observatories on extreme poverty. Some evidences from a study on the homeless. In: Società Italiana di Statistica (a cura di) *Rischio e Previsione, Università Ca' Foscari, Venezia, 6-8 giugno 2007, Relazioni invitate*: 173-181 (+CD Rom).

Fabbris, L. and Sguotti, I. (2013). Measuring chronic poverty in Italy, *Rivista Italiana di Economia, Demografia e Statistica*, **LXVI(2)**: 99-122.

European Union (2010). *Income Poverty and Material Deprivation in European Countries*. Luxembourg: Publications Office of the European Union.

Guio, A.-C. (2009). *What Can be Learned from Deprivation Indicators in Europe?* Luxembourg: Office for Official Publications of the European Communities.

Guio, A.-C., Gordon, D., Najera, H. and Pomati, M. (2017). *Revising the EU Material Deprivation Variables*. Brussels: European Union.

Istat (2009). *La misura della povertà assoluta*. Metodi e norme n. 39. Istat, Roma (https://ebiblio.istat.it/digibib/Metodi%20e%20norme/MOD1546628Ed2009N39.pdf).

Istat (2015). *Le persone senza dimora*. Available from: http://www.ista.it/archivio/71958.

Istat (2017). *La povertà in Italia. Anno 2017*. Statistiche Report, Istat (https://www.istat.it/it/files//2018/06/La-povert%C3%A0-in-Italia-2017.pdf).

Martini, M.C., Fabbris, L., Vanin, C. (2007). Dimensions of extreme poverty: Results from a survey on homeless in the Veneto region**,** *Bulletin of the International Statistical Institute, 56th Session Proceedings, 21-29 August 2007,* Lisbon (CD Rom).

OECD (2008). *Growing Unequal? Income Distribution and Poverty in OECD Countries.* Paris: OECD.

Ravallion, M. (2011). On multidimensional indices of poverty. *Economic Journal*, 106: 1328-1343.

Ravallion, M., Chen, S. and Sangruala, P. (2008). *A Dollar a Day Revisited.* Washington, DC: World Bank.

Scapin, E. (2015). *La misura della povertà multidimensionale in Italia.* Graduation Thesis, Department of Statistical Sciences, University of Padua.

Sen, A. (1983). Poor, relatively speaking, *Oxford Economic Papers*, 35(2): 153-169.

Stephens, M., Fitzpatrick, S., Elsinga, M., van Steen, G. and Chzhen, Y. (2010). *Study on Housing Exclusion: Welfare Policies, Housing Provision and Labour Markets.* Brussels: European Commission.

Townsend, P. (1954). Measuring poverty. *British Journal of Sociology*, 5(2): 130-137.

Townsend, P. (1979). *Poverty in the United Kingdom.* London: Allen Lane and Penguin.

United Nations (1995). *Report of the World Summit for Social Development, Copenhagen, 6 to 12 March*. Available at: www.un.org/esa/socdev/wssd/index.html.

Walker, R. (2015). Multidimensional poverty. *GSDRC, Applied Knowledge Services* (ww.gsdrc.org).

Whelan, C.T., Nolan, B. and Maitre, B. (2008). *Measuring Material Deprivation in the Enlarged EU*, ESRI Working Paper no. 249 (www.esri.ie).

Whelan, C.T., Layte, R. and Maître, B. (2002). Multiple deprivation and persistent poverty in the European Union. *Journal of European Social Policy*, 12: 91-105.

doi.org/10.26398/IJAS.0035-014

FROM PLAIN TO SPARSE CORRESPONDENCE ANALYSIS: A GENERALIZED SVD APPROACH[†]

Hervé Abdi
ORCID: 0000-0002-9522-1978
School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Vincent Guillemot
ORCID: 0000-0002-7421-0655
Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France,

Ruiping Liu
ORCID: 0000-0001-8591-7712
School of Applied Science, Beijing Information Science and Technology University, Beijing, China,

Ndèye Niang
ORCID: 0000-0002-6109-9935
Cedric Lab, Conservatoire national des arts et métiers, Paris, France,

Gilbert Saporta[*]
ORCID: 0000-0002-3406-5887
Cedric Lab, Conservatoire national des arts et métiers, Paris, France,

Ju-Chi Yu
ORCID: 0000-0002-6360-1861
Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Canada,

[*]***Corresponding author:*** Gilbert Saporta, `gilbert.saporta@cnam.fr`

[†] The order of the authors reflects only the alphabetical order. All authors contributed equally to this paper.

Abstract *Correspondence Analysis (CA)—the method of choice to analyze contingency tables—is widely applied in text analysis, psychometrics, chemometrics, etc. But CA becomes difficult to interpret when items load on several dimensions, when dimensions comprise items whose loadings are of intermediate values, or when the number of rows or columns is large—a configuration routinely found in contemporary statistical practice. For principal component analysis (PCA), this interpretation problem has been traditionally handled with rotation and more recently with sparsification methods often inspired by the LASSO. Curiously, despite the strong connections between CA and PCA, sparsifying correspondence analysis remains essentially unexplored. In this paper, we extend the penalized matrix Decomposition (a relatively recent method based on the singular value decomposition) to sparsify CA. We present some theoretical results and properties of the resulting sparse correspondence analysis and illustrate this method with the analysis of a large textual data set.*

Keywords: *Sparsity, Correspondence analysis, Generalized singular value decomposition, LASSO, Penalized matrix decomposition.*

1. INTRODUCTION

Correspondence analysis (CA)—the method of choice to analyze contingency tables—becomes difficult to interpret when 1) the data structure is complex as opposed to the *simple structure* (formalized by early psychometricians, such as, e.g., Thurstone, 1935, 1947) where each component is characterized by few items and each item contributes only to few—ideally one—components) or 2) when the number of rows or columns is large—a configuration routinely found in contemporary statistical practice. This interpretation problem, not specific to CA, also occurs in related multivariate methods such as principal component analysis (PCA) where it has been traditionally addressed with methods such as rotation and more recently with sparsification methods mostly derived from the LASSO (Hastie et al., 2001; Tibshirani, 1996). These sparsification methods are also commonly used in fields where the data comprise large numbers of variables (Jenatton et al., 2011) or observations that can include tens of thousands (e.g., in genomics, Chun and Keleş, 2010) to millions (as in neuroimaging, see, e.g., Le Floch et al., 2012; Silver et al., 2012).

But these recent sparsification methods have not yet been widely adapted for CA and its variants. In fact, so far, mostly multiple correspondence analysis (MCA)—which can be seen as an extension of PCA for qualitative variables, as well as an extension of CA to more than two qualitative variables—has benefited from such a (precious) few of these approaches (specifically, see Bernard et al., 2012; Guillemot et al., 2020; Mori et al., 2016).

It is only recently that sparsification for CA per se has been proposed (see, Liu et al., 2023). This approach uses the fact that CA can be interpreted 1) as a double weighted PCA of both rows and columns of the data matrix, or, equivalently, 2) as a generalized singular value decomposition (GSVD, see, e.g., Abdi, 2007; Greenacre, 1984) that incorporates metric constraints on the rows and columns of the data matrix. Within this framework, sparsification is implemented by adding additional constraints on the optimization problem solved by the singular value decomposition (SVD). This constrained SVD still decomposes the data matrix into ("pseudo") singular vectors and ("pseudo") singular values, but this decomposition seeks a compromise between concurrently maximizing explained variance and sparsity. Liu et al. (2023) distinguish two cases depending on whether sparsity is required for either rows or columns, or both.

This paper replicates and extends the approach of Liu et al. (2023) in particular by proposing in lieu of their sequential algorithm, a global algorithm for the simultaneous optimization of the dimensionality of the sparsified space and the sparsification parameters of the rows and columns of the data.

Although the theory of sparsity-inducing constraints is well documented (especially for PCA), the extension to CA is not as straightforward given its special properties. In this paper, we introduce a general formulation of sparsification which can generalize PCA to other related multivariate methods and, specifically, to CA.

We begin with the definition and main properties of CA, followed by a short exposition of the relevant approaches to sparsify PCA. We then show how to extend the concepts from sparse PCA to obtain a sparse version of CA, and describe how sparsifying CA conflicts with some of its key properties that are therefore lost in the process. Finally, we illustrate sparse CA with an analysis of an example of textual analysis extracted from Project Gutenberg (Gerlach and Font-Clos, 2020).

2. BACKGROUND

2.1. NOTATIONS

Matrices are denoted in upper case bold letters, vectors are denoted in lowercase bold letters, and their elements are denoted in lowercase italic letters (note that, by default, vectors are column vectors). Matrices, vectors and elements from the same matrix all use the same letter (e.g., $\mathbf{A}$, $\mathbf{a}$, a). The transpose operation is denoted by the superscript $^{\mathsf{T}}$, the inverse operation is denoted by $^{-1}$. The identity matrix is denoted $\mathbf{I}$, vectors or matrices of ones are denoted $\mathbf{1}$, matrices or vectors of zeros are denoted $\mathbf{0}$ (by default, $\mathbf{I}$, $\mathbf{0}$, and $\mathbf{1}$ are conformable with the other

terms in a formula). The standard product between two matrices is indicated by juxtaposition (i.e., $\mathbf{XY}$ means $\mathbf{X}$ times $\mathbf{Y}$); the Hadamard product (i.e., element-wise) is denoted by $\odot$ (e.g., $\mathbf{X} \odot \mathbf{Y}$), note that the Hadamard product is defined only between matrices with the same dimensions.

When provided with a square matrix, the diag operator gives a vector that contains the diagonal elements of this matrix. When provided with a vector, the diag operator gives a diagonal matrix with the elements of the vector as the diagonal elements of this matrix. A diagonal matrix is denoted $\mathbf{D}$, when a subscript is attached, it denotes the vector that stores the diagonal elements ot he matrix; for example, $\mathbf{D_a} = \mathsf{diag}\,(\mathbf{a})$. When provided with a square matrix, the trace operator gives the sum of the diagonal elements of this matrix. For an I by J matrix $\mathbf{X}$ and for $\mathbf{M}$ being a J by J symmetric positive definite matrix, the squared $\mathbf{M}$-norm of $\mathbf{X}$ is denoted $\|\mathbf{X}\|^2_{\mathbf{M}}$ and is computed as:

$$\|\mathbf{X}\|^2_{\mathbf{M}} = \mathsf{trace}\left(\mathbf{XMX}^{\mathsf{T}}\right) . \tag{1}$$

When $\mathbf{M}$ is the identity matrix, the $\mathbf{M}$-norm is equal to the square root of the sum of squares of the entries of the matrix and is called the *Frobenius* norm denoted $L_2 = \|\mathbf{X}\|^2_2$. Another useful norm is the sum of the absolute values of the matrix called the L_1 norm.

A probabilistic matrix (i.e., a matrix with non-negative elements whose sum is equal to 1) is denoted $\mathbf{Z}$, its row (respectively column) sums are stored in vector $\mathbf{r}$ (respectively $\mathbf{c}$): $\mathbf{r} = \mathbf{Z1}$ (respectively $\mathbf{c} = \mathbf{Z}^{\mathsf{T}}\mathbf{1}$). The matrix of row (respectively column) profiles is denoted $\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{Z}$ (respectively $\mathbf{C} = \mathbf{ZD_c}^{-1}$).

When describing an optimization problem, the operator $\underset{\mathbf{x}}{\arg\min} f(\mathbf{x})$ [respectively $\underset{\mathbf{x}}{\arg\max} f(\mathbf{x})$] gives the value of $\mathbf{x}$ that minimizes (respectively maximizes) the function $f(\mathbf{x})$.

2.2. SVD AND GENERALIZED SVD

The singular value decomposition (SVD) and its extension the generalized singular value decomposition (GSVD, see, e.g., Abdi, 2007; Allen et al., 2014; Greenacre, 1984; Holmes, 2008; Takane, 2002) are the foundations of most contemporary multivariate statistical approaches.

The SVD of an $I \times J$ matrix $\mathbf{X}$ solves the following maximization problem (Eckart and Young, 1936): Find a matrix (denoted $\widehat{\mathbf{X}}_L$) of rank L [with $L \leq$

$\min(I,J)$], computed as

$$\widehat{\mathbf{X}}_L = \sum_{\ell=1}^{L} \delta_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\top = \mathbf{U}_L \boldsymbol{\Delta}_L \mathbf{V}_L^\top \text{ with } \mathbf{U}_L^\top \mathbf{U}_L = \mathbf{V}_L^\top \mathbf{V}_L = \mathbf{I} \text{ and } \boldsymbol{\Delta}_L = \text{diag}\,(\boldsymbol{\delta}_L) \quad (2)$$

such that $\widehat{\mathbf{X}}_L$ is the matrix of L rank closest to $\mathbf{X}$ (in the metric defined by the L_2 norm):

$$\underset{\mathbf{U}_L, \boldsymbol{\Delta}_L, \mathbf{V}_L}{\arg\min} \|\mathbf{X} - \widehat{\mathbf{X}}_L\|_2^2 \quad (3)$$

The SVD of a matrix can be computed by first computing its rank one approximation [i.e., the singular triplet (δ_1, $\mathbf{u}_1$, $\mathbf{v}_1$)] and then subtracting this rank one approximation from $\mathbf{X}$—a procedure called *deflation*. The first singular triplet of the deflated matrix $\mathbf{X}$ is then the second singular triplet of $\mathbf{X}$, etc.

The generalized SVD (GSVD), differs from the plain SVD by incorporating different orthogonality constraints on the singular vectors. Specifically, with $\mathbf{M}$ being an $I \times I$ positive definite matrix (called the row *metric* matrix) and $\mathbf{W}$ a $J \times J$ positive definite matrix (called the column metric matrix), the GSVD of $\mathbf{X}$ solves the following optimization problem (compare with Equation 3):

$$\underset{\mathbf{P}_L, \boldsymbol{\Delta}_L, \mathbf{Q}_L}{\arg\min} \|\mathbf{X} - \widehat{\mathbf{X}}_L\|_2^2 = \underset{\mathbf{P}_L, \boldsymbol{\Delta}_L, \mathbf{Q}_L}{\arg\min} \|\mathbf{X} - \mathbf{P}_L \boldsymbol{\Delta}_L \mathbf{Q}_L^\top\|_2^2 \quad (4)$$

with

$$\mathbf{P}_L^\top \mathbf{M} \mathbf{P}_L = \mathbf{Q}_L^\top \mathbf{W} \mathbf{Q}_L = \mathbf{I}, \text{ and } \boldsymbol{\Delta}_L = \text{diag}\,(\boldsymbol{\delta}_L)\,, \quad (5)$$

where $\mathbf{P}_L$ is the $I \times L$ matrix containing the *generalized* left singular vectors and $\mathbf{Q}_L$ the $J \times L$ matrix containing the generalized right singular vectors. In practice, the GSVD of a matrix $\mathbf{X}$ can be obtained from the plain SVD of a matrix denoted $\widetilde{\mathbf{X}}$ obtained by pre- and post-multiplying $\mathbf{X}$ by the square root of the row and column metric matrices:

$$\widetilde{\mathbf{X}} = \mathbf{M}^{\frac{1}{2}} \mathbf{X} \mathbf{W}^{\frac{1}{2}}\,. \quad (6)$$

More details are given in Appendix A.

2.3. BASICS OF PLAIN CORRESPONDENCE ANALYSIS

Correspondence analysis was originally developed to analyze the pattern of deviations from independence (as measured by a χ^2 statistic) in a contingency table (see Abdi and Béra, 2018). CA provides, for both rows and columns, a set of factor scores whose total inertia is proportional to the independence χ^2 computed on the original contingency table. The factor scores are obtained from the following

generalized singular value decomposition (cf. Equation 4) where $\mathbf{D_r}^{-1}$ and $\mathbf{D_c}^{-1}$ are called χ^2-metric matrices (Greenacre, 2010):

$$\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \text{ with } \mathbf{P}^{\mathsf{T}}\mathbf{D_r}^{-1}\mathbf{P} = \mathbf{Q}^{\mathsf{T}}\mathbf{D_c}^{-1}\mathbf{Q} = \mathbf{I}\,. \tag{7}$$

Correspondence analysis can also be obtained from the plain SVD of :

$$\widetilde{\mathbf{Z}} = \mathbf{D_r}^{-\frac{1}{2}}\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}}\right)\mathbf{D_c}^{-\frac{1}{2}} \tag{8}$$

(For further properties refer to Appendix C).

3. SPARSE SVD WITH PROJECTED PENALIZED MATRIX DECOMPOSITION

Because correspondence analysis is a particular PCA (and therefore a specific SVD, see Equations 8, above, as well as Equations 45 to 47 in Appendix B) a straightforward approach to the sparsification of CA is to adapt an already sparse version of PCA or SVD.

PCA being the oldest and most well-known multivariate method, it is no surprise that several sparse methods have been developed for PCA since the pioneering papers of Vines (2000) and Jolliffe et al. (2003). Case in point, in their—already old—review paper, Ning-min and Jing (2015) count about twenty algorithms for sparsifying PCA.

Recently, several authors have proposed sparse variants of the SVD (see, for reviews, e.g., Allen et al. 2014; Guillemot et al. 2019; Hastie et al. 2015; Jolliffe and Cadima 2016; Witten et al. 2009; Zou et al. 2006), or, specifically, of PCA (Benidis et al. 2016; Mattei et al. 2016). For most of these sparse variants, sparsification is obtained by adding sparsity constraints on both **P** and **Q**, or on **Q** alone. We decided to use the *penalized matrix decomposition* method (PMD) developed by Witten et al. (2009) because it is well-known and is implemented in R (with the `PMA` package).

3.1. PENALIZED MATRIX DECOMPOSITION: BACKGROUND

The penalized matrix decomposition (PMD) method (Witten et al., 2009) generalizes the plain SVD by adding sparsification constraints on the right and left singular vectors. Specifically, the PMD methods solves the following optimization

problem:

$$\underset{\substack{\delta_\ell,\mathbf{u}_\ell,\mathbf{v}_\ell \\ \ell=1,\ldots,L}}{\arg\min} \left\| \mathbf{X} - \sum_{\ell=1}^{L} \delta_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\top \right\|_2^2 \text{ subject to } \begin{cases} \mathbf{u}_\ell^\top \mathbf{u}_\ell = 1 \\ \mathbf{v}_\ell^\top \mathbf{v}_\ell = 1 \\ \|\mathbf{u}_\ell\|_1 \leq s_{1,\ell} \\ \|\mathbf{v}_\ell\|_1 \leq s_{2,\ell} \end{cases} \tag{9}$$

where $s_{1,\ell}$ and $s_{2,\ell}$ are positive constants, provided by the user as two vectors of length L denoted $\mathbf{s}_1$ and $\mathbf{s}_2$ that will drive the sparsity of the solution. The solution to this optimization problem denoted $(\dot{\boldsymbol{\delta}}, \dot{\mathbf{U}}, \dot{\mathbf{V}})$ is called a pseudo-singular triplet (containing respectively the pseudo-singular values, left pseudo-singular vectors, and right pseudo-singular vectors).

In PMD, the first pseudo-singular triplet is estimated by solving Equation 9 for $\ell = 1$. The next pseudo-singular triplets are estimated by approximating each subsequent deflated matrix by a rank one matrix. At each iteration $\ell > 1$, the deflated matrix is equal to

$$\mathbf{X}_\ell = \mathbf{X}_{\ell-1} - \dot{\delta}_{\ell-1} \dot{\mathbf{u}}_{\ell-1} \dot{\mathbf{v}}_{\ell-1}^\top, \tag{10}$$

where, by convention, $\mathbf{X}_1 = \mathbf{X}$. This procedure is very similar to the standard (i.e., Hotelling's) deflation for the SVD, but in Equation 10, the deflated matrix is not guaranteed to be orthogonal to the previous rank one optimal matrix (as noted, e.g., by Mackey, 2009).

To (partially) palliate this problem, Witten et al. (2009) and Mackey (2009) independently proposed a heuristic to handle the non-orthogonality of the row (i.e., left) factor scores (in the context of sparse PCA) where the left pseudo-singular vectors are not required to be sparse. In this case, Hotelling's deflation is replaced by the so-called *projection deflation*

$$\mathbf{X}_\ell = \left(\mathbf{I} - \dot{\mathbf{u}}_{\ell-1} \dot{\mathbf{u}}_{\ell-1}^\top\right) \mathbf{X}_{\ell-1} \ . \tag{11}$$

3.2. PROJECTED PENALIZED MATRIX DECOMPOSITION

In our case, we want to be able to obtain both sparse left and right singular vectors. To do so, we propose to extend the updating step from Equation 11 to the left and right pseudo-singular vectors. This way, for each Dimension ℓ, the projected deflated matrix is obtained as:

$$\mathbf{X}_\ell = \left(\mathbf{I} - \dot{\mathbf{u}}_{\ell-1} \dot{\mathbf{u}}_{\ell-1}^\top\right) \cdots \left(\mathbf{I} - \dot{\mathbf{u}}_1 \dot{\mathbf{u}}_1^\top\right) \mathbf{X} \left(\mathbf{I} - \dot{\mathbf{v}}_1 \dot{\mathbf{v}}_1^\top\right) \cdots \left(\mathbf{I} - \dot{\mathbf{v}}_{\ell-1} \dot{\mathbf{v}}_{\ell-1}^\top\right). \tag{12}$$

With this deflation scheme, PMD is applied iteratively to the data matrix after the projection deflation. Combining PMD and the projected deflation, gives the *projected Penalized Matrix Decomposition* (pPMD, see Liu et al. 2023). It should be noted, however, that pPMD does not yield perfect orthogonality but (according to Witten et al., 2009) as for projection deflation, the solutions are unlikely to be highly correlated.

4. SPARSE CORRESPONDENCE ANALYSIS

In this section, we present a new way to select optimal values for the sparsity parameters, as well as choosing the optimal number of dimensions for sparse CA. Finally, we discuss the effect of introducing sparsity on the properties of CA.

4.1. SPARSE CA WITH pPMD

Because CA is obtained from the plain SVD of $\widetilde{\mathbf{Z}}$ (see Equation 8), the current version of sparse CA is obtained by applying pPMD to $\widetilde{\mathbf{Z}}$. This procedure generates (see Equation 9) pseudo-singular values (denoted $\dot{\delta}$), left pseudo-singular vectors (denoted $\dot{\mathbf{U}}$), and right pseudo-singular vectors (denoted $\dot{\mathbf{V}}$). These pseudo vectors and values are then used to compute sparse weight and contribution matrices for rows and columns as :

- row weight matrix $\dot{\mathbf{P}} = \mathbf{D}_{\mathbf{r}}^{\frac{1}{2}}\dot{\mathbf{U}}$,
- column weight matrix $\dot{\mathbf{G}} = \mathbf{D}_{\mathbf{c}}^{\frac{1}{2}}\dot{\mathbf{V}}$,

Contribution matrices are obtained from the weight matrices:

- row contributions $\dot{\mathbf{T}}_{\mathbf{I}} = \dot{\mathbf{U}} \odot \dot{\mathbf{U}}$,
- column contributions $\dot{\mathbf{T}}_{\mathbf{J}} = \dot{\mathbf{V}} \odot \dot{\mathbf{V}}$.

Note that a zero weight implies a null contribution.

In this paper, we define, (in a manner reminiscent of the transition formula from CA), the factor scores of sparse CA as linear combinations of the profiles with the (sparse) weights:

- row factor scores $\dot{\mathbf{F}} = \mathbf{R}\mathbf{D}_{\mathbf{c}}^{-1}\dot{\mathbf{Q}}$,
- column factor scores $\dot{\mathbf{G}} = \mathbf{C}\mathbf{D}_{\mathbf{r}}^{-1}\dot{\mathbf{P}}$,

and (just like in plain CA) for each dimension, the variance of the factor scores is equal to the squared pseudo-singular value. Note, also, that, while weights are sparse, factor scores may not be sparse.

Even though the computation of the factor scores for sparse CA bears some resemblance to the transition formula (and would be equivalent to the transition formula in plain CA), the transition formulas (from Equation 43 in the Appendix) no longer hold: row (respectively column) factor scores are not barycenters anymore of the column (respectively row) factor scores: Transition formulas hold only for plain CA.

4.2. TWO TYPES OF SPARSITY

Using a sparse version of CA is especially useful when the data are high-dimensional. However, large data sets come in two types: 1) both row and column sets are high-dimensional and sparsifying both dimensions makes sense, or 2) only one of the row or column sets is high-dimensional—and the data table is a "flat" or a "tall" contingency table—and then, it makes more sense to sparsify only the larger set of items. Therefore, two types of sparsity need to be considered for the sparse solution of CA: 1) both-side (or double) sparse CA or 2) one-side (or simple) sparse CA.

Both-side sparsity looks for underlying dimensions that are explained by sparse combinations of both rows and columns. Sparse SVD provides this solution. The easiest way to implement both-side sparsity is to sparsify rows and columns weights in the same proportion, an approach which leads to choose similar degrees of sparsity for $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$. Following Witten et al. (2009), for $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$ to have a similar level of sparsity, η is set constant (with $\eta < 1$) and the sparsity parameters are obtained as $s_{1,\ell} = \eta\sqrt{I}$, and $s_{2,\ell} = \eta\sqrt{J}$ (with I and J being the number of rows and columns of the data matrix). But, if the rows and columns contingency table correspond to essentially different types of variables, then it makes more sense to choose different degrees of sparsity for rows and columns. In this case, the parameter settings "grid" can be used, in order to restrict the L_1 norms of $\dot{\mathbf{P}}$ and $\dot{\mathbf{Q}}$ at different values: $s_{1,\ell}$ and $s_{2,\ell}$ will be restricted to take values (respectively) in the intervals $[1, \sqrt{I}]$ and $[1, \sqrt{J}]$.

One-side sparsity is suitable in asymmetrical situations when, for example, only the rows (or the columns) of the contingency table are relevant. In this case, only the relevant set needs to be sparsified. Interestingly, one-way sparsity is a special case of both-side sparsity when there is no penalty on one side (which is then left un-sparsified) and setting the sparsity parameter of the side to be sparsi-

fied equal to the square root of the cardinal of this set (e.g., to sparsify only the rows, set $s_{1,\ell} = \sqrt{I}$).

4.3. CHOOSING AN APPROPRIATE VALUE FOR THE SPARSITY PARAMETERS:THE SPARSITY INDEX

An essential decision when using sparse CA is the choice of the values for the sparsity parameters $\mathbf{s}_1$ and $\mathbf{s}_2$ and the number of dimensions L. Various methods have been proposed: cross-validation (Witten et al., 2009), AIC or BIC (Shen et al. 2013; Zou et al. 2006), and compromise between the goodness of fit and sparsity (see, e.g., Trendafilov 2014; Trendafilov et al. 2017).

Among these procedures, we chose the sparsity index presented by Trendafilov et al. (2017). This sparsity index denoted, here, $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$, is the product of a "fit ratio" and a "zero ratio."

The *fit ratio* is computed as the ratio of the sum of the pseudo-eigenvalues to the sum of the eigenvalues of the non-sparse solution, specifically

$$\text{fit ratio} = \frac{\sum_{\ell=1}^{L} \dot{\lambda}_\ell}{\sum_{\ell=1}^{L} \lambda_\ell} . \tag{13}$$

The fit ratio takes values between 0 and 1 with larger values indicating a better fit. The *zero ratio* is the ratio of the number of zero weights to the total number of weights, specifically:

$$\text{zero ratio} = \frac{\#0(\dot{\mathbf{P}}) + \#0(\dot{\mathbf{Q}})}{(I+J)L}, \tag{14}$$

where $\#0(\dot{\mathbf{P}})$ [resp. $\#0(\dot{\mathbf{Q}})$] is the total number of zeros in $\dot{\mathbf{P}}$ (resp. $\dot{\mathbf{Q}}$). The zero ratio takes values between 0 and 1 with larger values indicating a sparser solution. The sparsity index $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$ is obtained as the product of the fit and the zero ratios, namely:

$$\varsigma(\mathbf{s}_1, \mathbf{s}_2, L) = \underbrace{\frac{\sum_{\ell=1}^{L} \dot{\lambda}_\ell}{\sum_{\ell=1}^{L} \lambda_\ell}}_{\text{"fit ratio"}} \underbrace{\frac{\#0(\dot{\mathbf{P}}) + \#0(\dot{\mathbf{Q}})}{(I+J)L}}_{\text{"zero ratio"}}, \tag{15}$$

To sum up, the sparsity index is a compromise between maximizing the explained variance (i.e., the fit ratio) and sparsifying the results (i.e., the zero ratio). In our

application of sparse CA, we will therefore seek for the value(s) of L, $\mathbf{s}_1$, and $\mathbf{s}_2$ that maximize $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$.

Our global optimization algorithm differs from the sequential algorithm of Liu et al. (2023) which searches for the optimal sparsity level for each dimension conditional on the sparsity levels obtained for the previous dimensions, but without searching for an optimal value of L (i.e., the dimensionality of the space). In contrast, we obtain a global optimum in a space of Dimension L with the additional constraints that all dimensions have identical levels of sparsity : $s_{11} = s_{12} = \ldots = s_{1L}$.

4.4. LOST PROPERTIES AND OTHER ISSUES

In addition to the usual (but still open) question of "How many components to keep?" sparse exploratory methods raise new specific issues such as—among others—loss of orthogonality and choice of the sparsity level.

The simultaneous orthogonality of the weight vectors and of the factor scores characterizes PCA (and SVD) because weight vectors and factor scores are both true eigenvectors. But this simultaneous orthogonality is lost in sparse PCA and similar methods: One cannot have both orthogonality for the weights and for the factor scores. For example, if we force successive sparse weight vectors to be orthogonal, as in SCoTLaSS (Jolliffe et al., 2003), the associated factor scores are no longer orthogonal.

This lack of orthogonality makes the interpretation of the factor scores somewhat difficult (in a way reminiscent of the issues linked to oblique rotation in traditional factor analysis) because conclusions about one dimension involve all correlated dimensions and because the same information is explained (to different degrees) by all correlated dimensions. When interpreting the factor scores, one could erroneously find the same information in different dimensions. In addition, with non-orthogonal factor scores, the variances explained by different dimensions are no longer additive (i.e., the sum of the variances explained by a set of non-orthogonal dimensions will over-estimatethe variance of the sub-space spanned by these dimensions).

As we have noticed before, the simultaneous pseudo-barycentric transition formulas (from Equation 43 in Appendix B) do not hold anymore because these formulas are a characteristic property of plain CA. Here $\dot{\mathbf{F}}$ is not proportional to $\dot{\mathbf{P}}$ and $\dot{\mathbf{G}}$ is not proportional to $\dot{\mathbf{Q}}$: In other words, the relationship between the weights and the factor scores is not linear anymore. As a consequence, graphics should be drawn using the factor scores $\dot{\mathbf{F}}$ and $\dot{\mathbf{G}}$ rather than the weights $\dot{\mathbf{P}}$ and

$\dot{\mathbf{Q}}$, because a graphic drawn from the weights is likely to have too many points stuck to the axes (these will be the items with zero weights). However, a graph drawn from the weights or even from the signed contributions could be of interest in some applications.

5. A REAL DATA EXAMPLE

We applied sparse CA to a data set—obtained from the Project Gutenberg (Gerlach and Font-Clos, 2020)—compiling common words used in 100 books each from 5 book categories: Biographies, Love stories, Mystery, Philosophy, and Science fiction. This created a contingency table (counting the number of occurrences of words per book) with 1502 rows (words) and 500 columns (books).

5.1. PLAIN CA RESULTS

Factor scores maps for plain CA for dimensions 1 and 2 are shown in Figures 1 (for the words) and 2 (for the books). The word factor map shows only a few words, whereas the book factor map does not show the names of the books but color them by genre and add, for each type of book, a 70% tolerance convex hulls (a *K% tolerance* interval comprises *K%* of a sample or a population, see, e.g., Abdi et al. 2009)[1]. Parallel to the partition of the vocabulary, dimension 1 (see Figure 2) differentiates the Philosophical genre from Love stories, Mystery, and Science fiction.

The second dimension of plain CA explains 7% of the inertia. As shown in Figure 1 and Table 1b), the row factor scores are characterized by the opposition of (on the negative side) words related to war (enemy, battle, war, government) or geography (e.g., south, north, west, east, miles, city), and verbs in the past tense (e.g., ordered, united, received, sent, arrived) versus on the negative side, words related to thoughts (e.g., understanding, experience, reason, meaning, conscious) and verbs in the present tense (e.g., does, miss, mean, thinks, makes, is, can). Parallel to the partition of the vocabulary, dimension 2 (see vertical axis Figure 2) differentiates the Biography genre from books of Philosophy.

In general, the Philosophy genre uses language closely aligned with scientific writing. In contrast, the Biography genre typically uses language focused on the events of a male character's past, often involving war. Meanwhile, the other three genres of fiction are more concerned with describing emotions and the experiences of female characters.

[1]The graph and convex hulls were created using functions `ggConvexHull` and `CreateFactorMaps4CA` from the `R-package PTCA4CATA`.

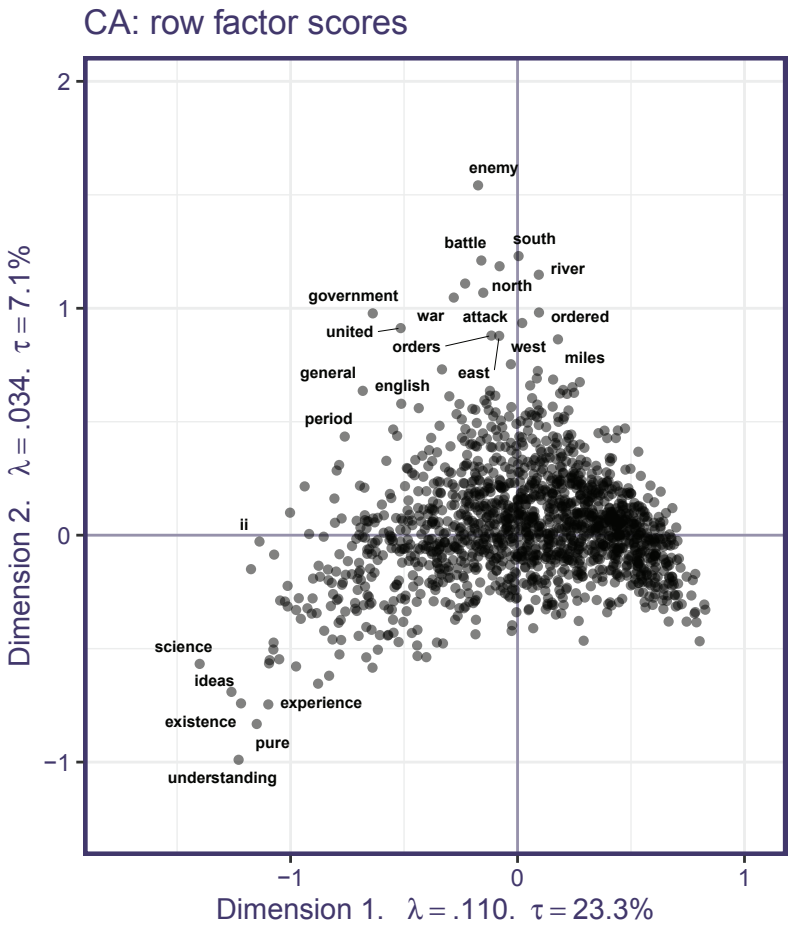


Figure 1: Plain CA: Row factor scores

The Philosophy genre is particularly distinctive in its use of language, because it preferably uses highly specialized terminology not commonly found in other genres. However, interpreting factor scores for this genre is challenging given the presence of many words with close-to-zero weights that are difficult to integrate in a coherent framework.

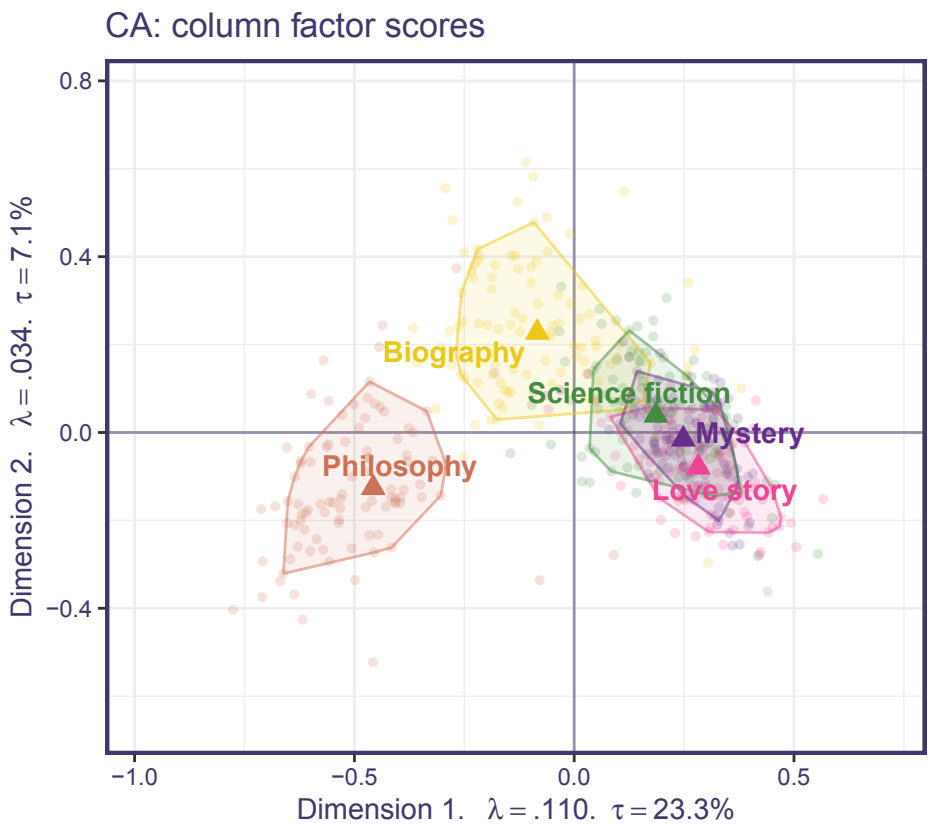


Figure 2: Plain CA: Column factor scores

5.2. CHOOSING AN OPTIMAL VALUE OF THE CONSTRAINTS PARAMETERS WITH THE SPARSITY INDEX

Recall that the sparsity index $\varsigma(\mathbf{s}_1, \mathbf{s}_2, L)$ is a function of $\mathbf{s}_1$, $\mathbf{s}_2$, and L. To find the sparsity index optimal value, we explored the (3-dimensional) space spanned by these parameters. We chose L (i.e., the number of dimensions) from all integers between 2 and 20, and we chose $\mathbf{s}_1$ (resp $\mathbf{s}_2$) from the 20 possible values evenly distributed between 1 and $\sqrt{I}$ (resp. $\sqrt{J}$). To speed up computations and provide the same sparsity value for each dimension, we decided to have identical values for $\mathbf{s}_1$ and $\mathbf{s}_2$. We then iteratively applied sparse CA to the Gutenberg Project

data set with a number of dimensions equal to L (with L varying from 1 to 20). Figure 3 shows the scatterplot of all combinations of the parameters with zero ratio (see Equation 15) on the horizontal axis and the fit ratio on the vertical axis. The points in the figure are colored according to the number of dimensions of these solutions. The isolines correspond to a fixed value of the sum of squared fit and zero ratios, for example, the thick isoline going from a fit ratio of 1 to the zero ratio of 1 is the locus of the sum of these two squared ratios equal to 1. The closest solutions to the upper right corner (which matches a fit and a zero ratio of 1) is the optimal one with the largest sparsity index. In Figure 3, this solution is indicated by the arrow and the value of its sparsity index. In this analysis, the optimal sparsity index is equal to .47 and occurs for an $L = 2$ factor solution with a fit ratio equal to .72, a zero ratio equals .66, and sparsity parameters for rows being $\mathbf{s}_1 \approx (10.94, 10.94)$ and for columns being $\mathbf{s}_2 \approx (13.37, 13.37)$.

Figure 4 shows the values taken by the sparsity parameter on a map where the horizontal axis corresponds to the 20 values chosen between $\frac{1}{I}$ and $\mathbf{s}_1$ and the vertical axis corresponds to the 20 values chosen between $\frac{1}{J}$ and $\mathbf{s}_2$. The values of $\mathbf{s}_1$ (resp. $\mathbf{s}_2$) are scaled by I (resp. J) so that the range of these possible sparsity parameters becomes between 0 and 1. In Figure 4, the optimal solution, which has the largest sparsity index is identified by the star.

5.3. SPARSE CA RESULTS

The results from sparse CA are shown in Figures 5 to 8. With sparsification, the words that have small contributions in plain CA now have zero contribution and the words with large contributions now have even larger contributions—a pattern shown in Figure 5 which plots, for the rows, the plain CA contributions (ordered from left to right by their factor scores) versus on the bottom the sparse CA.

A similar pattern, but with with a smaller effect of sparsity, is found for the contributions of the books (see Figure 6). The factor scores are shown in Figures 7 and 8 with words and books that have zero contributions on both dimensions indicated by hollow dots. The first dimension of sparse CA explains 17% of the inertia. Similar to plain CA, the first dimension differentiates neutral pronouns (e.g., itself, their, us, human, this) and words (e.g., understanding, pure, science, ideas) from words of feminine figures (e.g., girl, she, herself, lady) and words describing emotions (e.g., miss, smiled, laughed, sorry; see Table 2a and the horizontal axis in Figure 7).

Patterns similar to plain CA are also found for the column factor scores of sparse CA, which differentiates the Philosophy genre from Love stories, Mystery,

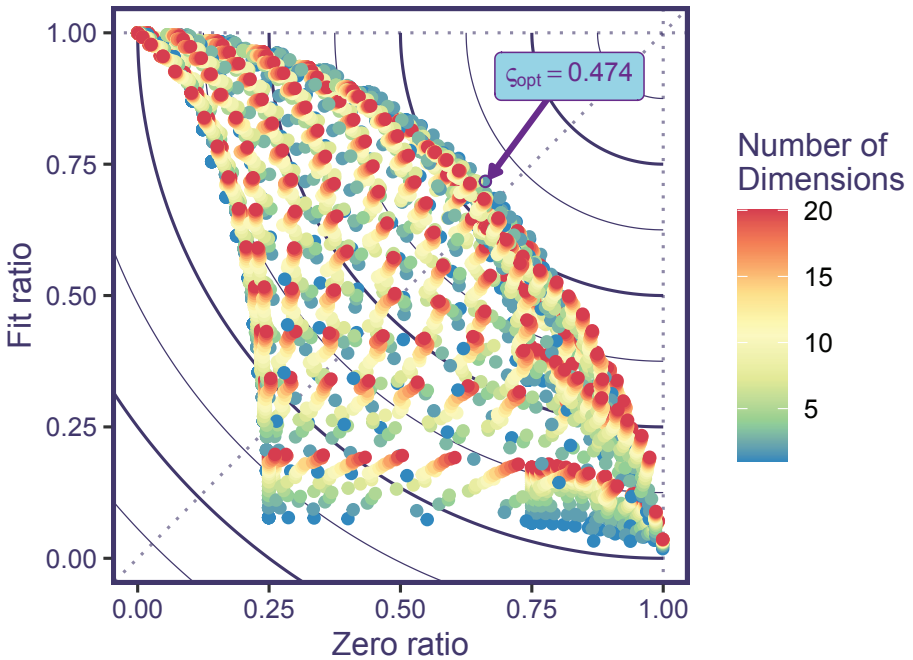


Figure 3: Fit ratio to zero ratio graph.

and Science fiction (see the horizontal axis in Figure 8). Books with a small contribution in plain CA have a zero contribution with sparse CA and books with a large contributions in CA now have even larger contributions in sparse CA—a pattern shown in Figure 6 which plots, for the books, the plain CA contributions (ordered from left to right by their factor scores) versus, on the bottom, the sparse CA.

The second dimension of sparse CA explains 5% of the inertia. The row factor scores are characterized by the opposition (on the positive side the dimension) of words related to war (e.g., enemy, battle, attack, war), geography (e.g., south,

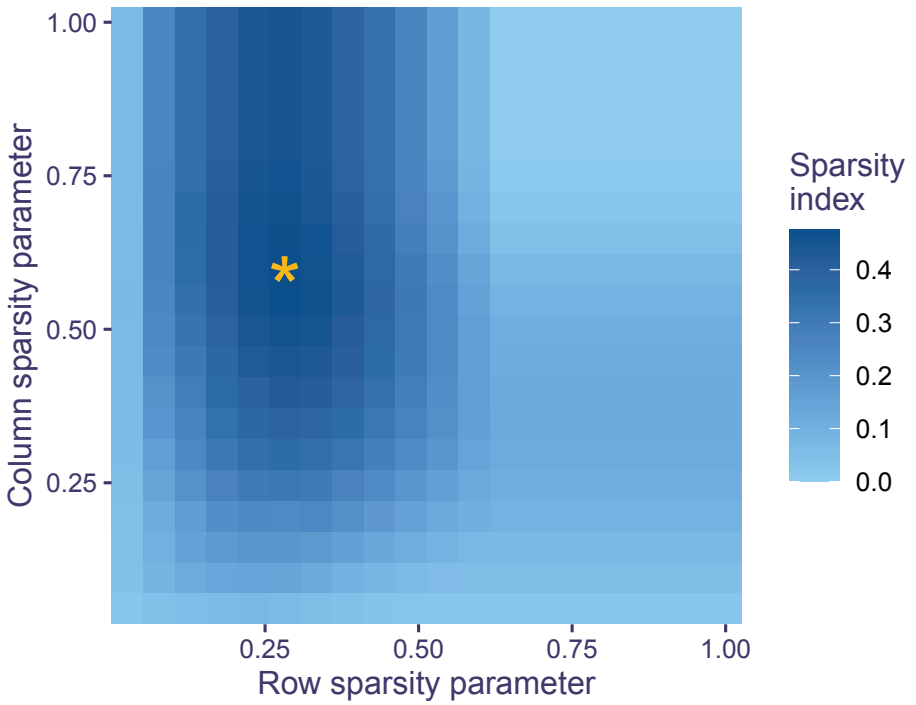


Figure 4: Sparsity Index map.

north, east, west, miles, city), and verbs in the past tense (e.g., ordered, united, received) to (on the negative side the dimension) words related to thoughts (e.g., understanding, experience, reason, meaning, think) and feminine figures (e.g., she, girl, herself, lady, her), and verbs in the present tense (e.g., miss, is, does, say, be, can, am) (see the vertical axis in Figure 7 and Table 2b).

The column factor scores differentiate books from the Biography genre from books of Philosophy, Love stories, and Mystery (see the vertical axis in Figure 8 and bottom panel of Figure 6).

Compared to the results from plain CA, because sparse CA shrunk some words

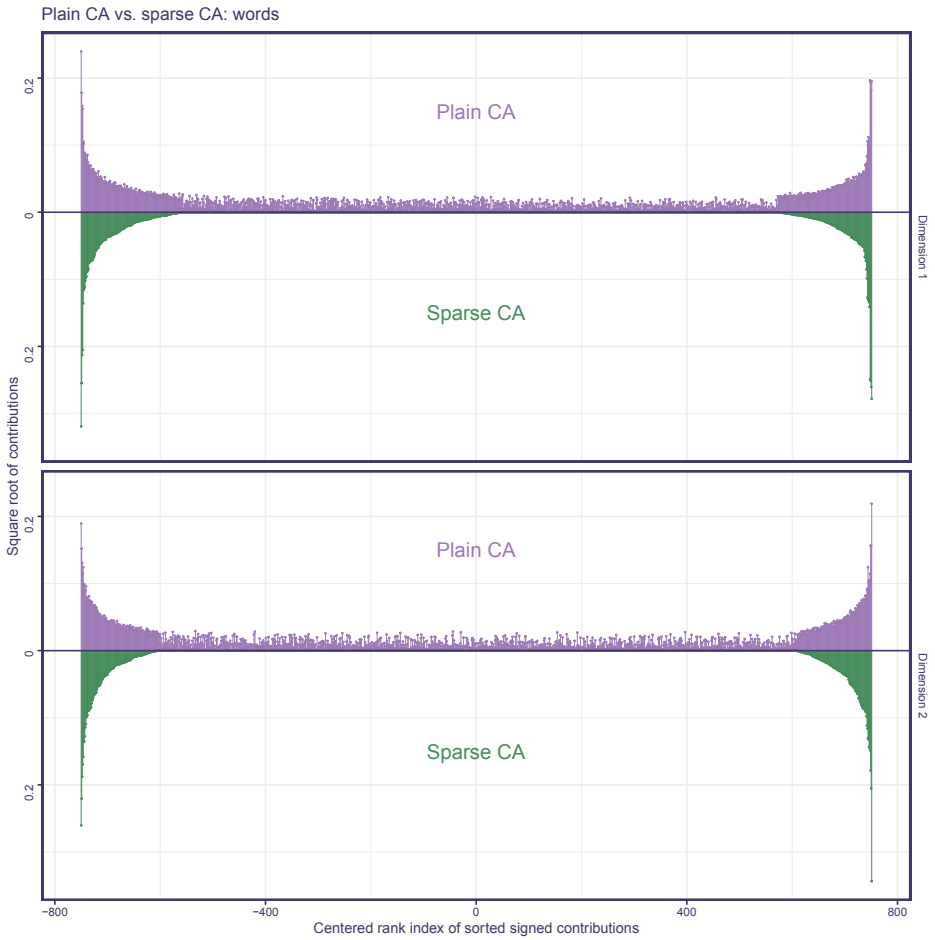


Figure 5: Plain CA vs. Sparse CA: Row contributions

while emphasizing others, the pattern opposing neutral versus feminine pronouns along Dimension 1 becomes more noticeable. Compared to the results from plain CA, because the transition formula is no longer valid in sparse CA, the sparsity of the contributions (derived from loadings; see Figures 5 and 6) does not propagate to give sparse factor scores. But, as demonstrated, the sparsity of contributions can be integrated to facilitate the interpretation of factor scores. Moreover, although the results from sparse CA do not have the optimal proportion of explained inertia, sparse CA gives the solution with the optimal trade-off between the inertia explained and sparsity. Finally, it is worth noting that because the sparse CA

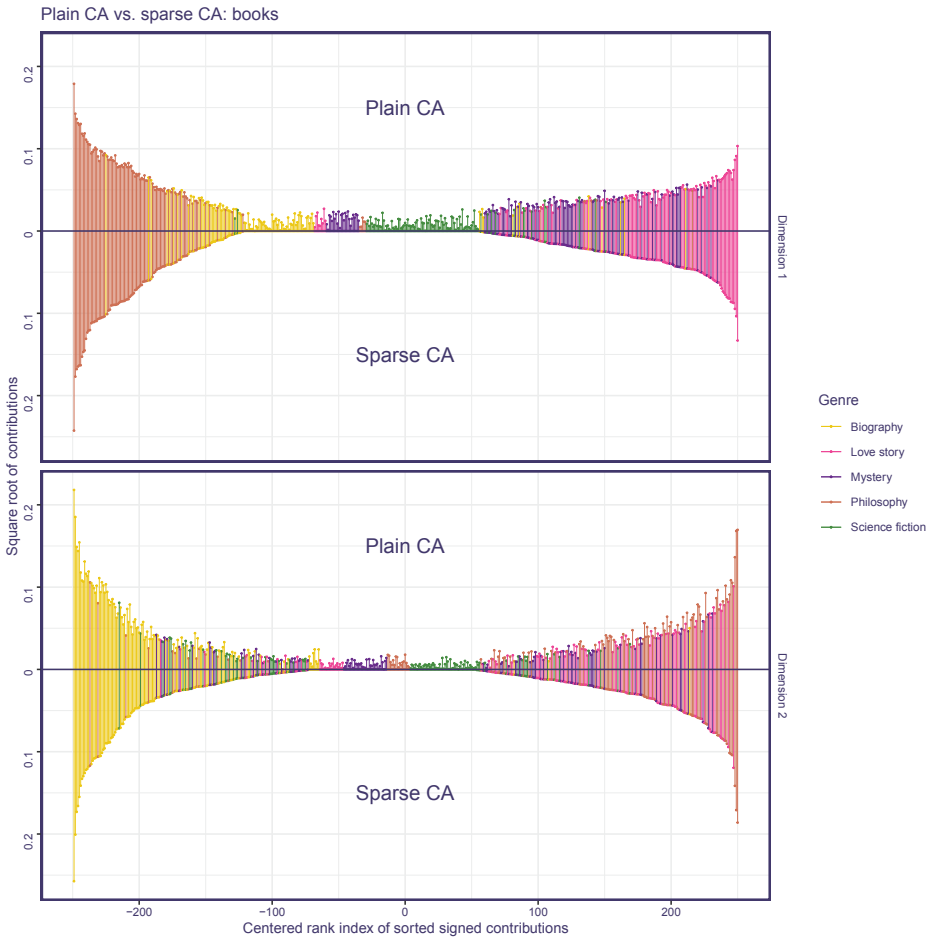


Figure 6: Plain CA vs. Sparse CA: Column contributions

factor scores of the two dimensions are not orthogonal, their percentages of explained inertia are not additive and need to be considered separately. However, here although the components from sparse CA are not orthogonal, the two selected dimensions are close-to-orthogonal with a correlation coefficient of .02.

words	loadings	contributions*	factor scores
science	−4.22	60	−1.40
ideas	−3.80	38	−1.26
understanding	−3.71	21	−1.23
existence	−3.68	39	−1.22
system	−3.54	28	−1.17
pure	−3.47	20	−1.15
ii	−3.43	27	−1.14
experience	−3.31	36	−1.10
physical	−3.30	16	−1.09
material	−3.30	15	−1.09
nature	−3.25	79	−1.08
knowledge	−3.24	49	−1.07
iii	−3.24	15	−1.07
itself	−3.17	64	−1.05
according	−3.16	22	−1.05
example	−3.09	14	−1.02
parts	−3.06	17	−1.01
series	−3.05	10	−1.01
progress	−3.02	15	−1.00
object	−2.94	26	−0.98
stepped	2.11	3	0.70
sat	2.12	14	0.70
box	2.13	5	0.70
sorry	2.13	6	0.71
smile	2.13	9	0.71
cried	2.13	16	0.71
door	2.15	27	0.71
window	2.17	10	0.72
yes	2.18	30	0.72
guess	2.25	5	0.74
nice	2.25	4	0.74
laughed	2.31	9	0.77
shook	2.33	8	0.77
she	2.35	380	0.78
ca	2.36	14	0.78
whispered	2.36	6	0.78
oh	2.38	31	0.79
miss	2.42	36	0.80
girl	2.48	31	0.82
smiled	2.50	9	0.83

(a) Dimension 1

words	loadings	contributions*	factor scores
understanding	−5.40	4	−0.99
pure	−4.54	3	−0.83
experience	−4.07	5	−0.75
existence	−4.05	4	−0.74
ideas	−3.77	3	−0.69
space	−3.57	2	−0.65
reason	−3.38	6	−0.62
sense	−3.19	4	−0.58
object	−3.16	2	−0.58
science	−3.10	3	−0.57
physical	−3.08	1	−0.56
material	−3.01	1	−0.55
itself	−2.99	5	−0.55
meaning	−2.94	1	−0.54
conscious	−2.94	0	−0.54
does	−2.91	5	−0.53
merely	−2.87	2	−0.53
soul	−2.76	2	−0.50
nature	−2.75	5	−0.50
absolutely	−2.65	0	−0.49
post	3.74	0	0.69
hundred	3.77	4	0.69
report	3.95	1	0.72
english	3.99	4	0.73
city	4.11	4	0.75
miles	4.71	4	0.86
east	4.79	2	0.88
orders	4.80	2	0.88
united	4.98	2	0.91
west	5.11	2	0.94
government	5.34	5	0.98
ordered	5.36	2	0.98
war	5.72	9	1.05
attack	5.83	3	1.07
command	6.05	4	1.11
river	6.27	8	1.15
north	6.47	5	1.19
battle	6.61	5	1.21
south	6.72	5	1.23
enemy	8.42	13	1.54

(b) Dimension 2

Table 1: The 20 most extreme words from each dimension of Plain CA.

Note: The contributions shown as 0 in 1b were too small to be displayed as integers; it is worth noting that this value does not indicate zero contributions. * indicates that the original values were multiplied by 10,000 and rounded to the nearest integer for display purposes.

words	loadings*	contributions*	factor scores
understanding	−0.67	2	−1.45
science	−1.75	9	−1.44
ideas	−1.16	5	−1.32
pure	−0.63	2	−1.30
existence	−1.24	5	−1.29
experience	−1.29	5	−1.19
system	−0.79	2	−1.18
ii	−0.81	2	−1.16
physical	−0.42	1	−1.13
knowledge	−1.82	7	−1.12
material	−0.36	0	−1.11
series	−0.25	0	−1.11
itself	−2.58	10	−1.11
nature	−3.09	12	−1.10
iii	−0.37	0	−1.09
object	−0.99	3	−1.07
space	−0.59	1	−1.06
according	−0.65	1	−1.05
parts	−0.49	1	−1.04
example	−0.33	0	−1.02
sat	0.44	0	0.64
lips	0.18	0	0.64
dear	0.71	1	0.65
smiling	0.00	0	0.65
sorry	0.04	0	0.65
yes	1.17	2	0.65
shook	0.07	0	0.65
ca	0.30	0	0.65
whispered	0.02	0	0.66
smile	0.20	0	0.66
lady	1.01	1	0.67
nice	0.00	0	0.68
laughed	0.15	0	0.68
her	22.60	67	0.70
herself	0.86	1	0.70
smiled	0.13	0	0.73
oh	1.17	2	0.73
girl	1.09	2	0.75
she	23.05	77	0.78
miss	1.65	4	0.82

(a) Dimension 1

words	loadings*	contributions*	factor scores
understanding	−0.95	5	−0.93
pure	−0.78	3	−0.76
experience	−1.58	7	−0.67
space	−0.75	2	−0.64
existence	−1.28	5	−0.63
ideas	−1.01	3	−0.57
reason	−2.28	9	−0.57
object	−0.97	3	−0.53
sense	−1.31	4	−0.52
miss	−1.46	3	−0.51
does	−2.27	7	−0.49
mean	−0.93	2	−0.48
meaning	−0.17	0	−0.48
conscious	−0.05	0	−0.46
physical	−0.23	0	−0.46
merely	−0.57	1	−0.45
material	−0.17	0	−0.45
things	−2.88	8	−0.44
nice	−0.00	0	−0.44
absolutely	−0.05	0	−0.43
advance	0.17	0	0.67
report	0.16	0	0.72
city	0.97	3	0.73
general	3.26	16	0.76
english	1.48	7	0.77
miles	0.92	4	0.80
east	0.36	1	0.88
orders	0.57	2	0.92
west	0.54	2	0.94
united	0.56	2	0.95
ordered	0.51	2	1.01
government	1.43	9	1.05
war	2.34	18	1.12
attack	0.69	4	1.13
river	1.65	13	1.17
command	1.05	8	1.22
north	1.04	8	1.23
south	1.11	9	1.29
battle	0.93	7	1.29
enemy	2.32	28	1.67

(b) Dimension 2

Table 2: The 20 most extreme words from each dimension of Sparse CA.

Note: The loadings and contributions shown as 0 in these tables were too small to be displayed as integers; it is worth noting that these values do not indicate sparsity. * indicates that the original values were multiplied by 10,000 and rounded to the nearest integer for display purposes.

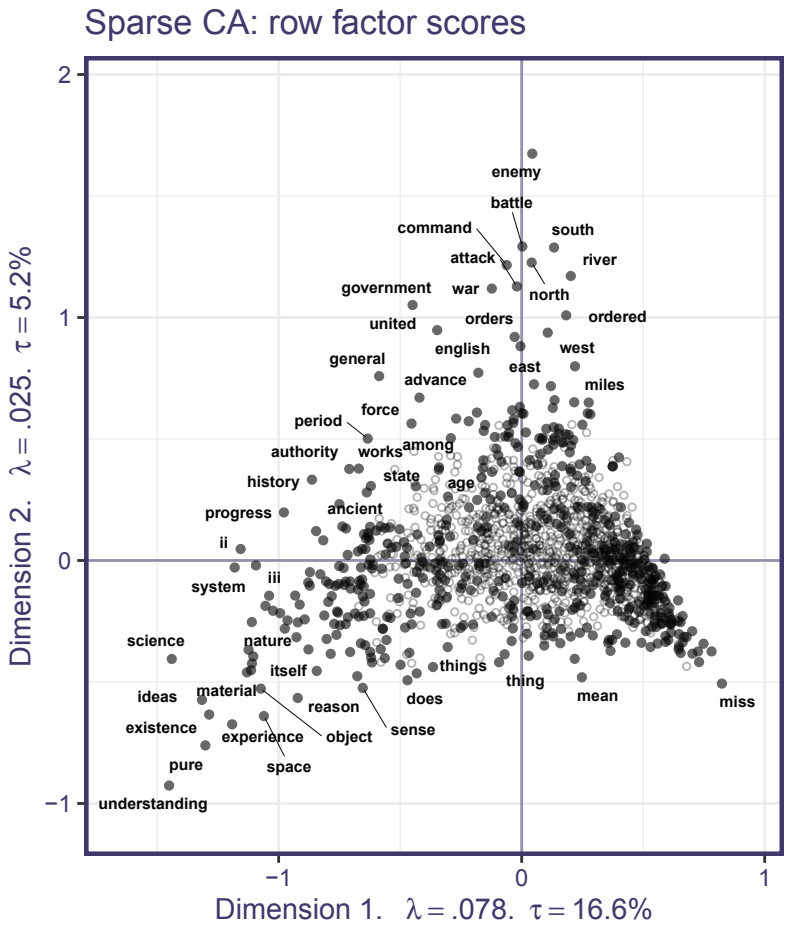


Figure 7: Sparse CA: Row factor scores

6. CONCLUSION AND PERSPECTIVES

In this paper, we extended sparse correspondence analysis developed by Liu et al. (2023) by adding a new global algorithm that identifies the optimal sparsity solution by determining both the optimal sparsity tuning parameters and the optimal number of kept dimensions. Specifically, by integrating this global algorithm, this new version of sparse CA estimates the optimal solution in a more analytic and objective way. Sparse correspondence analysis simplifies the interpretation in the analysis of large tables by highlighting important categories and obtaining simple

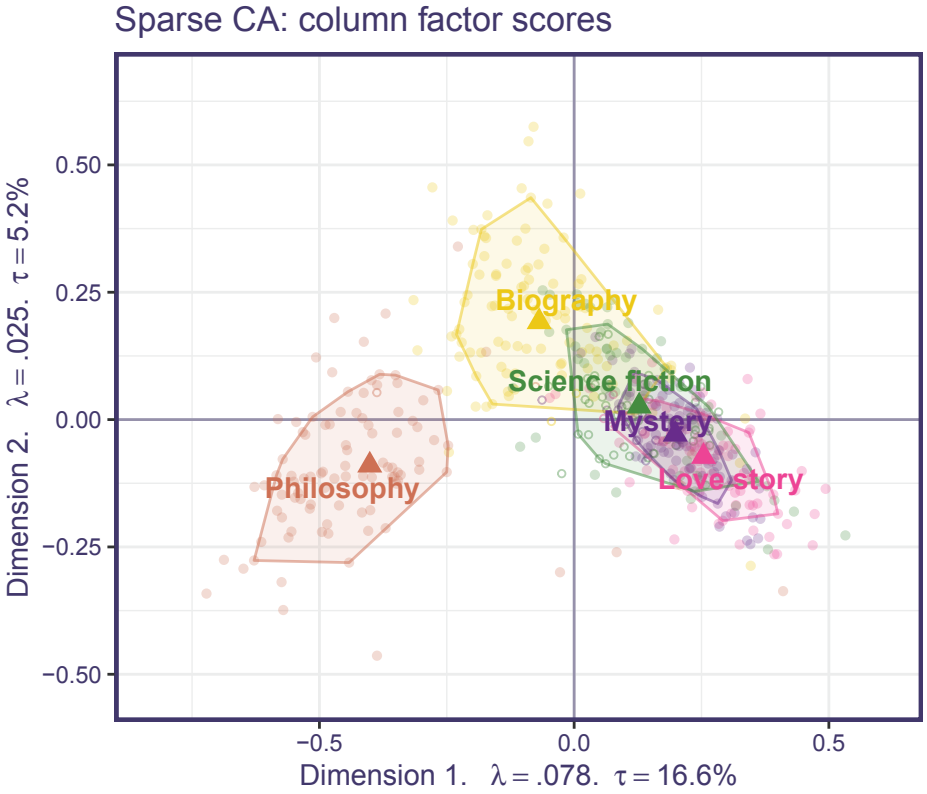


Figure 8: Sparse CA: Column factor scores

successive dimensions in the spirit of the simple structure of factor analysis. Its practical application raises new problems such as the choice of the optimal level of sparsity for rows and or columns, which could be different according to each dimension.

Another concern is the loss of orthogonality of successive dimensions—An issue that should be explored in future work.

Sparse CA remains basically a symmetrical method where rows and columns play the same role. In future work, we also plan to develop sparse variants of the non symmetric correspondence analysis introduced by Lauro and D'Ambra

(1984) and explored by Balbi (1998).

Code and data are available at:
`https://github.com/vguillemot/sparseCorrespondenceAnalysis`.

References

Abdi, H. (2007). Singular value decomposition (SVD) and generalized singular value decomposition (GSVD). In N. Salkind, ed., *Encyclopedia of Measurement and Statistics*, 907–912. Sage Publications, Thousand Oaks.

Abdi, H. and Béra, M. (2018). Correspondence analysis. In R. Alhajj and J. Rokne, eds., *Encyclopedia of Social Networks and Mining (2nd Edition)*, 1–12. Springer, New York. doi:10.1007/978-1-4614-7163-9_140-2.

Abdi, H., Dunlop, J., and Williams, L.J. (2009). How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the bootstrap and 3-way multidimensional scaling (DISTATIS). In *NeuroImage*, 45 (1): 89–95. doi:10.1016/j.neuroimage.2008.11.008.

Abdi, H. and Williams, L.J. (2010). Principal component analysis. In *WIREs Computational Statistics*, 2 (4): 433–459. doi:https://doi.org/10.1002/wics.101.

Allen, G.I., Grosenick, L. and Taylor, J. (2014). A generalized least-square matrix decomposition. In *Journal of the American Statistical Association*, 109 (505): 145–159. doi:10.1080/01621459.2013.852978.

Balbi, S. (1998). Graphical displays in non-symmetrical correspondence analysis. In J. Blasius and M.J. Greenacre, eds., *Visualization of Categorical Data*, 297–309. Academic Press, San Diego. doi:10.1016/B978-012299045-8/50023-1.

Beaton, D. (2020). Generalized eigen, singular value, and partial least squares decompositions: The GSVD package. doi:10.48550/ARXIV.2010.14734.

Beh, E.J. and Lombardo, R. (2021). *An Introduction to Correspondence Analysis*. Wiley, London. doi:10.1002/9781119044482.

Benidis, K., Sun, Y., Babu, P., and Palomar, D.P. (2016). Orthogonal sparse PCA and covariance estimation via procrustes reformulation. In *IEEE Transactions on Signal Processing*, 64 (23): 6211–6226. doi:10.1109/TSP.2016.2605073.

Bernard, A., Guinot, C., and Saporta, G. (2012). Sparse principal component analysis for multiblock data and its extension to sparse multiple correspondence analysis. In A. Colubi et al., eds., *Proceedings of the 20th International Conference on Computational Statistics (COMPSTAT 2012)*, 99–106. International Association for Statistical Computing.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. In *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 72 (1): 3–25. doi: 10.1111/j.1467-9868.2009.00723.x.

Dray, S. (2014). Analysing a pair of tables: coinertia analysis and duality. In J. Blasius and M.J. Greenacre, eds., *Visualization and Verbalization of Data*, 289–300. CRC Press, London.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. In *Psychometrika*, 1 (3): 211–218. doi:10.1007/BF02288367.

Escoufier, Y. (2006). Operators related to a data matrix: a survey. In A. Rizzi and M. Vichi, eds., *Proceedings of the 17th International Conference on Computational Statistics (COMPSTAT 2006)*, 285–297. Physica-Verlag, Heidelberg. doi:10.1007/978-3-7908-1709-6_22.

Gerlach, M. and Font-Clos, F. (2020). A standardized project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. In *Entropy*, 22 (1). doi:10.3390/e22010126.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, New York.

Greenacre, M.J. (2010). Correspondence analysis. In *Wiley Interdisciplinary Reviews: Computational Statistics*, 2 (5): 613–619.

Guillemot, V., Beaton, D., Gloaguen, A., Löfstedt, T., Levine, B., Raymond, N., Tenenhaus, A. and Abdi, H. (2019). A constrained singular value decomposition method that integrates sparsity and orthogonality. In *PLOS ONE*, 14 (3): e0211463. doi:10.1371/journal.pone.0211463.

Guillemot, V., Le Borgne, J., Gloaguen, A., Tenenhaus, A., Saporta, G., Chollet, S., Beaton, D. and Abdi, H. (2020). Sparse multiple correspondence analysis. In *52èmes Journées de Statistique*, 830–835. URL `https://pasteur.hal.science/pasteur-03037346/`.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, New York. doi:10.1007/978-0-387-84858-7.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press, Boca Raton. doi:10.1201/b18401.

Holmes, S. (2008). Multivariate data analysis: The French way. In *Probability and Statistics: Essays in Honor of David A. Freedman*, vol. 2, 219–234. Institute of Mathematical Statistics. doi:10.1214/193940307000000455.

Jenatton, R., Audibert, J.Y. and Bach, F. (2011). Structured variable selection with sparsity-inducing norms. In *The Journal of Machine Learning Research*, 12: 2777–2824. doi:10.5555/1953048.2078194.

Jolliffe, I.T. and Cadima, J. (2016). Principal component analysis: A review and recent developments. In *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 374 (2065). doi: 10.1098/rsta.2015.0202.

Jolliffe, I.T., Trendafilov, N.T. and Uddin, M. (2003). A modified principal component technique based on the LASSO. In *Journal of Computational and Graphical Statistics*, 12 (3): 531–547. doi:10.1198/1061860032148.

Lauro, N. and D'Ambra, L. (1984). L'analyse non symétrique des correspondances. In E. Diday, ed., *Data Analysis and Informatics III*, 433–446. Elsevier, North–Holland.

Le Floch, E., Guillemot, V., Frouin, V., Pinel, P., Lalanne, C., Trinchera, L., Tenenhaus, A., Moreno, A., Zilbovicius, M., Bourgeron, T., Dehaene, S., Thirion, B., Poline, J.B. and Duchesnay, E. (2012). Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse partial least squares. In *NeuroImage*, 63 (1): 11–24. doi:10.1016/j.neuroimage.2012.06.061.

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York.

Liu, R., Niang, N., Saporta, G. and Wang, H. (2023). Sparse correspondence analysis for large contingency tables. In *Advances in Data Analysis and Classification*, 17: 1–20. doi:10.1007/s11634-022-00531-5.

Mackey, L. (2009). Deflation methods for sparse PCA. In D. Koller, D. Schuurmans, Y. Bengio and L. Bottou, eds., *Advances in Neural Information Processing Systems*, 21, 1017–1024. Curran Associates, Inc.

Mattei, P.A., Bouveyron, C., and Latouche, P. (2016). Globally sparse probabilistic pca. In A. Gretton and C.C. Robert, eds., *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, vol. 51 of *Proceedings of Machine Learning Research*, 976–984. PMLR, Cadiz, Spain.

Mori, Y., Kuroda, M. and Makino, N. (2016). Sparse multiple correspondence analysis. In Y. Mori, M. Kuroda and N. Makino, eds., *Nonlinear Principal Component Analysis and Its Applications*, 47–56. Springer, New York. doi: 10.1007/978-981-10-0159-8_5.

Ning-min, S. and Jing, L. (2015). A literature survey on high-dimensional sparse principal component analysis. In *International Journal of Database Theory and Application*, 8 (6): 57–74. doi:10.14257/ijdta.2015.8.6.06.

Saporta, G. and Niang-Keita, N. (2006). Correspondence analysis and classification. In M.J. Greenacre and J. Blasius, eds., *Multiple Correspondence Analysis and Related Methods*, 371–392. CRC Press, London. doi:10.1201/9781420011319-19.

Shen, D., Shen, H. and Marron, J.S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. In *Journal of Multivariate Analysis*, 115: 317–333. doi:10.1016/j.jmva.2012.10.007.

Silver, M., Janousova, E., Hua, X., Thompson, P.M., Montana, G. and Alzheimer's Disease Neuroimaging Initiative, T.A.D.N. (2012). Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. In *NeuroImage*, 63 (3): 1681–1694. doi:10.1016/j.neuroimage.2012.08.002.

Takane, Y. (2002). Relationships among various kinds of eigenvalue and singular value decompositions. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano and J. Meulman, eds., *New Developments in Psychometrics*, 45–56. Springer Verlag, Tokyo.

Thurstone, L.L. (1935). *The Vectors of Mind: Multiple Factor Analysis for the Isolation of Primary Traits.* University of Chicago Press. doi:10.1037/10018-000.

Thurstone, L.L. (1947). *Multiple Factor Analysis.* University of Chicago Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. In *Journal of the Royal Statistical Society. Series B (Methodological)*, 58 (1): 267–288. URL http://www.jstor.org/stable/2346178.

Trendafilov, N.T. (2014). From simple structure to sparse components: a review. In *Computational Statistics*, 29 (3–4): 431–454. doi:10.1007/s00180-013-0434-5.

Trendafilov, N.T., Fontanella, S. and Adachi, K. (2017). Sparse exploratory factor analysis. In *Psychometrika*, 82 (3): 778–794. doi:10.1007/s11336-017-9575-8.

Vines, S. (2000). Simple principal components. In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49 (4): 441–451. doi:10.1111/1467-9876.00204.

Witten, D.M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. In *Biostatistics*, 10 (3): 515–534. doi:10.1093/biostatistics/kxp008.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. In *Journal of Computational and Graphical Statistics*, 15 (2): 265–286. doi:10.1198/106186006X113430.

Appendix

A. NOTATIONS

Matrices are denoted in upper case bold letters, vectors are denoted in lowercase bold letters, and their elements are denoted in lowercase italic letters (note that, by default, vectors are column vectors). Matrices, vectors and elements from the same matrix all use the same letter (e.g., $\mathbf{A}$, $\mathbf{a}$, a). The transpose operation is denoted by the superscript $^{\mathsf{T}}$, the inverse operation is denoted by $^{-1}$. The identity matrix is denoted $\mathbf{I}$, vectors or matrices of ones are denoted $\mathbf{1}$, matrices or vectors of zeros are denoted $\mathbf{0}$ (by default, $\mathbf{I}$, $\mathbf{0}$, and $\mathbf{1}$ are conformable with the other terms in a formula). The standard product between two matrices is indicated by juxtaposition (i.e., $\mathbf{XY}$ means $\mathbf{X}$ times $\mathbf{Y}$); the Hadamard product (i.e., element-wise) is denoted by $\odot$ (e.g., $\mathbf{X} \odot \mathbf{Y}$), note that the Hadamard product is defined only between matrices with the same dimensions.

When provided with a square matrix, the diag operator gives a vector that contains the diagonal elements of this matrix. When provided with a vector, the diag operator gives a diagonal matrix with the elements of the vector as the diagonal elements of this matrix. A diagonal matrix is denoted $\mathbf{D}$, and the subscript denotes the vector that stores the diagonal elements, for example, $\mathbf{D}_{\mathbf{a}} = \mathsf{diag}\,(\mathbf{a})$. When provided with a square matrix, the trace operator gives the sum of the diagonal elements of this matrix. For an I by J matrix $\mathbf{X}$ and for $\mathbf{M}$ being a J by J symmetric positive definite matrix, the squared $\mathbf{M}$-norm of $\mathbf{X}$ is denoted $\|\mathbf{X}\|^2_{\mathbf{M}}$ and is computed as:

$$\|\mathbf{X}\|^2_{\mathbf{M}} = \mathsf{trace}\left(\mathbf{X}\mathbf{M}\mathbf{X}^{\mathsf{T}}\right) . \tag{16}$$

When $\mathbf{M}$ is the identity matrix, the $\mathbf{M}$-norm is equal to the square root of the sum of squares of the entries of the matrix and is called the *Frobenius* norm denoted $L_2 = \|\mathbf{X}\|^2_2$. Another useful norm is the sum of the absolute values of the matrix called the L_1 norm.

When describing an optimization problem, the operator $\underset{\mathbf{x}}{\arg\min} f(\mathbf{x})$ searches for the value of $\mathbf{x}$ that minimizes the function $f(\mathbf{x})$, and the operator $\underset{\mathbf{x}}{\arg\max} f(\mathbf{x})$ searches for the value of $\mathbf{x}$ that maximizes the function $f(\mathbf{x})$.

B. THE PLAIN AND GENERALIZED SVD

The singular value decomposition (SVD) and its extension—the generalized singular value decomposition (GSVD, for details on the generalized singular value

decomposition see Abdi 2007; Greenacre 1984; Takane 2002)—are the foundations of most contemporary multivariate statistical approaches.

B.1. THE (PLAIN) SVD

The SVD of an $I \times J$ matrix $\mathbf{X}$ solves the following maximization problem (Eckart and Young, 1936): Find a matrix, denoted $\widehat{\mathbf{X}}_L$, of rank L [with $L < \min(I,J)$] equal to

$$\widehat{\mathbf{X}}_L = \sum_{\ell=1}^{L} \delta_\ell \mathbf{u}_\ell \mathbf{v}_\ell^\mathsf{T} = \mathbf{U}_L \boldsymbol{\Delta}_L \mathbf{V}_L^\mathsf{T} \text{ with } \mathbf{U}_L^\mathsf{T}\mathbf{U}_L = \mathbf{V}_L^\mathsf{T}\mathbf{V}_L = \mathbf{I} \text{ and } \boldsymbol{\Delta}_L = \mathsf{diag}\,(\boldsymbol{\delta}_L) \quad (17)$$

where $\mathbf{U}$ is the $I \times L$ matrix containing the left singular vectors, $\mathbf{V}$ is the $J \times L$ matrix containing the right singular vectors, and $\boldsymbol{\Delta}$ the $L \times L$ diagonal matrix containing the singular values $\delta_1 \geq \cdots \geq \delta_L \geq 0$, and such that $\widehat{\mathbf{X}}_L$ is the L rank matrix closest to $\mathbf{X}$. Specifically, $\widehat{\mathbf{X}}_L$ solves the following minimization problem:

$$\underset{\mathbf{U}_L,\boldsymbol{\Delta}_L,\mathbf{V}_L}{\arg\min} \|\mathbf{X} - \widehat{\mathbf{X}}_L\|_2^2 = \underset{\mathbf{U}_L,\boldsymbol{\Delta}_L,\mathbf{V}_L}{\arg\min} \|\mathbf{X} - \mathbf{U}_L\boldsymbol{\Delta}_L\mathbf{V}_L^\mathsf{T}\|_2^2 \quad \text{with} \quad \mathbf{U}_L^\mathsf{T}\mathbf{U}_L = \mathbf{V}_L^\mathsf{T}\mathbf{V}_L = \mathbf{I}, \quad (18)$$

When L is equal to the rank of $\mathbf{X}$, the SVD of $\mathbf{X}$ is called the *complete* SVD (when unspecified, the SVD is the complete SVD), in this case, matrices $\mathbf{U}$ and $\mathbf{V}$ are written without their L index. When L is smaller than the rank of $\mathbf{X}$, its SVD is called the *truncated* SVD of $\mathbf{X}$.

The SVD of a matrix can be computed by first computing its rank one approximation [i.e., the singular triplet (δ_1, $\mathbf{u}_1$, $\mathbf{v}_1$)] and then subtracting this rank one approximation from $\mathbf{X}$—a procedure called a *deflation*. The first singular triplet of the deflated matrix $\mathbf{X}$ is obtained then the second singular triplet of $\mathbf{X}$. These procedure can then be continued till completion of the SVD of $\mathbf{X}$.

B.2. GENERALIZED SVD

The generalized SVD (GSVD), differs from the plain SVD by incorporating different orthogonality constraints on the singular vectors. Specifically, with $\mathbf{M}$ being an $I \times I$ positive definite matrix (called the row *metric* matrix) and $\mathbf{W}$ a $J \times J$ positive definite matrix (called the column metric matrix); the GSVD of $\mathbf{X}$ solves the following problem (compare with Equation 18): Specifically, $\widehat{\mathbf{X}}_L$ solves

$$\underset{\mathbf{P}_L,\boldsymbol{\Delta}_L,\mathbf{Q}_L}{\arg\min} \|\mathbf{X} - \widehat{\mathbf{X}}_L\|_2^2 = \underset{\mathbf{P}_L,\boldsymbol{\Delta}_L,\mathbf{Q}_L}{\arg\min} \|\mathbf{X} - \mathbf{P}_L\boldsymbol{\Delta}_L\mathbf{Q}_L^\mathsf{T}\|_2^2 \quad (19)$$

with

$$\mathbf{P}_L^{\mathsf{T}}\mathbf{M}\mathbf{P}_L = \mathbf{Q}_L^{\mathsf{T}}\mathbf{W}\mathbf{Q}_L = \mathbf{I}, \boldsymbol{\Delta}_L = \mathsf{diag}\left(\delta_L\right). \tag{20}$$

where $\mathbf{P}_L$ is the $I \times L$ matrix containing the *generalized* left singular vectors, $\mathbf{Q}_L$ is the $J \times L$ matrix containing the generalized right singular vectors, and $\boldsymbol{\Delta}_L$ is the diagonal matrix of the generalized singular values.

Similarly to the plain SVD, the optimal rank-L approximation of $\mathbf{X}$ is obtained by $\widehat{\mathbf{X}}_L$ (i.e., the L-truncated GSVD of $\mathbf{X}$) as:

$$\widehat{\mathbf{X}}_L = \sum_{\ell=1}^{L} \delta_\ell \mathbf{p}_\ell \mathbf{q}_\ell^{\mathsf{T}} = \mathbf{P}_L \boldsymbol{\Delta}_L \mathbf{Q}_L^{\mathsf{T}}. \tag{21}$$

B.3. GENERALIZED SVD AND FROM PLAIN SVD

In practice, the GSVD matrix $\mathbf{X}$ can be obtained from a plain SVD of a matrix denoted $\widetilde{\mathbf{X}}$ obtained by first pre- and post-multiplying $\mathbf{X}$ by the square root of the row and column metric matrices:

$$\widetilde{\mathbf{X}} = \mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}^{\frac{1}{2}} . \tag{22}$$

Matrix $\widetilde{\mathbf{X}}$ is then decomposed with a plain SVD as:

$$\widetilde{\mathbf{X}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathsf{T}} \quad \text{such that} \quad \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}. \tag{23}$$

The generalized singular vectors of $\mathbf{X}$ are then obtained from the (plain) singular vectors of $\widetilde{\mathbf{X}}$ as

$$\mathbf{P} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U} \quad \text{and} \quad \mathbf{Q} = \mathbf{W}^{-\frac{1}{2}}\mathbf{V}. \tag{24}$$

The constraints from Equations 18 and 19 are equivalent because

$$\begin{aligned} \mathbf{P}^{\mathsf{T}}\mathbf{M}\mathbf{P} &= \mathbf{U}^{\mathsf{T}}\mathbf{M}^{-\frac{1}{2}}\mathbf{M}\mathbf{M}^{-\frac{1}{2}}\mathbf{U} \\ &= \mathbf{U}^{\mathsf{T}}\mathbf{U} \\ &= \mathbf{I}, \end{aligned} \quad \text{and} \quad \begin{aligned} \mathbf{Q}^{\mathsf{T}}\mathbf{W}\mathbf{Q} &= \mathbf{V}^{\mathsf{T}}\mathbf{W}^{-\frac{1}{2}}\mathbf{W}\mathbf{W}^{-\frac{1}{2}}\mathbf{V} \\ &= \mathbf{V}^{\mathsf{T}}\mathbf{V} \\ &= \mathbf{I}. \end{aligned} \tag{25}$$

Finally, the decomposition of $\mathbf{X}$ from Equations 18 and 19 are also equivalent because

$$\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathsf{T}}\mathbf{W}^{-\frac{1}{2}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{M}^{\frac{1}{2}}\mathbf{X}\mathbf{W}^{\frac{1}{2}}\mathbf{W}^{-\frac{1}{2}} = \mathbf{X} . \tag{26}$$

C. PLAIN CORRESPONDENCE ANALYSIS

Just like most multivariate methods, CA can be interpreted as an optimization problem (actually, as several equivalent optimization problems). But, in order to sparsify CA, we will need to add more constraints to its standard GSVD optimization problem. These new constraints can, in some cases, conflict with the original optimization problem and therefore, as a trade-off, some of the essential properties of CA could be relaxed or even lost. To facilitate the evaluation of this trade-off, we list below the relevant basic equations for CA along with its essential properties (for more details see, e.g., Abdi and Béra, 2018; Abdi and Williams, 2010; Beh and Lombardo, 2021; Greenacre, 1984; Lebart et al. 1984; or Saporta and Niang-Keita, 2006).

C.1. THE BASIC EQUATIONS OF CORRESPONDENCE ANALYSIS

Correspondence analysis was originally developed to analyze the pattern of deviations from independence (as measured by a χ^2 statistic) in a contingency table (see Abdi and Béra, 2018). CA provides, for both rows and columns, a set of factor scores whose total inertia is proportional to the independence χ^2 computed on the original contingency table.

The contingency table to be analyzed is stored in an I rows by J columns matrix denoted $\mathbf{X}$, whose generic element $x_{i,j}$ gives the number of observations that belongs to the ith level of the first nominal variable (i.e., the rows) andthe jth level of the second nominal variable (i.e., the columns). The grand total of the table is denoted N.

The matrix $\mathbf{X}$ is first transformed into a probability matrix (i.e., a matrix comprising non-negative numbers and whose sum is equal to one) denoted $\mathbf{Z}$ and computed as $\mathbf{Z} = N^{-1}\mathbf{X}$. We denote: $\mathbf{r}$ the I by 1 vector of the row totals of $\mathbf{Z}$ and by r_i the ith element of $\mathbf{r}$ (i.e., $\mathbf{r} = \mathbf{Z1}$, with $\mathbf{1}$ being a conformable vector of 1's); $\mathbf{c}$ the J by 1 vector of the columns totals, by c_j the jth element of $\mathbf{c}$ (i.e., $\mathbf{c} = \mathbf{Z}^{\mathsf{T}}\mathbf{1}$); and $\mathbf{D_c} = \mathsf{diag}\,(\mathbf{c})$, $\mathbf{D_r} = \mathsf{diag}\,(\mathbf{r})$ the diagonal matrices obtained from (respectively) $\mathbf{r}$ and $\mathbf{c}$; these two diagonal matrices are called (respectively) row and column *mass* matrices. We denote by $\mathbf{R} = \mathbf{D_r}^{-1}\mathbf{Z}$ (respectively $\mathbf{C} = \mathbf{D_c}^{-1}\mathbf{Z}^{\mathsf{T}}$) the row (respectively column) profile matrix (i.e., all elements are not negative, rows of $\mathbf{R}$ and columns of $\mathbf{C}$ sum to 1).

The factor scores are obtained from the following generalized singular value decomposition (cf. Equation 19) where the metric matrices $\mathbf{D_r}^{-1}$ and $\mathbf{D_c}^{-1}$ are

called χ^2-metric matrices (Greenacre, 2010)

$$\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \text{ with } \mathbf{P}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P} = \mathbf{Q}^{\mathsf{T}}\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q} = \mathbf{I}\,. \tag{27}$$

Here, by subtracting the (rank one) matrix $\mathbf{r}\mathbf{c}^{\mathsf{T}}$ from the probability matrix $\mathbf{Z}$, the decomposed matrix: $\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)$ is *double-centered* because now all rows and columns have zero means. In addition, this double centering of matrix $\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}$ propagates to the singular vectors.

For example, to show that matrix $\mathbf{Q}$ has zero mean, we compute the column means of $\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}$ and replace it by its GSVD from Equation 27 to get

$$\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)\mathbf{1} = \mathbf{0} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}}\mathbf{1} \Longrightarrow \mathbf{P}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}\Delta\mathbf{Q}^{\mathsf{T}}\mathbf{1} = \mathbf{0} \Longrightarrow \Delta\mathbf{Q}^{\mathsf{T}}\mathbf{1} = \mathbf{0} \Longrightarrow \mathbf{Q}^{\mathsf{T}}\mathbf{1} = \mathbf{0}\,, \tag{28}$$

where $\mathbf{1}$ is a J by 1 vector of 1s.

The squared singular values are called *eigenvalues* (denoted λ_k) and are stored into the diagonal matrix $\boldsymbol{\Lambda}$. The sum of the eigenvalues gives the total inertia (denoted $\mathscr{I}$ or φ^2 and equal to χ^2/N) of $\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)$. With the so-called "triplet notation," (Dray, 2014; Escoufier, 2006; Holmes, 2008)—sometimes used as a general framework to formalize multivariate techniques—CA is equivalent to the analysis of the triplet $\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}, \mathbf{D}_{\mathbf{c}}^{-1}, \mathbf{D}_{\mathbf{r}}^{-1}\right)$. From this GSVD, the principal row and (respectively) column factor scores are obtained as

$$\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}\boldsymbol{\Delta} \qquad \text{and} \qquad \mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}\boldsymbol{\Delta}\,. \tag{29}$$

Note that the inertia of each dimension (i.e., each column of $\mathbf{F}$ and $\mathbf{G}$) is equal to its eigenvalue and that factor scores corresponding to different eigenvalues are orthogonal (under the constraints imposed by their masses). Specifically:

$$\lambda_\ell = \sum_{i=1}^{I} r_i f_{i,\ell}^2 = \sum_{j=1}^{J} c_j g_{j,\ell}^2 \text{ and } \sum_{i=1}^{I} r_i f_{i,\ell} f_{i,\ell'} = \sum_{j=1}^{J} c_j g_{j,\ell} g_{j,\ell'} = 0 \quad \forall \ell \neq \ell' \tag{30}$$

or, in matrix notations:

$$\mathbf{F}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}\mathbf{F} = \boldsymbol{\Lambda} \text{ and } \mathbf{G}^{\mathsf{T}}\mathbf{D}_{\mathbf{c}}\mathbf{G} = \boldsymbol{\Lambda}\,, \tag{31}$$

where $\mathbf{D}_{\mathbf{r}}$ and $\mathbf{D}_{\mathbf{c}}$ are called (respectively) the row and column *mass* matrices. This equality can be directly derived from Equations 27 and 29 (here illustrated for $\mathbf{F}$):

$$\mathbf{F}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}\mathbf{F} = \boldsymbol{\Delta}\mathbf{P}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{D}_{\mathbf{r}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}\boldsymbol{\Delta} = \boldsymbol{\Delta}\mathbf{P}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}\boldsymbol{\Delta} = \boldsymbol{\Delta}^2 = \boldsymbol{\Lambda}\,. \tag{32}$$

By contrast with the *principal* factor scores whose $\mathbf{D}_{\mathbf{r}}$ and $\mathbf{D}_{\mathbf{c}}$ norms are equal to the singular values, the *standard* factor scores (indicated by a superscript *) have $\mathbf{D}_{\mathbf{r}}$ and $\mathbf{D}_{\mathbf{c}}$ norms equal to one, and are computed as

$$\mathbf{F}^* = \mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P} \qquad \text{and} \qquad \mathbf{G}^* = \mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}\,. \tag{33}$$

C.2. CORRESPONDENCE ANALYSIS FROM A PLAIN SVD

Correspondence analysis can also be obtained from both the plain SVD and the GSVD (for details, see, e.g., Abdi 2007, and Beaton 2020). Specifically, generalized singular vectors and values and factor scores can be obtained by the following plain SVD:

$$\widetilde{\mathbf{Z}} = \mathbf{D_r}^{-\frac{1}{2}} \left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}}\right) \mathbf{D_c}^{-\frac{1}{2}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{V}^{\mathsf{T}} \tag{34}$$

which, in turn, gives the generalized singular vectors as

$$\mathbf{P} = \mathbf{D_r}^{\frac{1}{2}}\mathbf{U} \quad \text{and} \quad \mathbf{Q} = \mathbf{D_c}^{\frac{1}{2}}\mathbf{V}\,. \tag{35}$$

Finally, transposing this last equation in Equation 29 gives:

$$\mathbf{F} = \mathbf{D_r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{D_r}^{-\frac{1}{2}}\mathbf{U}\boldsymbol{\Delta} \qquad \text{and} \qquad \mathbf{G} = \mathbf{D_c}^{-1}\mathbf{Q}\boldsymbol{\Delta} = \mathbf{D_c}^{-\frac{1}{2}}\mathbf{V}\boldsymbol{\Delta}\,. \tag{36}$$

As indicated by Equation 30, the inertia of a dimension is the sum of the inertia of either all the rows or all the columns, therefore a convenient way of evaluating the importance of a row (respectively a column) is to compute the proportion accounted by a given row (respectively column) into this total. This index, called the *contribution* of a row (respectively a column) is denoted $t_{i,\ell}$ (respectively $t_{j,\ell}$ for a column) and is computed as

$$t_{i,\ell} = \frac{r_i f_{i,\ell}^2}{\sum_{i'=1}^{I} r_{i'} f_{i',\ell}^2} = r_i f_{i,\ell}^2 \lambda_\ell^{-1} \quad \text{and} \quad t_{j,\ell} = \frac{c_j g_{j,\ell}^2}{\sum_{j'=1}^{J} c_{j'} g_{j',\ell}^2} = c_j g_{j,\ell}^2 \lambda_\ell^{-1} \tag{37}$$

In matrix notations, the row (respectively columns) contributions are stored in the matrix $\mathbf{T_I}$ (respectively $\mathbf{T_J}$) computed as:

$$\mathbf{T_I} = \mathbf{D_r}\left(\mathbf{F} \odot \mathbf{F}\right)\boldsymbol{\Lambda}^{-1} = \mathbf{D_r}^{-1}\left(\mathbf{P} \odot \mathbf{P}\right) \text{ and } \mathbf{T_J} = \mathbf{D_c}\left(\mathbf{G} \odot \mathbf{G}\right)\boldsymbol{\Lambda}^{-1} = \mathbf{D_c}^{-1}\left(\mathbf{Q} \odot \mathbf{Q}\right). \tag{38}$$

Note that contributions can be obtained in two equivalent ways: from the factor scores or from the generalized singular vectors.

To facilitate the interpretation of a given dimension, to interpret a dimension we, traditionally, use only the items whose contribution is larger than their mass (i.e., r_i or c_j). The contributions are also often plotted according to the sign of their corresponding factor scores and are then called *signed* contributions.

C.3. IMPORTANT PROPERTIES OF CORRESPONDENCE ANALYSIS

In this section we list some important properties of correspondence analysis relevant for sparsification.

C.3.1. INERTIA AND χ^2

The inertia (i.e., $\mathscr{I}$ or equivalently φ^2) of the centered matrix $\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)$—as obtained from Equation 27—is equal to the independence χ^2 divided by N. Recall that, with the present notations, $\chi^2/N = \varphi^2$ is computed as

$$\varphi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(z_{i,j}-r_i c_j\right)^2}{r_i c_j} = \mathsf{trace}\left(\mathbf{D_c}^{-\frac{1}{2}}\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)^{\mathsf{T}}\mathbf{D_r}^{-1}\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)\mathbf{D_c}^{-\frac{1}{2}}\right). \quad (39)$$

To show that φ^2 is equal to the sum of the eigenvalues from Equation 27, suffice to plug the singular values decomposition from Equation 27 into Equation 39 to get:

$$\varphi^2 = \mathsf{trace}\left(\mathbf{D_c}^{-\frac{1}{2}}\left(\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}}\right)^{\mathsf{T}}\mathbf{D_r}^{-1}\left(\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}}\right)\mathbf{D_c}^{-\frac{1}{2}}\right). \quad (40)$$

Using the properties of the trace operator and re-arranging shows that φ^2 is equal to the sum of the eigenvalues of $\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}$, namely that:

$$\varphi^2 = \mathsf{trace}\left(\boldsymbol{\Delta}\mathbf{P}^{\mathsf{T}}\mathbf{D_c}^{-1}\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}}\mathbf{D_c}^{-1}\mathbf{Q}\right) = \mathsf{trace}\left(\boldsymbol{\Delta}^2\right) = \mathsf{trace}\left(\boldsymbol{\Lambda}\right) = \sum_{\ell=1}^{L}\lambda_\ell\,. \quad (41)$$

which shows, as stated, that $\varphi^2 = \sum\lambda_\ell$.

C.3.2. FACTORS ARE CENTERED

The centering of the singular vectors propagates to the factor scores when the means are computed using the masses stored in the diagonal matrices $\mathbf{D_r}$ (for the rows) and $\mathbf{D_c}$ (for the columns). So, $\overline{\mathbf{F}}$ (respectively $\overline{\mathbf{G}}$), denoting the average row (respectively column) factor scores, is computed as

$$\overline{\mathbf{F}} = \mathbf{1}\mathbf{D_r}\mathbf{F} = \mathbf{1}\mathbf{D_r}\mathbf{D_r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{0} \text{ and } \overline{\mathbf{G}} = \mathbf{1}\mathbf{D_c}\mathbf{G} = \mathbf{1}\mathbf{D_c}\mathbf{D_c}^{-1}\mathbf{Q}\boldsymbol{\Delta} = \mathbf{0} \quad (42)$$

(where $\mathbf{1}$ and $\mathbf{0}$ are conformable vectors of 1s and 0s).

C.3.3. TRANSITION FORMULAS: FROM ROW TO COLUMN FACTOR SCORES AND BACK

In CA the factor scores of one set (e.g., the rows) can be obtained from the profiles of this set and the factor scores of the other set (e.g., the columns). Specifically we have

$$\mathbf{F} = \mathbf{D}_\mathbf{r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{R}\mathbf{G}\boldsymbol{\Delta}^{-1} \text{ and } \mathbf{G} = \mathbf{D}_\mathbf{c}^{-1}\mathbf{Q}\boldsymbol{\Delta} = \mathbf{C}\mathbf{F}\boldsymbol{\Delta}^{-1} . \tag{43}$$

These formulas called *transition formulas* are obtained from Equations 27 and 29; for example, the transition formula for the row factor scores (i.e., from the column factor scores) is derived as (taking into account that $\mathbf{Q}$ is centered)

$$\begin{aligned}
\mathbf{F} &= \mathbf{D}_\mathbf{r}^{-1}\mathbf{P}\boldsymbol{\Delta} = \mathbf{D}_\mathbf{r}^{-1}\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}\right)\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q} \quad \left[\text{because } \mathbf{P}\boldsymbol{\Delta} = \left(\mathbf{Z} - \mathbf{r}\mathbf{c}^\mathsf{T}\right)\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q}\right] \\
&= \mathbf{D}_\mathbf{r}^{-1}\mathbf{Z}\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q} - \mathbf{D}_\mathbf{r}^{-1}\mathbf{r}\mathbf{c}^\mathsf{T}\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q} \\
&= \mathbf{D}_\mathbf{r}^{-1}\mathbf{Z}\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q} \quad \left[\text{because } \mathbf{c}^\mathsf{T}\mathbf{D}_\mathbf{c}^{-1}\mathbf{Q} = \mathbf{1}^\mathsf{T}\mathbf{Q} = \mathbf{0}\right] \\
&= \mathbf{D}_\mathbf{r}^{-1}\mathbf{Z}\mathbf{D}_\mathbf{c}^{-1}\mathbf{D}_\mathbf{c}\mathbf{G}\boldsymbol{\Delta}^{-1} \quad \left[\text{because } \mathbf{Q} = \mathbf{D}_\mathbf{c}\mathbf{G}\boldsymbol{\Delta}^{-1}\right] \\
&= \mathbf{D}_\mathbf{r}^{-1}\mathbf{Z}\mathbf{G}\boldsymbol{\Delta}^{-1} \\
&= \mathbf{R}\mathbf{G}\boldsymbol{\Delta}^{-1}.
\end{aligned} \tag{44}$$

Note that, together, the two transition formulas imply that the eigenvalues in CA cannot be larger than 1.

The transition formulas can be interpreted as a two step process. Using the formula above (for computing $\mathbf{F}$ from $\mathbf{G}$) The first step corresponds to the term $\mathbf{RG}$ and computes the row factor scores as the weighted average (i.e., the *barycenter*) of the column factor scores; the second step corresponds to the term $\boldsymbol{\Delta}^{-1}$ and is an expansion that is inversely proportional to the singular value of each factor (this is an expansion because the singular values being no larger than 1, their inverse is no smaller than 1).

C.3.4. CORRESPONDENCE ANALYSIS AS A DOUBLE PRINCIPAL CORRESPONDENCE ANALYSIS

The row and column factor scores of CA can also be obtained from two different GSVD (or equivalently two weighted PCA), one performed on the row profiles (i.e., the matrix $\mathbf{R}$) and the other one on the column profiles (i.e., the matrix $\mathbf{C}$).

This way, the factor scores are obtained from the GSVD of the matrix of the row profiles matrix (i.e., $\mathbf{R}$) as:

$$\mathbf{D}_{\mathbf{r}}^{-1}\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)=\left(\mathbf{R}-\mathbf{1}\mathbf{c}^{\mathsf{T}}\right)=\mathbf{P}_{\mathbf{R}}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \text{ with } \mathbf{P}_{\mathbf{R}}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}\mathbf{P}_{\mathbf{R}}=\mathbf{Q}^{\mathsf{T}}\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}=\mathbf{I} \quad (45)$$

where $\mathbf{P}_{\mathbf{R}}$ contains the left generalized singular vectors of the row profile matrix $\mathbf{R}$. We can link the decomposition of the row profile matrix to the centered data as:

$$\mathbf{P}_{\mathbf{R}}=\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}, \quad \mathbf{F}=\mathbf{P}_{\mathbf{R}}\boldsymbol{\Delta}=\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}\boldsymbol{\Delta}, \quad \text{and} \quad \mathbf{G}=\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}\boldsymbol{\Delta}\ ; \quad (46)$$

But these factor scores can also be obtained from the GSVD of the matrix of the column profiles (i.e., matrix $\mathbf{C}$) as:

$$\left(\mathbf{Z}-\mathbf{r}\mathbf{c}^{\mathsf{T}}\right)\mathbf{D}_{\mathbf{c}}^{-1}=\left(\mathbf{C}-\mathbf{1}\mathbf{c}^{\mathsf{T}}\right)=\mathbf{P}\boldsymbol{\Delta}\mathbf{Q}_{\mathbf{C}}^{\mathsf{T}} \text{ with } \mathbf{P}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}=\mathbf{Q}_{\mathbf{C}}^{\mathsf{T}}\mathbf{D}_{\mathbf{c}}\mathbf{Q}_{\mathbf{C}}=\mathbf{I} \quad (47)$$

where $\mathbf{Q}_{\mathbf{C}}$ contains the right generalized singular vectors of the column profile matrix $\mathbf{C}$. We can link the decomposition of the column profile matrix to the centered data as:

$$\mathbf{F}=\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{P}\boldsymbol{\Delta}, \quad \mathbf{Q}_{\mathbf{C}}=\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q} \quad \text{and} \quad \mathbf{G}=\mathbf{Q}_{\mathbf{C}}\boldsymbol{\Delta}=\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}\boldsymbol{\Delta}. \quad (48)$$

Within the framework of generalized PCA, Equations 44, 46, and 48 show, together, that the matrices of the principal factor scores can be obtained as linear combinations of the row (respectively column) profile matrix as (respectively):

$$\mathbf{F}=\mathbf{R}\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q} \quad \text{and} \quad \mathbf{G}=\mathbf{C}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}\ . \quad (49)$$

In this framework, the matrix $\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}$ (respectively $\mathbf{D}_{\mathbf{r}}^{-1}\mathbf{P}$) that stores the coefficients of the linear combinations of the columns of $\mathbf{R}$ (respectively $\mathbf{C}^{\mathsf{T}}$) is called the matrix of the *row-weights* (respectively *column-weights*). Note that the weight matrix for one set (e.g., matrix $\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}$ for the rows) is the matrix of the standard coordinates for the other set (i.e., $\mathbf{G}^{*}=\mathbf{D}_{\mathbf{c}}^{-1}\mathbf{Q}$, see Equation 33).

doi.org/10.26398/IJAS.0035-015

MULTILINGUAL TEXTUAL DATA: AN APPROACH THROUGH MULTIPLE FACTOR ANALYSIS

Belchin Kostov
Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Barcelona, Spain

Ramón Alvarez-Esteban[1]
Department of Economics and Statistics, Universidad de León, León, Spain

Mónica Bécue-Bertaut
Department of Statistics and Operational Research, Universitat Politècnica de Catalunya, Barcelona, Spain

François Husson
Institut Agro, Université Rennes 1, CNRS, IRMAR, Rennes, France

Abstract *This paper focuses on the analysis of open-ended questions answered in different languages. Closed-ended questions, called contextual variables, are asked to all respondents in order to understand the relationships between open-ended and closed-ended responses across samples, as the latter are likely to influence word choice. We have developed "Multiple Factor Analysis on Generalised Aggregated Lexical Tables" (MFA-GALT) to examine together open-ended responses in different languages through the relationships between word choice and the variables that drive that choice. MFA-GALT investigates whether the variability between words is structured in the same way as the variability between variables, and vice versa, from one sample to another. An application to an international satisfaction survey shows the easy-to-interpret results proposed.*

Keywords: *Correspondence analysis, Lexical tables, Textual and contextual data, Multiple factor analysis, Generalised aggregated lexical table*

1. INTRODUCTION

Socio-economic surveys benefit from the introduction of open-ended questions alongside closed-ended questions because they are mutually enriching. Closed-ended questions may inform the interpretation of open-ended questions, as the meaning of words is related to the speaker's characteristics or opinions. For example, customers in a satisfaction survey are asked to rate certain aspects of the

[1]Ramón Alvarez-Esteban, ramon.alvarez@unileon.es. ORCID 0000-0002-4751-2797

product and then to give their free opinion on which aspects could be improved, which is clearly linked to the ratings. In a survey that includes the question "What does health mean to you?", closed-ended questions such as gender, age, education and health status are very helpful for exploring how definitions of health vary with these variables. In the case of international surveys, which is our framework, these open-ended questions raise the issue of analysing responses from different samples in different languages.

For a single language, textual statistics (Benzécri, 1981; Lebart et al., 1998) provide multidimensional tools for processing free responses. Separately for each sample, the free responses are coded in the form of respondents $\times$ words, called a lexical table (LT). A standard methodology is to apply correspondence analysis to this LT (CA-LT; direct analysis) and to use the closed information as a complement. It is also common to group the responses of the categories of a closed question (e.g. age crossed with gender or education level, called a contextual variable), and to create a frequency table of words $\times$ categories, known as an aggregated lexical table (ALT), which can also be analysed by CA (CA-ALT).

These approaches are extended to multiple quantitative or qualitative contextual variables by using linearly constrained CA methods (Takane et al., 1991). Balbi and Giordano (2001) deal with textual data including external information; Balbi and Misuraca (2010) propose a double projection strategy by involving external information on both documents and words; while Spano and Triunfo (2012) apply canonical correspondence analysis (CCA; ter Braak (1986, 1987)) to textual data. In line with these works, Bécue-Bertaut et al. (2014) and Bécue-Bertaut and Pagès (2015) propose the CA method on a generalised aggregated lexical table (CA-GALT). The GALT is analysed by means of a CCA adapted to textual data. In CA-GALT, as in any CA, the variability of the vocabulary is explained by the variability of the variables, and the variability of the variables is explained by the variability of the vocabulary. This fits perfectly with the perspective we have chosen here.

In the case of multilingual surveys, we propose to analyse simultaneously the different GALTs, one for each monolingual sample, using a multiple factor analysis (MFA; (Escofier and Pagès, 2016; Pagès, 2014)) adapted to processing a multiple GALT. This produces the Multiple Factor Analysis for Generalised Aggregate Lexical Tables (MFA-GALT). This paper outlines how to adapt MFA reasoning to handle a multiple GALT, and details its properties and graphical representations.

The aim of MFA-GALT is to jointly study the open-ended responses from several samples in different languages through the relationships between the choice

of words and the variables that motivate this choice. These relationships may or may not have similar structures. In other words, MFA-GALT examines whether the variability between words is structured in the same way as the variability between variables, and vice versa, across samples.

The paper is organised as follows: Section 2 presents the data structure and notation; Section 3 recalls the principles of CA-GALT and MFA, the methods that form the basis of our approach; Section 4 is devoted to MFA adapted to multiple GALTs (MFA-GALT); and Section 5 presents the properties of the method. Finally, MFA-GALT is used in a full-scale application (Section 6) to demonstrate its capabilities. The main conclusions are presented in Section 7.

2. DATA STRUCTURE AND NOTATION

L samples answered a questionnaire with closed questions, either quantitative or categorical, all of the same type; these constitute the contextual data. They also answered an open-ended question in different languages, the answers to which are the source of the textual data set. The l sample has I_l respondents who all together use J_l different words in the l language. From these answers we build the $(I_l \times J_l)$ table $\mathbf{Y_l}$, respondents $\times$ words; N_l is the grand total for this table.

The closed questions are common to all samples. The answers are coded in the $(I_l \times K)$ table $\mathbf{X_l}$, whose columns correspond to either quantitative or dummy variables encoding the categories of one or more categorical variables. Regardless of the type, k and K denote the column-variable k and the total number of column-variables respectively. The term *variable* is henceforth used for both types. From $\mathbf{Y_l}$, the proportion table $(I_l \times J_l)$ is calculated $\mathbf{P_l} = \mathbf{Y_l}/N_l$.

If we consider only the sample l, the respondents' weights are taken from the margin of the rows of $\mathbf{P_l}$ — thus proportional to the number of occurrences of words in their free answers — and stored in the $(I_l \times I_l)$ diagonal matrix $\mathbf{D_l}$. The total weight of the respondents belonging to the same sample is equal to 1. The weights of the words are similarly obtained from the margin of the columns of $\mathbf{P_l}$, thus proportional to their counts, and stored in the $(J_l \times J_l)$ diagonal matrix $\mathbf{M_l}$. The total weight of the words used by the same sample is equal to 1. $\mathbf{X_l}$ is centred and possibly normalised in the case of quantitative variables, using the weighting system $\mathbf{D_l}$. The $(J_l \times K)$ table $\mathbf{Q_l} = \frac{\mathbf{Y_l^T X_l}}{\mathbf{N_l}} = \mathbf{P_l^T X_l}$ is the data structure containing the relations between words and variables. $\mathbf{Q_l}$ is called a generalised aggregated lexical table.

Note. The name **Generalised Aggregated Lexical Table** and the acronym **GALT** are used to emphasise the close similarity between this table and the classi-

cal **Aggregated Lexical Table** (**ALT**) developed in the case of a single categorical variable (Lebart et al., 1998).

The calculation is exactly the same in both cases. What changes is only the expression of the matrix **X** itself. An ALT consists of the dummy variables corresponding to the categories of a single categorical variable.

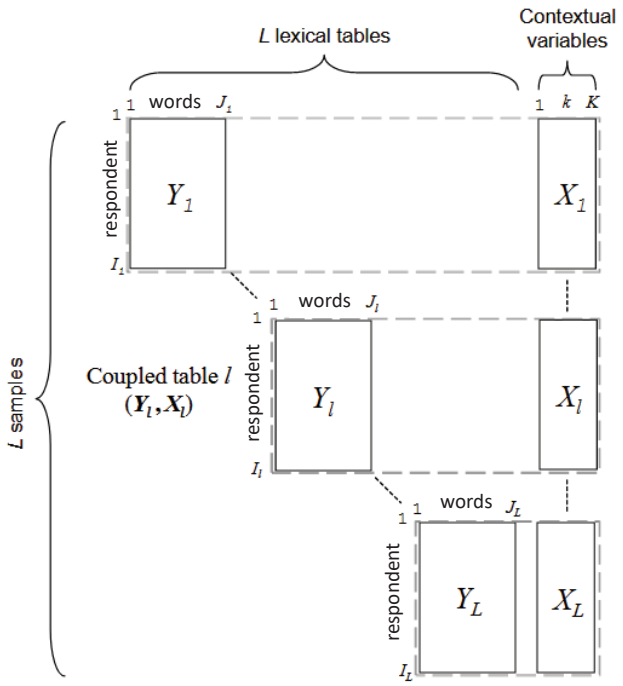


Figure 1: Sequence of L coupled tables

In the global analysis of the L samples, we have to deal with $I = \sum_l I_l$ respondents who have used $J = \sum_l J_l$ different words in the $N = \sum_l N_l$ occurrences that they have pronounced in all their free responses. The respondent and word weights are rescaled so that both totals are equal to 1 for the I respondents and J words respectively. To do this, the respondent and word weights in sample l are multiplied by N_l/N. The global weights of the respondents are stored in the $(I \times I)$ diagonal matrix **D**. The global weights of the words are stored in the $(J \times J)$ diagonal matrix **M**.

The $(I \times K)$ global table **X** is obtained combining by rows the L tables $\mathbf{X_l}$, centred by set. Table **X** is therefore also centred for weighting system **D**.

We assume $K < J$. The symbols I, I_l, J, J_l, K, L henceforth refer to both the

set and its cardinality.

3. METHODS USED AS THE BASIS OF OUR APPROACH

3.1. DEALING WITH ONE SAMPLE

In this section, we deal with only one sample and therefore consider it unnecessary to use index l.

3.1.1. CA-GALT method

We want to analyse the GALT matrix $\mathbf{Q}$ following a CA-like approach as far as possible. We therefore use the CA-GALT method (Bécue-Bertaut and Pagès, 2015; Bécue-Bertaut et al., 2014), as summarised below.

Let the $(K \times K)$ matrix $\mathbf{C} = (\mathbf{X^T D X})$ be the weighted correlation/covariance matrix of the variable-columns of the matrix $\mathbf{X}$. We compute the $(J \times K)$ matrix $\mathbf{Z}$, the double standardised form of the matrix $\mathbf{Q}$:

$$\mathbf{Z} = \mathbf{M^{-1} Q C^{-1}}. \tag{1}$$

If $\mathbf{C}$ is not invertible, $\mathbf{C^{-1}}$ is replaced by the Moore-Penrose pseudoinverse $\mathbf{C}^-$.

CA-GALT is then performed by principal component analysis (PCA) in two metrics: $\mathbf{C}$ in the row space, and $\mathbf{M}$ in the column space, i.e. PCA$(\mathbf{Z}, \mathbf{C}, \mathbf{M})$. This involves computing the S $(S \leq K)$ eigenvalues and eigenvectors of

$$\mathbf{Z^T M Z C}. \tag{2}$$

The eigenvalues are stored in the $(S \times S)$ diagonal matrix Λ, and the eigenvectors in the $(K \times S)$ matrix $\mathbf{U}$.

CA-GALT is a dual-projected analysis (Bécue-Bertaut and Pagès, 2015) that explains the variability of the words according to the variability of variables and, the variability of the variables according to the variability of the words.

Note. Metric $\mathbf{C^{-1}}$ (or $\mathbf{C}^-$) performs a multivariate standardisation that not only standardises the columns of $\mathbf{X}$ separately, but also makes them uncorrelated (Brandimarte, 2011; Härdle and Simar, 2012).

3.2. MFA GENERAL SCHEME

Multiple factor analysis (Escofier and Pagès, 2016; Pagès, 2014) analyses the multiple table combining by columns either quantitative or categorical tables. It has also been extended to frequency tables (Bécue-Bertaut and Pagès, 2004). This

method analyses a set of rows described by different sets of columns. The core of MFA is a PCA with specific weights and metrics applied to the multiple table containing quantitative tables as in PCA, categorical tables as in Multiple Correspondence Analysis (MCA), and frequency tables, especially lexical tables, as in CA. The specific approach to each type of table is obtained by coding the initial data and choosing the appropriate weights and metrics.

In order to balance the influence of the sets on the first factorial dimension, the initial weights of the columns in a given set are divided by the first eigenvalue resulting from the separate analysis of the corresponding table (PCA, MCA or CA depending on its type). The highest axial inertia of each set is thus normalised to 1. MFA identifies the main directions of variability in the data from a description of the rows by all the different sets of columns but balancing the importance of these sets, and provides the classic results of principal component methods. The characteristics and interpretation rules of PCA, MCA and CA are preserved for the quantitative, categorical and frequency sets. MFA also offers graphical tools for comparing the different sets, such as the superimposed global and partial representation of the rows as induced by all the sets or by each set separately, as well as a synthetic representation of the sets where each one is represented by only one point. These graphical results allow us to compare the typologies provided by each set in a common reference space.

4. MFA ON MULTIPLE GALT

Below, we adapt MFA to the case where the separate tables are GALTs built from the various samples, i.e. from the different coupled tables $(\mathbf{Y_l}, \mathbf{X_l})$ $(l = 1, .., L)$. The GALTs and their analysis are integrated into this approach by means of CA-GALT.

As described above, MFA is usually applied to a set of rows described by several sets of columns. We now need to analyse several sets of row-words described by one set of column-variables. However, here we are in a CA-like context where the roles of rows and columns are interchangeable. We could do this without changing the results. In the following sections we present the MFA-GALT method in a direct way.

MFA-GALT is performed in two steps exactly like a classic MFA. First, each sub-table — here a GALT — is analysed separately by applying the appropiate factorial method for its type, here CA-GALT. In the second step, a global factorial analysis is performed on all sets of multiple tables, treating each set as in the separate analyses, but taking into account the reweighting used to balance the

influence of the sets so the different sets of rows have a similar influence on the first global axis. This reweighting consists of dividing the weights of the rows of set l by the first eigenvalue obtained in the separate analyses of this set, so the highest axial inertia of each set is standardised to 1. Among the properties of this reweighting of the rows, it should be noted that the within-sets structures are not modified and that except for very special cases, the first axis of the global analysis is common to several sets and cannot therefore be generated by a single table. These two steps are described in more detail below.

First step: separate analyses

Separate CA-GALTs are performed in each set on the GALT $\mathbf{Q_l}$ according to the method in Section 3.1.1, with the exception of the metric used in the row space (and the weighting system in the column space). In this case, the covariance/correlation matrix computed from all the respondents, i.e. $\mathbf{C} = (\mathbf{X^T D X})$, is used in all the separate analyses instead of the matrices $\mathbf{C_l} = (\mathbf{X_l^T D_l X_l})$ as all row sets must be located in the same metric space. $\mathbf{C^{-1}}$ (or $\mathbf{C}^-$, if $\mathbf{C}$ is not invertible) is therefore used to standardise $\mathbf{Q_l}$. In this first step $\mathbf{Z_l} = \mathbf{M_l^{-1} Q_l C}^-$ is analysed by means of PCA($\mathbf{Z_l}$,$\mathbf{C}$,$\mathbf{M_l}$). The L first eigenvalues λ_1^l are used in the second step.

Second step: global analysis

The row weighting system is updated to balance the influence of each set in the global analysis. By construction, the matrix $\mathbf{M}$ is divided into L blocks. Block l corresponds to the J_l words used in sample l. The weights of the words in block l are divided by λ_1^l, the first eigenvalue of the separate analysis of sub-table l. The resulting weights are stored in the $(J \times J)$ matrix $\mathbf{M}_\lambda$.

The $(J \times K)$ multiple table GALT $\mathbf{Q}$ combines by rows the L matrices $\mathbf{Q_l}$ but resized by multiplying them by coefficient N_l/N ($\mathbf{Q_l} \times N_l/N$). A double standardisation of $\mathbf{Q}$ on the rows and the columns produces the $(J \times K)$ table $\mathbf{Z} = \mathbf{M}_\lambda^{-1} \mathbf{Q C^{-1}}$. If $\mathbf{C}$ is not invertible, $\mathbf{C^{-1}}$ is replaced by the Moore-Penrose pseudoinverse $\mathbf{C}^-$. MFA-GALT is then performed by a non-standardised weighted PCA on the multiple table $\mathbf{Z}$, with $\mathbf{M}_\lambda$ as row weights and metric in the column space and $\mathbf{C}$ as column weights and metric in the row space, i.e. PCA($\mathbf{Z}$,$\mathbf{C}$,$\mathbf{M}_\lambda$).

5. MAIN PROPERTIES OF MFA-GALT

MFA-GALT provides the classic outputs of the principal components methods, in this case a specific MFA performed on the double standardised GALT multiple table crossing words (in rows) and categories or quantitative variables (in columns). In particular, we obtain:

- coordinates, contributions and qualities of representation of word-rows
- coordinates and qualities of representation of category-columns or quantitative variable-columns.

Respondents could be reintroduced into the analysis either as supplementary rows (and thus represented on the basis of the values they take for the contextual variables) or as supplementary columns (and thus represented on the basis of the words they use). This is not further explored in this paper.

In addition, MFA outputs are provided as a partial representation of the variables, a synthetic representation of the sets, and a measure of the similarity between the sets.

5.1. REPRESENTATION OF ROW-WORDS AND COLUMN-VARIABLES

PCA($\mathbf{Z}$,$\mathbf{C}$,$\mathbf{M}_\lambda$) involves the diagonalisation of the matrix $\mathbf{Z^T M}_\lambda \mathbf{ZC}$. The principal axis of rank s corresponds to the eigenvector $\mathbf{u_s}$ ($\|\mathbf{u_s}\|_\mathbf{C}$=1) associated with the eigenvalue λ_s:

$$\mathbf{Z^T M}_\lambda \mathbf{ZCu_s} = \lambda_s \mathbf{u_s}. \tag{3}$$

The eigenvalues λ_s are stored in the $(S \times S)$ diagonal matrix Λ and the eigenvectors u_s — the dispersion axes — are stored in the columns of the $(K \times S)$ matrix $\mathbf{U}$.

By factor s we mean the vector of coordinates on axis s of either the word-rows (denoted F_s) or the variable-columns (denoted G_s) (Benzécri, 1973; Pagès, 2014). The values of the S row factors are stored in the columns of the $(J \times S)$ matrix $\mathbf{F}$, calculated as follows:

$$\mathbf{F} = \mathbf{ZCU}. \tag{4}$$

The row factors place the words in the direction of either the categories of respondents who use them frequently or the quantitative variables for which the respondents who use them have high values.

The values of the S column factors are stored in the columns of the $(K \times S)$ matrix $\mathbf{G}$. The matrix $\mathbf{G}$ is computed by using the transition relations between the

row and column factors, as in any PCA:

$$\mathbf{G} = \mathbf{Z}^{\mathbf{T}}\mathbf{M}_{\lambda}\mathbf{F}\Lambda^{-1/2}. \tag{5}$$

Thus, these scores are equal to the weighted covariances, or weighted correlation coefficients, between the standardised row factors and the doubly standardised columns of the multiple GALT.

5.2. SUPERIMPOSED REPRESENTATION OF THE *l* CLOUDS OF VARIABLES

According to the L sets of row-words, the column-variable k of $\mathbf{Z}$ can be divided into L sub-columns, called partial variables and denoted k^l. It is useful to represent simultaneously the L partial scatterplots, each made up of the corresponding K partial variables, on the same axes of reference. We successively consider the L matrices $\mathbf{Z_l}$ of dimension $(J_l \times K)$ issued from the matrix $\mathbf{Z}$ by retaining only the row-words belonging to the set l. From these matrices, the $(J \times K)$ matrices $\tilde{\mathbf{Z}}_l$ are built by completing $\mathbf{Z_l}$ with zeroes to have the same dimension as $\mathbf{Z}$. In order to be represented on the global axes, the K partial variables corresponding to the set l are considered as supplementary columns in the global analysis. Their coordinates are calculated using the transition relations and stored in the $(K \times S)$ matrix $\mathbf{G^l}$:

$$\mathbf{G^l} = \tilde{\mathbf{Z}}_l{}^{\mathbf{T}}\mathbf{M}_{\lambda}\mathbf{F}\Lambda^{-1/2}. \tag{6}$$

Therefore, the coordinates of the partial variables corresponding to set l can be calculated from the coordinates of the words used by sample l only. Thanks to the structure of the matrix $\tilde{\mathbf{Z}}_l$ which contains only 0 except for the rows belonging to set l, this relation for the partial variable k^l is expressed very simply :

$$G_s(k^l) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\lambda_1^l}} \sum_{j \in J_l} z_{jk} m_{jj} F_s(j). \tag{7}$$

In Eq.7, $[z_{jk}]$ denotes the generic term of $\mathbf{Z}$ and $\frac{1}{\sqrt{\lambda_1^l}} m_{jj}$ denotes the generic term of the matrix $\mathbf{M}_{\lambda}$, where m_{jj} is the initial weight of the word j (see Section 2).

According to Eq.7, the partial variables relative to set l are on the side of the words in this sample that are overused by respondents with high values for these contextual variables.

All the "partial" variables can usually be represented on the same scatterplot, thus providing information about the similarities/dissimilarities between the samples.

5.3. GLOBAL REPRESENTATION OF THE SETS

Another result is the representation of the L groups on the same graph, each of which is represented by a point (Pagès, 2014). To this end, the Lg coefficient (the formula of wich will be reminded below), the linkage measurement between one variable and one set of variables, is applied here to measure the linkage between each axis retained and each set of variables. First, the $(K \times K)$ matrix of scalar products $\mathbf{W_l}$ between the K column-variables of set l is computed as

$$\mathbf{W_l} = \mathbf{Z_l^T M_{\lambda l} Z_l}. \tag{8}$$

where the diagonal matrix $\mathbf{M_{\lambda l}}$, as block l of the matrix $\mathbf{M_\lambda}$, contains the weights of the variables of set l, equal here to $\frac{1}{\lambda_l^1}$.

$Lg(l, \mathbf{u_s})$ is then calculated as follows:

$$Lg(l, \mathbf{u_s}) = \langle \mathbf{W_l C}, \mathbf{u_s C} \rangle = trace(\mathbf{W_l C u_s u_s^T C}). \tag{9}$$

$Lg(l, \mathbf{u_s})$ will be used as a coordinate to place set l on the axis of rank s. This coordinate always has a value between 0 and 1. This produces a map of all the sets, each represented by one point. This map also shows the similarity, i.e. the proximity between the structures in the L sets.

5.4. MEASURE OF THE ASSOCIATION BETWEEN VOCABULARY AND CONTEXTUAL VARIABLES

Our proposal also includes the measurement of the association between vocabulary and contextual variables, firstly to select the variables that actually play a role, and secondly to interpret the results. The measures carried out successively for each sample are described in detail in Bécue-Bertaut and Pagès (2015).

Briefly, vocabulary is said to be associated with a variable if words differ significantly in the values taken by the people using them. The association between a categorical variable and vocabulary is evaluated with the classic chi-square test on the frequency table crossing words and categories (=lexical table).

A one-way analysis of variance (ANOVA) is considered in the case of a quantitative variable. The data table is reorganised as shown in Figure 2 before computing the one-way ANOVA: each row corresponds to one occurrence of a word (there are as many rows as the total number of occurrences in the corpus). The score variable and the words variable have as many values as occurrences. The one-way ANOVA is then performed between the score and the words to detect any relationships between vocabulary and scores.

Individuals	Score variable	Words		
		word A	word B	word C
ind 1	4	2	0	1
ind 2	6	1	0	0
ind 3	3	0	1	1

Individuals	Words	Score
ind 1	word A	4
ind 1	word A	4
ind 1	word C	4
ind 2	word A	6
ind 3	word B	3
ind 3	word C	3

Figure 2: Reorganization of the data for the one-way ANOVA measuring the association between vocabulary and a contextual variable.

It should be noted that since the occurrences are not independent, the usual assumptions of ANOVA are not satisfied and it is better to use permutation tests.

6. REAL DATA APPLICATION: INTERNATIONAL SURVEY

A railway company conducted a survey to determine how satisfied passengers were with its night trains. Passengers were asked to rate their satisfaction with 13 aspects related to comfort (general, cabin, bed, seat), cleanliness (common areas, cabin, toilet), staff attention (welcome attention, trip attention, language skills) and others (cabin room, air conditioning, general aspects). Each aspect was rated on a 11-point Likert scale, from 0 (very poor) to 10 (excellent). An open-ended question was added asking about the aspects that needed improvement. This question required spontaneous answers, in this case expressed in English or Spanish. The data are stored in the data structure shown in Figure 3.

The pre-processing of the data includes a careful correction of the spelling of the free answers. Stop-words are removed and the words used at least ten times are then kept for the Spanish corpus (=all answers given in Spanish), while the threshold for the English corpus is five (Lebart et al., 1998; Murtagh, 2005). Finally, 977 respondents from the Spanish sample and 283 from the English sample have no empty answers. The average length of free answers is 3.1 occurrences in both cases. The Spanish corpus contains 3029 occurrences corresponding to 88 different words and the English corpus has 871 occurrences corresponding to 68 different words.

Missing values have been imputed for the score variables. It should be noted that the rating scale has been inverted to make the graphs easier to read. The highest scores correspond to the highest levels of dissatisfaction.

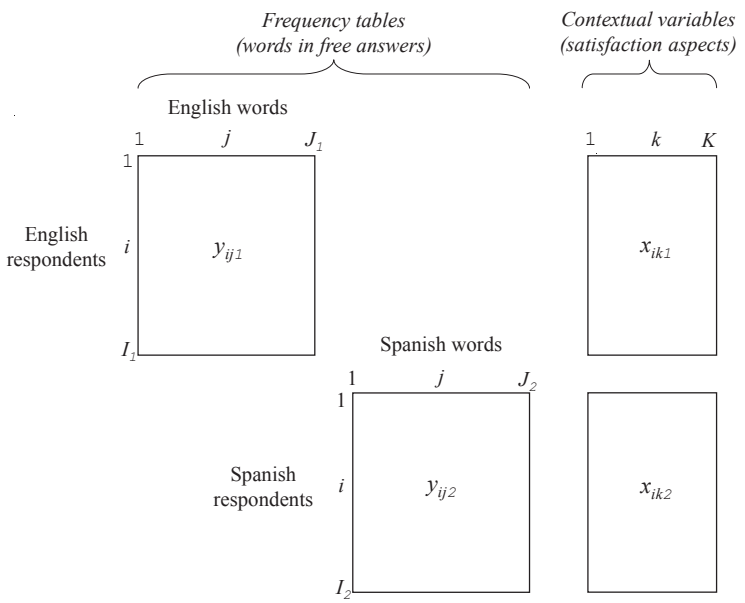


Figure 3: The dataset. On the left, the lexical tables; on the right, the contextual variables. In the example, I_1 = 283 (English respondents), I_2 = 977 (Spanish respondents), J_1 = 68 (English words), J_2 = 88 (Spanish words), K = 13 (satisfaction aspects).

Table 1: Mean satisfaction scores and association with vocabulary ratios

	Spanish respondents		English respondents	
Satisfaction aspects	mean (SD)	ass.ratio (p-value)	mean (SD)	ass.ratio (p-value)
General comfort	6.82 (1.80)	0.062 (<0.001)	6.66 (1.95)	0.091 (0.148)
Cabin comfort	6.37 (2.07)	0.063 (<0.001)	6.37 (2.03)	0.121 (0.010)
Cabin room	5.33 (2.43)	0.089 (<0.001)	5.71 (2.35)	0.136 (<0.001)
Bed comfort	6.70 (1.98)	0.050 (<0.001)	6.72 (2.03)	0.063 (0.918)
Seat comfort	6.10 (2.20)	0.059 (<0.001)	5.99 (2.38)	0.123 (0.010)
Air conditioning	6.55 (2.55)	0.107 (<0.001)	6.51 (2.71)	0.226 (<0.001)
Common areas cleanliness	7.41 (1.92)	0.043 (<0.001)	7.54 (1.86)	0.082 (0.548)
Cabin cleanliness	7.59 (1.88)	0.056 (<0.001)	7.59 (1.81)	0.116 (0.036)
Toilet cleanliness	6.21 (2.55)	0.090 (<0.001)	6.29 (2.40)	0.150 (<0.001)
Staff welcome attention	7.99 (1.92)	0.040 (0.018)	7.29 (2.45)	0.108 (0.062)
Staff trip attention	8.07 (1.85)	0.038 (0.048)	7.34 (2.29)	0.092 (0.294)
General aspects	7.77 (1.65)	0.038 (0.034)	7.48 (1.91)	0.079 (0.590)
Staff language skills	7.72 (2.08)	0.052 (<0.001)	7.14 (2.52)	0.154 (<0.001)

6.1. INITIAL FINDINGS

The most frequent words give a preliminary overview of the complaints, which are similar in both languages and expressed with homologous words. *Espacio/space* is too reduced, no place for *maletas/luggages*. *Cabinas/cabins* and *asientos/seats* lack *comodidad/comfort*, while *aseos/toilets* would benefit from more *limpieza/cleanliness*. The *Aire acondicionado/Air conditioning* seems to be causing problems. In the English sample, the words *staff* and *English* are frequently mentioned. Aspects that were not asked about are mentioned, such as *precio/price*.

Table 1 provides a first insight with the means and standard deviations of the satisfaction scores. *Staff trip attention* obtains the highest score (8.07) from Spanish-speaking respondents while English-speaking respondents gave the highest score to *Cabin cleanliness* (7.59). The lowest score is for *Cabin room* for both Spanish (5.33) and English-speaking respondents (5.71). It is worth noting that the three aspects related to staff (*Staff welcome*, *Staff trip attention* and *Staff language skills*) are significantly less valued by English-speaking than by Spanish-speaking respondents.

The association between vocabulary and a contextual variable (see Table 1, columns *ass.ratio (p-value)*) shows that *Air conditioning* receives the highest ratio for both Spanish (0.107) and English-speaking respondents (0.226). *Toilet cleanliness* is the second most important indicator for Spanish-speaking respondents (0.090), while *Staff language skills* is second with 0.154 for English respondents, although closely followed by *Toilet cleanliness* in third place (0.150). It should

be noted that Spanish-speaking respondents rank the *Staff language skills* only eighth. *Cabin room* is ranked third for Spanish-speaking and fourth for English-speaking respondents.

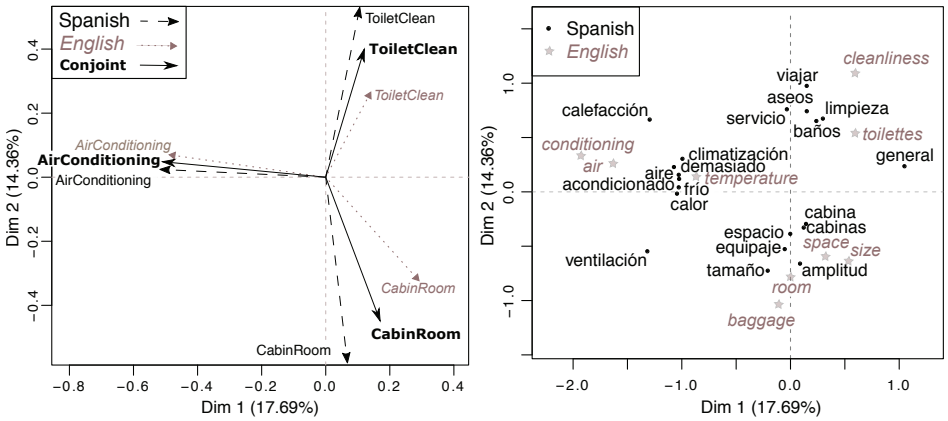
6.2. MFA-GALT ON THE MULTILINGUAL DATASET

The ranking aspects from the association-with-vocabulary ratio do not coincide with the score-average ranking. This shows that the information from the free comments differs from the closed questions, and that these two types of information are complementary, implying that the aspects the passengers believe should be improved do not match the aspects with which they are less satisfied. This justifies the interest in collecting information through open-ended questions, as this information is different and complementary.

MFA-GALT is applied on the multiple generalised aggregated lexical table. The total inertia is equal to 9.91. The first eigenvalue (1.75 corresponding to 17.69% of the total inertia) is close to the number of sets, which means that the two sets share the dispersion direction corresponding to the first global axis. The second (1.42, 14.36% of the total inertia) and third eigenvalue (1.23, 12.39% of the total inertia) are close, but the following eigenvalues are much smaller, so we focus only on the first three axes. To avoid over-emphasising the example, we will only interpret the first two axes. For a more detailed description of the results, and particularly the third dimension, the reader can refer to the thesis of Kostov (2015).

6.2.1. Global representation of the satisfaction scores and words

MFA-GALT provides graphical results in which each variable (each score) points to the words associated with it. It thus indicates the shortcomings of the scored aspect, whether or not they are common to both languages. Figure 4a shows the best represented satisfaction scores on the first MFA-GALT principal plane through their covariances with the axes. To avoid overloading the graphs, only the scores that are well represented are shown (in this case, those that have a square cosine sum on the two axes over 0.5). We first look at only the global representations of the scores, which has a three-polar structure. The three poles refer to inconveniences associated with *Air conditioning*, lack of *Toilet cleanliness* and problems related to *Cabin room*. This is in line with what the association-with-vocabulary ratios suggested. Figure 4b shows the Spanish and English words that contribute more than twice to the average contribution. We can then see words that are strongly associated with *air conditioning*, showing its shortcomings: *air*/*aire*,



(a) Best represented satisfaction scores (b) Most contributive words

Figure 4: MFA-GALT: Representation of variables (i.e. scores) and words on the plane (1,2)

conditioning/*acondicionado*, *temperature*, *frío* (=cold) *climatización* (=air conditioner), *ventilación*/*ventilation* and *calefacción* (=heating). On the positive part of the second axis, the lack of *Toilet cleanliness* is characterised by *cleanliness*/*limpieza*, *toilettes*/*aseos*/*baños*. On the negative part, the problems with *Cabin room* are described using the words *size*/*espacio* and *cabins*/*cabina(s)*.

In this example, axis 3 is specific to only the English set, which points to problems with the staff speaking poor English. This may seem like trivial information, but it shows that the method works and offers the possibility of highlighting information specific to only one sub-population, and also makes the transport company aware that language difficulties are a real problem highlighted by English-speaking respondents, unlike, for example, in air transport.

6.2.2. Partial representation of the satisfaction scores

Figure 4a shows the superimposed representation of the global and partial representations of the satisfaction scores on the plane (1,2) and highlights the similarities and differences between the two sets in terms of the association between words and scores. *Air conditioning* behaves similarly in both sets on the first axis. On the second axis, *Toilet cleanliness* and *Cabin room* are more strongly associated with Spanish than with English vocabulary, which translates into higher covariances; the complaints using English vocabulary appear more accentuated

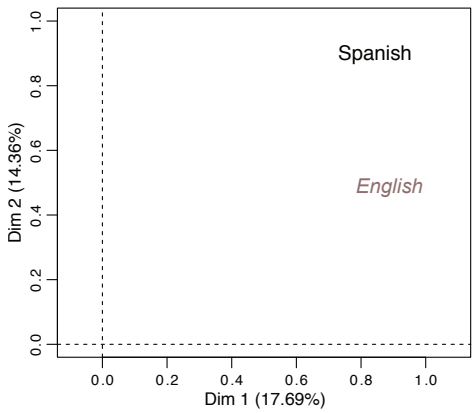


Figure 5: Representation of the sets.

and give rise to more words. In the third dimension, only English-speaking respondents complain about the lack of *Staff language skills*.

6.2.3. Representation of the sets

Similarity measures confirm that both sets share some dispersion directions. The value of the RV coefficient, multivariate generalisation of the squared Pearson correlation coefficient, equal to 0.74 ($p<0.001$), confirms that the partial configurations are relatively close but not homothetic.

According to the representation of the sets on the first dimension, the coordinate of the Spanish sample is 0.85 while the coordinate of the English sample has a slightly higher value (0.91) (Figure 5). This means that the first axis provided by MFA-GALT is of major importance for both sets and is therefore a common axis dispersion, while the Spanish set has a much larger coordinate on the second axis (0.91 vs. 0.51). The second MFAGALT axis is thus very important for Spanish-speaking respondents, and not so much for the English-speakers, while the opposite is observed for the third axis (0.42 for Spanish vs. 0.81 for English).

7. CONCLUSION

This paper proposes an original principal component method to deal with open-ended questions answered in different languages. This type of textual and contextual data produces a sequence of coupled tables, each comprising one frequency table (=lexical table) and one quantitative/qualitative table. We approach these

data through the relationships between the words and the contextual variables. Two methods — CA-GALT and MFA — are combined, hence the name of the new method: *Multiple Factor Analysis on Generalised Aggregated Lexical Tables* (MFA-GALT). The first places the words of the different sets in the same space generated by the variables, resulting in the construction of the GALTs; while the second allows the simultaneous analysis of these tables in a way that preserves the MFA properties.

An international survey with open questions answered in different languages was analysed with MFA-GALT, making it possible to study similarities among words from the same language, similarities among homologous words from different languages, associations between words and satisfaction scores, similarities between satisfaction score structures (partial representations) and similarities between groups. The results of this application show that MFA-GALT provides a good synthesis of the data and produces outputs that are easy to interpret.

The R package `Xplortext` includes the `LexGalt` function, which enables the implementation of the CA-GALT and MFA-GALT methods.

References

Balbi, S. and Giordano, G. (2001). A factorial technique for analysing textual data with external information. In S. Borra, R. Rocci, M. Vichi and M. Schader, eds., *Advances in Classification and Data Analysis*, 169–176. Springer, Berlin, Heidelberg.

Balbi, S. and Misuraca, M. (2010). A doubly projected analysis for lexical tables. In C.H. Skiadas, ed., *Advances in Data Analysis: Theory and Applications to Reliability and Inference, Data Mining, Bioinformatics, Lifetime Data, and Neural Networks*, 13–19. Birkhäuser, Boston.

Bécue-Bertaut, M. and Pagès, J. (2004). A principal axes method for comparing contingency tables: MFACT. In *Computational Statistics and Data Analysis*, 45 (3): 481–503.

Bécue-Bertaut, M. and Pagès, J. (2015). Correspondence analysis of textual data involving contextual information: CA-GALT on principal components. In *Advances in Data Analysis and Classification*, 9: 125–142.

Bécue-Bertaut, M., Pagès, J. and Kostov, B. (2014). Untangling the influence of several contextual variables on the respondents' lexical choices. A statistical approach. In *Statistics and Operations Research Transactions*, 38: 285–302.

Benzécri, J.P. (1973). *Analyse des Données*. Bordas, Paris.

Benzécri, J.P. (1981). *Pratique de l'Analyse des Données. Tome 3, Linguistique & Lexicologie*. Bordas, Paris.

Brandimarte, P. (2011). *Quantitative Methods: an Introduction for Business Management*. John Wiley & Sons, New Jersey.

Escofier, B. and Pagès, J. (2016). *Analyses Factorielles Simples et Multiples*. Dunod, Paris, 5th edn.

Härdle, W. and Simar, L. (2012). *Applied Multivariate Statistical Analysis*. Springer Verlag, Heidelberg, Berlin.

Kostov, B. (2015). *A principal Component Method to Analyse Disconnected Frequency Tables by Means of Contextual Information.* Ph.D. dissertation, UPC, Departament d'Estadística i Investigació Operativa. URL `https://upcommons.upc.edu/handle/2117/95759`.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data*. Kluwer Academic Publishers, Dordrecht.

Murtagh, F. (2005). *Correspondence Analysis and Data Coding with Java and R*. Chapman & Hall / CRC Press, New York.

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. Chapman and Hall/CRC, New York.

Spano, M. and Triunfo, N. (2012). La relazione sulla gestione delle società italiane quotate sul mercato regolamentato. In A. Dister, D. Longrée, and G. Purnelle, eds., *Actes de 11ème Journées d'analyse de données textuelles*. URL `http://lexicometrica.univ-paris3.fr/jadt/jadt2012/tocJADT2012.htm`.

Takane, Y., Yanai, H. and Mayekawa, S. (1991). Relationships among several methods of linearly constrained correspondence analysis. In *Psychometrika*, 56: 667–684.

ter Braak, C.J.F. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. In *Ecology*, 67: 1167–1179.

ter Braak, C.J.F. (1987). *Canoco–A FORTRAN Program for Canonical Community Ordination by [Partial] [Detrended] [Canonical] Correspondence Analysis (Version 2.1).* ITI-TNO Institute of Applied Computer Sciences, Wageningen.

doi.org/10.26398/IJAS.0035-016

VISUALIZATION OF TEXTUAL DATA: A COMPLEMENT TO AUTHORSHIP ATTRIBUTION

Ludovic Lebart[1]
Centre National de la Recherche Scientifique (CNRS), Paris, France

Abstract. *In textual data analysis, authorship attribution is precisely a leading case of statistical decision. While analyzing a large corpus of 50 French novels of the 20th century, we investigate the frontiers between descriptive (or unsupervised) methods, and confirmatory (or supervised) methods. It will be shown that additive trees applied to the coordinates of a preliminary correspondence analysis (CA) can provide both a description and an help for a decision. Our results aim at showing the complementarity between exploratory techniques and AI. in that field.*

Keywords: *Textual data visualization, Authorship attribution, Additive trees, CA.*

1. INTRODUCTION

If artificial intelligence (AI) methods often give excellent results in terms of authorship attribution, the specialist of the concerned texts sometimes remains frustrated by the binary and blind nature of the decision. In the framework of a problem of (literary) matching (50 novels written by 24 authors) and in the spirit of "Deep learning" which introduces the "unsupervised" in AI, we will show that the joint use of correspondence analysis (CA) in a mixed supervised/unsupervised framework (technique of supplementary variables/ visualized regression) makes it possible to both obtain satisfactory results and understand the context of these results. It will also be recalled in passing that CA (like regression) is also a particular case of neural networks, and fully deserves to be included in the panoply of AI techniques.

[1] Ludovic Lebart, ludovic@lebart.fr

2. A PROBLEM OF LITERARY MATCHING: 50 novels/ 24 authors

Text The following analyzes relate to a large corpus consisting of 50 novels from 24 francophone writers, selected and provided by Etienne Brunet (Brunet *et al.* 2021). Corpus Size: 3,501,883 words (tokens) among which 82,914 distinct words (types) containing 31,503 hapaxes (words appearing once). The corresponding Type/Token ratio (TTR), a decreasing function of the size of the corpus, is: TTR = 0.024.

By "analysis" we mean here a sequence of correspondence analysis (CA) of lexical tables, followed by an additive tree (AT) computed on a subspace of CA principal axes. The CA phase involves the chi-square distance (and its property of distributional equivalence) and gives the possibility to select the dimension of the subspace, allowing for a regularization of the data (see for example: Author *et al.*, 1977, 1984; Author, 1992). Starting the processing with a principal axes analysis brings this approach within the framework of deep learning, which recommends preliminary structural analyzes and a possible regularization of the data. But the tools remain geometric and transparent.

Table 1: List of 25 "authors" with their two selected titles [and their corresponding symbols for figures 1 and 2] (* = Nobel prize)

Ajar: Gros-Câlin & La vie devant soi [aja.grosca & aja.viedev]
Aragon: Les Beaux Quartiers & Blanche ou l'oubli [ara.beauxq & ara.blanch]
Breton: Nadja & L'Amou Fou [bre.Nadja & bre.amour]
Camus*: L'étranger & La Chute [cam.etrang & cam.chute]
Colette: Sido & La Vagabonde [col.sido & col.vagabo]
Duras: Barrage au Pacifique & L'Amant [dur.barag & dur.amant]
Ernaux*: La Honte & Les Années. [ern.honte & ern.annees]
Gary1: La Promesse de l'Aube & Les Racines du Ciel [gar.promes & gar.racine]
Gary2: Clair de Femme & Au_delà de cette limite [gar.clair & gar.delali]
Gide*: La Symphonie Pastorale & L'Immoraliste [gide.sympho & gide.immora]
Giono: Le Grand Troupeau & Le Hussard sur le toit [gio.grand & gio.hussar]
Giraudoux: Simon le Pathétique & Bella [gir.Simon & gir.Bella]
Gracq: Le Rivage des Syrtes & Un Balcon en forêt [gra.rivage & gra.balcon]
Le Clézio*: Hasard & Le Désert [cle.hasard & cle.desert]
Malraux: L'Espoir & Les Conquérants [mal.espoir & mal.conque]
Mammeri: La Colline oubliée & La Traversée [mam.colli & mam.traver]
Mauriac*: Le Baiser... & Le Mystère Frontenac [mau.baiser & mau.myster]
Montherlant: Les Célibataires & Les Bestiaires [mon.celiba & mon.bestia]
Pérec: L'Homme qui dort & Les Choses [pere.hommed & per.choses]
Proust: Du côté de chez Swann & Le Temps retrouvé [pro.cote & pro.temps]
Queneau: Le Chiendent & Zazie dans le métro [que.chiend & que.zaziem]
Saint-Exupéry: Courrier Sud & Terre des Hommes [exu.courri & exu.terreh]

Tournier: Vendredi ou les limbes... & Eléazar [tou.vendre & tou.eleaza]Vian: L'Ecume des jours & L'Automne à Pékin [via.ecum & via.auto]
Yourcenar: Mémoires d'Hadrien & L'Oeuvre au noir [you.memoi & you.oeuvre]

Note that one author appears three times in that list. "Romain Gary" (Gary1, Gary2, Ajar). At the origin of a famous literary deception, Gary managed to win the most prestigious French literary prize twice (Prix Goncourt) by hiding behind the name of "Emile Ajar". The double presence of Gary (triple, with Ajar) in the list aims to analyze more finely this oddity. However, all the results presented here remain still valid without this over-representation.

3. BRIEF REMINDER ABOUT THE TOOLS

3.1 SUPERVISED AND UNSUPERVISED MODELS

Let us remind that the "unsupervised approach" (exploratory or descriptive) is the counterpart of the "supervised approach (confirmatory or explanatory approach). Factor analysis, PCA, CA and clustering are unsupervised whereas discriminant analysis or regression methods are supervised.

External validation is the standard procedure in the case of supervised learning. Once the model parameters are estimated (learning phase), external validation is used to evaluate the model (generalization phase), usually with cross validation methods. External validation occurs in the context of correspondence analysis in two practical circumstances:

a) when the data set may be divided into two or more parts, one part being used to estimate the model, the other part used to verify the suitability of this model,

b) when certain metadata or external information are available to supplement the description of items.

We assume that external information is in the form of "supplementary elements". Note that a statistical validation (mostly bootstrap) is the indispensable complement of these technique.

3.2 ADDITIVE TREES (AT): THE PHYLOGENETIC EXPLOSION

AT technique will be extensively and exclusively used in the paper. These trees were originally proposed by Buneman (1971), then studied by Sattah and Tverski (1977). The concept of hierarchy at the base of the ascending classification was to approximate the initial distances by an ultrametric distance. Additive trees are less demanding. More flexible than the Minimum Spanning Tree which depends on *n-1* parameter, the AT implies *2n-3* parameters. It remains to find an approximation of the initial distances which satisfies these conditions. With AT distance, a tree can be drawn with the objects as nodes, such that the distance between two objects is the length of the path joining these two objects on the tree.

Stimulated by the works of Barthélémy and Guénoche (1988), tree analysis methods have been widely used in the field of text analysis. However, the first proposed algorithms required a prohibitive computation volume for large numbers of objects to classify. Saitou and Nei (1987) proposed an algorithm called Neighbor Joining which approximately reduces the search for the additive tree to a classical ascending classification procedure. This heuristic which was implemented by Huson and Bryant (2006) [SplitsTree] and used here, had a huge impact on the rapidly expanding world of phylogenetic research. Saitou and Nei's article has been cited more than 68,000 times since its publication. Theoretical justifications for the algorithm's efficiency were presented by Mihaescu *et al.* (2009).

3.3 CORRESPONDENCE ANALYSIS AS A NEURAL NETWORK

The links between Singular Value Decomposition (SVD) and Principal Components Analysis (PCA) with some particular neural networks have been stressed by Bourlard and Kamp (1988), Baldi and Hornik (1989), Asoh and Otsu (1989). Correspondence Analysis (Benzécri, 1969; or its non-symmetrical version, Lauro and D'Ambra, 1984; Balbi, 1994; Balbi and Triunfo, 2013) is at the meeting point of many techniques. It can be described as both supervised and unsupervised multilayer perceptrons (Author, 1997). In the supervised case, the input and the output layers are respectively the rows and the columns of the contingency table. In the unsupervised case, both the input layer and the output

layer could be the rows, whereas the observations could be the columns of the table. In both situations, the networks make use of the identity function as a transfer function. More general transfer functions might lead to interesting non-linear extensions of the method.

3.4 SUPPLEMENTARY VARIABLES AND REGRESSION

Adding supplementary elements in a principal axes technique (SVD, PCA, CA) constitutes a descriptive variant of the multiple regression (being itself a simple form of perceptron). From a geometrical point of view, the two situations are indeed similar (see, e.g.: Lebart *et al.*, 1984, 2019):

Regression: The p explanatory variables generate a subspace having at most p dimensions on which is projected the variable y to explain.

CA or PCA: the p active variables of the analysis also generate a subspace with at most p dimensions that we reduce to q factors to visualize it. It is on this subspace reduced to q dimensions that we project afterwards the supplementary variables to locate them with respect to the active variables. A visualization is then possible in the space spanned by every pair of axes.

All the following results have been obtained with the help of the freely downloadable software DtmVic (www.dtmvic.com).

4. MAIN RESULTS

The analysis will focus on vocabulary, more in the spirit of content analysis than in the context of stylometry. We do not seek to discriminate between authors, and the pairings of texts observed in the forthcoming graphical displays will come somewhat as a surprise, a "statistical fact". We will lemmatize the corpus, using the free software *TreeTagger* (Schmid, 1994) discarding function words, proper nouns or personal pronouns (to eliminate, for instance, the effects of narrative at first person) (Section 4.1). Then we study the subset of verbs alone (Section 4.2). Finally, Section 4.3 analyzes directly and blindly the pages (here: sequences of 50 lines), without reference to a novel (the novels being positioned *a posteriori* as centroids of their pages). The approach followed is then similar, in a descriptive framework, to the so-called *Word2vec* approaches of AI.

4.1 BASIC GLOBAL ANALYSES ON LEMMAS

Figure 1 displays the first visualization of the whole lemmatized corpus, for a frequency threshold of 100 (2364 lemmas). The results are satisfactory: only one author escapes the matchings: Montherlant. The divergence between his two novels "Bestiaires" and "Célibataires" will be the big exception for many approaches. These two novels are indeed from the same author, but they call upon pools of exceptionally different vocabularies for one and the same author: we will see later that this difference is detectable even on verbs alone. Note that the tree of figure 1 remains similar with identical conclusions for a smaller frequency threshold of 50 corresponding to 4018 lemmas.

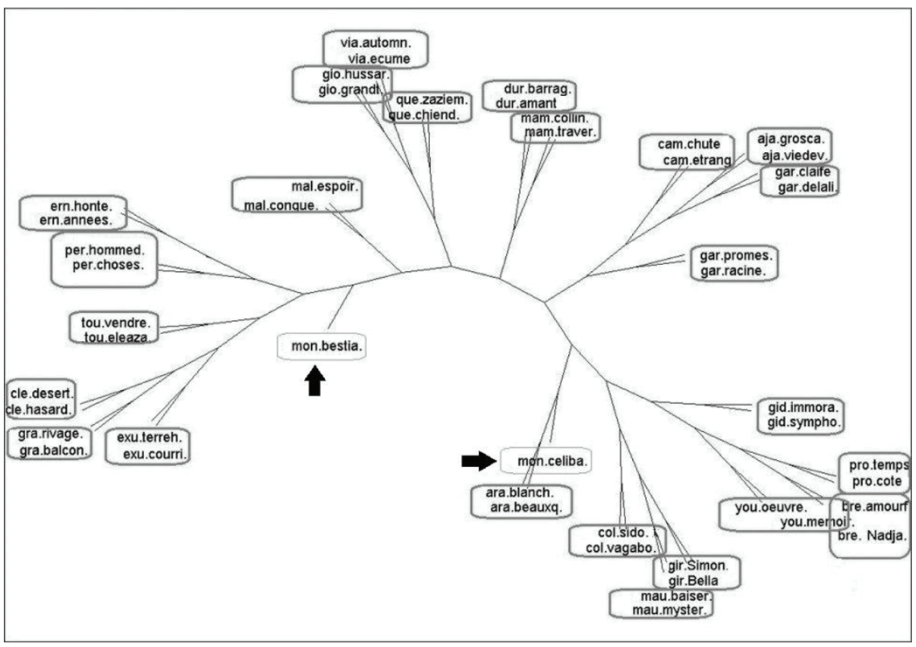


Figure 1: Additive tree for 50 novels de scribed by 2364 words (lemmas) appearing at least 100 times in the corpus. All CA axes (49) are kept. Only one author corresponds to unmatched novels (black arrows): Montherlant (novels: *Bestiaires* and *Les Célibataires*).

4.2 LIMITING THE VOCABULARY TO VERBS

The second approach concerns only verbs. Verbs are much less characteristic of a specific novel than nouns or adjectives, obviously more linked to the content of the text. Figure 2, however, gives us a surprising good result: 23 authors have been correctly matched (except Camus and again Montherlant).

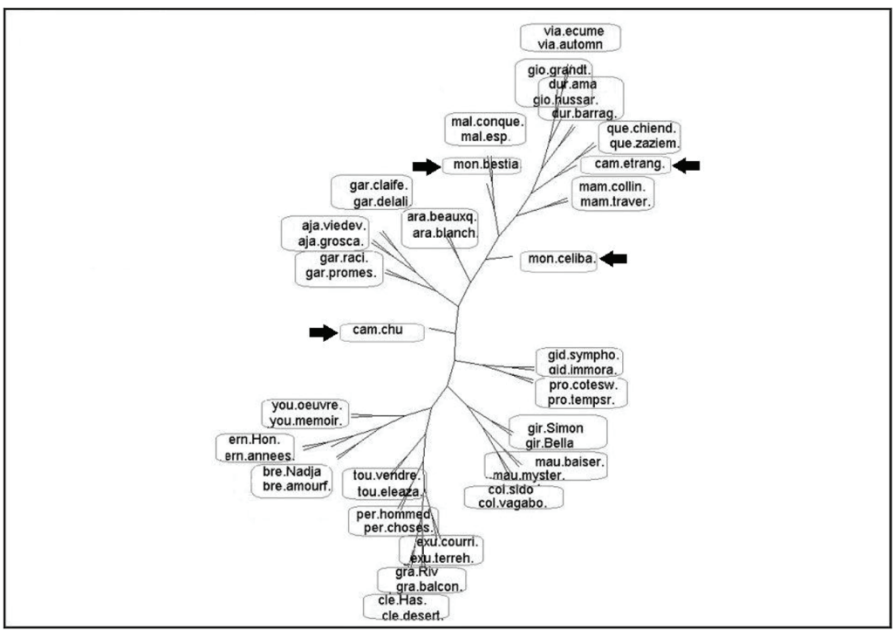


Figure 2: Additive tree for the 50 novels described only by their 726 verbs (without auxiliary verbs such as "to be", "to have". Frequency threshold for verbs: 84). Misclassified: Camus, Montherlant (black arrows).

4.3 FRAGMENTED NOVELS: ANALYSING THE 3547 PAGES

Finally, the third approach presented here is radically different. This time, the basic analysis is completely unsupervised. New "artificial observations" can be created in a text corpus, generalizing to large fragments the *context units* of the original approach proposed by Reinert (1983) at the basis of a procedure known as ALCESTE methodology.

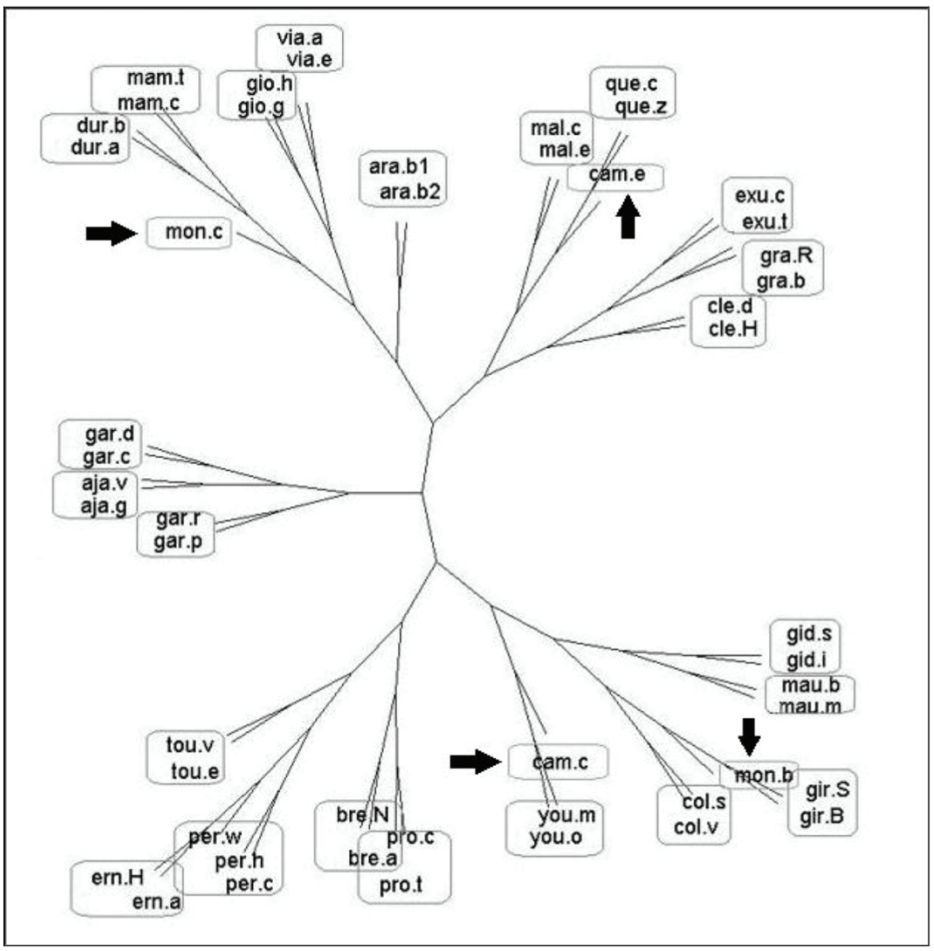


Figure 3: Verbs only. Additive tree built from the coordinates of a totally unsupervised CA of 3547 pages of 50 lines. Frequency threshold for the 726 verbs: 84. *A posteriori* projections of the centroids of pages belonging to a same novel. Misclassified: 2 authors out of 25: Camus, Montherlant (black arrows).

The advantages of the fragmentation of the corpus are the following:

- The structure of the text **inside** each novel is now taken into account, a piece of information overlooked in the classical approach to the single aggregate table of Sections 4.1 and 4.2. This entails a deeper understanding of the internal structure of each text, a finer granularity.

- An external validation evidence can then be achieved using the partition of the initial corpus of texts (the classes of which being the novels).

We are now dealing with an analysis of the 3547 pages of 50 lines (which rather correspond to printed double pages) of the lemmatized file, which can be shuffled like playing cards. Once the typology of these pages has been obtained, the novels are positioned as the average points of their pages. The analysis does not seek to contrast the novels, but to contrast the pages. It is therefore a very severe test.

If we fragment into pages the lemmatized corpus of Section 4.1, 17 authors (out of 25) are well matched. That result will be improved by the fragmentation into pages of the corpus limited to verbs used in Section 4.2.

We mentioned above that verbs were more evenly distributed in the texts than nouns and adjectives. We will now continue working on verbs (pages of verbs) (Figure 3) to observe that verb pages allow better prediction than word pages (lemmas). Indeed, despite the severity of the test, 23 out of 25 authors are characterized by their pages of verbs.

Only the two writers Montherlant and Camus are left to make an exception to this new endeavor to match novels. Note that "Bestiaires" is an autobiographic novel written by the young Montherlant passionate with bullfighting, whereas the second novel "Les célibataires", is dedicated to the sad end of life of two elderly bachelors. For Camus, "L'étranger" is his first novel, and "La chute" his last one.

Evidently, these remarks inspired by external pieces of information are only sketches and hypotheses that can be improved and enriched with all the available tools and parameters of these exploratory phases: levels of fragmentation (paragraphs, pages, chapters), grammatical units (function words, nouns, adjectives), size of the subspace of coordinates, thresholds of frequencies for lexical units.

CONCLUSIONS

At this stage, we have combined several techniques. Regularization through Principal Axes Techniques (here: CA), fragmentation (related or similar to *Word2vec* approach), projection of supplementary (or illustrative) variables, nearest neighbors prediction (through Additive trees representation) that can be expressed either in terms of Neural Networks and Machine Learning, or more aptly in terms of deep learning (Vanni *et al.*, 2018).

About Deep Learning, let us quote the inspiring remark of Le Cun, Bengio and Hinton (Le Cun et al., 2015): "…we expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object".

In the field of textual data analysis, the priority is not systematically "recognition" but discovery, description, comparison, understanding. Such approach remains partially supervised in the sense that both the available external information and the discovered structures are used to enhance the exploration.

But within Machine Learning toolbox, we have selected transparent procedures, interpretable at each step, whose results could be either visualized (planes, trees), or assessed via statistical procedures (bootstrap). Obviously, the selected methods are only a part of the potential of machine learning. But this was the price to pay for the transparency and the algebraic simplicity of the process. Using the arsenal of black boxes available, the machine learns. Using the subset of selected visualization techniques, the researcher learns, we learn.

REFERENCES

Asoh, H. and Otsu, N. (1989). Nonlinear data analysis and multilayer perceptrons. *IEEE, IJCNN*, 89 (2), 411-415.

Balbi, S. (1994). *L'Analisi Multidimensionale dei dati negli anni '90*. Dipartimento di Matematica e Statistica. (Univ. Federico II), Rocco Curto Editore, Napoli.

Balbi, S. and Triunfo, N. (2013). Statistical tools in the joint analysis of closed and open-ended questions. In: *Survey data Collection and Integration*. Davino C., Fabbris L. (eds). Springer Verlag, Berlin. 61-74.

Baldi, P. and Hornik, K. (1989): Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks*, 2: 52-58.

Barthélémy, J.-P. and Guénoche, A. (1988). *Les Arbres et les Représentations de Proximité*. Masson, Paris.

Benzécri, J.-P. (1969): Statistical analysis as a tool to make patterns emerge from clouds. In: *Methodology of Pattern Recognition*, S. Watanabe, (ed.) Academic Press: 35-74.

Bourlard, H. and Kamp, Y. (1988): Auto-association by Multilayers perceptrons and singular value decomposition. *Biological Cybernetics*, 59: 291-294.

Brunet, E., Lebart. and Vanni, L. (2021) Littérature et intelligence artificielle. In: Mayaffre D., Vanni L., (eds) *L'Intelligence Artificielle des Textes*. Honoré Champion, Paris : 73-128.

Buneman, P. (1971). The recovery of trees from measurements of dissimilarity. In: Hodson F. R. D. Kendall G., and Tautu P., (Editors). *Mathematics in the Archeological and Historical Sciences*. Edinburgh University Press: 387-395.

Huson, D.H. and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution*, vol. (23), 2: 254-267.

Lauro, N. C and D'Ambra, L. (1984): L'Analyse non-symétrique des correspondances. In: *Data Analysis and Informatics*, III, Diday *et al.* (eds.), North-Holland: 433-446.

Le Cun, Y., Bengio, Y. and Hinton G. (2015). Deep Learning, *Nature*, 521, 436-444.

Lebart, L., Morineau, A. and Tabard N. (1977). *Technique de la Description Statistique*. Dunod, Paris.

Lebart, L., Morineau A. and Warwick K. (1984). *Multivariate Descriptive Statistical Analysis*. John Wiley and Sons, New York.

Lebart, L. (1992). Discrimination through the regularized nearest cluster method. In: *Computational Statistics*, Y. Dodge *et al*. (eds.), Springer Verlag, Berlin, Heidelberg: 103-118.

Lebart, L. (1997). Correspondence analysis, discrimination and neural networks. In: *Data Science, Classification and Related Methods*. Hayashi C., Ohsumi N., Yajima K., Tanaka Y., Bock H.- H. and Baba Y. (eds), Springer, Berlin, 423-430.

Lebart, L., Pincemin, B. and Poudat, C. (2019). *Analyse des Données Textuelles*, PUQ, Québec, Canada.

Luong, X. (1988). *Méthodes d'Analyse Arborée. Algorithmes, Applications*. Thèse pour le doctorat des sciences. Université Paris V.

Mihaescu, R., Levy, D. and Pachter, L. (2009). Why Neighbor-Joining works? *Algorithmica*, vol. (54): 1-24.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique: Application à l'analyse lexicale par contexte. *Cahiers de l'Analyse des Données*, vol. (3): 187-198.

Saitou, N. and Nei, M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, vol. 4 (4): 406-425.

Sattah, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika*, vol. 42 (3), 319-345.

Schmid, H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Vanni, L., Mayaffre, D. and Longrée, D. (2018). ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables, *JADT 2018, Universitalia,* Rome, ⟨hal-01823560⟩.