ARTICLE IN PRESS

Socio-Economic Planning Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

journal homepage: http://www.elsevier.com/locate/seps

A study of the family service expenditures and the socio-demographic characteristics via fixed marginals correspondence analysis

Luigi D'Ambra [a], Pietro Amenta [b,*], Antonello D'Ambra [c], Jules S. de Tibeiro [d]

[a] *Department of Economics, Management and Institutions, University of Naples, Italy*
[b] *Department of Law, Economics, Management and Quantitative Methods, University of Sannio, Benevento, Italy*
[c] *Department of Economics, University of Campania "Luigi Vanvitelli", Capua, Italy*
[d] *Secteur Sciences, University of Moncton, Canada*

ARTICLE INFO

Keywords:
Contingency table
Family service expenditures
Generalized singular value decomposition
Constrained analysis

ABSTRACT

Our study focuses on two data set, the former provides the expenditures for several services for each family and the latter contains socio-demographic variables for the same statistical units. The main aim is to analyze, in a Correspondence Analysis context, the service expenditure of families based on the whole given data-set under two types of constraints: the global relative expenses for a given service and the global relative expenses for a given socio-demographic category. The purpose of measuring the relationship between expenditure on social services and the socio-demographic characteristics of families is conducted in an exploratory and predictive perspective. A new approach is then introduced which ensures compliance with the required constraints. Moreover, through a procedure, we have obtained a table of regression coefficients. This table shows interesting properties and it is easy to interpret. Finally, the performance of the results has been evaluated using computer-based resampling techniques.

1. Introduction

The consumption expenditure of households (the amount spent on goods and services at household level) becomes more and more significant worldwide. As revenues increase, people now have access to a multitude of consumer goods and services related to greater prosperity. In this context, service costs tend to be an important part of the family budget. The effects of these costs in several contexts (like in tourism, health, instruction) likewise their related factors, are at the center of many studies in the literature (e.g. Refs. [6,22,31,35]). Our study focuses on a data set that provides the expenditures for several services (transport, communication services, environment and tributes for the territory, health, education, culture, tourism, utilities and social assistance services) for each individual (the head of the given family household). Moreover, another data set is defined, which contains identifier or socio-demographic variables (sex, age, income, level of education and profession) for the same individuals (observations). We thus have two tables crossing the same individuals but with different types of variables. The main task is to analyze the service expenditure of families based on the whole given data set under two type of constraints: the global relative expenses for a given service and the global relative expenses for a given socio-demographic category. Indeed, the core matrix of this new approach cross-tabulates the service expenditure categories with the socio-demographic categories, keeping the same original column margins of the original matrices. These margins then become the constraints of our analysis and cannot be modified.

Given the quantitative nature of the data, it is possible to analyze the family expenses with respect to the socio-demographic characteristics using several methodologies. Indeed, the asymmetrical relationships between two sets of quantitative variables have been studied extensively in the literature. In addition to the ordinary multivariate regression model, after the seminal Rao's paper [33], several authors considered the problem in different contexts and with different approaches: we can mention, among others, lzenman [25], Davies and Tso [11], Robert and Escoufier [34] in the framework of the linear multivariate model, Van den Wollenberg [45] in the context of Redundancy Analysis, D'Ambra and Lauro [9] constraining the PCA solutions to a reference subspace, Rao [33] and later Sabatier et al. [36] combining PCA and multivariate regression analysis. Core matrix of all these approaches is the image of the dependent variables onto the subspace spanned by the explanatory data sets and it is given by the matrix $\boldsymbol{\Xi}_Z\mathbf{X}$, where $\boldsymbol{\Xi}_Z$ is the orthogonal projector onto the subspace spanned by the columns of matrix $\mathbf{Z}$. On

* Corresponding author.
E-mail address: amenta@unisannio.it (P. Amenta).

https://doi.org/10.1016/j.seps.2020.100833
Received 15 November 2019; Received in revised form 12 March 2020; Accepted 16 March 2020
Available online 2 April 2020

closer inspection, this matrix is a part of the Takane and Shibayama's decomposition [39,40] of the matrix $\mathbf{X}$

$$\mathbf{X} = \Xi_Z\mathbf{X}\Xi_H + \Xi_Z^{\perp}\mathbf{X}\Xi_H + \Xi_Z\mathbf{X}\Xi_H^{\perp} + \Xi_Z^{\perp}\mathbf{X}\Xi_H^{\perp} \quad (1)$$

with $\mathbf{H}$ column external information and $\Xi_Z^{\perp} = \mathbf{I} - \Xi$ projector onto the ortho-complement subspace spanned by $\mathbf{Z}$. Important property of this decomposition is that the sum of squares of $\mathbf{X}$ (global inertia) can be decomposed according to the sum of inertia of each submatrix (sums of squares) in the right hand of (1) with specific statistical meanings [39, 41]. In the absence of column external information ($\mathbf{H} = \mathbf{I}$), decomposition (1) amounts to $\mathbf{X} = \Xi_Z\mathbf{X} + \Xi_Z^{\perp}\mathbf{X}$ and so $\Xi_Z\mathbf{X}$ is a part of $\mathbf{X}$. This implies that matrix $\Xi_Z\mathbf{X}$ cannot have the same margins of $\mathbf{X}$ and so all these approaches are not appropriate in our study not ensuring compliance with the required constraints. A different approach must therefore be followed. We point out that variables of matrix $\mathbf{X}$ have the same unit of measurement and so the sums by each row and column have both a precise meaning: the expense for each service supported by a given family and the global expense for a given service, respectively. Matrix $\mathbf{X}$ can then be assimilated to a contingency table èpossessing the same characteristics and we can consider $\mathbf{Z}$ as a matrix of external information that defines a row partition of matrix $\mathbf{X}$. Severaòl authors incorporate external information, through linear constraints on the row and/or column scores of Correspondence Analysis (CA) (e.g. Refs. [3,4, 19,24,30,38,41,44,46]) to obtain a more parsimonious and easier-to-understand representation of the data. Nothing guarantees however that the core matrices of all these constrained approaches ensure compliance with the required constraints. Cazes et al. [7] and Escofier [16] instead optimally decomposed the between group inertia of the row profile cloud of a pseudo-contigency table $\mathbf{X}$ using matrix $\mathbf{Z}$ (binary matrix of a single variable) as row partition matrix. We point out that some core matrices of both approaches keep the original row and column margins as we require for our core matrix. Moreover, Lebart et al. [26] faced geometrically the problem of the asymmetrical relationships through a sort of "visualized regression" by means of the supplementary points technique. We can then study, in a CA context, the service expenditure table with respect to additional socio-demographics characteristics under the required constraints by using a new strategy developed on the main theoretical results of [7,16,26].

The paper is organized as follows. In section 2, we specify the notation and the basic definitions. The new approach is introduced in Section 3. An empirical study is shown in section 4.

2. Notation and basic definitions

Let I rows and R columns be collected into the $I\times R$ dependent matrix $\mathbf{X}$ (the consumption expenditure of household) with entries x_{ir} ($i = 1\dots I$ and $r = 1\dots R$). Let $x_{i+} = \sum_{r=1}^{R} x_{ir}$ and $x_{+r} = \sum_{i=1}^{I} x_{ir}$ denote the sums of the i-th row and r-th column, respectively, and $x_{++} = \sum_{i=1}^{I}\sum_{r=1}^{R} x_{ir}$ denotes the grand total of $\mathbf{X}$. Moreover, we define the matrix $\mathbf{P}_X = x_{++}^{-1}\mathbf{X}$ (called correspondence matrix) with entries p_{ir}^X, $\mathbf{D}_I^{P_X} = diag(p_{i+}^X)$ and $\mathbf{D}_R^{P_X} = diag(p_{+r}^X)$ which are diagonal matrices with entries $p_{i+}^X = \sum_{r=1}^{R} p_{ir}^X$ and $p_{+r}^X = \sum_{i=1}^{I} p_{ir}^X$, respectively. Let I rows and C columns be collected into the $I\times C$ independent matrix $\mathbf{Z}$ (socio-demographic data) with entries z_{ic} ($i = 1,\dots,I$ and $c = 1,\dots,C$) (we have adopted for each qualitative variable a complete disjunctive form (0,1)). Let $z_{+c} = \sum_{i=1}^{I} z_{ic}$ and $x_{i+} = \sum_{c=1}^{C} z_{ic}$ denote the sums of the i-th row and r-th column, respectively, and $z_{++} = \sum_{i=1}^{I}\sum_{c=1}^{C} z_{ic}$ denotes the grand total of $\mathbf{Z}$. Furthermore, we define the matrix $\mathbf{P}_Z = z_{++}^{-1}\mathbf{Z}$ with entries p_{ic}^Z, $\mathbf{D}_I^{P_Z} = diag(p_{i+}^Z)$ and $\mathbf{D}_C^{P_Z} = diag(p_{+c}^Z)$ be the diagonal matrices with entries $p_{i+}^Z = \sum_{c=1}^{C} p_{ic}^Z$ and $p_{+c}^Z = \sum_{i=1}^{I} p_{ic}^Z$ respectively. Finally, we define the matrix $\mathbf{Y} = x_{++}\mathbf{D}_I^{P_X}\mathbf{Z}$ with entries y_{ic}, $y_{++} = \sum_{i=1}^{I}\sum_{c=1}^{C} y_{ic}$ denotes the grand total of $\mathbf{Y}$, such that $\mathbf{P}_Y = y_{++}^{-1}\mathbf{Z}$, $\mathbf{D}_I^{P_Y} = diag(p_{i+}^Y)$ and $\mathbf{D}_C^{P_Y} = diag(p_{+c}^Y)$ be the diagonal matrices with entries $p_{i+}^Y = \sum_{c=1}^{C} p_{ic}^Y$ and $p_{+c}^Y = \sum_{i=1}^{I} p_{ic}^Y$, respectively.

The Generalized Singular value Decomposition (hereafter GSVD) [21,47] is a matrix decomposition which subsumes the most famous Singular value Decomposition (SVD). The GSVD of a real matrix $\mathbf{A}$ of order $(n\times p)$ is defined as $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ such that $\mathbf{U}^T\mathbf{\Omega}\mathbf{U} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{\Phi}\mathbf{V} = \mathbf{I}$, and where $\mathbf{\Omega}$ and $\mathbf{\Phi}$ are given positive definite symmetric matrices of order $(n\times n)$ and $(p\times p)$, respectively, with $\mathbf{\Lambda}$ diagonal matrix containing the generalized singular values in descending order. Further, if $\mathbf{\Omega} = \mathbf{I}$ and $\mathbf{\Phi} = \mathbf{I}$ then the GSVD reduces to SVD, explaining the name. GSVD refers so to an SVD under non-identity metrics and it is noted $\mathrm{GSVD}(\mathbf{A})_{\mathbf{\Omega},\mathbf{\Phi}}$.

Lastly, the characterizing set of a statistical multivariate study [5,17, 18] is formed by a matrix $\mathbf{X}$ of order $(n\times p)$, an $(n\times n)$ weights matrix $\mathbf{D} = \mathrm{diag}(d_1,\dots,d_n)$ and a non negative definite metric $\mathbf{Q}_{\mathbf{X}}$ of order $(p\times p)$ so that the squared distance between two statistical units $\mathbf{x}_j$ and $\mathbf{x}_k$ is given by $(\mathbf{x}_j - \mathbf{x}_k)^T\mathbf{Q}_{\mathbf{X}}(\mathbf{x}_j - \mathbf{x}_k)$. From a geometrical point of view, it is like searching the principal axes $\mathbf{w}_k$ ($k = 1,\dots,p$) of maximum variance directions of a cloud of n points (rows of $\mathbf{X}$) of $\Re^p$ or looking for the principal components $\mathbf{t}_k = \mathbf{X}\mathbf{Q}_{\mathbf{X}}\mathbf{w}_k$ ($\mathbf{X}$ scores) of a cloud of p points in $\Re^n$ (columns of $\mathbf{X}$). The solutions involve the eigen-decomposition of the matrices $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}_{\mathbf{X}}$ and $\mathbf{X}\mathbf{Q}_{\mathbf{X}}\mathbf{X}^T\mathbf{D}$, respectively. They are then the basic matrices of a ($\mathbf{D}$ and $\mathbf{Q}_{\mathbf{X}}$) metric-based Principal Component Analysis (PCA) of matrix $\mathbf{X}$ [5,17,18]. It is noted by $\mathrm{PCA}(\mathbf{X},\mathbf{Q}_{\mathbf{X}},\mathbf{D})$ and it is equivalent to $\mathrm{GSVD}(\mathbf{X})_{\mathbf{D},\mathbf{Q}_{\mathbf{X}}}$. Further, if $\mathbf{D} = \frac{1}{n}\mathbf{I}$ and $\mathbf{Q}_{\mathbf{X}} = \mathbf{I}$ then it corresponds to PCA in the original scales and to a straightforward SVD on $\mathbf{X}$.

3. A constrained method with fixed marginals (FMNSCA)

First step of our strategy is based on a CA applied to the matrix $\mathbf{S}_Y$

$$\mathbf{S}_Y = \left(\mathbf{P}_Y - \mathbf{D}_I^{\mathbf{P}_Y}1_I1_C^T\mathbf{D}_C^{\mathbf{P}_Y}\right).$$

CA solutions are then computed by using the GSVD of matrix $\mathbf{S}_Y$ with metric $(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}$ and $(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}$. This is written as:

$$\mathrm{GSVD}(\mathbf{S}_Y)_{(\mathbf{D}_I^{\mathbf{P}_Y})^{-1},(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}} \Rightarrow \mathbf{S}_Y = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$$

with $\tilde{\mathbf{U}}^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{U}} = \mathbf{I}_I$ and $\tilde{\mathbf{V}}^T(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{V}} = \mathbf{I}_C$, where $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are singular vectors of dimension $I\times M$ and $C\times M$, respectively. $\tilde{\mathbf{\Lambda}}$ is diagonal matrix of the singular values $\tilde{\lambda}_m$ (arranged in descending order such that $\tilde{\lambda}_1 \geq \tilde{\lambda}_2 \geq \dots \geq \tilde{\lambda}_m \geq 0$) with $m = 1,2,\dots,M$ and $M = rank(\mathbf{S}_Y)$. The profile coordinates are given by:

$$\tilde{\mathbf{F}}_{CA} = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}} \qquad \tilde{\mathbf{G}}_{CA} = (\mathbf{D}_C^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{V}}\tilde{\mathbf{\Lambda}}$$

Therefore, according to Ref. [26], we denote with $\tilde{\mathbf{G}}_{\mathbf{S}_Y}^{\mathbf{P}_X}$ the supplementary coordinates profile column of the correspondence matrix $\overline{\mathbf{P}}_X = \mathbf{P}_X - \mathbf{D}_I^{\mathbf{P}_X}1_I1_R^T\mathbf{D}_R^{\mathbf{P}_X}$

$$\tilde{\mathbf{G}}_{\mathbf{S}_Y}^{\mathbf{P}_X} = (\mathbf{D}_R^{\mathbf{P}_X})^{-1}\overline{\mathbf{P}}_X^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{U}}$$

Using the classical formula for recreating the matrix by means the coordinates and singular values, it's possible to build the matrix $\mathbf{Q}$ of dimension $R\times C$:

$$\mathbf{Q} = \mathbf{D}_R^{\mathbf{P}_X}\tilde{\mathbf{G}}_{\mathbf{S}_Y}^{\mathbf{P}_X}\tilde{\mathbf{\Lambda}}^{-1}\tilde{\mathbf{G}}_{CA}^T\mathbf{D}_C^{\mathbf{P}_Y} + \mathbf{D}_R^{\mathbf{P}_X}1_R1_C^T\mathbf{D}_C^{\mathbf{P}_Y} = \check{\mathbf{Q}} + \mathbf{D}_R^{\mathbf{P}_X}1_R1_C^T\mathbf{D}_C^{\mathbf{P}_Y}$$

where $\check{\mathbf{Q}} = \mathbf{D}_R^{\mathbf{P}_X}\tilde{\mathbf{G}}_{\mathbf{S}_Y}^{\mathbf{P}_X}\tilde{\mathbf{\Lambda}}^{-1}\tilde{\mathbf{G}}_{CA}^T\mathbf{D}_C^{\mathbf{P}_Y}$ is a regression coefficient matrix [27] of service expenditures explained by the socio-demographics characteristics in a PCA context.

The regression coefficient matrix $\check{\mathbf{Q}}$ shows interesting properties: it is a doubly centred matrix with respect to the metric weights. Also the matrix $\mathbf{S}_Y\check{\mathbf{Q}}^T$ is a doubly centred matrix with respect to matrices $\mathbf{D}_I^{\mathbf{P}_Y}$ and

$\mathbf{D}_R^{\mathbf{P}_X}$. This implies that the sum of the predicted values of service expenditures, given by the matrix $(\mathbf{S}_Y\check{\mathbf{Q}}^T + \mathbf{D}_I^{\mathbf{P}_X}1_I1_R^T\mathbf{D}_R^{\mathbf{P}_X})$ of order $(I \times R)$, will be equal to the sum of those observed as in the usual regression model.

Moreover, rows and column margins of matrix $\mathbf{Q}$ are equal to the column margins of matrices $\mathbf{P}_X$ and $\mathbf{P}_Y$, respectively. Due to these characteristics, matrix $\mathbf{Q}$ can be equated to a pseudo contingency table which cross classifies the service expenditure categories with respect to the socio-demographic characteristics of the family, ensuring compliance with the required constraints of our problem.

Consider now the previous GSVD decomposition of matrix $\mathbf{S}_Y = \tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$ and left multiply it by $(\mathbf{D}_I^{P_Y})^{-1/2}$. We obtain $(\mathbf{D}_I^{P_Y})^{-1/2}\mathbf{S}_Y = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\tilde{\mathbf{U}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$. It can also be written $(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\mathbf{S}_Y = \tilde{\mathbf{U}}^*\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$ with $\tilde{\mathbf{U}}^* = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\tilde{\mathbf{U}}$. $\tilde{\mathbf{U}}^*$ is the left singular vector of the GSVD of matrix $(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\mathbf{S}_Y$

$$\text{GSVD}\left((\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\mathbf{S}_Y\right)_{\mathbf{I},(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}} \Rightarrow (\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\mathbf{S}_Y = \tilde{\mathbf{U}}^*\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T$$

such that $\tilde{\mathbf{U}}^{*T}\tilde{\mathbf{U}}^* = \mathbf{I}_I$ and $\tilde{\mathbf{V}}^T(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{V}} = \mathbf{I}_C$, which shares the same right singular vector $\tilde{\mathbf{V}}^T$ of $\text{GSVD}(\mathbf{S}_Y)_{(\mathbf{D}_I^{\mathbf{P}_Y})^{-1},(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}}$.

We highlight that, by straightforward algebra, it is now possible to write again the matrix $\check{\mathbf{Q}}$ of order $R \times C$ in a different way. Indeed, replacing the definitions of $\tilde{\mathbf{G}}_{\mathbf{S}_Y}^{\mathbf{P}_X}$ and $\tilde{\mathbf{G}}_{CA}$ in $\check{\mathbf{Q}}$ we can write

$$\begin{aligned}
\check{\mathbf{Q}} &= \mathbf{D}_R^{\mathbf{P}_X}\tilde{\mathbf{G}}_{\mathbf{S}_Y}^{\mathbf{P}_X}\tilde{\mathbf{\Lambda}}^{-1}\tilde{\mathbf{G}}_{CA}^T\mathbf{D}_C^{\mathbf{P}_Y} \\
&= \mathbf{D}_R^{\mathbf{P}_X}\left[(\mathbf{D}_R^{\mathbf{P}_X})^{-1}\overline{\mathbf{P}}_X^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{U}}\right]\tilde{\mathbf{\Lambda}}^{-1}\left[\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}^T(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}\right]\mathbf{D}_C^{\mathbf{P}_Y} \\
&= \overline{\mathbf{P}}_X^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\tilde{\mathbf{U}}\tilde{\mathbf{V}}^T \\
&= \overline{\mathbf{P}}_X^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\widetilde{\mathbf{U}^*}\tilde{\mathbf{V}}^T \\
&= \left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}^*\tilde{\mathbf{V}}^T
\end{aligned}$$

and so $\check{\mathbf{Q}} = \left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}\tilde{\mathbf{V}}^T$. This matrix allows us to clarify how to address our problem.

To analyze the structure of the association between row and column categories, we carry out then a CA of matrix $\check{\mathbf{Q}}$ with metrics $(\mathbf{D}_R^{\mathbf{P}_X})^{-1}$ and $(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}$ or, equivalently, a PCA of matrix $\check{\mathbf{Q}}$ by using matrices $\mathbf{I}_R$ and $(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}$ as row and column metrics, respectively, if we consider $\check{\mathbf{Q}}$ as a quantitative matrix

$$\text{PCA}\left[\check{\mathbf{Q}}, \mathbf{I}_R, (\mathbf{D}_C^{\mathbf{P}_Y})^{-1}\right] \tag{2}$$

As clarified in the last paragraph of Section 2, this PCA amounts to

$$\text{GSVD}\left(\left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}^*\tilde{\mathbf{V}}^T\right)_{\mathbf{I}_R,(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}}$$

and solutions (right singular vectors $\mathbf{v}$) are given by the eigen-decomposition

$$\left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}^*\tilde{\mathbf{U}}^{*T}(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\overline{\mathbf{P}}_X\mathbf{v} = \lambda\mathbf{v}$$

Since $\widetilde{\mathbf{U}^*}\widetilde{\mathbf{U}^*}^T$ is the orthogonal projector onto the subspace spanned by the columns of matrix $\mathbf{S}_Y$, the previous PCA is then equivalent to a PCA applied to matrix $\widetilde{\mathbf{U}^*}\widetilde{\mathbf{U}^*}^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\overline{\mathbf{P}}_X$. We highlight that this last eigen-decomposition is also equivalent to

$$\left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}\tilde{\mathbf{U}}^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X\mathbf{v} = \lambda\mathbf{v}$$

and then $\text{PCA}[\check{\mathbf{Q}}, \mathbf{I}_I, (\mathbf{D}_C^{\mathbf{P}_Y})^{-1}]$ amounts to the study of the inertia of the family expenditures $(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X$ with respect to the subspace spanned by the socio-demographic characteristics.

We point out that the solution of our approach (2) can be written in a different way using the pseudo contingency $\mathbf{Q}$. Indeed, solution (2) is also equivalent to

$$\text{GSVD}\left[\mathbf{Q}(\mathbf{D}_C^{\mathbf{P}_Y})^{-1} - \mathbf{D}_R^{\mathbf{P}_X}1_R1_C^T\right]_{\mathbf{I}_R,\mathbf{D}_C^{\mathbf{P}_Y}} \tag{3}$$

amounting then to perform a CA with suitable metrics which describes the association where one categorical variable is considered as logically explanatory of the other, that is the aim of the Non Symmetrical Correspondence Analysis [10]. Due to the non symmetrical role played by the categorical variables, we name our approach “Fixed Marginals Non Symmetrical Correspondence Analysis" (FMNSCA). This approach analyze the difference between the columns profile respect to rows. Moreover, it is possible to show that this approach amounts also to a PCA of matrix $\mathbf{B} = \left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}\tilde{\mathbf{V}}^T(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}$ by using matrices $\mathbf{I}_R$ and $\mathbf{D}_C^{\mathbf{P}_Y}$ as row and column metrics, respectively

$$\text{PCA}\left[\mathbf{B}, \mathbf{I}_R, \mathbf{D}_C^{\mathbf{P}_Y}\right] \tag{4}$$

thus obtaining another form of solution to our problem.

According to Ref. [26] it is possible to use the above results to compute the correlations between the coordinates of the service expenditure categories and of a new profile of household $\mathbf{y}^+$. It can be achieved by the scalar product of the projected and normalized new profile $\mathbf{y}^+$ and row profiles $(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X$ onto the subspaces spanned by matrix $\mathbf{S}_Y$ in supplementary way

$$\left\{\frac{\left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}}{\left\|\left[(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X\right]^T\tilde{\mathbf{U}}\right\|}\right\} \times \left\{\frac{\tilde{\mathbf{V}}^T(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}\mathbf{y}^+}{\left\|\tilde{\mathbf{V}}^T(\mathbf{D}_C^{\mathbf{P}_Y})^{-1}\mathbf{y}^+\right\|}\right\}$$

3.1. Links with other approaches

FMNSCA belongs to the family of constrained approaches defined with respect to co-inertia analysis (COA) [13] and named Generalized Constrained Co-Inertia Analysis (GCCOA) [1]. This approach incorporates external information in COA. COA has been introduced to study symmetrical interdependence relationships between two sets of variables $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ obtained in experimental applications. COA is based on the covariance criterion and it improves the Canonical Correlation Analysis (CCA) [23] that may lead to very high correlated and uninteresting pairs of canonical variables. After decomposing each matrix according to (1) GCCOA applies COA to any pair of derived matrices to study relationships between them. GCCOA subsumes several constrained approaches proposed in literature [1]. We point out that FMNSCA is a special case of GCCOA with $\tilde{\mathbf{X}} = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X$, $\tilde{\mathbf{Y}} = \mathbf{I}$ and $\mathbf{S}_Y$ as row external information matrix.

FMNSCA is also linked to other two COA-based approaches: Co-Correspon-dence Analysis (Co-CA) [43] and Canonical Correspondence Analysis (C-CA) [42]. Co-CA combines the ideas of co-inertia analysis with the unimodal response model familiar to CA or C-CA methods. The aim is to relate two species abundance or occurrence matrices such that the resulting decomposition into axes are those combinations that best explain the covariation between species and observations in the two matrices. By straightforward algebra, it can be shown that symmetric co-correspondence analysis is the co-inertia analysis of the statistical triplets $(\mathbf{Q}_X, \mathbf{K}_X, \mathbf{R}_X)$ and $(\mathbf{Q}_Y, \mathbf{K}_Y, \mathbf{R}_Y)$ with

Table 1
Categories of the explicative variables.

Sex	Age	Income (€)	Education	Profession
M−male	Age1 - < 35 years	Inc1 - < 20000	Edu1 - Primary	Free - Freeland
F - female	Age2 - 35–45 years	Inc2 - 20000-50000	Edu2 - Secondary	Emp - Employee
	Age3 - > 45 years	Inc3 - > 50000	Edu3 - University	Ret - Retiree

Source: our elaboration.

$\mathbf{Q}_X = \mathbf{R}_X^{-1}\mathbf{X}\mathbf{K}_X^{-1} - 1_n 1_p^T$ and $\mathbf{Q}_Y = \mathbf{R}_Y^{-1}\mathbf{Y}\mathbf{K}_Y^{-1} - 1_n 1_q^T$. It is possible to show that (predictive) Co-CA subsumes FMNSCA (see Eqns. (1)–(3)), only for the axis solution of order 1, with the settings $\mathbf{Q}_X = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\overline{\mathbf{P}}_X$, $\mathbf{K}_X = \mathbf{I}$, $\mathbf{R}_X = \mathbf{I}$, $\mathbf{Q}_Y = \tilde{\mathbf{U}}\tilde{\mathbf{U}}^T$, $\mathbf{K}_Y = \mathbf{I}$ and $\mathbf{R}_Y = \mathbf{I}$. The solutions of both methods diverge starting from the second axis due to the different calculation method. All the FMNSCA solutions are computed in a single step using definition (3) while (predictive) Co-CA obtains the remaining axes placing them in a PLS framework [12,28] with additional constraint to the maximization of co-inertia criterion.

Canonical Correspondence Analysis is formally the co-inertia of the triplets $(\mathbf{Q}_X, \mathbf{K}_X, \mathbf{R}_X)$ and $(\mathbf{Z}, (\mathbf{Z}^T\mathbf{R}_X\mathbf{Z})^{-1}, \mathbf{R}_X)$ with $\mathbf{Z}$ being the matrix containing the environmental data. By straightforward algebra, it is possible to show that C-CA subsumes FMNSCA (see Eqns. (1)–(3)) with the settings $\mathbf{Q}_X = \overline{\mathbf{P}}_X$, $\mathbf{K}_X = \mathbf{I}$, $\mathbf{R}_X = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1}$, $\mathbf{Q}_Y = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1/2}\mathbf{Y}$, $\mathbf{K}_Y = \left[\mathbf{Y}^T(\mathbf{D}_I^{\mathbf{P}_Y})^{-1}\mathbf{Y}\right]^{-1}$ and $\mathbf{R}_Y = (\mathbf{D}_I^{\mathbf{P}_Y})^{-1}$.

Lastly, we point out that it is possible to include C–Ca and Co-CA in the family of constrained approaches GCCOA by using suitable settings.

4. Case study

In this study we consider a selection of a large data set provided by the Project "Covi - the Vesuvius coast" (funded by the Campania region in 2015) aimed to point out the main characteristics of family expenses in the neighborhoods of Naples by a survey. The selected subset, which was the subject of our case study, was made up of homogeneous records of expenses in public service incurred in 2015 by 1660 families on the following items: Transport, Communication Services (C.S.), Environment and Tributes for the Territory (E.&T.T.), Health, Education, Culture, Tourism, Utilities and Social Assistance Services (S.A.S.). The survey was conducted using stratified sampling with respect to the neighborhoods of Naples (Italy) such that the sample size of each stratus was proportional to the population of each neighborhood. The sample size was obtained using a confidence level of 95% and a sampling error of 2.5%. Additional socio-demographic variables of the survey in our subset were: gender, age, income, education level and profession. The survey then provided an additional table of heterogeneous set of variables expressed in different physical units. We adopted a complete disjunctive form (0, 1) for all the variables after recoding the quantitative ones in groups (Table 1). The list of additional variables of the dataset of the heads of family has been structured as follows:

Main aim of our study was to analyze the family expenses in public service in the neighborhoods of Naples under two type of constraints: the global relative expenses for a given service and the global relative expenses for a given socio-demographic category. The amount of these expenses must not change when analyzed statistically. FMNSCA ensures these required conditions and so it has been used to analyze these data.

Table 2 shows the percentages of FMNSCA inertia according to formula (3).

The two first axes indicate an acceptable proportion of explained inertia (about 88%). The break (from axis 2 to axis 3) between the steep slope and a leveling off among the rectangles of the histogram (Scree Plot) in Table 2 suggests the same number of meaningful factors to retain [8], different from random error. Since the eigenvalues of Table 2 are a decomposition of the inertia of the family expenditures with respect to the subspace spanned by the socio-demographic characteristics and their sum is proportional [10] to the Goodman and Kruskal's predicability index [20], then the first plane highlights an high part of the association between family expenditures and the socio-demographic

Table 2
FMNSCA inertia.

		Explained	Cumulated	
Axis	Inertia	Inertia (%)	Inertia (%)	Scree Plot - Histogram
1	0.226	74.485	74.485	□□□□□□□□□□□□□□□□□□□□□□□□□
2	0.040	13.313	87.798	□□□□
3	0.022	7.131	94.929	□□
4	0.010	3.144	98.073	□
5	0.004	1.217	99.290	|
6	0.002	0.616	99.906	|
7	0.000	0.087	99.993	|
8	0.000	0.007	100.000	
Total	0.303	100.000	–	

Source: our elaboration.

Table 3
Regression coefficients.$\check{\mathbf{Q}}$

Label	M	F	Age1	Age2	Age3	Income1	Income2	Income3	Educ1	Educ2	Educ3	Freelance	Employee	Retiree
Culture	0.118	−0.118	0.349	−0.750	0.400	**−1.536**	0.347	1.189	−0.207	0.283	−0.076	−0.316	1.228	−0.912
E.&T.T.	0.483	−0.483	−0.453	0.740	−0.287	0.346	−0.011	−0.336	−0.385	−0.146	0.531	−0.351	0.090	0.262
C.S.	−0.615	0.615	0.326	0.394	−0.720	**1.364**	−0.443	−0.922	**0.844**	−0.134	−0.710	−0.094	−1.026	1.120
Transport	0.217	−0.217	0.234	0.497	−0.732	**1.949**	−0.685	−1.264	0.032	−0.334	0.302	−0.436	−1.167	1.603
Health	−0.132	0.132	0.653	0.070	−0.723	1.287	−0.370	−0.917	0.698	0.263	−0.961	−0.128	**−2.053**	**2.181**
Tourism	−0.438	0.438	−0.717	−0.723	**1.440**	−0.407	−0.849	**1.256**	**−1.558**	−0.983	**2.540**	−0.719	2.084	−1.365
Utilities	−0.815	0.815	−0.623	0.915	−0.293	0.376	1.075	−1.450	0.718	0.705	−1.423	1.589	−1.776	0.187
S.A.S.	0.992	−0.992	0.492	−0.392	−0.099	**1.451**	−0.557	−0.895	0.611	0.249	−0.860	1.059	−2.047	0.988
Education	0.259	−0.259	−0.771	1.051	−0.280	−0.281	0.951	−0.670	**0.944**	0.031	−0.975	**1.905**	−0.791	−1.114

Source: our elaboration.

Table 4
Bootstrap p-values of the regression coefficients.

Label	M	F	Age1	Age2	Age3	Income1	Income2	Income3	Educ1	Educ2	Educ3	Freelance	Employee	Retiree
Culture	0.930	0.931	0.487	0.070	0.367	**0.000**	0.397	**0.013**	0.763	0.406	0.610	0.617	**0.017**	0.068
E.&T.T.	0.135	0.135	0.245	**0.021**	0.459	0.283	0.815	0.245	0.309	0.781	0.224	0.514	0.918	0.457
C.S.	0.121	0.131	0.386	0.345	0.071	**0.001**	0.204	**0.042**	**0.025**	0.564	0.186	0.584	**0.046**	**0.008**
Transport	0.660	0.662	0.642	0.329	0.186	**0.000**	0.137	**0.026**	0.992	0.381	0.361	0.239	**0.029**	**0.001**
Health	0.868	0.869	0.154	0.899	0.146	**0.008**	0.235	0.152	0.134	0.821	0.177	0.560	**0.000**	**0.000**
Tourism	0.424	0.423	0.165	0.222	**0.032**	0.545	0.142	0.088	**0.000**	0.068	**0.000**	0.398	**0.014**	**0.005**
Utilities	0.070	0.080	0.242	0.083	0.789	0.697	**0.051**	0.059	0.199	0.223	0.055	**0.011**	**0.037**	0.819
S.A.S.	0.062	**0.049**	0.551	0.518	0.939	0.081	0.507	0.444	0.605	0.977	0.638	0.304	0.052	0.465
Education	0.472	0.472	0.086	**0.010**	0.736	0.284	0.029	0.313	0.049	0.922	0.118	**0.000**	0.272	**0.003**

Source: our elaboration.

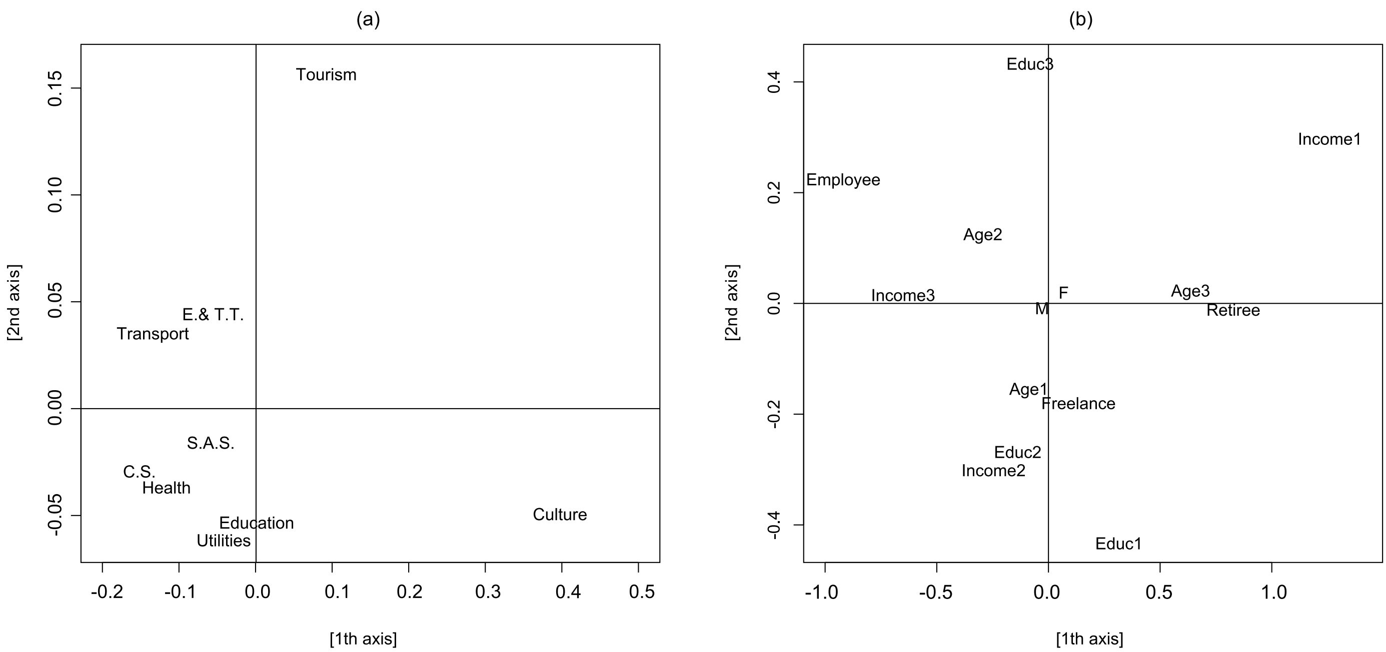

Fig. 1. FMNSCA plots. First plane. Source: our elaboration.

characteristics.

A first interpretation of the links between all the variables of the survey is given by Table 3 that shows the regression coefficients $\check{\mathbf{Q}}$. These coefficients give the links between the expenses in public services and the socio-demographic categories (sex, age, income, level of education and profession) about each family (for all factors). Reading these values, we can appreciate the different positive or negative impacts (in terms of spending propensity), in predicting new service expenditures for a new given category membership of a family. For example we can highlight the very low spending propensity for the "Culture" if the family belongs to category *Income 1* (−1.536). In the meantime, if a family belongs to category *Income 1* then it tends to allocate more money for "Transport" costs (1.949), "S.A.S." (1.451), "C.S." (1.364) and "Health" (1.287). With the same reading key, we can point out the negative impacts of *Employee* for "Health" (−2.053) and "S.A.S." (−2.047); and of *Educ 1* for "Tourism" (−1.558), whereas the coefficients associated to "*Age 2″*, is important, in positive, for "E&T.T." (0.740), "Utilities" (0.915) and "Education" (1.051). Moreover, the coefficients associated to "*Educ 1″* is important in positive for "C.S." (0.844) and "Education" (0.944), whereas the category "*Employee*" is important, in positive, for "Culture" (1.228), while families with an high education or with an high income or older than 45 years tends all to allocate more money mainly for "Tourism" (2.540, 1.256 and 1.440, respectively). Finally, categories "Freelance" and "Retiree" result to be positive related to "Education" (1.905) and to "Health" (2.181), respectively.

We point out that, unfortunately, we do not know the exact probability distribution of the regression coefficients and consequently we cannot perform the statistical tests. For this reason to verify if the relation between the matrix **X** and **Z** is significance we use the bootstrap procedure [14]. The bootstrap method provides a way of assessing the stability of parameter estimates. A form of nonparametric bootstrapping is used to derive empirical distributions of parameter estimates. We draw a large number of samples (e.g. 1000) with replacement from $\mathbf{K} = [\mathbf{X}|\mathbf{Z}]$. Each sample produce a set matrix $\mathbf{K}^{boot}$, following our approach we compute the $\check{\mathbf{Q}}^{boot}$, we count the number of times that $\check{\mathbf{Q}}$ estimate is greater than a given significance level (e.g. 0.05 or 0.01). Table (4) shows the bootstrap p-values of the regression coefficients where we can appreciate which coefficients values (in bold) for the socio-demographic categories have a non-random nature in explaining the propensity to spend.

Let consider the FMNSCA plots of Fig. 1. We note that the socio-demographic categories "M″", "F″" are very close to the origin of the axes while "Age1″", "Age2″ and "Age3″" are not very far from it (Fig. 1 (b)). This means that these categories have no effect on the distribution of the service expenditures, while several categories of variables Income, Education and Profession, which are located on the ends of both axes, seem to influence the service expenditures (Fig. 1(b)).

Table 5 shows the contribution of each service expenditure to the first factorial plane (axes 1–2) and expresses the proportion of variability

Table 5
Absolute contributions.

Label	Variable	Axis 1	Axis 2
Culture	Culture	**0.724**	0.052
E.&T.T.	Environment and Tributes for the Territory	0.011	0.043
C.S.	Communication Services	0.069	0.020
Transport	Transport	0.095	0.032
Health	Health	0.042	0.028
Tourism	Tourism	0.043	**0.646**
Utilities	Utilities	0.008	0.098
S.A.S.	Social Association Services	0.008	0.005
Education	Education	0.000	0.076

Source: our elaboration.

of that factor (determined by its eigenvalue) accounted for by that variable. These contributions are an aid in interpreting the dimensions, and points with relatively large contributions are most important to the dimension concerned. It takes values between zero (without contribution) and one (which would indicate that axis α is determined only by that element). We observe the important role of "Culture" and "Tourism"in determining the variance of both axes and therefore providing the higher contributions to the orientation of the first factorial plane (Fig. 1(a)). The other variables show low contributions to orientation and so they do not assist in the interpretation of dimensions.

Finally, one important aspect in this method is low dimensional displays of predictive relationships. In this case it is not possible to use the confidence circles [2] by means of the Catanova test statistic. For this reason we constructed 95% confidence regions [32] by means of bootstrap method [14], repeatedly drawing random bootstrap samples from the original data set with replacement. We analyzed each of the bootstrap samples by FMNSCA to obtain parameter estimates. We then computed means and variances of the estimates, from which biases and standard errors have been achieved. Starting of the standard errors we have plotted the confidence region where we have appreciated if the socio-demographic categories can be considered significant. If a confidence region includes the origin of axes then we can declare that the category is not statistically significant. Fig. 2 shows the 95% confidence regions, where to facilitate the reading of the graph, the edges of the not significant regions were drawn using a "dotted" line and a "solid" for the significant.

5. Conclusion

In this paper, we proposed a procedure to analyze two data sets that provided the expenditures for several services for each individual and the identifier or socio-demographic variables for the same individuals, respectively. Both data sets have been provided within the Project "Covi - the Vesuvius coast" (funded by the Campania region in 2015). The main task was to analyze the service expenditure of neapolitan families under two type of constraints: the global relative expenses for a given service and the global relative expenses for a given socio-demographic category. The amount of these expenses were not changed when analyzed statistically. Given the quantitative nature of the data and the asymmetrical relationships among the family expenses with respect to the socio-demographic characteristics, it was possible to analyze these two sets of variables using several methodologies (e.g. Refs. [9,11,25,33,34,36,45]) or incorporate the latter set in the former as external information (e.g. Refs. [3,4,19,24,30,38–41,44,46]). However, nothing guaranteed that the core matrices of all these approaches ensure compliance with the required constraint: the expenses had to remain unchanged in the application of the methodologies. We then used our approach for its associated benefits. First FMNSCA advantage is that the core matrix cross-tabulates the service expenditure categories with the socio-demographic categories keeping the same original column margins of the original matrices. Then it ensures the required constraints. Moreover, our approach allows a graphical representation the

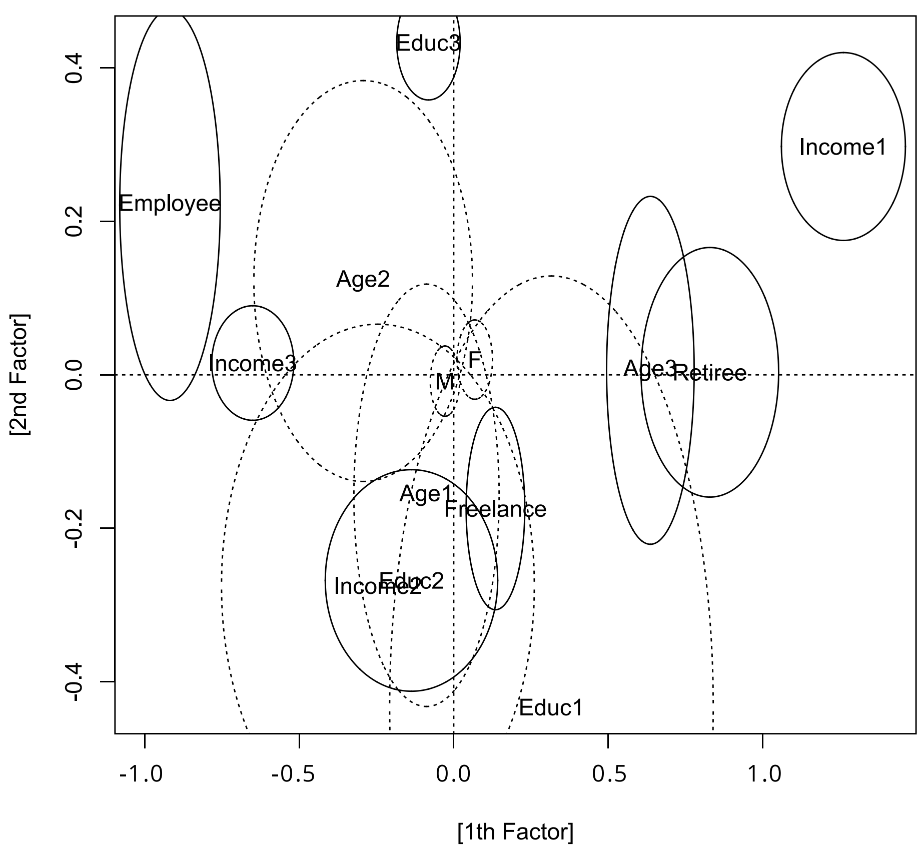

Fig. 2. Confidence regions plot. Source: our elaboration.

relationships between the service expenditure categories with the socio-demographic categories. Bootstrap based confidence regions are also drawn to verify the significant relationships.

Additional benefit of our approach is to obtain a table of regression coefficients which are easily interpreted with interesting properties: the matrix of these regression coefficients results to be rows and columns centred using the weights of original data (this condition is usually not meet in the multivariate regression model) and allows to equal the sum of the observed values to the predicted ones. Moreover, bootstrap based confidence regions can be also used to verify the significance of the coefficient regressions. Finally, according to Ref. [26] it is possible to compute of the correlations between the coordinates of the service expenditure categories and those of a new profile of household.

Using FMNSCA, it was then possible to study the family expenditures with respect to the subspace spanned by the socio-demographic characteristics leaving the expenses for a given service and for a given socio-demographic category as constants. The results seem to underline how variables "Sex" and "Age" have no effect on the distribution of the service expenditures, while they ("Culture" and "Tourism" mainly) seem to be influenced by several categories of variables Income, Education and Profession. More precise indications come also from the reading of the weighted regression coefficients that highlight the different positive or negative impacts (in terms of spending propensity), in predicting new service expenditures for a new given category membership of a family. For example, category "*Employee*" results to be important for "Culture", while families with an high education or with an high income or older than 45 years tends all to allocate more money mainly for "Tourism".

We point out that FMSNSCA may also be useful in other studies involving homogeneous transferable additive variables where it is necessary to keep the sums constant. For example, it could be interesting to study the balance sheet items of several companies with respect to their different characteristics that can be considered influential on their performance.

Moreover, FMSNSCA treats all the explicative socio-demographic variables as nominals. Future work could concern the possibility to take into account the ordinal nature of several categorical variables directly in the FMSNCA criterion. This extension could be based on two lines of research. The first extension could be developed using the Emerson's orthogonal polynomial [15] and the latter based on the cumulative frequencies of the ordinal variable [29,37]. Both extensions are under investigation.

Credit author statement

The authors contributed equally to the development of this research work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.seps.2020.100833.

References

[1] Amenta P. Generalized constrained Co-inertia analysis. Adv Data Anal Classif 2008; 2:81–105.
[2] Beh EJ, D'Ambra L. Non-symmetrical correspondence analysis with concatenation and linear constraints. Aust N Z J Stat 2010;52(1):27–44.
[3] Böckenholt U, Böckenholt I. Canonical analysis of contingency tables with linear constraints. Psychometrika 1990;5**5**:633–9.
[4] Böckenholt U, Takane Y. Linear constraints in correspondence analysis. In: Greenacre M, Blasius J, editors. Correspondence analysis in the social sciences: recent developments and applications. New York: Academic Press; 1994. p. 70–111.
[5] Cailliez F, Pagès JP. Introduction à l'analyse des donnèes. Paris: SMASH; 1976.
[6] Carr-Hill RA, Rice N, Smith PC. The determinants of expenditure on children's personal social services. Br J Soc Work 1999;29(5):679–706.
[7] Casez P, Chessel D, Doledec S. L'analyse des correspondances internes d'un tableau partitionné: son usage en hydrobiologie. Rev Stat Appl 1988;36:39–54.
[8] Cattell RB. The scree test for the number of factors. Multivariate Behav Res 1966;1: 245–76.
[9] D'Ambra L, Lauro NC. Analisi in componenti principali in rapporto ad un sottospazio di riferimento. Rivista di Statistica Applicata 1982;**15**.
[10] D'ambra L, Lauro NC. Non-symmetrical correspondence analysis for three-way tables. In: Coppi R, Bolasco S, editors. Multiway data analysis. Amsterdam: North Holland; 1989. p. 301–15.
[11] Davis PT, Tso MKS. Procedures for reduced rank regression. Appl Stat 1982;31: 244–55.
[12] De Jong S. SIMPLS: an alternative approach to partial least squares regression. Chemometr Intell Lab Syst 1993;18:251–63.
[13] Dolédec S, Chessel D. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw Biol 1994;31:277–94.
[14] Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, Florida: CRC Press; 1993.
[15] Emerson PL. Numerical construction of orthogonal polynomials from a general recurrence formula. Biometrics 1968;24:696–701.
[16] Escofier B. Analyse de la différence entre deux mesures dèfinies sur le produit de duex mêmes ensembles. Cah Anal Données 1983;8(3):325–9.
[17] Escoufier Y. The duality diagram: a means of better practical applications. In: Legendre P, Legendre L, editors. Development in numerical ecology. Berlin: NATO advanced Institute, Springer Verlag; 1987.
[18] Escoufier Y. Operator related to a data matrix: a survey. In: Rizzi A, Vichi M, editors. Compstat 2006 - proceedings in computational statistics. Physica-Verlag HD; 2006. p. 285–97.
[19] Gilula Z, Haberman SJ. The analysis of multivariate contingency tables by restricted canonical and restricted association models. J Am Stat Assoc 1988;83: 760–71.
[20] Goodman LA, Kruskal WH. Measures of association for cross classification. Berlin: Springer Verlag; 1954.
[21] Greenacre MJ. Theory and applications of correspondence analysis. London: Academic Press; 1984.
[22] Hong GS, Morrison AM, Cai LA. Household expenditure patterns for tourism, products and services. J Trav Tourism Market 1996;4(4):15–40.
[23] Hotelling H. Relations between two sets of variates. Biometrika 1936;28:321–77.
[24] Hwang H, Takane Y. Generalized constrained multiple correspondence Analysis. Psychometrika 2002;67:215–28.
[25] Izenman AJ. Reduced-rank regression for the multivariate linear model. J Multivariate Anal 1975;5(2):248–64.
[26] Lebart L, Morineau A, Piron M. Statistique exploratoire multidimensionnelle. Paris: Dunod; 1995.
[27] Mandel J. Use the singular value decomposition in regression analysis. Am Statistician 1982;**36** 1:15–24.
[28] Martens H, Naes T. Multivariate calibration. Chicester, UK: Wiley; 1992.
[29] Nair VN. Chi-squared type tests for ordered alternatives in contingency tables. J Am Stat Assoc 1987;82:283–91.
[30] Nishisato S. Analysis of categorical data: dual scaling and its applications. Toronto: University of Toronto Press; 1980.
[31] Okpala CO, Okpala AO, Smith FE. Parental involvement, instructional expenditures, family socioeconomic attributes, and student achievement. J Educ Res 2001;95(2):110–5.
[32] Ramsay JO. Confidence regions for multidimensional scaling analysis. Psychometrika 1978;43:145–60.
[33] Rao CR. The use and interpretation of principal component analysis in applied research. Sankhya 1964;25:329–58.
[34] Robert P, Escoufier Y. A unifying tool for linear multivariate statistical methods: the RV–coefficient. Appl Stat 1976;25(3):257–65.
[35] Sarti S, Terraneo M, Bordogna MT. Poverty and private health expenditures in Italian households during the recent crisis. Health Pol 2017;121(3):307–14.
[36] Sabatier R, Lebreton JD, Chessel D. Principal component analysis with instrumental variables as a tool for modelling composition data. In: Coppi R, Bolasco S, editors. Multiway data analysis. Amsterdam: Elsevier; 1989. p. 341–52.
[37] Taguchi G. A new statistical analysis for clinical data, the accumulating analysis, in contrast with the chi-square test. Saishin Igaku 1974;29:806–13.
[38] Takane Y, Jung S. Regularized non symmetric correspondence Analysis. Comput Stat Data Anal 2009;53:3159–70.
[39] Takane Y, Hunter MA. Constrained principal component analysis: a comprehensive theory. Appl Algebra Eng Commun Comput 2001;12:391–419.
[40] Takane Y, Shibayama T. Principal component analysis with external information on both subjects and variables. Psychometrika 1991;56:97–120.
[41] Takane Y, Yanai H, Mayekawa S. Relationships among several methods of linearly constrained correspondence analysis. Psychometrika 1991;56:667–84.
[42] Ter Braak CJF. Canonical Correspondence Analysis: a new eigenvector multivariate direct gradient analysis. Ecology 1986;67(5):1167–79.
[43] Ter Braak CJF, Schaffers AP. Co-Correspondence Analysis: a new ordination method to relate two community compositions. Ecology 2004;85(3):834–46.
[44] Van de Velden M, Groenen PJF, Poblome J. Seriation by constrained correspondence analysis: a simulation study. Comput Stat Data Anal 2009;53(8): 3129–38.
[45] Van den Wollenberg AL. Redundancy analysis: an alternative for canonical correlation analysis. Psychometrika 1977;42:207–19.
[46] Van der Heijden PGM, De Leeuw J. Correspondence analysis used complementary to loglinear analysis. Psycometrika 1985;50:429–47.
[47] Van Loan CF. Generalizing the singular value decomposition. SIAM J Numer Anal 1976;13(1):76–83.

Luigi D'Ambra is Full Professor in Statistics, Department of Economics, Management Institutions (University of Naples- Italy) Universit degli Studi di Napoli Federico II, Naples (Italy). Prof. Luigi D'Ambra has carried out an intense didactic and scientific activity in the University of Naples. He has followed numerous theses of statistics with focusing to the customer satisfaction in the context of Public services (transport, education, health). Its activity of research is centralized on the multivariate analysis, three-way data matrices, multidimensional scaling and the not symmetrical correspondence analysis proposing the solutions characterized from originality elements. In the context of the PhD of Computational Statistics in many Italian Universities.

Pietro Amenta is Associate Professor of Statistics at the Department of Law, Economics, Management and Quantitative Methods of the University of Sannio (Piazza Arechi II, 82,100, Benevento, IT; e-mail: amenta@unisannio.it). His main research interests concern: Multidimensional Data Analysis (two-three ways array and Multiblock data analysis of categorical/quantitative groups of variables showing symmetrical or non-symmetrical relationships; Multidimensional data analysis of complex structure data with linear constraints; cluster analysis); - Multidimensional statistical methods for soil study and pedaphology using vis-NIR spectroscopy; - Multicriterial decision making analysis. He is Editor in Chief of the International Journal of Multivariate Data Analysis (Inderscience), Member of the Advisory Board of the Journal of Applied Quantitative Methods (JAQM) and Guest Editor of the special issue on "Statistical methods & models for the evaluation systems for the public sector" of the Journal Socio-Economic Planning Sciences (Elsevier). He is also full member of the scientific associations: International Association for Statistical Computing (IASC), European Network for Business and Industrial Statistics (ENBIS), Italian Statistical Society (SIS), Statistics for the Evaluation and Quality of Services - SIS Group (SVQS), Classification and Data Analysis - SIS Group (CLADAG).

Antonello D'Ambra is Associate Professor in Statistic at the University of Campania Luigi Vanvitelli. Author of many publications concerning Multivariate Analysis, Poisson Factor Models, Log Ratio Analysis, Odds ratio Analysis, Analysis of Variance for categorical data, Decomposition in main effects and interaction term in multiple contingency tables and Correspondence Analysis in presence the ordinal variables by means orthogonal polynomials and/or cumulative analysis.

Jules Joseph Sunda de Tibeiro is full Professor in Statistic and Mathematics at the University of Moncton, Department of Sciences, Campus of Shippagan in New-Bruswick (UMCS). His research and publications focus on the following fields of specialization: Functional Model of Continuous Correspondence, factorial Regression, Qualitative Discriminant Factor Analysis, Supervised Classification, Optimal Coding, Semi-linear Barycentric Coding and Personal Equation, Analysis and Modelling of probabilistic Data by Decomposition of Copula Mixture and Bayesian Analysis