Journal of Data Analysis and Information Processing, 2016, 4, 81-89
Published Online May 2016 in SciRes. http://www.scirp.org/journal/jdaip
http://dx.doi.org/10.4236/jdaip.2016.42007

Subset Multiple Correspondence Analysis as a Tool for Visualizing Affiliation Networks

Achilles Dramalidis, Angelos Markos

School of Education, Democritus University of Thrace, Alexandroupolis, Greece
Email: amarkos@eled.duth.gr

Received 4 March 2016; accepted 16 May 2016; published 19 May 2016

Abstract

In this paper we investigate the potential of Subset Multiple Correspondence Analysis (s-MCA), a variant of MCA, to visually explore two-mode networks. We discuss how s-MCA can be useful to focus the analysis on interesting subsets of events in an affiliation network while preserving the properties of the analysis of the complete network. This unique characteristic of the method is also particularly relevant to address the problem of missing data, where it can be used to partial out their influence and reveal the more substantive relational patterns. Similar to ordinary MCA, s-MCA can also alleviate the problem of overcrowded visualizations and can effectively identify associations between observed relational patterns and exogenous variables. All of these properties are illustrated on a student course-taking affiliation network.

Keywords

Affiliation Networks, Two-Mode Networks, Subset Correspondence Analysis

1. Introduction

Social Network Analysis (SNA; see [1], for a comprehensive overview) provides a valuable approach to the analysis of relational data. As such, SNA draws on various concepts from graph theory and structural theory and has a wide range of applications in fields like sociology and education. In this paper we deal with the analysis of affiliation networks, a special case of two-mode networks, which contrasts with the more widely used one-mode networks. Affiliation networks consist of two types of nodes: a set of actors and a set of events. The actors are linked to events and each link denotes the participation of that actor in that particular event. Given an $n \times m$ matrix $\boldsymbol{A}$ with actors in rows and events in columns, α_{ij} equals to 1 if a row actor is affiliated with a column event, and 0 otherwise. Some examples of affiliation networks include the actor-movie networks, where two actors are linked if they co-appear in the same movie [2], the co-authorship networks, where two scientists are linked if they co-author a paper [3], occurrence networks, where two words are linked if they co-occur in the same

How to cite this paper: Dramalidis, A. and Markos, A. (2016) Subset Multiple Correspondence Analysis as a Tool for Visualizing Affiliation Networks. *Journal of Data Analysis and Information Processing*, **4**, 81-89.
http://dx.doi.org/10.4236/jdaip.2016.42007

sentence [4], the interlocking networks of board of directors, where two directors are linked together if they are members of the same board [5] and student course-taking networks, where students are registered for courses [6].

There are primarily two main approaches to the analysis of affiliation networks [7]. One approach is to convert the data into two (or more) one-mode networks, either as an actor-by-actor or an event-by-event network. In particular, these one-mode networks can be constructed by either multiplying the affiliation matrix $\boldsymbol{A}$ by its transpose (*i.e*, $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$) or multiplying the transpose of the affiliation matrix with itself (*i.e*., $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$). The resulting $n \times n$ or $m \times m$ matrices are symmetric and record the co-membership relation for each actor or the overlapping relations for each event, respectively. This process can be described as the projection of the bipartite graph onto the unipartite space of actors or events only. The second approach considers the joint analysis of the two modes. This approach provides a richer understanding than the analysis of one-mode data, since much information in the original bipartite structure is lost as a result of the projection process [8].

Using either of the two approaches, a fundamental issue is the direct visualization of the affiliation structure (e.g., [8]). A common strategy is to apply factorial methods, such as multidimensional scaling [1] and correspondence analysis [9]-[13]. In particular, Correspondence Analysis (CA) is not constrained by strict model assumptions or distributional requirements and leads to a low-dimensional representation of actors and/or events in order to evaluate relational patterns [10]. In this context, the affiliation matrix can be treated as a two-way contingency table on which CA can be applied. The pros and cons of this approach have been thoroughly discussed [7] [9] [11]. The CA representation has certain advantages, such as it provides an easy interpretation of the similarities among actors/events, and makes it possible to add covariates (exogenous variables) to the analysis [11]. Another option is to treat the affiliation matrix as a case-by-variable matrix, which leads to Multiple CA, the extension of CA to the case of more than two categorical variables [14]. This option has been recently described in [11] as an effective approach to analyze and graphically represent affiliation networks. A recent simulation study has indicated that MCA tends to be more stable than CA with respect to some network characteristics such as density and the presence of structural equivalent blocks [11].

Although it is common to apply CA/MCA to the complete data set, there are cases when the analysis of a subset of the original data may be more appropriate or desirable. For instance, when analyzing a large number of events (columns), the interpretation may be obscured by the large number of points or vectors in the map, so that interpretation and conclusions are limited to broad generalities. The basic problem is that CA/MCA visualizes many different types of relationships simultaneously so that the factorial maps may not be easily conductive to visualizing those relationships of particular interest to the researcher [15]-[17]. In other data analytic scenarios, events might subdivide naturally into groups and it would be interesting to analyze each group separately, taking also into consideration the broad relational patterns that exist between groups. Therefore, it is often desirable to restrict attention to a subset of the affiliation matrix, in order to facilitate the interpretation as well as to make the conclusions substantively richer and more interesting.

A further analysis of interest would be in the case when some edges (participation of actors in specific events) are missing from the dataset, e.g. due to survey non-response. It is generally accepted that the analysis of social networks is hampered by missing values, because the visualization of the network structure is especially sensitive to missing data. Recent studies have shown the negative effects of missing actors and edges on the structural properties of social networks [18]. The most popular strategy to overcome the problems created by non-responses is missing value imputation. An alternative strategy, adopted in this paper, is to focus on the subset of observed edges alone, *i.e*. to partial out the influence of non-responses, or even to analyze just the non-responses by themselves, in order to understand how these are related to external variables. This could narrow down the interpretation to the more substantive relational patterns.

All of the aforementioned issues can be addressed through the use of Subset CA (s-CA), a simple variant of CA. The idea in s-CA, as the name suggests, is to visualize a subset of the rows or a subset of the columns (or both) in subspaces of the same full space as the original complete set [14]. s-MCA maintains the geometry of the complete MCA, with the difference that the elements of the subset are not expressed with respect to their own totals, but maintain their profile values with respect to the totals of the complete data set. Thus, in the context of two-mode networks, s-CA can be used to investigate relational patterns of subgroups of actors or events to be directly analyzed and visualized while preserving all the information contained in the original affiliation matrix. From a mathematical perspective, the ordinary CA algorithm is modified so that the marginal frequencies of the original matrix are retained in the analysis of the subset. A generalization of the algorithm to the multiple case, *i.e*. Subset MCA (s-MCA), is straightforward [14]. The aim of this paper is to describe s-MCA and highlight its properties as a method for visualizing subsets of affiliation networks.

The paper is organized as follows. In Section 2 we introduce the mathematical background of s-MCA in the context of affiliation networks. Section 3 illustrates three important properties of the method: 1) the handling of missing edges in a student-course taking affiliation network, 2) the visualization of relational patterns in interesting subsets of the network and 3) the incorporation of external information in a subset analysis to facilitate interpretation of observed relational patterns. Section 4 offers some concluding remarks and future directions.

2. Subset Multiple Correspondence Analysis for Affiliation Networks

2.1. Definitions

Using a similar notation to that of [11], let $\mathcal{G}(V_1, V_2, \mathcal{R})$ be an affiliation network consisting of two sets of relationally connected units, actors and events, where $V_1 = \{a_1, a_2, \cdots, a_n\}$ represents the set of n actors whereas $V_2 = \{e_1, e_2, \cdots, e_n\}$ represents the set of m events. Note that $V_1 \bigcap V_2 = \varnothing$. $\mathcal{R}$ is a set of edges or arcs, $\mathcal{R} \subseteq V_1 \times V_2$, where the edge $r_{ij} = (a_i, e_j), r_{ij} \in \mathcal{R}$, is an ordered couple, and indicates whether an actor a_i attends an event e_j. The set $V_1 \times V_2$ can be fully represented by the binary affiliation matrix $\boldsymbol{A} = (a_{ij}), i = 1, \cdots, n, j = 1, \cdots, m$, with $a_{ij} = 1$ if $(a_i, e_j) \in \mathcal{R}$ and 0 otherwise. The row and column marginals of $\boldsymbol{A}$ coincide with the degree d_i of the i^{th} actor and the size s_j of the j^{th} event, respectively. The $n \times m$ affiliation matrix $\boldsymbol{A}$ can be transformed into an indicator matrix $\boldsymbol{Z}$ with n rows and 2 m columns, where each event e_j is described by a dummy variable with categories e_j^+ (participation in the event) and e_j^- (non-participation in the event). Thus, the indicator matrix $\boldsymbol{Z}$ turns out to be a doubled matrix. In case of missing information on event participation, a third column, $e_j^?$, could be added to the matrix $\boldsymbol{Z}$ for each event. The matrix will have$n$ rows and 3 m columns (**Figure 1**). Then the ordinary CA algorithm can be applied to this "expanded" matrix $\boldsymbol{Z}$, and this is known in the literature as the Multiple CA (MCA) on $\boldsymbol{A}$ [17]. The row totals of $\boldsymbol{Z}$ are constant and equal to the number of events m, while the column totals are equal to the event size s_j when associated to e_j^+, $n - s_j - k$ when associated to e_j^-, or k when associated to $e_j^?$, and k indicates the total number of missing edges.

Now suppose that the interest is to analyze and visualize a subset of events only (a column subset of $\boldsymbol{Z}$). For instance, in a student course-taking affiliation network, where actors are students and events are university courses students are registered for, the interest may be to a subset of courses with similar content (e.g. science) or a subset of courses of a specific semester or year of study. Another interesting case may be to focus attention on existing edges only, that is ignoring students who provided with no information about a course or groups of courses (rows of $\boldsymbol{Z}$ with 1s in columns $e_j^?$). The most obvious approach in such cases, would be simply to apply MCA to the corresponding submatrix of $\boldsymbol{Z}$ or, in other words, to delete the unnecessary columns or rows. However, one or both of the margins of the submatrix would differ from those of the original data matrix and the corresponding geometric structures will consequently differ (see [16] for a thorough discussion). A remedy to this, is a simple variant of MCA, Subset MCA (s-MCA), which maintains the geometry of the original MCA and allows relational patterns of subsets of events to be directly analyzed and visualized in subspaces of the same full space as the original complete set.

2.2. The s-MCA Algorithm

Let $\boldsymbol{Y} = \frac{1}{m}\boldsymbol{Z} - \frac{1}{nm}\boldsymbol{1}\boldsymbol{1}^{\mathrm{T}}\boldsymbol{Z}$, where $\boldsymbol{Z}$ is divided by the number of events m and centered with respect to the averages of its columns ($\boldsymbol{1}$ is $an \times m$ matrix of ones). The averages of the columns are the column totals of $\boldsymbol{Z}$ divided by $\boldsymbol{Z}$'s grand total nm, where n is the number of actors and, hence, are exactly the proportions of actors participating (or not participating) to the corresponding events, divided by the number of events, m. Let $D_a(n \times n)$

	e_1	e_2	e_3	$\cdots$	e_m
a_1	1	0	?	$\cdots$	0
a_2	?	1	0	$\cdots$	0
a_3	1	?	1	$\cdots$	1
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
a_n	?	0	0	$\cdots$	0

→

	e_1			e_2			e_3			$\cdots$	e_m		
	+	−	?	+	−	?	+	−	?	$\cdots$	+	−	?
a_1	1	0	0	0	1	0	0	0	1	$\cdots$	0	1	0
a_2	0	0	1	1	0	0	0	1	0	$\cdots$	0	1	0
a_3	1	0	0	0	0	1	1	0	0	$\cdots$	1	0	0
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$	$\cdots$	$\cdots$	$\cdots$
a_n	0	0	1	0	1	0	0	1	0	$\cdots$	0	1	0

Figure 1. An affiliation matrix $\boldsymbol{A}$ with missing participation information and the corresponding indicator matrix $\boldsymbol{Z}$.

and $D_e\left(m\times m\right)$ be diagonal matrices of row averages and column averages (or masses in CA terminology), respectively, which give differentiated importance to the actors and the events. Suppose that $\boldsymbol{H}$ is a selected subset of events (columns of $\boldsymbol{Y}$) and that the corresponding subset of column weights is denoted by $\boldsymbol{h}$. Then Subset MCA (s-MCA) is defined as the principal component analysis of $\boldsymbol{H}$ with row masses $\boldsymbol{a}$ in $\boldsymbol{D}_a$ and column weights $\boldsymbol{D}_h$. Therefore, the original relative frequencies of the categories are maintained and are not re- expressed relative to totals within the subset, as would normally be done in a regular MCA of the subset. The solution can be obtained using the generalized singular value decomposition (GSVD) of the matrix $\left(\boldsymbol{I}-\boldsymbol{1}\boldsymbol{a}^{\mathrm{T}}\right)\boldsymbol{H}$ which corresponds to the ordinary SVD of $\boldsymbol{D}_a^{-1/2}\left(\boldsymbol{I}-\boldsymbol{1}\boldsymbol{a}^{\mathrm{T}}\right)\boldsymbol{H}\boldsymbol{D}_h^{-1/2}$. The steps can be summarized as follows [16]:

$$\text{Step 1: } \boldsymbol{S}=\boldsymbol{D}_a^{-1/2}\left(\boldsymbol{I}-\boldsymbol{1}\boldsymbol{r}^{\mathrm{T}}\right)\boldsymbol{H}\boldsymbol{D}_h^{-1/2} \tag{1}$$

$$\text{Step 2: Obtain the SVD of } \boldsymbol{S}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{2}$$

$$\text{Step 3: Standard coordinates of actors: } \boldsymbol{\Gamma}=\boldsymbol{D}_a^{-1/2}\boldsymbol{U}=\sqrt{n}\boldsymbol{U} \tag{3}$$

$$\text{Principal coordinates of actors: } \boldsymbol{F}=\boldsymbol{\Gamma}\boldsymbol{\Delta}=\sqrt{n}\boldsymbol{U}\boldsymbol{\Sigma} \tag{4}$$

$$\text{Step 4: Standard coordinates of events: } \boldsymbol{\Delta}=\boldsymbol{D}_h^{-1/2}\boldsymbol{V} \tag{5}$$

$$\text{Principal coordinates of events: } \boldsymbol{G}=\boldsymbol{\Delta}\boldsymbol{\Sigma} \tag{6}$$

Step 2 is the SVD, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the left and right singular vectors, respectively, and $\boldsymbol{\Sigma}$ is a diagonal matrix of singular values in decreasing order of magnitude.

The main output of s-MCA is the joint representation of actors and events in a two-dimensional map, with coordinates in the first two columns of $\boldsymbol{F}$ and Δ, respectively. In this biplot, actors are usually represented as points in the event space spanned by the axes with principal coordinates in $\boldsymbol{F}$ and events are usually represented as vectors in the actor space spanned by the axes with standard coordinates in Δ [11]. There are various ways to scale the biplot, and its interpretation depends on this scaling (see [19] for available options). The first two dimensions or factorial axes account for most of the variance (or inertia) in the data, and in most cases they offer the most accurate and interpretable representation of the main oppositions and/or associations. However, the analyst could seek to identify patterns which lie into subsequent dimensions, e.g. the factorial maps formed by dimensions 1 and 3, 2 and 3 or 3 and 4.

The distance between two actors in a factorial map best approximates the chi-square distance among the corresponding actor profiles in the original space, and represents the actors' relative positions in the network [11]. This distance is equal to zero when the actors participate in the same events. The distance is greater than zero when the participation patterns differ. In addition, each event e_j in the actor space is represented by two opposite vectors, corresponding to the two poles e_j^+ and e_j^-, lying on the same direction and passing through the origin. This is the so-called *doubling approach* and its advantages in the case of affiliation networks are discussed in [11]. According to this approach, the cosine of the angle between two event vectors is the correlation between participation patterns in the events. A small angle between two event segments with positive poles indicates events with similar participation patterns, whereas if the poles are opposite, the events will have opposite participation patterns. Moreover, the length of each segment joining the two poles is proportional to the inverse of the product between the participation and the non-participation rates. If one distance (e_j^+ or e_j^-) is much larger than the other, the event is either rare or common. Also note that in the analysis of affiliation matrices with a large number of rows (actors), assessing the role of single actors on s-MCA maps is difficult. In that case, s-MCA can be applied to the one-mode matrix of events, $\boldsymbol{Z}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$, which in CA terminology is referred to as the Burt matrix.

Another important aspect of s-MCA is that it allows the projection of supplementary rows or columns as points in the factorial map in order to investigate the association between existing relational patterns and actor or event covariates. Supplementary rows or columns do not participate in the creation of the factorial axes as they have zero masses and their relative positions can be evaluated to facilitate the interpretation. The coordinates of supplementary columns can be calculated as the weighted average of the actor standard coordinates with weights equal to the event profiles of the original affiliation matrix, using the so-called transition formula:

$$\boldsymbol{G}_{\mathrm{sup}}=\boldsymbol{D}_e^{-1}\frac{\boldsymbol{Z}}{nm}\boldsymbol{F}\boldsymbol{\Sigma}^{-1}=\boldsymbol{D}_e^{-1}\frac{\boldsymbol{Z}}{nm}\boldsymbol{\Gamma} \tag{7}$$

Finally, the quality ofrepresentation of eachindividualpoint (actor) orvector (event) on a factorialmap, could-beassessedviaaset of appropriateindices, suchascontribution (COR), correlation (CTR) and quality (QLT), which arepart of the standardoutput ofthesoftwarepackagesimplementing CA/MCA.

3. Main Properties of s-MCA for Affiliation Networks

In order to demonstrate the important aspects of s-MCA in the context of affiliation networks, we consider a binary network of student enrollment in elective courses as part of their undergraduate studies in a primary education university department, located in a city of Northern Greece. The training of elementary pre-service teachers has been established as 4-year studies with eight semesters. In each semester, the department offers a wide range of elective courses in science, language, psychology, computer science, mathematics, statistics, social studies, music, art, physical education, and a miscellaneous group of courses. The student course-taking data were collected as part of a larger cross-sectional study aiming to associate students' course enrollment with the reasons behind their choices and a variety of background characteristics. The affiliation network under study consists of 193 students and 67 elective courses offered by the department, in which participation has been recorded over four academic years (2011/12 through 2015/16), along with student-related attributes of gender, educational background in high school, perception towards post-graduate studies and the reasons for taking these specific courses. The sample is composed of 90% female and 10% male students. Approximately 81% of the students reported that they had a theoretical educational background in high school, 10% had a technological background, 7% had a scientific background and 2% did not provide any data.

Part of the 193 × 67 affiliation matrix used in our analysis is shown in **Table 1** for the eleven elective courses offered in the 5th semester. During all four years of their undergraduate studies, students had to take ten elective courses in total (one course in each one of the 1st, 2nd, 3rd, 4th and 6th semester, two courses in the 5th and three courses in the 8th semester). Therefore, the total degree of a student with no missing data in the network equals to ten (last column in **Table 1**). The marginal row of the table corresponds to the total number of students who have selected each one of the eleven courses. The courses were taken by 22 to 45 students in total, with the exception of Cosmography (taken by nine students only). Note that student 3did not provide any data for any of the courses of the 5th semester (all values in the corresponding row are missing-?), but suppose he/she has done so for the courses in the other seven semesters; hence his/her total degree is less by two (eight). Therefore, the corresponding indicator matrix $\boldsymbol{Z}$ will be of size 193 × 142, where 134 of its columns correspond to the 67 original courses (two columns per course, e_j^+ and e_j^-) and the remaining 8 columns correspond to the missing course-taking patterns in each semester ($e_j^?$). Note that, for this dataset, values could be missing for a semester as a whole and not for a specific course, as would be the case described earlier for **Figure 1**. s-MCA was applied to the data using the **CA** package in R [20].

3.1. Handling Missing Data

In practice, missing response categories often dominate the CA/MCA factorial map because of high association,

Table 1. Part of the full 193 × 67 affiliation matrix describing the student-course enrollment data in the 5th semester, with actor degree and event size.

Student	Courses (5th semester)											
	Voc	EUEdu	MPhil	EUHist	Bio	CDiv	TheHist	Cosmo	Geo	CProg	GrThe	*Degree*
1	0	0	0	0	0	0	0	1	1	0	0	10
2	0	1	1	0	0	0	0	0	0	0	0	10
3	?	?	?	?	?	?	?	?	?	?	?	8
4	0	1	0	0	1	0	0	0	0	0	0	10
5	0	0	1	0	0	0	0	0	0	0	0	10
...	...	...	...	...	...	...	...	...	...	...	...	...
Event size	44	25	24	45	44	19	22	9	37	44	33	

Voc: Vocabulary: description and pedagogy, EUEdu: Trends in European Education, MPhil: Modern Philosophy, EUHist: Modern European History, Bio: Topics in Biology, CDiv: Cultural Diversity in the Classroom, TheHis: History of Theatre, Cosmo: Cosmography, Geo: Geography Education, CProg: Computer Programming, GrThe: Group Theory, "?" indicates a missing edge.

forcing a contrast between these and the substantive categories. In order to motivate our approach, we first consider the usual MCAmap of the full indicator matrix in **Figure 2(a)**, showing course enrolment and non-enrolment vectors for all 67 courses (indicated by course name followed by "1" for enrolment or "0" for non-enrolment), as well as8 vectors corresponding to non-responses for each semester (indicated by a letter from A to H followed by "mis"). The variance explained by the first axis is 5.2%, and by the second axis is 4.3% (about 10% both axes). Note that this overall percentage of variability in the data that is explained by the first two dimensions appears to be small, but its calculation is pessimistic since it was based on the analysis of the indicator matrix [17].

The map of **Figure 2(a)** is typical of analyses of survey data such as these where missing values have been introduced into the analysis: the non-response categories are highly associated across semesters and are aligned with the first principal axis (vectors pointing to the left). This resulted in an elongation of the scale along this axis that, in turn, resulted in a clumping together of course vectors near the origin and along the second principal axis. As a consequence, non-response categories mask more relevant or substantive relational patterns and it is very difficult to distinguish between the "substantial" course vectors.

For the current data, event non-response rates varied between 4% and 11%. To address the effect of missing values, s-MCA was applied to the 193 × 134 subset of $\boldsymbol{Y}$ (see Section 2.2), thus focusing on the subset of observed edges that lie, however, in the subspace of the same full space as the complete set. The corresponding s-MCA map is shown in **Figure 2(b)**. The elimination of the effects of non-responses resulted in a more simplistic and a less crowded display. Three main groups of courses can now be distinguished in the map: art-related courses of the 1st semester pointing to the top-left of the map, theoretical courses related to the teaching profession pointing to the bottom and bottom-right, and science, psychology and mathematics courses pointing to the top-right. However, the interpretation is still complicated because of over-crowding.

3.2. Visualization of Relational Patterns in Subsets of Events

Another case where s-MCA can be useful is when the purpose is to visualize the relational patterns of a specific subset of events. **Figure 3(a)** depicts in the first two dimensions the enrolment patterns between the eleven courses which were offered in the 5th semester only. Recall that in this semester each student should enroll to two out of the eleven courses available. The map is a result of an s-MCA on the corresponding subset of 22 columns of the indicator matrix $\boldsymbol{Z}$. In this map, one can now easily identify groups of courses that were usually chosen together or others that were rarely chosen together. For instance, along the first (horizontal) axis of course-taking, Vocabulary: description and pedagogy (VOC), Modern Philosophy (MPhil) and Trends in European Education (EUEdu) form a group of courses with similar enrolment patterns, whereas courses in this group are rarely chosen together with Geography (Geo) or Modern European History (EUHist). Geo and EUHist are

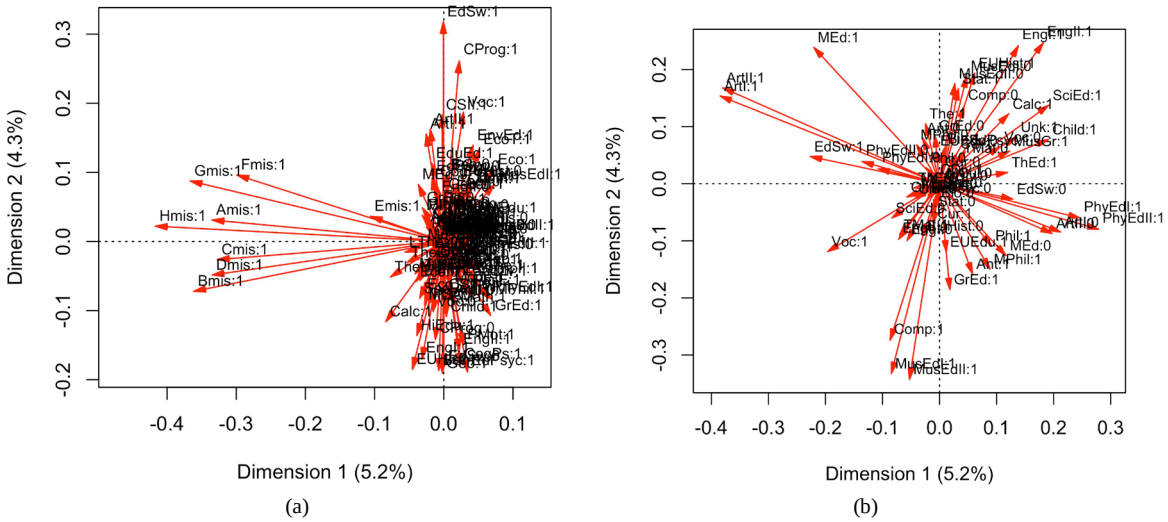

Figure 2. (a) MCA map of the courses (vectors) in the student space for the full 193 × 142indicator matrix (b) s-MCA map of the courses (vectors) in the student space for the 193 × 134 indicator matrix (omitting missing values).

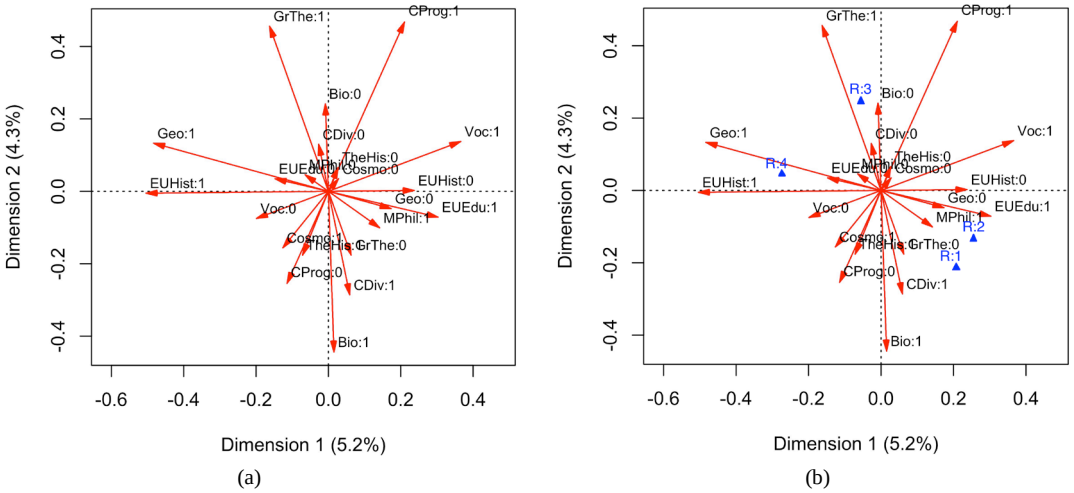

Figure 3. s-MCA map of the 5th semester courses (vectors) in the student space (a) without and (b) with supplementary variables.

usually chosen together, as the corresponding vectors form a small angle and point to the opposite direction from VOC, MPhil and EUEdu along the first axis. A different story takes place along the second (vertical) axis. To the top, Computer Programming (CProg) and Group Theory (GrThe) form a small group with similar enrolment patterns, in contrast with Topics in Biology (Bio), Cosmography (Cosmo), History of Theatre (TheHis) and Cultural Diversity in the Classroom (CDiv), which form another group to the bottom. In addition, the length of the segment joining the two poles is indicative of the popularity of each course. Thus, the most popular courses in this semester are GrThe, Cprog, Geo, Voc, EUHist and Bio, whereas the least popular are Cosmo, TheHis, CDiv, MPhil and EUEdu. A quick look at the frequency of enrolment in **Table 1** confirms this finding.

At this point, one could ask how the map of a direct application of MCA to this subset of courses, ignoring the rest, would be different from that of s-MCA in **Figure 3(a)**. An answer is that in the case of s-MCA, where a part of the affiliation matrix is analyzed, preserving the information contained in the full affiliation matrix, the vectors of this subset of eleven courses are positioned not only according to student enrolment patterns in those specific courses—as it is expected to happen with a direct MCA—but also with regard to information about student enrolment in all other courses.

3.3. Adding Covariates to the s-MCA Representation

An emerging question concerns the reasons which could potentially explain the student enrolment patterns observed in **Figure 3(b)**. Here comes into play another important feature of s-MCA and CA/MCA in general: the option to project exogenous (or supplementary) variables in an existing two-dimensional space so as to facilitate interpretation. **Figure 3(b)** is different in that it also plots the supplementary categories of an external variable (points indicated by a triangle), that corresponds to the following question: "*What is the main reason for taking these specific courses*? *R*1. *Interesting content*, *R*2. *Expected to be easy*, *R*3. *Presence in the course is not mandatory*, *R*4. *Left it to chance*". This variable has been added as an additional column to the indicator matrix $\boldsymbol{Z}$ and the coordinates of the four corresponding points were obtained via Equation (7). The group of courses located to the right part of the map (EUEdu, MPhil, CDiv, Voc) were mostly taken due to their interesting content or the students' expectation of a good grade. The two courses located to the top (CProg, GrThe) did not require the student to be present in the lectures and this seems to have been the main reason they were chosen. Finally, the choice of the two courses located to the left (Geo and EUHist) seems to be the result of random choice by the majority of students that took them. Additional student background variables could be plotted to the factorial map so as to explain the reasons behind course-taking, such as gender, perception towards post-graduate studies and educational background in high school.

4. Concluding Remarks

In this paper we have discussed the use of an extension of MCA, Subset MCA, to visually explore relational

patterns in subsets of two-mode networks. The application of s-MCA for social network analysis can serve a four-fold purpose: 1) to partial out the influence of missing data in an affiliation matrix, 2) to visualize relational patterns that lie in interesting subsets of the matrix in subspaces of the same full space as the original complete set, 3) to alleviate the problem of crowded representations of large affiliation networks and 4) to identify associations between observed relational patterns and exogenous variables (covariates).

The application of s-MCA to an affiliation matrix with missing data showed that it provided a meaningful approach to reveal substantive relational patterns while ignoring the non-substantive ones. In this context, s-MCA can be applied irrespective of the missing data mechanism present, it is computationally simple and it is able to handle large affiliation matrices. We argue that this exploratory method is easier to apply than the existing multiple imputation methods in which many complexities need to be considered.

When visualizing relatively large affiliation matrices, it is almost always true in practice that the interpretation of the maps is degraded by the large number of points and vectors analyzed, all of which load to a greater or lesser extent on every dimension, thereby limiting the interpretation and conclusions to broad generalities. Once the broad picture is seen in the complete analysis, there is value in a subsequent division of the events into numerous smaller, sensibly selected, mutually exclusive and exhaustive subsets. The actors-by-events structure of a two-mode network fits the structure that is assumed in s-MCA, which can provide a summary of the relationships within each subset.

Finally, we would like to highlight that s-MCA belongs to a large family of exploratory techniques that allow the data analyst to observe the patterns of associations in the data and to generate hypotheses that could be tested in a subsequent stage of research. Other methods of the same family that are worth investigating for the analysis of affiliation networks are [21] [22], which combine CA/MCA with *k*-means clustering in a single step. The potential of these methods for community detection and visualization of affiliation matrices can be left as future work.

References

[1] Wasserman, S. and Faust, K. (1994) Social Network Analysis: Methods and Applications (Vol. 8). Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9780511815478

[2] Watts, D. and Strogatz, S. (1998) Collective Dynamics of Small-World Networks. *Nature*, **393**, 440-442. http://dx.doi.org/10.1038/30918

[3] Newman, M. (2001) The Structure of Scientific Collaboration Networks. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 404-409. http://dx.doi.org/10.1073/pnas.98.2.404

[4] Ferrer, I, Cancho, R. and Solé, R.V. (2001) Two Regimes in the Frequency of Words and the Origin of Complex Lexicons: Zipf's Law Revisited. *Journal of Quantitative Linguistics*, **8**, 165-173. http://dx.doi.org/10.1076/jqul.8.3.165.4101

[5] Conyon, M.J. and Muldoon, M.R. (2006) The Small World of Corporate Boards. *Journal of Business Finance & Accounting*, **33**, 1321-1343. http://dx.doi.org/10.1111/j.1468-5957.2006.00634.x

[6] Ferrare, J.J. (2013) The Duality of Courses and Students: A Field-Theoretic Analysis of Secondary School Course-Taking. *Sociology of Education*, **86**, 139-157. http://dx.doi.org/10.1177/0038040712456557

[7] Borgatti, S.P. and Halgin, D.S. (2011) On Network Theory. *Organization Science*, **22**, 1168-1181. http://dx.doi.org/10.1287/orsc.1100.0641

[8] Borgatti, S.P. and Everett, M.G. (1997) Network Analysis of 2-Mode Data. *Social Networks*, **19**, 243-269. http://dx.doi.org/10.1016/S0378-8733(96)00301-2

[9] Roberts, J.M. (2000) Correspondence Analysis of Two-Mode Network Data. *Social Networks*, **22**, 65-72. http://dx.doi.org/10.1016/S0378-8733(00)00017-4

[10] Faust, K. (2005) Using Correspondence Analysis for Joint Displays of Affiliation Networks. In: Carrington, P.J., Scott, J. and Wasserman, S., Eds., *Models and Methods in Social Network Analysis*, Cambridge University Press, New York, 117-147. http://dx.doi.org/10.1017/cbo9780511811395.007

[11] D'Esposito, M.R., De Stefano, D. and Ragozini, G. (2014) On the Use of Multiple Correspondence Analysis to Visually Explore Affiliation Networks. *Social Networks*, **38**, 28-40. http://dx.doi.org/10.1016/j.socnet.2014.01.003

[12] Zhu, M., Kuskova, V., Wasserman, S. and Contractor, N. (2016) Correspondence Analysis of Multirelational Multilevel Networks. In: Lazega, E. and Snijders, T., Eds., *Multilevel Network Analysis for the Social Sciences*, Springer International Publishing, Cham, 145-172. http://dx.doi.org/10.1007/978-3-319-24520-1_7

[13] De Nooy, W. (2003) Fields and Networks: Correspondence Analysis and Social Network Analysis in the Framework of Field Theory. *Poetics*, **31**, 305-327. http://dx.doi.org/10.1016/S0304-422X(03)00035-4

[14] Blasius, J. and Greenacre, M. (2006) Correspondence Analysis and Related Methods in Practice. In: Greenacre, M. and Blasius, J., Eds., *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London, 4-40. http://dx.doi.org/10.1201/9781420011319.ch1

[15] Greenacre, M. and Pardo, R. (2006) Subset Correspondence Analysis Visualizing Relationships Among a Selected Set of Response Categories from a Questionnaire Survey. *Sociological Methods & Research*, **35**, 193-218. http://dx.doi.org/10.1177/0049124106290316

[16] Greenacre, M. and Pardo, R. (2006) Multiple Correspondence Analysis of Subsets of Response Categories. In: Greenacre, M. and Blasius, J., Eds., *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London, 197-217. http://dx.doi.org/10.1201/9781420011319.ch8

[17] Greenacre, M. (2006) From Simple to Multiple Correspondence Analysis. In: Greenacre, M. and Blasius, J., Eds., *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London, 41-76. http://dx.doi.org/10.1201/9781420011319.ch2

[18] Kossinets, G. (2006) Effects of Missing Data in Social Networks. *Social Networks*, **28**, 247-268. http://dx.doi.org/10.1016/j.socnet.2005.07.002

[19] Greenacre, M. (2013) Contribution Biplots. *Journal of Computational and Graphical Statistics*, **22**, 107-122. http://dx.doi.org/10.1080/10618600.2012.702494

[20] Nenadić, O. and Greenacre, M. (2007) Correspondence Analysis in R, with Two- and Three-dimensional Graphics: The ca Package. *Journal of Statistical Software*, **20**, 1-13.

[21] D'Enza, A.I. and Palumbo, F. (2013) Iterative factor clustering of binary data. *Computational Statistics*, **28**, 789-807. http://dx.doi.org/10.1007/s00180-012-0329-x

[22] Hwang, H., Dillon, W.R. and Takane, Y. (2006) An Extension of Multiple Correspondence Analysis for Identifying Heterogeneous Subgroups of Respondents. *Psychometrika*, **71**, 161-171.