MULTIPLE CORRESPONDENCE ANALYSIS AS EXPLORATORY TECHNIQUE FOR EEG SIGNALS

J.C.G.D. Costa*, P.J.G. Da_Silva*, R.M.V.R. Almeida*and A.F.C. Infantosi*

* Programa de Engenharia Biomédica, COPPE – Universidade Federal do Rio de Janeiro
email: afci@peb.ufrj.b

Abstract: An exploratory technique named Multiple Correspondence Analysis (MCA) was used to analyze symmetries between EEG-attributes and gender under two protocol conditions: seated / eyes open and upright / eyes open. A total of 31 subjects were included in the analysis (21 male/10 female), and only the *power* of the alpha (8-13 Hz), theta (4-8 Hz) and beta (13-30 Hz) bands of the right parietal derivation (P4) were considered in this study. Two MCA dimensions were retained by the Scree plot, with a total of 91.7% explained inertia (76.2% and 15.5% for the first and second dimension, respectively). Results suggested an association between females and higher levels of power independently of the protocol condition and also between upright position / eyes open and medium levels of power.
Keywords: EEG, Exploratory Data Analysis, Multiple Correspondence Analysis.

Introduction

Many scientific fields use quantitative and qualitative variables together in order to model a phenomenon. For example, in the staging of neurological disorders, attributes of power spectrum derived from EEG signals or other continuous measurements (quantitative variables), can be analyzed together with the patient's symptoms and / or personal features (qualitative variables).

However, techniques which use both types of variables are not common, since the usual approach is carried out separately by applying specific techniques for each of the sets. This option cannot be considered optimal, because unrealistic assumptions about probability distributions are necessary, and because this approach can result in information loss [1]. Nevertheless, since the exploratory analysis of mixed data (with quantitative and qualitative variables) is thus made easier, some information loss can be acceptable.

Multiple Correspondence Analysis (MCA) is the multivariate version of simple Correspondence Analysis (CA) technique [2], in which categorical variables are displayed in an orthogonal space, allowing for an exploratory analysis of the data [3,4]. Here, we propose the use of MCA for the analysis of categorized attributes derived from an EEG periodogram together two dichotomous variables, gender and the experimental protocol conditions (*eyes open / seated* and *eyes open / upright*).

Background

Multiple Correspondence Analysis is a multivariate technique in which, through an appropriate scaling, variables' categories are displayed as points in a multidimensional space (the Category-Points, CP), allowing for the visualization of patterns that emerge from the bulk of data [5]. A disjunctive matrix, the Superindicator Matrix, or its cross-product (the Burt matrix), is used to this end. Correspondence Analysis and Multiple Correspondence analysis are comprehensively presented in many texts such as in [2, 3, 4], thus only a brief introduction is given here.

The Superindicator Matrix ($\boldsymbol{Z}$) is built through a juxtaposition of variables in a binary coding scheme, with observations (patients, cases) in rows (n) and categories (attributes, features) in columns (p). Thus $\boldsymbol{Z} = \left[\boldsymbol{Z_1} \,|\, \boldsymbol{Z_2} \,|\, \ldots \,|\, \boldsymbol{Z_Q} \right]$ where $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ and $\boldsymbol{Z_Q}$ are indicator matrices for the first, second and the Q^{th} variable (for instance, if a patient belongs to a specific variable category, the respective column is assigned the value "one", and "zero" otherwise). Therefore, z_{ij} is an element of $\boldsymbol{Z}$, $1 \le i \le n$, $1 \le j \le p$, and for Q categorical variables, $p = \sum_{q=1}^{Q} p_q$ is the total number of categories and p_q the number of categories of the q^{th} variable. An example of a Superindicator Matrix with two categorical variables is shown in Table 1.

Table 1: An example of a MCA Superindicator Matrix with n subjects and two categorical variables; Varb-1 with three categories (ct-1, ct-2 and ct-3) and Varb-2 (dichotomous).

Subjects	**Varb-1**			**Varb-2**		**Total**
	ct-1	**ct-2**	**ct-3**	**ct-1**	**ct-2**	
#1	0	0	1	0	1	2
#2	1	0	0	1	0	2
#3	1	0	0	0	1	2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
#n	0	1	0	0	1	2
Total		n			n	2n

Let $\boldsymbol{F} = \boldsymbol{Z} \cdot N^{-1}$ be the correspondence matrix, with f_{ij} as elements, $r_{i+} = \sum_{j=1}^{p} f_{ij}$ and $c_{+j} = \sum_{i=1}^{n} f_{ij}$ being the row and column totals, respectively. N is the grand total of

the correspondence matrix, or: $N = \sum_{i=1}^{n}\sum_{j=1}^{p} z_{ij} = n \cdot Q$. The F matrix for the example displayed in Table 1 is shown in Table 2.

Table 2: The same example as in Table 1, showing the **F** matrix.

Subjects	Varb-1			Varb-2		$\mathbf{c_0}$
	ct-1	ct-2	ct-3	ct-1	ct-2	
#1	0	0	1/2n	0	1/2n	r_{1+}=1/n
#2	1/2n	0	0	1/2n	0	r_{2+}=1/n
#3	1/2n	0	0	0	1/2n	r_{3+}=1/n
⋮	⋮	⋮	⋮	⋮	⋮	⋮
#n	0	1/2n	0	0	1/2n	r_{n+}=1/n
$\mathbf{r_0}$	c_{+1}	c_{+2}	c_{+3}	c_{+4}	c_{+5}	

Computation of MCA starts from the standardized matrix:

$$\boldsymbol{S} = \boldsymbol{D}_r^{-0.5} \cdot \boldsymbol{F} \cdot \boldsymbol{D}_c^{-0.5}, \quad (1)$$

where $\boldsymbol{D}_r = diag(\boldsymbol{r_0})$ and $\boldsymbol{D}_c = diag(\boldsymbol{c_0})$ are diagonal matrices with vectors $\mathbf{r_0} = \left[c_{+1}, c_{+2}, \ldots, c_{+p}\right]$ and $\mathbf{c_0} = \left[r_{1+}, r_{2+}, \ldots, r_{n+}\right] = \left[\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right]$, being the average row and column profiles. Each element of the average row profile vector defines the *mass* of a category (all cases have the same mass).

The Singular Value Decomposition Algorithm (SVD) applied to $\boldsymbol{S}$ results in three matrices [7]:

$$\boldsymbol{S} = \boldsymbol{U} \cdot \boldsymbol{\Sigma} \cdot \boldsymbol{V}^{\boldsymbol{T}} \quad (2)$$

Above, $\boldsymbol{U}$ and $\boldsymbol{V}$ are the matrices of right and left singular vectors, with constraints $\boldsymbol{U} \cdot \boldsymbol{U}^{\boldsymbol{T}} = \boldsymbol{I}$ and $\boldsymbol{V}^{\boldsymbol{T}} \cdot \boldsymbol{V} = \boldsymbol{I}$, $\boldsymbol{I}$ the identity matrix with an adequate dimension. Matrix $\boldsymbol{\Sigma}$ is $n \times n$ diagonal with the non-negative singular values.

Since $\boldsymbol{S}$ is positive semi-definite, with all $s_{ij} \geq 0$, the singular values are displayed in decreasing order, $\sigma_0 \geq \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k > \sigma_{k+1} > \ldots \sigma_n = 0$ in which $k = n - Q - 1$ is the $\boldsymbol{S}$'s rank. Standard coordinates for rows are given by

$$\boldsymbol{\Phi} = \boldsymbol{D}_r^{-0.5} \cdot \boldsymbol{U}, \quad (3)$$

while for columns are:

$$\boldsymbol{\Gamma} = \boldsymbol{D}_c^{-0.5} \cdot \boldsymbol{V}. \quad (4)$$

The analysis is performed on principal coordinates, which are, for rows

$$\boldsymbol{R} = \boldsymbol{\Phi} \cdot \boldsymbol{\Sigma}, \quad (5)$$

and for columns

$$\boldsymbol{C} = \boldsymbol{\Gamma} \cdot \boldsymbol{\Sigma}. \quad (6)$$

This procedure yields a trivial solution ($\sigma_0 = 1$), but since no structure emerges from such a solution, the first columns of $\boldsymbol{U}$, $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are usually discarded. The squared singular values $\sigma_j^2 = \lambda_j$ are the eigenvalues of the Burt matrix, ($\boldsymbol{B} = \boldsymbol{Z}^{\boldsymbol{T}} \cdot \boldsymbol{Z}$) and the total inertia is given by $\sum_{j=1}^{k} \lambda_j = tr(\boldsymbol{D}_\lambda)$, given by n/Q -1 and dependent only on the disjunctive table format [2,3,4].

Since MCA inertias applied to the Superindicator Matrix are underestimated, a correction procedure [4] for the eigenvalues has been suggested:

$$\lambda_j^{adj} = \left(\frac{Q}{Q-1}\right)^2 \cdot \left(\lambda_j - \frac{1}{Q}\right)^2 \quad (7)$$

where λ_j^{adj} is the adjusted inertia. This adjusted inertia can, then, be used for adjusting principal coordinates, and is a measure of goodness-of-fit., while the matrix **F** for the same example is shown in Table 2.

Materials and Methods

Experimental Protocol and Data - EEG signals *from a* postural control protocol dataset was used in this study. Thirty-one subjects (21 male and 10 female), ages 21 to 45 (31.0 ± 6.6) years, *heights* 154 to 187 (172.7 ± 9.4) cm and body weights 46 to 107 (73.3 ± 17.3) kg participated in the study. All subjects presented no history of neurological pathologies, osseous, muscles or joints diseases, or equilibrium disorders. An anamnesis was performed to obtain information about headaches, illnesses, vertigo, eyestrain. Subjects using contact lenses or glasses were included when no problem with their use was reported. The study was approved by a Local Institutional Review Board (IESC/UFRJ – Ref. 100/2011). None of the authors participated as volunteer.

The complete experimental protocol consisted of acquiring the EEG simultaneously with the stabilometric signals, but, as mentioned, only EEG signals were analyzed here. The experiments were performed in an electro-magnetically shielded room, under controlled environmental conditions (23ºC, attenuated sound and light control), with the subject seated in a comfortable armchair and barefooted on a force platform placed 1 meter apart from a white wall.

The EEG signals for the complete protocol were acquired during five-minutes for each protocol's condition, with subjects *seated* or *in upright* position and with *eyes open* or *closed*.

Since the aim was to study differences between seated and upright position over the right hemisphere, only the P4 derivation was used in this study (International 10/20 System: monopolar derivation, bilateral ear-lobe reference and ground in FPz, impedance less than 5kΩ). EEG signals were first segmented into 1 second zeroed-mean epochs of 400 samples. An artifact rejection methodology [7] was applied, resulting in distinct number of epochs for each volunteer and condition (min = 20, max = 300). For computational convenience (all epochs were stored in an array) volunteers with a minimum of 150 artifacts free epochs were retained in the study.

Signal Processing and Variables - Since one signal was collected (P4 derivation) for each condition, there were 31 volunteers x one derivations x two conditions = 62 EEG signals. A rectangular window was applied to

each epoch, and the averaged periodogram was calculated by the Bartlett method [8]. Three variables were extracted from the periodogram: *power* of the alpha (8-13 Hz), theta (4-8 Hz) and beta (13-30 Hz) bands in log10 μV^2 (defined as the trapezoidal area centered in the maximum peak of respective bands). These continuous variables were categorized by quartiles. The nomenclature here adopted was XY, where X is the band power and Y is the number of the quartile, for example, *t1* means the power of the first quartile of theta band. *Gender* was included as a dichotomous variable coded "M" (male) and "F" (female).

Protocol conditions were characterized as a nominal /categorical variable: (i) seated and eyes closed (spontaneous EEG with room lights off, denoted as "A"); (ii) seated and eyes open ("B"); (iii) standing up with eyes open ("C") and (iv) standing up with eyes closed ("D"). The trials carried out with eyes open were conducted with room lights on and with the subject observing a white wall. An interval of three minutes was taken between each condition, and the subject remained resting in the chair during this period. Conditions analyzed in this study were only "B" and "C" (Figure 1). Therefore, a data matrix 62 x 5 (signals x variables) was converted in a 62 x 16 Superindicator Matrix.

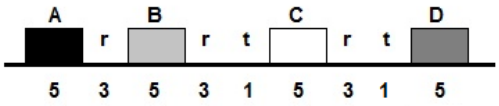

Figure 1: Protocol conditions, where *r* e *t* are rest and transition, respectivaly. The numbers 1, 3 and 5 are the duration of each condition, in minutes.

Multiple Correspondence Analysis - MCA is mainly carried out on CP's positions in the symmetric map (given by equation (6), adjusted by (7)), and on obtained statistics such as explained inertia [2,3,4]. In general, proximity between CP's can suggest associations between them, while contrasts in relation to the origin can suggest dissimilarities. Also, projections onto each dimension can indicate similarity. The Scree plot was applied to determine the number of dimensions to be analyzed.

All computations and calculations were performed by the "R" software (www.r-project.org), freely available on the web.

Results

All 31 volunteers could be included in the study. Mean ± standard deviation for frequency domain variables were 6.4 ± 1.0 log10 μV^2 (alpha band power, range 4.2-8.3 log10 μV^2); 5.3 ± 0.7 log10 μV^2 (beta band power, range 4.0-6.4 log10 μV^2) and 6.1 ± 0.6 (theta band power, range 5.0-8.3 log10 μV^2). In Figure 2, the PSD of one volunteer (conditions "B" and "C", male) is shown, with the area corresponding to power centered at its maximum alpha band highlighted.

The Scree plot is shown in Figure 3, where two dimensions are suggested, with a percent of explained inertia of 91.7% (76.2% and 15.5% for the first and second dimensions; total adjusted inertia: 0.17). Inertias for the third, fourth and fifth dimensions are 7.7%, 0.4% and 0.3, respectively (maximum dimension: 5).

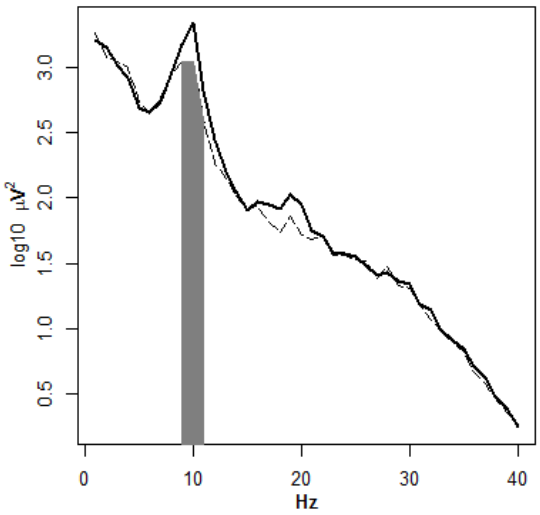

Figure 2: PSD of one volunteer in the postural control protocol. Solid line: eyes open-upright, dotted line: eyes open-seated position. In grey, the power of alpha band.

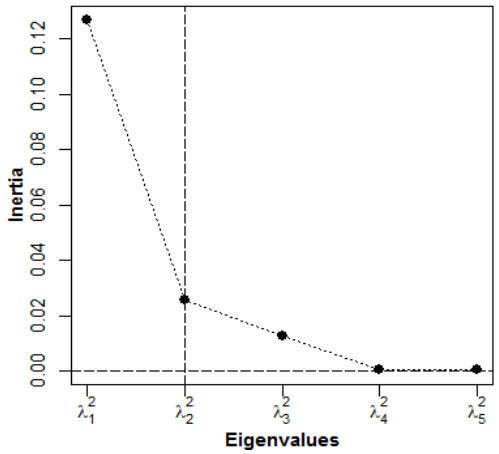

Figure 3: Scree plot in a MCA for EEG signal analysis. The vertical dashed line is the number of dimensions selected.

The general analysis of the symmetric map (Figure 4) shows a "horseshoe effect" for the categorized variables, from the first quartiles to the fourth ones. The *Gender* and *Condition* variables show a contrast between each dimension (since these are dichotomous variables). A total of 21 of the subjects were males, thus the CP "M" is closer to the centroid than "F". The horseshoe effect indicates a gradient of change according to the quartiles of the continuous variables, producing four main clusters, one in each quadrant. Most quartiles of alpha and theta power ("a" and "t") are very close, suggesting that both variables should be merged in the same variable. In the first quadrant, higher levels of alpha ("a4"), theta ("t4") and

beta ("b3") power seem to be more related to females than males, which are, mostly, related to medium levels of power ("a2", "b2", "t2", "t3"). Condition "B" seems to be related to lower levels of power ("a1", "b1", "t1"), in contrast to condition "C" ("a3", "b4").

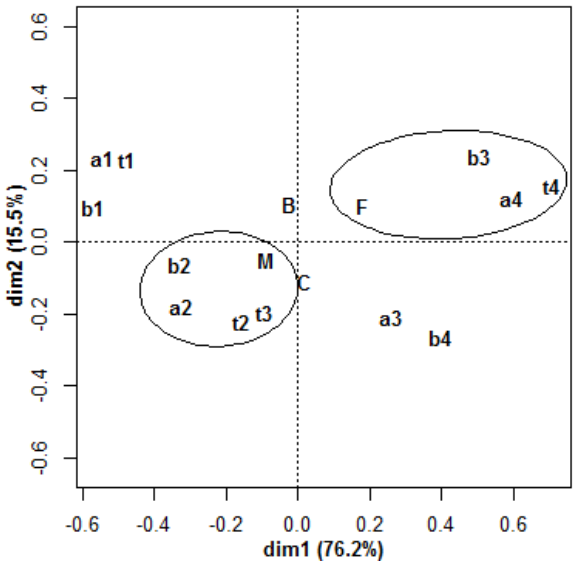

Figure 4: Symmetric map for a MCA analysis with EEG signals from a postural control protocol dataset. Brackets: the explained inertia for each dimension

The first dimension was characterized by the contrast between the lower and higher quartiles of power (with *males* associated to the former). Most second and third quartiles of power are closer to the centroid. Conditions "B" and "C" are also close the centroid. On the other hand, the second dimension was characterized mainly by the contrast between conditions "B" (associated to the lower and higher levels of power, together with females) and condition "C" (medium levels of power, together with males).

Discussion

Dealing with mixed data implies in either treating quantitative and qualitative variables separately or in categorizing continuous data. However, none of these options are optimal, and a trade-off between the number of assumptions needed for inference and categorization must be considered in order to ease exploratory analysis.

In this context, MCA is an interesting technique for simultaneously analyzing continuous and categorical variables. Therefore, since mixed variables are very common in the biomedical field, the technique can be used as hypothesis-generating mechanism, for instance, relating pathological conditions or gender (nominal variables) and spectrum characteristics or other quantitative variables.

This work was concerned with EEG signals collected through the P4 derivation. The parietal cortex is an associative area for sensorial integration, and, therefore, it is expected that, in parietal areas, alpha power should be higher for the seated position as opposed to upright stand [9]. In Figure 4, it can be seen that for the condition *seated* / *eyes open* ("B") there are opposite levels of power ("a1", "b1", "t1" against "b3", "a4", "t4") onto dimension 2. Since the volunteers remained seated during 5 minutes facing a white wall, it is possible that some of them were developing mental activities such as “planning”, what would explain the lower levels of alpha power in P4.

Furthermore, it could be observed that the beta band activity behaves similarly to the alpha / theta bands (Figure 4). The somatosensory cortex (P3 and P4) is associated to the motor cortex in self-paced tasks [10]. Hence, it can be hypothesized that, in this specific setting of balance tasks (quasi-static or quiet-standing), part of muscle activity control is medullar, and thus, even in an upright, relaxed position, beta activity remains low. Indeed, the EEG signals were acquired after the transition period for the seated / standing up positions, and, according to [11], beta activity in the motor cortex (C3 e C4) before and after the transition time is similar. Further studies using statistical tests, such as the F-spectral, should be performed to confirm these findings.

Concerning gender differences, it is well-known that females have higher levels of activity over the right hemisphere than males. These findings were also present in Figure 4, given that projections of "F" onto the first dimension were associated to the higher levels of all band activities (except "t3"), in contrast to "M".

Conclusion

Results suggest that for P4 and eyes open females are related to higher levels of power in both positions (seated and standing up) and that the protocol condition eyes open / upright position is related to medium levels of power.

Acknowledgements

This work received partial support from FAPERJ, CAPES (PROEx program) and the Brazilian Research Council (CNPq), to which we thank

References

[1] Krzanowski WJ. Mixtures of continuous and categorical variables in discriminant analysis. Biometrics. 1980; 36 (3):493-499.

[2] Infantosi AFC, Costa JCGD, Almeida RMVR. Correspondence analysis: a theoretical basis for categorical data interpretation in health sciences. Cadernos de Saúde Pública. 2014; 30 (3): 473-486. *In Portuguese.*

[3] Greenacre M, Hastie T. The geometric interpretation of correspondence analysis. Journal of the American Statistical Association.1987; 82(398):437-447.

[4] Greenacre MJ. Interpreting multiple correspondence analysis. Applied Stochastic Models and Data

Analysis. 1991; 7:195-210.

[5] Benzecri JP. Statistical analysis as a tool to make patterns emerge from data. In: Proceedings of The International Conference on Methodologies of Pattern Recognition; 1969, Honolulu, Hawaii, USA, 1969, pp. 35-74.

[6] Golub GH, van Loan CF. Matrix Computations. 2nd ed. The John Hopikins University Press, Baltimore; 1996.

[7] Simpson DM, Tierra-Criollo CJ, Leite RT, Zayen EJB, Infantosi AFC. Objective response detection in a electroencephalogram during somatosensory stimulation. Annals of Biomedical Engineering. 2000; 28: 691-698.

[8] Kay SM, Marple-Jr SL. Spectrum analysis – a modern perspective. Proceedings of the IEEE. 1981; 69: 1380-1489.

[9] Da_Silva PJG. Análise eletroencefalográfica do controle postural ortostático em ambiente de realidade virtual [Thesis]. Rio de Janeiro: COPPE – Universidade Federal do Rio de Janeiro (COPPE-UFRJ); 2010.

[10] Baker SN. Oscillatory interations between sensorimotor cortex and the periphery. Current Opinion in Neurobiology. 2007; 17: 649-655.

[11] Da_Silva PJG, Infantosi AFC. O sincronismo cortical durante o movimento voluntário. In: Anais do XXIII Congresso Brasileiro de Engenharia Biomédica; 2012 Out 1-5; Porto de Galinhas, Brasil. 2012. p. 1948-52.