Principal Components Clustering through a Variance-Defined Metric

J.C.G.D. Costa, D.B. Melges, R.M.V.R. Almeida, and A.F.C. Infantosi

COPPE-Federal University of Rio de Janeiro /Biomedical Engineering Program, Rio de Janeiro, Brazil

Abstract— **This work aims at proposing a clustering procedure through a new metric, a weighted Euclidean distance, in which the weights are the ratio of corresponding eigenvalues and the largest eigenvalue found after a Principal Components Analysis. In order to illustrate the method, the procedure was carried out on twenty-one newborn EEG segments, classified as TA (Tracé Alternant) or HVS (High Voltage Slow) patterns. The observed clustering structure was assessed by the cophenetic and agglomerative coefficients. Results showed that, despite its unlikely existence, a clustering structure was suggested by the traditional approach. This structure, however, was not confirmed by the proposed method.**

Keywords— **Cluster Analysis, EEG, Principal Components Analysis.**

I. Introduction

Cluster Analysis (CA) and Principal Components Analysis (PCA) are multivariate methods commonly used in studies of biomedical signals. Despite their limitations, they provide an exploratory, informative insight of data structure. On the other hand, biomedical signals are high dimensional data difficult to interpret due to the presence of artifacts and noise. Despite of that, CA and PCA are frequently carried out in order to extract features of interest, especially for diagnosis purposes [1, p.449].

CA through Euclidean distance is usually performed on a raw data matrix **M** (for example, a matrix with *n* individuals and *p* variables), but when PCA is applied to **M** all *n* individuals displayed in the p-dimensional space can be re-positioned in a new coordinate system, so that data variability is considered in the analysis. With this simple procedure, the information on variance is added to the analysis, without the need of discarding PCA dimensions. Furthermore, the distance between two points projected onto the axis with the lowest variance has the same value than that the same distance onto the highest one, because only an axes translation / rotation was carried out.

The aim of this study was to introduce a clustering procedure through a new metric, a weighted Euclidean distance, in which the weights are the ratio of estimated eigenvalues and the largest eigenvalue found after PCA. In order to illustrate the method, the procedure was carried out on twenty-one newborn EEG segments, classified as TA (Tracé Alternant) or HVS (High Voltage Slow) patterns.

II. Background

A. Clustering Algorithm

One of the most common types of clustering algorithm is the Single Linkage Hierarchical Algorithm (SLHA) [2], which, starting from a dissimilarity measure, assumes that the individuals (or signals) have been merged to the nearest neighbor point. Individuals are characterized by points in the Euclidean space and each one is grouped subsequently to the others, obeying some clustering rule (for example, to decrease variance within clusters and increasing variance between clusters). The most used dissimilarity measure is the Euclidean distance.

A clustering strategy frequently begins with the raw matrix **M**, with *n* individuals in rows and *p* variables in columns, from which a dissimilarity matrix is built. Since SLHA is a monotone admissible strategy, and since any monotone transformation to the dissimilarity matrix does not alter clustering results [2], this represents an interesting property for the study of dissimilarity measures.

B. Principal Components Analysis

One of the most used methods to display individuals as points in the Euclidean space is the mentioned Principal Components Analysis. PCA is based on a singular value decomposition (SVD) algorithm of the covariance (or correlation) matrix [3]. Used as an exploratory tool, PCA can reveal clusters graphically, in which case the Gaussianity assumption for data distribution is not mandatory [3, p. 49].

For graphical cluster representation, the relationship between individuals can be highlighted, using principal components (PC's) as axes plotting the individual points (IPs). Thus, after suitable scaling of **M**, one has:

$$\boldsymbol{Q} = svd(\boldsymbol{S}) = \boldsymbol{U} \cdot \boldsymbol{D} \cdot \boldsymbol{V}^{\boldsymbol{T}} \qquad (1)$$

where **U** and **V** are orthogonal matrices and the diagonal matrix **D** has eigenvalues in descending order. Since the covariance matrix **S** is a square matrix, the coordinates of individuals can be found as **Z=MV**. This approach yields eigenvector-eigenvalue pairs with higher retention of variance, and the PCs become the new uncorrelated variables.

P.D. Bamidis and N. Pallikarakis (Eds.): MEDICON 2010, IFMBE Proceedings 29, pp. 45–48, 2010.
www.springerlink.com

C. The New Metric D*

Hierarchical cluster algorithms are frequently employed to identify associations between IPs. However, one of the drawbacks of these algorithms is that their results *always* generate clusters, even if these are actually unstable and non-meaningful, thus demanding extra validation strategies for result assessment [2].

To deal with this drawback in defining actual "clusters" in PCA plots, a new metric (D*) is proposed, in line with the tolerance distance statistic suggested in a previous work [4]. This metric takes into account the explained variance pertinent to each PCA axis through weighted Euclidean distances. The idea is that if an axis has lower variability, then the distance between two clusters projected onto that axis should have less "importance" than the same distance in a higher variance axis. Thus, D* has the ordinary Euclidean distances weighted by the ratio between all eigenvalues and the eigenvalue of the first axis (which is the highest variance axis):

$$D^*(x,y)=\sqrt{\frac{(x_1-y_1)^2}{\tau_1}+\frac{(x_2-y_2)^2}{\tau_2}+\ldots} \qquad (2)$$

where **x** and **y** are the IP coordinates and τ_i are the ratio between the eigenvalue corresponding to the *i*-th axis and the first one. Equation (2) can be re-written as:

$$D^*(x,y)=\sqrt{\left(\frac{x_1}{\sqrt{\tau_1}}-\frac{y_2}{\sqrt{\tau_1}}\right)^2+\left(\frac{x_2}{\sqrt{\tau_2}}-\frac{y_2}{\sqrt{\tau_2}}\right)^2+\ldots} \qquad (3)$$

which is an ordinary Euclidean distance onto a new coordinate system. However, since τ_1=1, only axes of lower explained variances are re-scaled, and, hence, D* decrease IPs' similarity for dimensions of lower variances. SLHA can then be carried out on the new Euclidean space.

III. Materials and Methods

A. EEG Acquisition and Pre-processing

EEG signals were collected from derivation F4-P4 of seventeen full-term newborns (gestational age of 37-42 weeks and APGAR $\geq$ 8 in the first and fifth minutes post-delivery) during physiologic sleeping at the Instituto Fernandes Figueira (FIOCRUZ, Rio de Janeiro, Brazil). The signals were band-filtered (0.5-70 Hz) and digitized at the sample rate of 200 Hz (for further details refer to [5]).

Firstly, the EEG recordings corresponding to the Quiet Sleep Stage were identified and classified in the sleep patterns High Voltage Slow (HVS) or Tracé Alternant (TA) by a clinical expert. Twenty-one artifact-free EEG segments with thirty seconds of duration were selected, being fourteen of the TA pattern and seven of HVS. Four newborns had two segments selected for analysis (S.2-S.3, S.4-S.5, S.9-S.10, and S.13-S.14, as listed in Table 1).

Finally, the power spectral density was calculated using the Bartlett's Periodogram with *M*=10 subsegments. Thereby, spectral resolution was set to 0.333 Hz, and, in order to minimize spectral leakage, a Hamming window was applied to each subsegment.

Seven real-valued parameters were selected for characterizing EEG signals: the maximum power spectral density (μV^2/Hz) for the bands slow delta (0.25-2 Hz), fast delta (2-4 Hz), theta (4-8 Hz), alpha (8-13 Hz) and beta (13-30 Hz), and also the standard deviation (SD) of the samples in the samples segments and the difference between maximum positive and minimum negative values (Mn) of the samples segment (both μV).

B. Comparison between Methods

A raw data matrix with segments in rows and the features in columns was column-scaled for zero mean and unity variance, and a distance matrix was calculated. After the correlation matrix was defined, all subjects were displayed in the multidimensional space spanned by PC's, and D* was calculated from the segments' coordinates obtained by PCA. Then, after re-scaling the coordinates, the SLHA was applied to both distance matrices.

For assessing clustering performance in the results obtained by applying the ordinary Euclidean distance and applying D*, the cophenetic correlation (CC) and the agglomerative coefficient (AC) were used. CC is the correlation coefficient between the set with elements of the cophenetic distance matrix and the set composed by corresponding elements of the dissimilarity matrix, where each element of the cophenetic matrix is the distance in the dendrogram at which the respective pair of segments is merged [2]. AC is an index that measures the quality of structure found by the clustering algorithm [6] and it is defined as:

$$AC=\frac{1}{n}\sum_{i=1}^{n}\left(1-m(i)\right) \qquad (4)$$

where *m(i)* is the ratio between the dissimilarity (distance) in which the segment *i* is merged at the first step to the distance achieved in the last step (when all *n* segments are joined together). Therefore, both indices are dimensionless and vary in the range 0-1. If CC and AC are close to unity a strong structure may be accepted as existing [6].

C. *Simulated Data*

To determine whether the proposed method can identify an actual clustering structure in the data, three well-defined clusters were generated and the two procedures outlined above were carried out on them. Forty-five points were grouped in three clusters with fifteen individuals and two Gaussian variables, one with unity variance (V1) and the other with variance equal to 4.0 (V2). The Pearson correlation coefficient between variables was set to 0.016 (Fig. 1).

Table 1 Data summary of 21 EEG subsegments

EEG	S. Delta	F. Delta	Theta	Alpha	Beta	SD	Mn
TA	1801	176	73	19	17	7.9	97
TA	17171	627	125	49	54	22.2	140
TA	12157	1491	361	119	48	23.4	216
HVS	34190	1776	638	83	112	29.0	184
TA	34473	1040	615	78	153	28.1	174
HVS	44793	2567	1479	175	48	32.8	236
TA	93349	2158	391	154	39	41.0	316
TA	24198	1195	589	130	58	25.7	179
TA	13528	839	187	55	13	21.3	168
TA	15289	1026	296	74	13	18.9	156
TA	11762	655	179	43	39	17.5	147
TA	22518	2395	585	103	26	23.0	168
TA	16190	1777	675	115	31	21.1	219
TA	10740	1019	403	67	11	20.3	126
HVS	10524	869	248	84	49	19.3	119
HVS	9544	1199	276	124	40	17.6	150
TA	10123	1586	321	63	30	17.1	135
TA	14210	377	169	72	17	24.2	154
HVS	13806	1596	358	58	41	19.9	154
HVS	4939	523	123	39	47	17.7	116
HVS	8420	1520	411	105	96	17.7	107

All computations were performed with the R software (version 2.8.1), freely available on the Internet at www.r-project.org. Filtered signals were obtained from a MATLAB™ environment through R.matlab package and post-processed by R's *signal* package.

IV. Results

The PCA plot for all 21 subsegments in the first two dimensions is showed in Figure 2, where it can be seen that no cluster structure between TA and HVS patterns was suggested. Explained variances were respectively 43.6% and 18.8%. The dendrogram for the raw matrix **M** is shown in Fig. 3, and the one for the proposed approach is shown in Fig. 4.

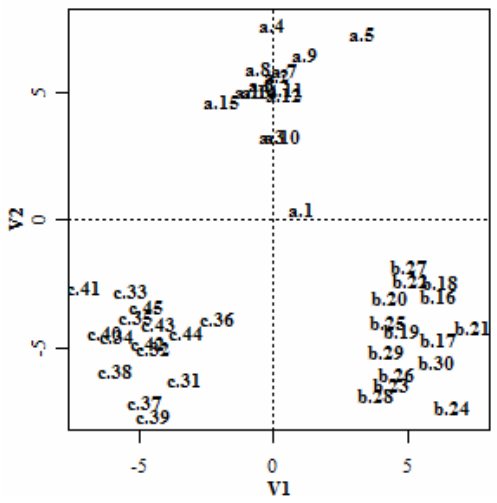

Fig. 1 Simulated data

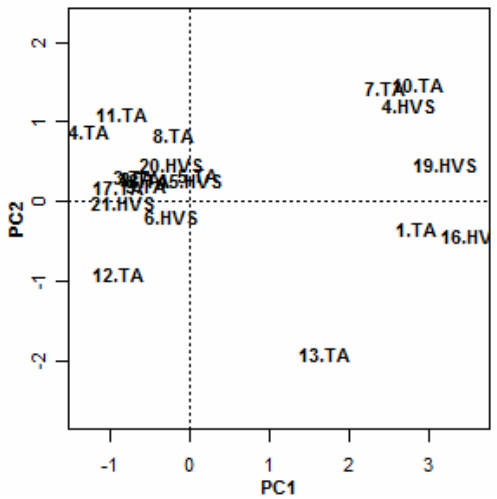

Fig. 2 Principal Components plot

CC and AC for the traditional approach were 0.91 and 0.62, respectively, suggesting a reasonable structure, while for the proposed approach the same indices found were 0.69 and 0.49, suggesting a weak structure, more accordingly to Fig. 2, 3 and 4. For the simulated data, both approaches showed the same values for CC and AC, 0.90 and 0.85, respectively, suggesting a genuine clustering structure, as expected.

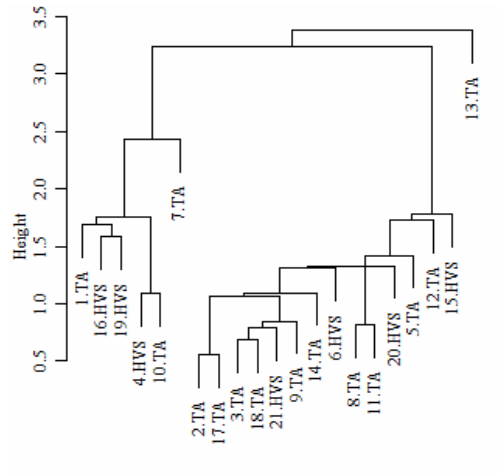

Fig. 3 SLHA for the raw matrix

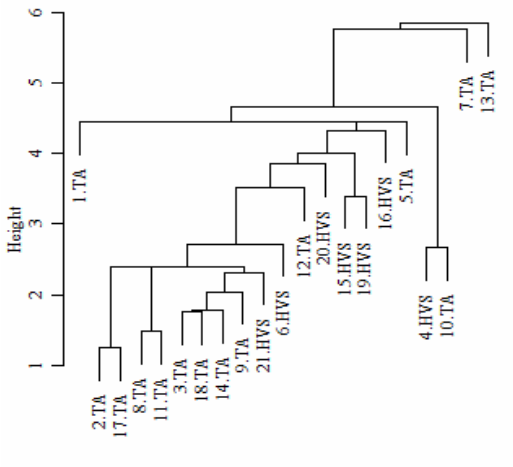

Fig. 4 SLHA for D*

V. Discussion

The objective of this work was to propose a new metric, based on a weighted Euclidean distance which considers data variance, instead of the ordinary Euclidean distance frequently used. This metric considers the concept of "variance means information", implying that less variance should have less "weight" in the analysis. This is not the same as discarding the less explained variance axes, since some important features can be extracted from these axes [3]. Thus, the proposed approach "stretches" the projected distances onto lower explained variances axes relatively to the original Euclidean space, since IPs coordinates are re-scaled by a factor larger than unity for all axes, since $\tau_i \leq 1$.

In addiction, although CA is often used after dimensionality reduction by PCA [7], some important data characteristics can be lost if few dimensions are retained for analysis. Thus, the use of additional knowledge about the data is recommended, and the data variance pertaining to each PCA axis was the statistic chosen to this end.

The results suggested that the new metric is more robust than the ordinary Euclidean distance, given that a clustering structure was suggested by the latter in an unlikely grouping representation (EEG), as opposed to the former. Further studies should include higher dimensional data in order to better explain the specific reasons for this discrepancy.

Acknowledgments

The National Council of Research of Brazil (CNPq) partially financed this research.

References

1. Rangayan, R M (2002) Biomedical Signal Analysis, a case Study Approach. IEEE Press, Piscataway, NJ, USA.
2. Gordon, A D (1987) A review of hierarchical classification. J Royal Stat Soc 150: 119-137.
3. Jolliffe, I.T (2004) Principal Component Analysis. Springer, NewYork, USA.
4. Costa, J C G D, Almeida, R M V R, Infantosi, A F C et al. (2008) A heuristic index for selecting similar categories in multiple correspondence analysis applied to living donor kidney transplantation. Comput Methods Programs Biomed 90:217-219.
5. Melges, D B, Infantosi, A F C, Ferreira, F R, Rosas, D B (2006) Using the discrete hilbert transform for the comparison between tracé alternant and high voltage slow patterns extracted from full-term neonatal EEG, IFMBE Proc. vol. 14, World Congress on Med. Phys. & Biomed. Eng., Seoul, 2006, pp 1003-1006.
6. Struyf, A, Hubert, M, Rousseeuw, P J (1996) Clustering in an object-oriented environment. J Stat Soft 1.
7. Mauldin-Jr, F W, Levy, J H, Behler, R H, Nichols, T C, Marron, J S, Gallippi, C M (2006) Blind source separation and k-means clustering for vascular ARFI image segmentation, in vivo and ex vivo. IEEE Ultrasonics Simposium 1:1666-1671.

Author: A.F.C. Infantosi
Institute: Biomedical Engineering Program / COPPE–Federal University of Rio de Janeiro
Street: Av. Horacio Macedo, 2030, Bl. H, Zip code 21941-972
City: Rio de Janeiro
Country: Brazil
Email: afci@peb.ufrj.br